A Pluralist Framework for the Philosophy of Social Neuroscience

Sergio Daniel Barberis, M. Itatí Branca, and A. Nicolás Venturelli

Abstract The philosophy of neuroscience has been a dynamic field of research in the philosophy of science since the turn of the century. As a result of this activity, a new mechanistic philosophy has emerged as the dominant approach to explanation and scientific integration in neuroscience. Rather surprisingly, the philosophy of social neuroscience has remained an almost uncharted territory. In this chapter, we advance a pluralistic framework for that field. Our framework seeks to ground the proliferation of modeling approaches, explanatory styles, and integrative trends within social neuroscience. First, we highlight the plurality of modeling approaches pursued by social neuroscientists by reviewing the distinctive features of mechanistic models, dynamical models, computational models, and optimality models. Second, we reject unitary explanatory perspectives and emphasize the plurality of explanatory styles that can emerge from those modeling approaches, considering their contents and vehicles. As regards their content, we present two kinds of information a model may provide, namely, causal/compositional or noncausal/structural information. As regards their vehicles, we examine and illustrate different guiding representational ideals (e.g., precision, generality, and simplicity). Third, we turn to integrative trends in social neuroscience, assessing the prospects of inter-theoretical reduction, mechanistic mosaic unity, and multilevel integrative analysis. We contend that the pluralist framework we develop is an adequate approach to scientific modeling, explanation, and integration in social neuroscience. We additionally address how this pluralistic perspective may shed light on the intersection between

S.D. Barberis (✉)
Universidad de Buenos Aires (UBA), Buenos Aires, Argentina

Agencia Nacional de Promoción Científica y Tecnológica (ANPCyT), Buenos Aires, Argentina
e-mail: sergiobarberis@gmail.com

M. Itatí Branca
Universidad Nacional de Córdoba (UNC), Córdoba, Argentina

Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Buenos Aires, Argentina
e-mail: itatibranca@gmail.com

A. Nicolás Venturelli
Instituto de Humanidades (UNC/CONICET), Buenos Aires, Argentina
e-mail: nicolasventurelli@gmail.com

© Springer International Publishing Switzerland 2017

A. Ibáñez et al. (eds.), *Neuroscience and Social Science*,
DOI 10.1007/978-3-319-68421-5_21

the neural and the social realms, in a context of greater interdisciplinary collaboration between neuroscientists and social scientists.

Keywords Social neuroscience • Models • Explanation • Pluralism • Integration

1 Introduction

The development of the philosophy of science from the second half of the twentieth century has been primarily characterized by an increasingly thorough focus on particular disciplinary areas, following the widespread recognition that each scientific area presents different philosophically relevant theoretical and methodological questions (see Bunge, this volume). To offer a panoramic view of recent philosophical work in neuroscience, here we address three prominent issues that are highly relevant for social neuroscience (SN), namely: modeling approaches, scientific explanation, and theoretical integration. Though this selection of topics is admittedly limited, it proves fairly representative of contemporary debates and results.

We will approach the issue of modeling in SN and its relation to the problems of scientific explanation and integration in the field, from the perspective of the working scientist who constructs, revises, and applies models under some specific, concrete objective. This aligns with a growing trend in the philosophy of science aiming to understand the dynamic aspect of scientific knowledge, including the processes underlying the emergence, change, and disappearance of research programs, disciplines, and whole fields, as well as the evolution of scientific instruments, experimental paradigms, models, and theories.

SN emerged only recently, around 1990, as a multilevel approach to the study of the neural bases of social behavior [1]. This approach intended to reject previous cognitive neuroscientific perspectives that primarily focused on the human brain considered in isolation; in this way, most research was overly indifferent to the inherently social nature of human beings, which in turn became the central subject of interest to SN [2, 3]. Since then, SN has experienced significant development, including the establishment of two journals in 2006, *Social Cognitive and Affective Neuroscience* (Oxford University Press) and *Social Neuroscience* (Taylor and Francis), and three societies, the Social and Affective Neuroscience Society (SANS), established in 2008; the European Society for Cognitive and Affective Neuroscience (ESCAN), founded in 2009; and the Society for Social Neuroscience (S4SN), established in 2010. Even though the development of the field has certainly accelerated, SN is still very much in its infancy, full as it is of programmatic questions to be approached and conceptual issues that need to be reviewed.

In what follows, we will focus on a subfield which, though important for SN, is not exclusive to this field. Specifically, we will deal with bottom-up approaches, that is, approaches which take as a general starting point the description of the structural and functional aspects of neuronal mechanisms and neuronal systems, which can in

turn be connected with socially relevant psychological phenomena. This is a necessary restriction of our focus, given the methodological distance and differences in scope between SN and other neuroscientific arenas, such as neuroanthropology, neurosociology, or neuroeconomics. Although these share multiple aspects with SN and cognitive SN, they approach research in an inverse direction, hinging on the economic, political, and cultural influences on brain function and development. One common motivation behind the diverse disciplines that adopted a top-down approach has been a growing dissatisfaction with the traditional, non-mentalistic understanding of social phenomena (cf. [5], p. 11) and the realization that neuroscience could provide social science with a rigorous basis for conceptualizing and measuring the mind. Another way of framing this is that while these disciplines, despite their individual differences, tackle the neural dimension of classical social science questions, the subfield of SN we address here applies traditional neuroscientific methods to social phenomena—although we acknowledge SN is by no means restricted to such bottom-up approaches.

The case of SN, as we restrict it here, is peculiar both on account of its complexity (as a hybrid field approaching social phenomenon through neuroscientific methods) and its relative youth within neuroscience. This partially explains why the philosophical reflection specifically directed to problems arising from SN and social cognitive neuroscience is still very incipient. In what follows we thus make an effort to extend some of the more developed themes within the philosophy of neuroscience to enlighten relevant aspects of contemporary social neuroscientific research. This situation makes the advancement of an established philosophy of SN a promising and compelling challenge for the years to come.

2 Models in Social Neuroscience

2.1 Theories and Models in Philosophy of Science

Mainly stemming from the mid-century historical turn led by Thomas Kuhn, scientific models have come to occupy a fundamental and pervasive role in recent philosophy of science. The variety of modeling strategies across disciplines and the varied functions they serve led to the general recognition that models exert major influence in the production of scientific knowledge. This realization contrasts sharply with the merely psychological, pedagogical, or at most heuristic function that logical empiricist philosophers had previously ascribed to models. Relevant as they may have been from a psychological, sociological, or historical point of view, models were relegated to the periphery of the philosophy of science by influential thinkers such as Rudolf Carnap, Carl Hempel, and Karl Popper.

Following Bailer-Jones [4], a scientific model can be seen as an interpretative description of an empirical phenomenon whose primary general function is to facilitate cognitive access to it. This access can be either perceptual or intellectual. In

order to grant this kind of otherwise unavailable access, models tend to focus on specific aspects of a phenomenon. This privileged access is achieved, on the one hand, by leaving aside a host of other aspects pertaining to the phenomenon and, on the other, by simplifying or idealizing those aspects considered to be essential for the depiction of the phenomenon or vis-à-vis some specific objective pursued through the modeling effort. Sometimes, modeling may also imply appealing to some unrealistic or fictional assumptions to meet ongoing requirements (which can and cannot be of a representational kind). In this sense, a model is always a partial description of its target phenomenon, under some particular problem context.

The distinction between scientific theories and models involves at least three points of contrast: generality, structure, and function. Although the nature, composition, and rate of change of scientific theories remain a hotly debated topic, here we will conceive them as articulate and wide-ranging constructions that represent and explain general characteristics of a set of phenomena (cf., [5]). A construction of this kind can take on different formats, such as linguistic or mathematical, but, in most cases, it would allow its expression in symbolic notation. From a comparative perspective, a scientific theory is taken to be the most exhaustive and far-reaching presentation of the particular way things are thought to be and function within a certain state of affairs and for a particular scientific community. In this sense, although models can help unleash their representational potential, theories occupy a somewhat distant position with regard to phenomena.

Now, from the point of view of the model's user, a scientific model may also be thought of as a complex tool for thought [6, 7]. As such, it can be directed toward a wide spectrum of related endeavors, such as representing data sets, exploring novel phenomena, orienting or directing experimental design, driving computational simulation efforts, theory application, construction, revision, and so forth. The often-highlighted central position of models, as mediators between theory and phenomena, is inherent to this multiplicity of roles and uses (a multiplicity that is accordingly absent in the case of theory). This also contrasts sharply with the above concept of theory, which can be conceived as a sort of end product of modeling and experimental efforts, though open to revision and adjustment. The model-as-tool notion is thus another important point of departure between both concepts.

One particularly notable aspect of this understanding of scientific models, especially regarding SN as well as other areas of contemporary neuroscience, is the fact that models maintain an important kind of autonomy vis-à-vis theory. This concerns how models are elaborated as well as how they are variously deployed. The philosophical tradition that we are following here (see, e.g., [8]) has emphasized several ways in which this autonomy can be found in the history of science. In particular, the cognitive profit brought about through modeling exceeds by far its representational capacity as derived by theory, and most importantly, it has to be acknowledged even in absence of a firm and fully developed theory within a particular discipline or area of research. This is a situation that fits perfectly with SN as practiced today. In the next subsection, we will consider modeling in neuroscience and particularly in SN, so as to then assess the particular kinds of models that are found in the field.

2.2 *Theoretical Principles and Models in Social Neuroscience*

As already anticipated, models in neuroscience have a preeminent role as well as a specific kind of autonomy regarding theory. The primacy of models in the field partly reflects the fact that theories of brain function are not dominant, as growingly acknowledged within the philosophy of neuroscience. For our purposes here, the main point is not that theories are nowhere to be found in SN (as we will shortly see); rather, we maintain that they do not define a high degree of agreement within relevant communities. In the recent literature, a certain widespread consensus can be found around this idea [9]. In a very early statement, Churchland and Sejnowski [10] defined neuroscience as a data-intensive field while remaining poor with regard to theory. While this premature recognition may not have been cautionary, the years to come have rapidly intensified this particular situation (as we will shortly see, to a degree recently deemed problematic by notable neuroscientists). The growth and sophistication of experimental approaches, greatly fueled by the resonant expansion of different kinds of structural and functional neuroimaging studies, certainly stands as a crucial factor contributing to this trend.

Although a systematic philosophical treatment of the theoretical status in neuroscience is still due, several philosophers have advanced considerations along these lines. The position defended by Valerie Hardcastle is worth considering in some detail. In a series of papers [11, 12], she portrays the theoretical dimension of neuroscience as a collection of loosely related and to some extent autonomous theoretical principles. The main point is that these principles are not (or so far have not been) articulated into a cohesive theory addressing some specific set of phenomena. On the other side, these general principles are used in order to guide experimental research and interpret experimental results. They can be thought of as contributing an interpretative framework that, in a moderate sense, drives research. Inasmuch as this is an accurate picture, theoretical frameworks of this kind may be in part responsible for the fragmentation inherent to almost all fields in cognitive and social neuroscience—and as already pinpointed in the very early moments of SN (e.g., [1]). Some principles that may be mentioned are, for example, the role of functional segregation as an organizing element in the cerebral cortex (e.g., [13]), the assumption that two or more sensory systems are anatomically overlapping (e.g., [14], one case considered by Hardcastle and Stewart), or the increasingly explored idea that neural networks learn statistical regularities from the natural world following Bayesian principles (e.g., [15]).

There have been some attempts to develop general theories in scientific fields that can be considered as part of the constellation of SN. Twenty years before Cacioppo and Berntson's contribution [1], Joseph Bogen and Warren TenHouten coined the term "neurosociology" to refer to "a confluence of neurologic and sociologic observations" and, in particular, to describe a series of studies of sociocultural variations in performance of lateralized cognitive tests ([16], p. 49). From the perspective of neurosociological analysis, the emphasis is on "the social production of thought and the social determination of brain organization and brain

function" ([17], p. 10). Crucially, in several works [18, 19], TenHouten has explored a general affect-spectrum theory of emotions. In what follows, we will briefly present it here for illustrative purposes.

The affect-spectrum theory is rooted in Plutchik's [20] psychoevolutionary theory of basic emotions (see TenHouten, this volume). According to Plutchik, there are four fundamental problems of life facing a wide range of species. For each existential problem, there is a negative aspect, or danger, and a positive aspect, or opportunity. The first problem is temporality, which refers to the finite life span of creatures and to the cycle of life, reproduction, and death [19]. The inevitability of separation and loss is definitive of sadness, while the possibility of social integration and support is definitive of joy. The second basic life problem is identity, which concerns membership in social groups. The opposed primary emotions surrounding the notion of identity are acceptance (incorporating) and rejection or disgust (expelling). The third problem is hierarchy, which involves power, authority, status, and prestige. The struggle for dominance defines anger, while the acceptance of lower status defines fear. The fourth problem is territoriality, which includes not only geographical space but commodities and all kinds of symbolic capital [19]. The control of territory defines exploration, while the violation of one's boundaries implies surprise.

TenHouten ([18], p. 55) proposes that each of these four existential problems has evolved into Fiske's [21] four elementary forms of sociality. In this way, the positive pole of Plutchik's temporality can be generalized into what Fiske [21] calls communal sharing, a social relationship of equivalence based on solidarity, unity, and identification with the collectivity, especially with the kinship system. Secondly, Plutchik's identity can be generalized into what Fiske calls equality matching, an egalitarian social relationship between distinct and coequal people in which each person receives roughly an equal share, regardless of the community's needs. Thirdly, hierarchy can be linked to authority ranking, a social relationship in which, according to Fiske, the superiors command and control the production and distribution of goods. Fourthly, since territoriality has been broadened to include all form of possessions, it can be assimilated to Fiske's notion of market pricing, a social relationship based on reciprocal exchanges mediated by values and determined by a market system.

Given these generalizations, Tenhouten [18, 19] defines a *quaternio*, a dynamically related double polarity in which there is an affinity between communal sharing and equality matching, on the one side, and authority ranking and market pricing, on the other. TenHouten's affect-spectrum theory predicts the spectrum of the 36 primary and secondary emotions using a generalization of Plateau's law, to wit: $\Psi_{ij} = kR_j^m iR_j^m j / f\left(d_{ij}\right)$, in which Ψ is the predicted level of the emotion, the Rs are two of the valenced Fiskeian social relations, and f is a function of the distance between the social relations on the quaternion [18]. In this way, the emotional experience is viewed as the product of social relationships: for example, love, which Plutchik defines as joy plus acceptance, is predicted as a product of communal sharing and equality matching. The theory has had certain impact on the sociology of emotions [22]. However, bearing in mind that many other influential theories have

been advanced and developed in the field, the mainstream in SN has tended to adopt a bottom-up approach that emphasizes the search for the neural and molecular correlates of emotional states and social relationships (compare, e.g., the bottom-up treatment of love and bondedness reviewed in Sects. 2.3, 3.1, and 4).

Additional examples of general conceptual principles guiding experimental research within the bottom-up approach to SN include views on whether human empathy is to be understood as a cognitive or an emotional process [23] or the extent to which one can define a functionally segregated neural system dedicated to a given social phenomenon as a guide for human brain mapping strategies (see, e.g., [24]). The generality, relevance, and testability of such principles vary greatly, also depending on the line of research and their specific role within it. They nevertheless define the theoretical profile of the field, opening up a sort of theoretical vacuum where models must operate: a mediating role for models which can, on the one hand, help clearly present and interpret experimental data in the light of a given principle or theoretical framework and, on the other hand, help specify the empirical relevance of a given principle or theoretical framework in order to guide or define experimental designs and protocols. As we will shortly appreciate, this middle ground where most modeling work is to be found offers a wide range of modeling strategies to connect theoretical and experimental research as well as a broad repertoire of types of models, which philosophers of neuroscience have identified and characterized.

Some neuroscientific positions must be highlighted which reinforce the picture presented. The contention that the field of neuroscience lacks strong, widely held theoretical constructions has been voiced by several influential neuroscientists. Marder et al. [25] underscore the idiosyncratic role of theoretical models, considering this lack of solid, structuring theories. Stevens (cf. [26], p. 177), in a brief review, goes so far as denying that any theory up to this moment can be considered to have made any fundamental contribution to neurobiology. We have then further reasons to reaffirm the idea that, more than properly neuroscientific theories, theoretical principles are variously deployed to guide the construction of different kinds of models and the development of experiments.

A concomitant fact to be mentioned is the growing trend of model-based cognitive neuroscience [27, 28]. The concurrent use of cognitive modeling to guide and complement different experimental strategies to explore brain functioning (such as electrophysiological and neuroimaging techniques) is a recent attempt to find unifying approaches and to face the dispersion of existing models and the diverse and data-intensive experimental results typical of the field. Although the complementary use of cognitive models and typical neuroscience techniques is not necessarily new, there is a marked and explicit recognition of the need for integrative efforts of this kind (see Sect. 4).

While the first advances toward establishing a bottom-up neuroscientific approach to social phenomena can be traced already to the first half of the 1990s (e.g., [1]), the methodological difficulties in the case of SN were even more comprehensive and more pressing than the ones faced by contemporary cognitive neuroscientists. On the one hand, there were the expectable hurdles accompanying the

application of complex areas of research (such as social psychology and social theory) to an already novel group of disciplines. On the other hand, the precise delimitation of target psychological phenomena together with the early realization that the neural systems implied are generally largely distributed entailed additional difficulties for both experimental and theoretical researchers alike.

Now, very specific descriptions are thus applied to the above theoretical anchors, stemming from the different kinds of experimental results obtained. It could be argued that, especially in the case of social cognitive neuroscience, this diversity is somewhat limited by the extended inclination to work with human experimental subjects, in part for obvious reasons concerning the kinds of phenomena under study and in part due to the rapid transformation and increasing availability of neuroimaging technology (and very specially, fMRI). Nevertheless, the distance between theoretical prescriptions and experimental descriptions is still very large, and, as already mentioned, it is within this gap where models come in and are most useful.

In what follows, we will present and analyze the different kinds of models and the associated modeling strategies that have been identified in the philosophical literature. The main aim is to offer a comprehensive picture of the theoretical mosaic which comprises contemporary SN, within the restricted group of bottom-up approaches we are considering. This will offer an outlook of this model-intensive field, inasmuch as it can then be tied to relevant explanatory and integrative efforts. Both of these endeavors will in turn be examined in the two following sections.

2.3 Kinds of Neuroscientific Models

Before presenting the main types of models that philosophers of neuroscience have discussed, it can be useful to introduce some standard distinctions commonly used in the neuroscientific literature. Some of the concepts below may overlap with some of the more philosophically oriented categories we will consider, that is, cognitive models, computational models, mechanistic models, and dynamical models. These categories are thoroughly debated in terms of the explanatory and integrative dimensions of neuroscience, in general, and its subfields, in particular. Their neuroscientific counterpart, on the other hand, will provide us with a platform to draw comparisons from and with a more comprehensive picture of modeling in SN.

In the preface of their remarkable 2001 book, Dayan and Abbott present a seemingly exhaustive distinction between descriptive, mechanistic, and interpretative models. While this categorization was proposed in reference to theoretical neuroscience, it can be easily extended to other areas of neuroscience, including SN. Such types of models are presented in terms of the differential questions that drive their construction: what it is that a particular neural system does (descriptive models), how it is that it does it (mechanistic models), and why (interpretative models). This is a very useful and at the same time very broad tripartite distinction that is silent on

issues such as the level of description, complexity, theoretical commitment, or explanatory scope of the models.

As Dayan and Abbott (cf., [29], p. 1) state, descriptive models summarize large amounts of experimental data under descriptive purposes. Mechanistic models describe how neural systems operate on the basis of known anatomical and physiological features. Finally, interpretative models focus on the behavioral and cognitive relevance of different aspects of brain function to define the computational principles behind it: the already mentioned efficient coding principle, according to which neural activity is minimized in order to transmit information along a processing stream, is a very general principle that can be used to elaborate specific computational models of brain function.

A related distinction, which goes well beyond the field of neuroscience and has also been thoroughly discussed by general philosophers of science, is the distinction between phenomenological and theoretical models (see, e.g., [30]). First, descriptive models are inherently phenomenological, inasmuch as they aim at representing phenomena—where a phenomenon is a scientifically relevant set of general and relatively stable features of the world. Second, interpretative models are inherently theoretical, positing as they do functional and operational principles that neural systems allegedly embody. Third, mechanistic models are more complex in the sense that they can be partially phenomenological and partially theoretical: to the extent that a model purporting to describe a system's mode of operation incorporates some kind of theoretical entity or hidden mechanism (not an uncommon situation in SN, as well as in other areas of neuroscience), then it exceeds this classical distinction (see Northoff, as well as Aristegui, this volume).[1]

A final related distinction is the one between quantitative and qualitative models. While most properly neuroscientific models are quantitative, or can be precisely expressed through mathematical or computational means, SN has benefited from qualitative models deriving from social psychology and cognitive science. Generally, when a set of phenomena is poorly understood or when its research is still in its infancy, qualitative modeling can be a possible, fruitful starting place. On a similar note, a model's complexity or its level of built-in biological detail can vary widely, according to the level of knowledge achieved on a particular neurobiological structure or neural system and, importantly, on the modeling purposes at hand. As we have already alluded, simplifying assumptions do not always depend on mere lack of knowledge and can instead be deliberately implemented (see, e.g., [32]).

SN, as most other areas of neuroscience, presents a vast range of models, stemming from ideal models designed to contrast intuitions on a conceptual matter (e.g., is empathy a genuine neuroscientific phenomenon, whose neural bases can be identified and described?) to very detailed models of oxytocin's neural pathways implied

[1] It can be pointed out that Craver [31] draws a distinction, not between phenomenological and theoretical models but between phenomenological and explanatory models. Theoretical enrichment, Craver would suggest, isn't necessary nor sufficient for a model to be explanatory. Similarly, a phenomenological model may theoretically enrich the description of the explanandum, as can be the case of LISP-based computational models.

in regulatory behavior related to stress outbursts. What can be called the level of granularity of a given model is certainly a very relevant feature to its assessment and has strong connections to a model's explanatory and integrative power. Below, we consider the different kinds of neuroscientific models that populate the philosophical literature and illustrate them with examples from SN.

At least four general kinds of models have been recently discussed in the philosophy of neuroscience. Although SN models do not figure prominently, the rapid growth of the field during the last decade will most probably be accompanied by an increase in the associated philosophical interest. It is also important to mention that the peculiarity of SN mainly comes from its problem domain, that is, the universe of neural and behavioral phenomena of an inherently social nature, and in this sense the kinds of models generally sought for and developed are on a continuum with other areas of neuroscience directed to cognitive phenomena (with some caveats that we will consider). These are cognitive models, computational models, mechanistic models, and dynamical models. Let us consider them in turn.

Cognitive models, sometimes also called functional models,[2] aim fundamentally at the specification of the operational stages necessary for a given psychological capacity to be carried out. Weiskopf [33] has described these kinds of models in terms of their epistemic aims and the array of techniques adopted to elaborate them. The purpose of cognitive models is to single out the functional properties of the neural system responsible for the psychological capacity under study. In terms of the well-known tripartite distinction between levels of analysis of an information processing system [34], cognitive models work at what Marr called the theory of calculus or computational level: they portray the activity of the system as a projection from one kind of information into another kind, within a series of necessary steps. Models of this kind posit a sequence of representational states and processes, needed for the performance of that particular capacity:

> Specifying such a model involves specifying the set of representations (primitive and complex) that the system can employ, the relevant stock of operations, and the relevant resources available and how they interact with the operations. It also requires showing how they are organized to take the system from its inputs to its outputs in a way that implements the appropriate capacity ([33], p. 323).

Although at first sight one may think these are not properly neuroscientific models, this would be an understatement: within a top-down approach, they can be very important to dismiss idle theoretical avenues and to direct further experimental efforts.

To illustrate this first kind of neuroscientific model, consider an early model of face recognition proposed by Bruce and Young [35]. This, explicitly presented as a functional model, centers on the sort of information (what the authors call "information codes") that has to be generated and accessed in order to recognize a familiar face, on the different stages involved in this process, and their organization. Hinging

[2] It should be noted that Weiskopf [33] understands cognitive models as a subtype of functional models. For reasons of clarity and considering the present context, we preferred to conflate both concepts.

on a host of reaction-time experimental results, data on typical patterns of error, and neuropsychological studies, Bruce and Young's model make clear-cut distinction between information processing operations, such as facial speech analysis and directed visual processing, and functional components of a face recognition system, such as face recognition units and person identity nodes. As is typical in this sort of modeling efforts, they stress the sequential order of relevant operations, claiming, for instance, that visual recognition necessarily precedes access to person knowledge. As we already stated, this kind of modeling work is not at all trivial and, although it may dominate the earliest stages in the study of a given phenomenon or neural system, this need not always be the case, as can be seen in Decety's [36] model of empathy.

Computational models can be likened to Dayan and Abbott's interpretative models as well as understood in terms of Marr's second, algorithmic, level of analysis. Predictably, models of this kind are generally computationally implemented, as this allows for their precise description and their valuable involvement in simulation studies. The growth of computational neuroscience, also due to the increasing level of neurobiological detail built into the models, has led to a proliferation of computational models, also in the field of SN. These models aim at uncovering the computational principles that guide the operation of neural systems, understood as information processing devices. Under this assumption, it is believed that manipulating models implemented in a computer can shed light on neural function, on a theoretical but also on an experimental basis.

In general, what computational models try to specify are the rules that need to be followed in order to produce the specific input-output transformations thought to be necessary for the execution of a given psychological capacity. Part of this endeavor is concerned with defining the computational constraints that govern neural systems, such as defining the computational tractability of an information processing problem or establishing time-related limits to the processing capacity of a given system. Part of the appeal and rationale behind the booming efficient coding research program is precisely a specification of the minimal resources to be employed on different computational operations (see [37] for a careful assessment of the explanatory profile of this sort of minimal models). Clearly, this is partly theoretical work but also a much-needed effort to channel laboratory research by making testable predictions and refining experimental questions.

A case of direct computational interpretation of neural activity can be seen in Behrens et al. [38], a rich review of computational roles attributed to different brain areas thought to be responsible for reward-guided behavior. An interesting example is the case of reinforcement learning algorithms, which state that "future expectations should be updated by the product of the prediction error and the learning rate" ([38], p. 1160). Midbrain dopamine neurons, projecting to the ventral striatum, have been attributed not only the role of predicting expected reward but also that of quantifying the associated deviation in observed reward. Specific model parameters and relative deviations have then been experimentally tested by recording neuronal activity via electrophysiological and neuroimaging methods.

Mechanistic models are the most common in neuroscience, and SN is no exception in this regard. They have recently received an unprecedented degree of attention by philosophers of neuroscience, especially concerning the problem of scientific explanation (see Sect. 3). While they can be understood in terms of Marr's level of implementation, "mechanistic" philosophers have construed this kind of models as part of a whole program of research in the field. For our present purposes, it suffices to say that mechanistic models ideally aim at specifying the set of relevant component parts, features, activities, and organization of the system causally responsible for a given neural or behavioral phenomenon. The identification and specification of a mechanism's structure can be realized on different spatiotemporal levels of the brain's structure, as mechanisms are thought to be hierarchically organized (at least according to the most popular versions such as Carl Craver's or William Bechtel's).

To exemplify, consider available research on oxytocin's role in social phenomena. Oxytocin has been strongly linked to attachment and maternal behavior. Insel and Young [39] review a number of mainly animal studies from molecular, cellular, and systems approaches, which jointly specify oxytocin's contribution to this special kind of selective behavior between a mother and her offspring. The model the authors present follows oxytocin receptors' activity along different pathways and in different cortical and non-cortical brain areas, while also assigning specific functional roles to this activity, both neutrally (such as increasing the activity of noradrenaline cells in the brainstem) and behaviorally (such as decreasing aggressive behavior toward the offspring).

Finally, dynamical models have also been discussed in the philosophical literature. These models focus on the temporal properties of a previously defined system—usually through systems of differential equations—analyzed through mathematical tools derived from general frameworks such as dynamical systems theory and graph theory. Typically, the modeled systems' parameters span the agent's brain and body, as well as relevant features of the environment, meeting a general rejection of the common strategy of partitioning cognitive systems into dedicated components. The research led by Ezequiel Di Paolo on different facets of social behavior is an example of this kind of highly interactive modeling (see, e.g., [40]). In the case of SN, this sort of models is at the moment still in its infancy, of a mostly qualitative nature, and hinging almost exclusively on behavioral parameters. Still, there is a tendency, specially stemming from systems neuroscience to model high-order parameters for large-scale neural systems. How this will unfold for specifically social phenomena will probably be seen in the short term.

3 Explanation in Social Neuroscience

Having reviewed the heterogeneity of modeling practices in SN, we can turn now to the issue of when these models explain. Scientific explanation has been a widely debated subject in the philosophy of neuroscience [33, 37, 41, 42]. In this section, we first introduce some "unitary" perspectives about explanation in neuroscience.

Scientific models can be analyzed considering two main features: (a) their contents or truth-conditions and (b) their vehicles, formats, or representational bearers. Unitary approaches to explanation may hold that explanatory models in neuroscience share the same kind of vehicle, the same kind of content, or both. Examples of unitary approaches are the deductive-nomological model [43] and mechanistic explanation [41, 44, 45]. We argue that these approaches seem to be inappropriate considering the diversity of explanatory practices in SN. Thus, we advance a pluralistic account for model-based explanation in SN. According to explanatory pluralism (EP), models in SN may be explanatory even when they do not exhibit the same kind of representational format nor the same kind of truth-conditions. Explanation in neuroscience, and particularly in SN, requires that modelers evaluate and selectively emphasize different representational ideals to represent different kinds of (causal and/or noncausal) structures in the brain. We think that SN provides an excellent case study for the development of a pluralistic perspective on the explanatory strategies and ideals that partially shape neuroscientific practice.

Concerns about the nature of explanation have a long history in philosophy of science. The first systematic treatment of this subject is Hempel and Oppenheim's classic "Studies in the logic of explanation" [46]. In that paper, they introduce the "deductive-nomological" (DN) model of explanation. The DN model conceives scientific explanation as an inference in which a sentence describing some aspect of an explanandum phenomenon is inferred as a logical consequence from premises describing true laws of nature and information about the antecedent conditions. The key feature of DN explanations is the nomic expectability of the explanandum phenomenon in light of the laws of nature (and the antecedent conditions) described in the explanans.

Several authors have raised serious conceptual concerns about the DN model of explanation. Just to mention some of the main problems, the account does not provide clear criteria to distinguish between true laws and accidental generalizations; it cannot account for the characteristic asymmetry of explanations, and it cannot exclude as non-explanatory inferences based on mere nomic covariations see, [41, 47]. In conjunction with these problems, the DN account does not seem to be representative of the kind of explanations employed in some special, "fragile" sciences, such as biology, neuroscience, or psychology, in which the search for universal laws of nature is at least peripheral. Attending to this feature of special sciences, some authors have claimed that in these disciplines where general laws are scarce and theoretical approaches are not as consolidated as in physics, explanations may adopt a different style.

It has been claimed that explanations in neuroscience and other biological sciences frequently do not address why questions (inquiring on the general conditions that determine the production of the explanandum phenomenon), but rather how questions (concerning the particular way in which the target system, be it cognitive or neuronal, subserves a given higher-level capacity) [41, 42, 48]. In these cases, explanations do not need to exhibit a clear propositional format and may instead involve presenting a scientific model of the underlying local "mechanism" that produces the phenomenon [49].

A scientific model provides a mechanistic explanation of an explanandum phenomenon to the extent that it identifies some aspects of the mechanism responsible for the phenomenon. In particular, a mechanistic model explanation usually involves decomposing the target mechanism into its parts or constituent entities, the activities of those entities, and their organization. This process of decomposition is iterative; thus, the parts identified in a first stage can be further decomposed into subparts. As a result, mechanistic explanations span multiple levels of a mechanism [41, 50]. Finally, this kind of explanation has a local scope, that is to say, mechanistic models are developed for explaining a particular phenomenon and do not extend beyond it. Therefore, the generalizations obtained by this type of explanation are often characterized as limited in scope, mechanistically fragile, and historically contingent ([41], pp. 66–70; [51]).

3.1 The Plurality of Model-Based Explanation in Social Neuroscience

The EP approach we will develop here recognizes both a plurality of representational ideals that may shape explanatory models in neuroscience and a plurality of different kinds of structures (i.e., causal and noncausal) that may be represented by those models. Specifically, we propose that the explanatory heterogeneity of SN can be fruitfully approached by differentiating two main aspects of scientific model explanations: (a) *their content* or truth-conditions, i.e., the kind of structures in the world a model must effectively represent in order to be explanatory, and (b) *their vehicle* or representational format, which may be embodying different representational ideals, like precision or accuracy. Evidently, these two aspects are intimately related in scientific practice. Nevertheless, the claim we want to advance here is that the distinction between them can provide a good framework for analyzing and assessing the explanatory credentials of scientific models in neuroscience.

The *content* of a model explanation is the information it provides about the phenomenon. Depending on the kind and the extent of information it provides, a model may be considered an acceptable explanation. We identify two kinds of content that an explanation may provide about its target system, namely, causal/compositional or noncausal/structural information. On the one hand, scientific models may provide causal explanations by identifying relations of causal dependence, either etiological or constitutive, among the explanandum phenomenon, antecedent conditions, and/or features of the mechanism underlying the phenomenon. This kind of content allows scientists to manipulate and control both the phenomenon and its mechanism in quite precise ways [41, 52]. On the other hand, scientific models may provide noncausal information about the target system. This kind of information includes, for example, the exhibition of counterfactual dependence relations between the design features of the target system and abstract environmental constraints [53]. It could also include purely mathematical relations between empirical phenomena or

information about the topological structure of the system. These dependence relations cannot be considered causal, since they are not diachronic nor do they necessarily ground experimental interventions. Furthermore, these relations may not be altered by changing the mechanistic realization of the target system in substantive ways: they are robust [54, 55]. Note that the two kinds of explanatory information a model may provide are perfectly compatible, and both make an important contribution to a thorough understanding of the phenomenon of interest.

Turning now to the *vehicle* of explanation, it may be characterized as the representational bearer of the explanation, that is, the kind of representational structure by which the explanatory information is conveyed, for example, linguistic statements, schematic diagrams, computational simulations, and mathematical equations. These vehicles allow scientists to represent different aspects of the phenomenon of interest and its underlying "mechanism," that is, to represent the intended content of the model. The choice of one representational vehicle over another is guided by several different representational ideals [56], and often modelers are forced to choose a particular vehicle considering the trade-off between different ideals. This is not a novel notion: Levins [57] had already pointed out that modelers often consider the trade-off among at least three representational ideals that cannot be maximized simultaneously: precision, generality, and realism. This trade-off may force some modelers to prioritize the precision and realism of a particular model, for example, in detriment of its generality. Taking into account the differences among the above representational ideals, one of us [58] has advanced a distinction between a mechanistic style, in which modelers tend to privilege structural details and realism, and a functionalist style, in which the ideal of generality is emphasized. The moral is that modelers have to find a preferred balance between the different representational ideals, selecting the most appropriate vehicle for representing the content they are interested in.

Some representational ideals in neuroscience and elsewhere in science are precision, simplicity, and generality. The ideal of precision involves the maximization of the representation's level of detail, either of structural features, component entities and activities, or temporal and spatial features of the system. The ideal of simplicity refers to the search of a model that maximizes the intelligibility of the phenomenon under study and its underpinnings. In many cases, meeting the ideal of simplicity may require scientists to abstract the model from irrelevant details and introduce idealizations. Finally, the ideal of generality refers to the model's ability to be applied across several domains and extrapolated to different target systems. Again, these representational ideals are intimately related to the kind of explanatory information that is conveyed. The analysis we propose might just provide a more complete toolbox for disentangling the varieties of explanation in neuroscience and SN.

With this framework for the analysis of a model's explanatory virtues in place, we now examine some representative cases in SN to exemplify usefulness of this approach. In this direction, we get back to two of the cases presented in the previous section exemplifying different kinds of models: the role of oxytocin in attachment [39] and the mathematical model of reinforcement learning of different patterns of activity related to decision-making processes [38].

Different mechanistic models have addressed the role of oxytocin in attachment [39]. These models have a causal content that pinpoints to a neurobiological mechanism including oxytocin as a major component: the models additionally attempt to determine its activities. The representational structure in these cases often involves diagrams and is guided by the representational ideals of precision and simplicity. The main objective consists in detailing the neural circuits, the different molecular components involved, and their organization related to behavioral expressions of attachment. Here lies the precision ideal displayed by these models. At the same time, the causal structure related to attachment is abstracted from other causal processes and different changes that may be induced in front of different contextual situations. Here we can appreciate the ideal of simplicity followed.

Consider one of the models presented in Insel and Young's review: oxytocin and the bonding behavior that sheep show toward their lambs. The selective and permanent bond appreciated within the 2 h of parturition has been explained by a neurobiological model that posits that:

> Afferent stimulation through the spinal cord from vaginocervical dilation during parturition increases the activity of noradrenaline-containing cells in the brainstem which project to the paraventricular nucleus (PVN) in the hypothalamus as well as to the olfactory bulb. Stimulation of oxytocin cells in the PVN facilitates maternal behaviour through coordinated effects on several regions in which oxytocin increases GABA (γ-aminobutyric acid) and noradrenaline release. Oxytocin in the olfactory bulb and medial preoptic area reduces aggressive or aversive responses to newborn lambs. Oxytocin in the mediobasal hypothalamus inhibits post-partum estrus ([39], p. 2).

This brief extract illustrates how maternal attachment in sheep is explained by a model that identifies different components involved (e.g., oxytocin, noradrenaline), their activities (oxytocin increases GABA release), and their organization.

In another direction, a structural content may be identified in the reinforcement learning model proposed for explaining different social phenomena [38]. In this case, an abstract mathematical structure is employed for expressing the main nuclear organization responsible for different patterns of activity. This model has a mathematical representational bearer (even though it could be represented in computational structures as well), to which two representational ideals may be related: generality and simplicity. Specifically, Behrens et al. [38] show how the simple structure *"Vt + 1 = Vt + atdt"*, which includes expectations of future reward ($Vt + 1$), current expectations (Vt), and their discrepancy from the actual outcome that is experienced—the prediction error (dt)—could be related to different patterns of activity observed in decision-making processes. In this case, social phenomena and the activity identified in different brain areas related to them are not explained in terms of precise component activity of neurotransmitters but instead in a more abstract equation that may relate expectancies, previous experience, and reward independently of the specific neurobiological structures that are involved in these functions in different cases. The authors have emphasized that the characteristic abstractness of these formal models makes them suitable for relating information about different neural activities involved in complex social phenomena from different species. In their own terms: "Such a mathematical formalism defines explicit

mechanistic hypotheses about internal computations underlying regional brain activity, provides a framework in which to relate different types of activity and understand their contributions to behavior" ([38], p. 1160).

3.2 *An Evaluation of the Mechanistic Unitary Approach and Explanatory Pluralism in Social Neuroscience*

For some mechanist philosophers, the ideal of mechanistic precision is a universal constraint on the vehicles of explanation (e.g., [45]). In this sense, more detail is always better. This kind of mechanistic approach does not recognize the diversity of ideals that may guide different models nor the trade-off among different representational ideals that is present in many modeling scenarios [33, 58, 59]. Other mechanists endorse [45] the idea that the same target system in neuroscience may be represented by a multiplicity of scientific models, each of them emphasizing a different aspect of the mechanism by selectively emphasizing some representational ideals more than others [60, 61]. However, virtually all mechanist philosophers endorse some kind of unitary approach concerning the content of model-based explanation. According to content unitary perspective, a scientific model provides explanatory information only to the extent that it identifies causal dependence relations underlying the phenomenon of interest [45, 60]. This unitary stance about content implies that cognitive or computational models in cognitive neuroscience, as well as in SN, are just incomplete sketches of mechanisms and that purely dynamical models are mere phenomenal, not explanatory models. We reject content unitary perspectives about explanation in neuroscience and SN.

What is explanatory pluralism? A first claim that should be made is that admitting a plurality of vehicles and contents for model-based explanation in SN should not be equated to the assumption that "anything goes" in explanation or to "the advocacy of retaining all, possibly inconsistent, theories that emerge from a community of investigators" ([62], p. 85). On the contrary, we think that the representational virtues proposed to contribute to a model's explanatory power should be clearly stated. In this sense, a fine balance must be achieved between admitting a plurality of explanatory vehicles and contents and the indistinctive inclusion of any proposed model in the set of explanatory models.

A second issue that we should take into consideration is that the notion of EP has been defined in multiple ways by different authors [33, 37, 62, 63]. To clarify the particular approach we propose here, it is useful to differentiate among three ways in which EP has been defined, to wit: (1) EP about *explanatory levels*, (2) EP about *representational structures*, and (3) EP about *explanatory styles*.

EP about explanatory levels emphasizes the existence of explanations at different levels of entities or size scales, a claim that contrasts with ruthless reductionist perspectives about explanation in neuroscience, like the stance advocated by Bickle [64–68]. The main thesis of EP concerning levels is that in order to explain some

phenomenon, entities at different compositional levels or size scales must be relevant. These entities usually are studied from different disciplines or fields, and all these perspectives at different levels of organization should be considered. In addition, it is usually claimed that all perspectives from different levels are complementary to each other and must be ideally integrated. The kind of integration that is expected ranges from complete autonomy to smooth mechanistic integration (see Sect. 4).

EP about *representational structures* admits the possibility and desirability that different scientific representations successfully pick out the same target system, i.e., "the same system in neuroscience can be represented and modelled in a variety of different ways, depending on the particular purposes of the investigation" ([37], p. 148). This conception implies that different representational bearers might be used in perfectly solid explanations of a given phenomenon. Nevertheless, EP about vehicles remains silent about the kind of informational content the different models must convey in order to be explanatory. A philosopher may adopt a unitary stance about the content of explanation, for example, endorsing a causal conception about the contents of explanation and nevertheless admit a plurality of representational structures for representing causes (mathematical equations, computational simulations, visual schemata, etc.).

Finally, EP about *explanatory styles* embraces the idea that different styles of explanation or explanatory virtues should be admitted as providing legitimate explanations [63]. The late Wesley Salmon has suggested this kind of pluralism, when he affirmed that:

> [I]t might be better to list various explanatory virtues that scientific theories might possess, and to evaluate scientific theories in terms of them. Some theories might get high scores on some dimensions, but low scores on others (…) I have been discussing two virtues, one in terms of unification, the other in terms of exposing underlying mechanisms. Perhaps there are others that I have not considered. ([69], p. 20)

Considering that EP about levels or representational structures is not incompatible with unitary accounts about explanatory styles, we consider this third kind of pluralism the most accurate for discriminating between unitary and pluralistic accounts of explanations.

According to our approach, the three kinds of EP are compatible and, in fact, we endorse them all. The idea of a single scientific representation that describes the behavior of the entities that are relevant for a phenomenon at the most fundamental level, that meets all the representational ideals that are appreciated by modelers, and that captures all the causal and noncausal features of the target system is a philosopher's fiction that covers our eyes to the diversity of explanation in neuroscience [70]. Considering model-based explanation in physics, Cartwright [71] has proposed a similar "patchwork" metaphor, according to which different models would be needed to account for the phenomenon under study. In the same direction, Weisberg [56] has highlighted a kind of "idealization of multiple models" which scientists are forced to resort to when dealing with highly complex phenomena. The idea is that there is a variety of explanatory styles in neuroscience and SN, each of

them emphasizing different explanatory virtues (in Salmon's sense); as a result, different types of vehicles are used by modelers to and explain to convey causal and noncausal information about different aspects of the target system, thus making very different assumptions about it.

4 The Unity of Social Neuroscience

SN, in the particular strand we are considering here, is an interdisciplinary research program that studies the neurobiological (neuronal, endocrine, and immune) processes that enable social cognition and behavior [72, 73]. The advancement of this scientific field requires the collaboration of researchers from many distinct disciplines, such as cognitive neuroscience, neuropsychology, cognitive science, neuroendocrinology, cellular and molecular neuroscience, social psychology, economics, and political science (see Salles and Evers, this volume). Many neuroscientists and philosophers of neuroscience see theoretical *unity* as a preeminent goal of neuroscience in general and SN in particular [41, 74–76]. How is the unity of SN achieved? In this section, we review three philosophical models of the unity of neuroscience and assess their validity vis-à-vis the modeling and experimental practices aimed at explanation in SN.

The first philosophical model we consider posits that the process of unification proceeds via a kind of reduction in practice [64, 77, 78]. The common experimental technique that grounds this reduction in practice is to intervene causally at lower levels of biological organization (e.g., cellular and molecular levels) in animal models and then to track the specific effects of these interventions on behavior in widely accepted experimental protocols for the target phenomena ([78], p. 230). The empirical success of this reductive experimental technique motivates a "ruthless reductionist" stance, i.e., one according to which, if a class of cognitive phenomena depends upon some molecular mechanisms that can be tracked experimentally, then the research on those molecular mechanisms assumes a kind of methodological priority ([78], p. 232).

Bickle's preferred exemplar of this kind of ruthless reductive unification in social neuroscience is the experimental work on the molecular basis of social recognition memory consolidation in mice [79]. Social recognition memory consists in the ability to remember and recall information tied to particular conspecifics after an initial episode of interaction with them. A standard behavioral protocol aimed to operationalize the concept of social recognition memory is based on Thor and Holloway's [80] idea that, "in the laboratory, social memory can be assessed reliably by measuring the reduction in investigation time of a familiar partner relative to a novel conspecific" ([81], p. 202). Furthermore, social recognition memory is considered to be dependent on the hippocampus, and, as many other forms of hippocampal-dependent long-term memory consolidation, it may be dependent on the activation of cyclic adenosine monophosphate (cAMP) responsive-element binding (CREB) proteins, especially two of its isoforms, α and δ (p. 232). To test this possibility, Kogan et al.

[79] obtained CREB$^{\alpha\delta}$ mutant mice—mice that show no expression of CREB α and δ isoforms—and trained a group of these mutants and a group of wild-type mice in a modified version of Thor and Holloway's [80] behavioral protocol for social recognition. They found that mutant mice CREB$^{\alpha\delta}$ engaged in social investigation (e.g., sniffing) of a given mouse to the same extent after 24 h as they did upon an initial encounter with the same individual. They interpret this finding as implying CREB$^{\alpha\delta}$ mutant mice are impaired in their social recognition abilities and, therefore, that long-term social memory is dependent on CREB function.

The main problem with Bickle's ruthless reductionism is that he seems to think that reduction in practice justifies global reductive claims concerning the molecular basis of some general phenomenon exhibited by organisms in the world (e.g., the molecular basis of social recognition memory *tout court*). However, it is not clear that the particular intervention undertaken by Kogan, Frankland, and Silva directly explains the data observed by another researcher in another laboratory studying the same phenomena but through distinct experimental designs and protocols [82]. What the "intervene molecularly and track behaviorally" technique brings about are "local within-experimental-protocol reductions," and it is not at all clear how these within-lab reductions will converge toward a global reductive claim concerning a general cognitive phenomenon ([82], p. 518). Furthermore, there is the problem of extrapolation. Bickle [78] emphasizes that the same molecular mechanisms for social recognition obtain across a wide variety of different species, from *Drosophila* to *Aplysia*. However, there are species-specific differences that question the generalizability of results obtained in mice to nonhuman primates and human beings [82]. For example, while in most non-primate mammals, social information is encoded via olfactory or pheromonal signals, in human and other primates, individual recognition relies on visual or auditory cues ([77], p. 201). Correspondingly, there are interspecies differences in the brain areas involved in the formation of social recognition memory. These differences cannot be neglected and prevent the sheer elimination of higher-level analyses concerning brain mechanisms that may underlie social cognition and behavior.

The second philosophical model of the unity of neuroscience (and, arguably, SN) incorporates the non-reductive and multilevel character of explanation as a central feature of the account. According to Kaplan and Craver ([45], p. 268), neuroscience is especially interesting to philosophers of science, among other reasons, because it is an interdisciplinary research community that "exemplifies a form of scientific progress in the absence of an overarching paradigm" (cf. [83]). How is this integration possible? Mechanist philosophers claim that the unity of neuroscience is effective when researchers from different scientific fields collaborate to build multilevel mechanistic explanations ([41], p. 18; see also [60, 84]). The product of this collaboration is an "explanatory mosaic" in which distinct scientific models "contribute piecemeal to the construction of a complex and evidentially robust mechanistic explanation" ([41], p. 19). Mechanistic explanations, in this sense, are built from the accumulation of constraints from different fields on the space of possible mechanisms for a given phenomenon. A constraint is a piece of information that shapes the

boundaries of the space of possible mechanisms or changes the probability distribution over that space, i.e., the probability that some region of the space describes the actual mechanism. The constraints from different scientific fields are used, like the tiles of a mosaic, to shape the space of possible mechanisms provided by mechanistic research programs.

Embracing mechanistic integration as a working hypothesis, many mechanists accept that modeling strategies from different fields are autonomous to the extent that each of these fields is free to choose which phenomena to explain, which experimental designs to apply, which conceptual resources to adopt, and the precise way in which they are constrained by scientific evidence from adjacent fields [41, 60, 84]. Against Bickle, they claim there is no methodological preeminence of molecular approaches to target phenomena in neuroscience. In fact, the capability of scientific fields to contribute novel constraints to a mechanistic research program demands their relative autonomy: "Because different fields approach problems from different perspectives, using different assumptions and techniques, the evidence they provide makes mechanistic explanations robust" ([41], p. 231). The ideal of a mosaic unity of neuroscience is congenial with Cacioppo and Decety's ([75], p. 166) emphasis on multilevel analysis in SN, that is, the idea that SN "necessitates the integration of multiple levels, and the explication of the mechanisms that link phenomena across these levels."

An example of mechanistic integration in SN comes from research on oxytocin and arginine vasopressin (AVP) as components of the mechanism for pair bonding in monogamous rodents [85–88]. The term "monogamy" refers to a social organization in which each member of a mating pair displays selective affiliation and copulation, nest sharing, and typically biparental care of offspring [87]. Voles provide valuable animal models for comparative studies on the neurobiological mechanisms of pair bonding [89]. Prairie voles (*Microtus ochrogaster*) exhibit a monogamous organization, forming enduring pair bonds following mating. Montane (*Microtus montanus*) and meadow (*Microtus pennsylvanicus*) voles, in contrast, are nonmonogamous species. The experimental protocol that is used in the lab in order to operationalize the concept of pair-bond formation is the partner-preference test. The experimental design includes an apparatus consisting of three chambers connected by tubes. The subject is allowed to move freely throughout the apparatus, while the "partner" and a novel "stranger" are confined to their own chambers. Pair bonding is considered to be present when the subject spends more time with the partner compared to the stranger [87]. The nonapeptides oxytocin and AVP emerged as constitutively relevant components of the mechanism for intense social attachment in voles. While oxytocin seems to be more important in females, AVP is more important in males. Thus, infusion of oxytocin into the cerebral ventricles of female prairie voles facilitates pair bonding, while AVP infusion facilitates pair bonding in male prairie voles. Furthermore, administration of selective oxytocin receptor and AVP receptor 1a (V1aR) antagonists blocks each of these behaviors in females and males, respectively. Considerations from systems neuroscience and evidence from anatomical and pharmacological studies are also relevant to constrain the space of possible pair bonding formation mechanisms. Compared to nonmonogamous

species, female prairie voles have higher densities of oxytocin receptors in the prefrontal cortex and nucleus accumbens, while male prairie voles have higher densities of AVP receptors in the ventral pallidum, medial amygdala, and mediodorsal thalamus [88]. These studies indicate that the prefrontal cortex, nucleus accumbens, and ventral pallidum are critical brain regions involved in pair-bond formation. Since these areas are also involved in the mesolimbic dopamine reward system, some researchers have hypothesized that pair bonding may be the result of conditioned reward learning. In this model, "the reinforcing, hedonic properties of mating may become coupled with the olfactory signatures of the mate, resulting in a conditioned partner preference," much in the way drugs of abuse work ([85], p. 1052).

There are two problems affecting the mechanistic ideal of a mosaic unity of (social) neuroscience. According to one criticism, mechanistic integration is too demanding. As mentioned when assessing the ruthless reductive account, within any field in neuroscience (and social neuroscience is not an exception), there is a multiplicity of experimental protocols associated with the "same phenomenon," so it is not at all clear how results obtained from different laboratories, using different experimental protocols, can fit together within a field, before the combined results of that field can be said to set constraints on the space of possible mechanisms for a phenomenon ([82], p. 525). Furthermore, even if a researcher identifies a working part or activity in the mechanism of pair bonding in rodents, it may not be immediately clear that that piece of evidence will constrain the space of possible mechanisms for pair bonding in humans [82]. In this sense, Young and Wang ([85], p. 1052) strongly emphasize that "there are no hard data demonstrating common physiological mechanisms for pair-bond formation in voles and man" and that "the emergence of the neocortex and its ability to modify subcortical function cannot be ignored." The two facts just mentioned are closely related: the multiplicity of experimental protocols concerning a target phenomenon arises in part because the phenomenon itself varies in different species, involving different mechanisms in different species [82].

Moreover, the philosophical issue concerning the level of discontinuity between human and nonhuman minds becomes relevant at this point. Against the dominant tendency in comparative cognitive psychology, Penn, Holyoak, and Povinelli [90] defend the hypothesis that there is a significant functional discontinuity in the degree to which human and nonhuman animals are able to approximate higher-order, abstract, relational capabilities of a physical symbol system. According to their *relational reinterpretation hypothesis* ([90], p. 111), although both humans and nonhumans are capable of learning and acting on the perceptual relations between different aspects of the world, only humans are capable of reinterpreting those relations in a systematic and productive way. For these researchers, the functional discontinuity between human and nonhuman minds pervades nearly every domain of cognition. Particularly, only humans can master general concepts based on structural criteria (beyond any particular source of stimulus control), find systematic analogies between disparate domains, draw logical inferences between higher-order relations, or postulate unobservable mental causes or physical forces as explanations of natural phenomena ([90], p. 110). If they are right and nonhuman

minds approximate the capabilities of a physical symbol system to a significantly lesser degree than human minds do, then the prospects of reductionistic or mechanistic integration *across* species are dim.[3]

The second criticism to mechanistic integration points in the opposite direction. Recently, Levy [76] has argued that the mosaic ideal of unity is too minimal, i.e., a version of unity that is "overly modest and for that reason not very attractive." In particular, what the mosaic ideal of unity does not require is the existence of shared theoretical content among the constraints on the space of possible mechanisms for a target phenomenon, that is, "general concepts, principles and explanatory schemas applying across a range of neuroscientific phenomena" ([76], p. 10). Levy compares Craver's "tiles" in the mosaic unity of neuroscience to members of an alliance, i.e., independent states joining efforts. He encourages a stronger, "federal" ideal of unity, in which a set of distinct states are united by general principles. Noticeably, Bickle's ruthless reductive account eschews this problem, since the general principles that unify the different fields of neuroscience are the principles and laws of physics and chemistry that determine molecular and cellular processes within the brain, since "to the extent that we have explained some 'higher level' phenomenon as a sequence (…) of molecular steps, we know that the only way for another 'higher level' process to employ it (…) is via molecular (or lower) mechanisms" ([78], p. 232). The common principles Levy [76] has in mind are not Bickle's physicochemical principles but abstract, recurrent patterns that, according to some recent (and rather speculative) theoretical work in neuroscience, transcend spatial and temporal scales and apply to a range of neural systems. As an example, Levy mentions Sterling and Laughlin's [91] principles of efficient design that apply to the brain as a whole and to different regions at different temporal and spatial scales. One of such principles of neural design is to "minimize wire" (i.e., axon length), which explains, for instance, the placement of ganglia in *Caenorhabditis elegans* nervous system [92] and the organization of neurons in cortical maps in the mammalian visual cortex [93]. Design explanations of this kind allow us to answer why questions such as: Why are neurons in the mammalian visual cortex organized in maps? Or why are neural circuits separate in layers, columns, stripes, or barrels? ([91], p. 446).

There is a very popular research program in SN that aims to provide answers, from the designer perspective to the kind of why questions just mentioned. Given the extraordinary cost of neural material [94], Dunbar [95] asks: why do primates (in particular) have unusually large brains for body size, compared to all other vertebrates? Dunbar's preferred proposal is the social brain hypothesis. According to this hypothesis, large brains are a consequence of natural selection for enhanced social skills, since "an individual's fitness is maximized by how well the group solves the problems that directly affect fitness, and this in turn is a consequence of how well bonded it is (this in turn being a consequence of the individual member's social cognitive skills)" [95]. The social brain hypothesis points to the bondedness of social groups as the intermediate step between brain size and the selective pressures driving brain evolution [96].

[3] We thank Warren TenHouten for bringing this issue to our attention.

There are some direct counterexamples to the social brain hypothesis. Lemurs, for example, live in relatively big social groups but have relatively small brains. Some authors have raised deeper concerns about the social brain hypothesis. Cachel ([97], p. 373) contends that if the social brain hypothesis were valid, then we would expect that our closest primate relatives, i.e., the chimpanzee and the bonobo, would exhibit the most complex primate sociality. However, the only truly eusocial nonhuman primates are some New World monkeys, like the tamarins, which exhibit cooperation in the care of their young, reproductive division of labor, and overlap of two or more generations contributing to social life ([98], p. 62), and monkeys are in several respects less intelligent than pongids. Furthermore, according to Cachel, there is a trade-off between social intelligence and natural history intelligence, and only the latter constitutes a principal factor contributing to the formation of a general human-like intelligence. Competitive social behavior is highly demanding in terms of attention and other cognitive resources and also discourages exploration of the natural world. Vervet monkeys, for example, exhibit acute social awareness but are "peculiarly obtuse or stupid about making associations and predictions about the external world" ([97], p. 165). TenHouten states [97]: "Freedom from hypersociality is necessary for the development of complex, symbolic models of the world that can then be subjected to abstract cognition and executive-level decision-making."[4]

In this section, we use the social brain hypothesis merely as an example of an abstract design principle on brain architecture, without endorsing it as a working hypothesis. The rationale behind the social brain hypothesis can be further specified as follows: "Members of social species, by definition, create organizations beyond the individual. These super-organismal structures evolved hand in hand with psychological, neural, hormonal, cellular, and genetic mechanisms to support them" ([75], p. 163). From the standpoint of the social brain hypothesis thus formulated, SN is not a mere alliance of disciplines gathered by the common goal of explaining some target phenomena but represents a broad theoretical paradigm in neuroscience, "a general perspective that underlies a range of theories and methodologies in the field," which presupposes that many central aspects of brain organization and function only make sense in the light of social organization and vice versa ([75], pp. 162–163).

We have reviewed three philosophical accounts of the unity of neuroscience in general and SN in particular. First, SN may become integrated by molecular reductions of social behavior. The challenge for this reductive approach is to account for the existence of a multiplicity of experimental protocols for a given phenomenon, given the different manifestations of that phenomenon across different species. Second, SN may become integrated by the piecemeal accumulation of constraints from autonomous fields on the space of possible mechanisms for the target phenomenon. The challenge for the mechanistic account is twofold. On the one hand, mechanist philosophers have to explain how different results from different laboratories become integrated within a field and how they can be extrapolated from one species to another. On the other hand, the ideal of a mosaic unity may be too minimal, since it does not require the existence of shared theoretical content. In the third place, the

[4] We thank Warren Tenhouten for drawing our attention to these concerns.

unity of SN may be achieved by common general principles and concepts, such as the social brain hypothesis. However, the debate concerning the principles of design and evolution of the social brain is still open. Experimental and modeling practices in SN seem to be quite independent from the development of that theoretical debate. In the absence of general design principles, the multiplicity of experimental protocols becomes the main feature of SN as a laboratory science. A kind of non-reductive pluralism [33, 37, 82, 99] in which that multiplicity is not neglected seems to be the most sensible position concerning the unity of SN at this stage of development.

According to Sullivan ([82], p. 534), there are two fundamental constraints on the experimental process that account for to the multiplicity of experimental protocols in neuroscience: reliability and external validity. These two constraints pull in opposite directions ([82], p. 535). Reliability prescribes simplifying measures in order to keep control in the laboratory and discriminate between competing hypotheses about a laboratory effect. External validity prescribes building into the experimental design as much complexity as possible in order to capture the phenomenon of interest, outside the laboratory. Thus, there is a trade-off between reliability and external validity. This trade-off sheds light at least on some points of intersection of the neural and the social. As emphasized by Callard and Fitzgerald ([100], p. 60), the need for more ecologically valid models (particularly regarding the social environment) in animal research is one of many arenas that would benefit from greater interdisciplinary collaboration between neuroscientists and social scientists.

Consider, for example, the physiological and psychological effects on rodents of laboratory housing conditions [101]. Practically all laboratory-housed rodents live in small "shoe-box" cages which afford little meaningful biological complexity. Physiological and behavioral studies strongly indicate that social isolation is detrimental for rats and mice and that company can be enriching and beneficial. In rodents, usual laboratory conditions may cause impairments in the neural and behavioral development and behavioral stereotypies. Stereotypies are uncommon in free-living wild animals, and they may be caused by the frustration of natural behaviors like finding food or mates, building nests, and avoiding predators. Since animals with stereotypies are poor models of normal behavior, implementing social environmental enrichment is needed in order to regain external validity. In fact, researchers using more naturalistic housing methods have detected deficits in transgenic mice that had been neglected in conventional laboratories [102]. A pluralistic approach predicts that such an increment of external validity will imply an attenuation of experimental reliability and the negotiation of a new equilibrium point between these two constraints of the experimental design.

5 Conclusion

In this chapter we introduced three general philosophical issues stemming from recent and actual research in the field of SN. These issues have become increasingly prominent in the literature and prove highly relevant for the present and near future

of SN. In particular, the philosophy of modeling has been intensely debated generally in the philosophy of science. Here, we addressed the philosophical problem of how scientific models and theories relate, while characterizing the different kinds of models and modeling approaches relevant in contemporary SN. Secondly, we presented the issue of scientific explanation, certainly a hot topic in recent philosophy of neuroscience: it can be argued that this problem is responsible for a great deal of the boost philosophy of neuroscience had during the last two decades. The third issue, scientific integration, is, in our opinion, a much pressing topic specifically for SN. The ways different aspects of SN research can be articulated and put into fruitful dialogue, considering specially the characteristic nature of this ambitious neuroscientific approach to social phenomena, are in need of detailed philosophical attention and, we think, will certainly be soon increasingly debated within the philosophical community.

Although we made an effort to present the issues without taking clear-cut sides, we defended a general pluralistic stance toward SN. We started from a resolute recognition of the diversity of modeling approaches today being developed and of the epistemic roles that models can and do play in SN. We then proposed a kind of EP that admits that models tackling different levels, representational bearers, and styles of explanation may be considered legitimately explanatory. In fact, we defended the idea that this plurality is desirable in order to reconstruct the "patchwork" picture of such a complex field as is SN. It is important to highlight, though, that this idea is not equivalent to an "anything goes" principle, and we here suggested a clear framework that might be useful when analyzing the explanatory virtues of different models in SN.

Finally, we reviewed a central question concerning SN: how to best approach the unity of the field. A philosophical account of integration in SN requires an explication of the way in which different empirical results from different laboratories can become integrated within the field and how they can be extrapolated from one model species to another. We have argued that non-reductive pluralism is the most adequate approach to these problems concerning extrapolation and the multiplicity of experimental protocols.

References

1. Cacioppo JT, Berntson GG. Social psychological contributions to the decade of the brain. Doctrine of multilevel analysis. Am Psychol. 1992;47(8):1019–28.
2. Dunbar RIM. Neocortex size and group size in primates: a test of the hypothesis. J Hum Evol. 1995;28(3):287–96.
3. Dunbar RIM. The social brain hypothesis. Evol Anthropol Issues News Rev. 1998;6(5):178–90.
4. Bailer-Jones DM. Scientific models in philosophy of science. Pittsburgh, PA: University of Pittsburgh Press; 2009.
5. Morrison M. Where have all the theories gone? Philos Sci. 2007;74(2):195–228.
6. Cartwright N, Shomar T, Suárez M. The tool box of science: tools for the building of models with a superconductivity example. Poznan Stud Philos Sci Humanit. 1995;44:137–49.

7. Harre R. Cognitive science: a philosophical introduction: SAGE Publications; 2002. p. 344.
8. Morgan MS, Morrison M. Models as mediators: perspectives on natural and social science. Cambridge: Cambridge University Press; 1999. p. 420.
9. Venturelli N. Un abordaje epistemológico de la integración neurocientífica. In: Rodriguez V, Velasco M, editors. Epistemología y prácticas científicas. Córdoba: Editorial Universitaria; 2015. p. 41–71.
10. Churchland PS, Sejnowski TJ. The computational brain. Cambridge, MA: MIT Press; 1994. p. 564.
11. Hardcastle VG, Stewart CM. What do brain data really show? Philos Sci. 2002;69(3):572–82.
12. Hardcastle VG. The theoretical and methodological foundations of cognitive neuroscience. Philos Psychol Cogn Sci. 2007:295–311.
13. Tononi G, Sporns O, Edelman GM. A measure for brain complexity: relating functional segregation and integration in the nervous system. Proc Natl Acad Sci U S A. 1994;91(11):5033–7.
14. Newlands SD, Perachio AA. Compensation of horizontal canal related activity in the medial vestibular nucleus following unilateral labyrinth ablation in the decerebrate gerbil. II. Type II neurons. Exp Brain Res. 1990;82(2):373–83.
15. Lee TS, Mumford D. Hierarchical bayesian inference in the visual cortex. J Opt Soc Am A Opt Image Sci Vis. 2003;20(7):1434–48.
16. Bogen JE, DeZure R, Tenhouten WD, Marsh JF. The other side of the brain. IV. The A-P ratio. Bull Los Angel Neurol Soc. 1972;37(2):49–61.
17. TenHouten WD. Neurosociology. J Soc Evol Syst. 1997;20(1):7–37.
18. TenHouten WD. Explorations in neurosociological theory: from the spectrum of affect to time consciousness. Soc Perspect Emot. 1999;5:41–80.
19. TenHouten WD. A general theory of emotions and social life. New York: Routledge; 2006.
20. Plutchik R. A general psychoevolutionary theory of emotion. In: Emotion: theory, research, and experience, Theories of emotion, vol. 1. New York: Academic; 1980. p. 3–33.
21. Fiske AP. The four elementary forms of sociality: framework for a unified theory of social relations. Psychol Rev. 1992;99(4):689–723.
22. Franks DD. The neuroscience of emotions. In: Stets JE, Turner JH, editors. Handbook of the sociology of emotions, Handbooks of sociology and social research. New York: Springer; 2006. p. 38–62. Available from: http://link.springer.com/chapter/10.1007/978-0-387-30715-2_3.
23. Preston SD, de Waal FBM. Empathy: its ultimate and proximate bases. Behav Brain Sci. 2002;25(1):1–20.
24. Eickhoff SB, Laird AR, Fox PT, Bzdok D, Hensel L. Functional segregation of the human dorsomedial prefrontal cortex. Cereb Cortex. 2016;26(1):304–21.
25. Marder E, Kopell N, Sigvardt K. How computation aids in understanding biological networks. In: PSG S, Grillner S, Selverston AI, Stuart DG, editors. Neurons, networks, and motor behavior. Cambridge, MA: MIT Press; 1997.
26. Stevens CF. Models are common; good theories are scarce. Nat Neurosci. 2000;3:1177.
27. Forstmann BU, Wagenmakers E-J, Eichele T, Brown S, Serences JT. Reciprocal relations between cognitive neuroscience and formal cognitive models: opposites attract? Trends Cogn Sci. 2011;15(6):272–9.
28. Palmeri TJ, Love BC, Turner BM. Model-based cognitive neuroscience. J Math Psychol. 2017;76:59–64.
29. Dayan P, Abbott LF. Theoretical neuroscience: computational and mathematical modeling of neural systems. Cambridge, MA: MIT Press; 2001. p. 576.
30. Frigg R, Hartmann S. Scientific models. In: Sarkar S, Pfeifer J, editors. The philosophy of science: an encyclopedia. New York: Routledge; 2006. p. 740–9.
31. Craver CF. When mechanistic models explain. Synthese. 2006;153(3):355–76.
32. Kronhaus DM, Eglen SJ. The role of simplifying models in neuroscience: modelling structure and function. In: Bio-inspired computing and communication. Berlin, Heidelberg: Springer; 2008. p. 33–44.

33. Weiskopf DA. Models and mechanisms in psychological explanation. Synthese. 2011;183(3):313.
34. Marr D. Vision: a computational investigation into the human representation and processing of visual information. San Francisco: W. H. Freeman and company; 1982. p. 432.
35. Bruce V, Young A. Understanding face recognition. Br J Psychol Lond Engl. 1986;77(Pt 3):305–27.
36. Decety J. A social cognitive neuroscience model of human empathy. In: Harmon-Jones E, Winkielman P, editors. Social neuroscience: integrating biological and psychological explanations of social behavior. New York: Guilford Press; 2007.
37. Chirimuuta M. Minimal models and canonical neural computations: the distinctness of computational explanation in neuroscience. Synthese. 2014;191(2):127–53.
38. Behrens TEJ, Hunt LT, Rushworth MFS. The computation of social behavior. Science. 2009;324(5931):1160–4.
39. Insel TR, Young LJ. The neurobiology of attachment. Nat Rev. Neurosci. 2001;2(2):129–36.
40. Di Paolo E, De Jaegher H. The interactive brain hypothesis. Front Hum Neurosci. 2012;6:163.
41. Craver CF. Explaining the brain: mechanisms and the mosaic unity of neuroscience. Oxford: Oxford University Press; 2007. p. 328.
42. Bechtel W. Mental mechanisms: philosophical perspectives on cognitive neuroscience. New York: Psychology Press; 2008. p. 322.
43. Hempel C. Aspects of scientific explanation and other essays in the philosophy of science. New York: The Free Press; 1965.
44. Wright CD, Bechtel W. Mechanisms and psychological explanation. In: Thagard P, editor. Philosophy of psychology and cognitive science. Amsterdam: Elsevier; 2007.
45. Kaplan DM, Craver CF. The explanatory force of dynamical and mathematical models in neuroscience: a mechanistic perspective. Philos Sci. 2011;78(4):601–27.
46. Hempel CG, Oppenheim P. Studies in the logic of explanation. Philos Sci. 1948;15(2):135–75.
47. Salmon WC. Four decades of scientific explanation. 1st ed. Pittsburgh: University of Pittsburgh Press; 1984. p. 240.
48. Cummins R. "How does it work" versus "what are the laws?": two conceptions of psychological explanation. In: Keil F, Wilson RA, editors. Explanation and cognition. Cambridge, MA: MIT Press; 2000. p. 117–45.
49. Bechtel W, Abrahamsen A. Explanation: a mechanist alternative. Stud Hist Phil Biol Biomed Sci. 2005;36(2):421–41.
50. Craver CF. Levels [Internet]. In: Open MIND. Frankfurt am Main: MIND Group; 2015. [cited 25 Dec 2016]. Available from: http://open-mind.net/papers/levels/getAbstract.
51. Illari PM, Williamson J. Mechanisms are real and local. New York: Oxford University Press; 2011.
52. Woodward J. Making things happen. New York: Oxford University Press; 2003.
53. Wouters AG. Design explanation: determining the constraints on what can be alive. Erkenntnis. 2007;67(1):65–80.
54. Irvine E. Models, robustness, and non-causal explanation: a foray into cognitive science and biology. Synthese. 2014:1–17.
55. Ross LN. Dynamical models and explanation in neuroscience. Philos Sci. 2015;81(1):32–54.
56. Weisberg M. Simulation and similarity: using models to understand the world. New York: Oxford University Press; 2013. p. 211.
57. Levins R. The strategy of model building in population biology. Am Sci. 1966;54(4):421–31.
58. Barberis SD. Functional analyses, mechanistic explanations and explanatory tradeoffs. J Cogn Sci. 2013;14(3):229–51.
59. Nolen S. In defense of dynamical explanation. Philosophical Theses [Internet]. 2013. Available from: http://scholarworks.gsu.edu/philosophy_theses/143.
60. Boone W, Piccinini G. The cognitive neuroscience revolution. Synthese. 2016;193(5):1509–34.
61. Levy A, Bechtel W. Abstraction and the organization of mechanisms. Philos Sci. 2013;80(2):241–61.

62. Mitchell SD. Why integrative pluralism? ECO Spec Double Issue. 2004;6(1–2):81–91.
63. Mantzavinos C. Explanatory pluralism. Cambridge: Cambridge University Press; 2016. p. 237.
64. Bickle J. Reducing mind to molecular pathways: explicating the reductionism implicit in current cellular and molecular neuroscience. Synthese. 2006;151(3):411–34.
65. Abney DH, Dale R, Yoshimi J, Kello CT, Tylén K, Fusaroli R. Joint perceptual decision-making: a case study in explanatory pluralism. Front Psychol. 2014;5:330.
66. Bouwel JV. Pluralists about pluralism? Different versions of explanatory pluralism in psychiatry. In: Galavotti MC, Dieks D, Gonzalez WJ, Hartmann S, Uebel T, Weber M, editors. New directions in the philosophy of science, The philosophy of science in a european perspective. Berlin: Springer; 2014. p. 105–19.
67. Gijsbers V. Explanatory pluralism and the (dis)unity of science: the argument from incompatible counterfactual consequences. Front Psych. 2016;7:32.
68. McCauley RN, Bechtel W. Explanatory pluralism and heuristic identity theory. Theory Psychol. 2001;11(6):736–60.
69. Salmon WC. Scientific explanation: causation and unification. Critica. 1990;22(66):3–23.
70. Venturelli AN. A cautionary contribution to the philosophy of explanation in the cognitive neurosciences. Mind Mach. 2016;26(3):259–85.
71. Cartwright N. The dappled world: a study of the boundaries of science. Cambridge: Cambridge University Press; 1999. p. 264.
72. Cacioppo JT, Berntson GG. Social neuroscience. In: Cacioppo JT, Berntson GG, Adolph R, Carter CS, Davidson RJ, McClintock MK, et al., editors. Foundations in social neuroscience. Cambridge, MA: MIT Press; 2002. p. 3–7.
73. Harmon-Jones E, Winkielman P, editors. A social cognitive neuroscience model of human empathy. New York: Guilford Press; 2007.
74. Bechtel W, Hamilton A. Reduction, integration, and the unity of science: natural, behavioral, and social sciences and the humanities. In: Kuipers T, editor. Philosophy of science: focal issues, Handbook of the philosophy of science, vol. 1. Amsterdam: Elsevier; 2007.
75. Cacioppo JT, Decety J. Social neuroscience: challenges and opportunities in the study of complex behavior. Ann N Y Acad Sci. 2011;1224:162–73.
76. Levy A. The unity of neuroscience: a flat view. Synthese. 2016;193(12):3843–63.
77. Bickle J. Philosophy and neuroscience: a ruthlessly reductive account. Dordrecht: Springer; 2003. p. 235.
78. Bickle J. Ruthless reductionism and social cognition. J Physiol Paris. 2007;101(4–6):230–5.
79. Kogan JH, Frankland PW, Silva AJ. Long-term memory underlying hippocampus-dependent social recognition in mice. Hippocampus. 2000;10(1):47–56.
80. Thor D, Holloway W. Social memory of the male laboratory rat. J Comp Physiol Psychol. 1982;96(6):1000–6.
81. Ferguson JN, Young LJ, Insel TR. The neuroendocrine basis of social recognition. Front Neuroendocrinol. 2002;23(2):200–24.
82. Sullivan JA. The multiplicity of experimental protocols: a challenge to reductionist and non-reductionist models of the unity of neuroscience. Synthese. 2009;167(3):511–39.
83. Kuhn TS. The structure of scientific revolutions. Chicago: University of Chicago Press; 1970. p. 228.
84. Piccinini G, Craver CF. Integrating psychology and neuroscience: functional analyses as mechanism sketches. Synthese. 2011;183(3):283–311.
85. Young LJ, Wang Z. The neurobiology of pair bonding. Nat Neurosci. 2004;7(10):1048–54.
86. Donaldson ZR, Young LJ. Oxytocin, vasopressin, and the neurogenetics of sociality. Science. 2008;322(5903):900–4.
87. Johnson ZV, Young LJ. Neurobiological mechanisms of social attachment and pair bonding. Curr Opin Behav Sci. 2015;3:38–44.
88. Insel TR. The challenge of translation in social neuroscience: a review of oxytocin, vasopressin, and affiliative behavior. Neuron. 2010;65(6):768–79.

89. Carter CS, DeVries AC, Getz LL. Physiological substrates of mammalian monogamy: the prairie vole model. Neurosci Biobehav Rev. 1995;19(2):303–14.
90. Penn DC, Holyoak KJ, Povinelli DJ. Darwin's Mistake: explaining the discontinuity between human and nonhuman minds. Behav Brain Sci. 2008;31(2):109–30.
91. Sterling P, Laughlin S. Principles of neural design. Cambridge, MA: MIT Press; 2015.
92. Cherniak C, Mokhtarzada Z, Rodriguez-Esteban R, Changizi K. Global optimization of cerebral cortex layout. Proc Natl Acad Sci U S A. 2004;101(4):1081–6.
93. Chklovskii DB, Koulakov AA. Maps in the brain: what can we learn from them? Annu Rev. Neurosci. 2004;27:369–92.
94. Aiello LC, Wheeler P. The expensive-tissue hypothesis: the brain and the digestive system in human and primate evolution. Curr Anthropol. 1995;36(2):199–221.
95. Dunbar RIM. Evolutionary basis of the social brain. In: Oxford handbook of social neuroscience. Oxford: Oxford University Press; 2011. p. 28–38.
96. Dunbar RIM, Shultz S. Evolution in the social brain. Science. 2007;317(5843):1344–7.
97. Cachel S. Primate and human evolution. Cambridge, UK: Cambridge University Press; 2006. p. 488.
98. TenHouten WD. Emotion and reason: mind, brain, and the social domains of work and love. New York: Routledge; 2013. p. 279.
99. Mitchell SD, Dietrich MR. Integration without unification: an argument for pluralism in the biological sciences. Am Nat. 2006;168(Suppl 6):S73–9.
100. Callard F, Fitzgerald D. Rethinking interdisciplinarity across the social sciences and neurosciences. Basingstoke, UK: Palgrave Macmillan; 2015. (Wellcome Trust–Funded Monographs and Book Chapters)
101. Balcombe JP. Laboratory environments and rodents' behavioural needs: a review. Lab Anim. 2006;40(3):217–35.
102. Vyssotski AL, Dell'Omo G, Poletaeva II, Vyssotsk DL, Minichiello L, Klein R, et al. Long-term monitoring of hippocampus-dependent behavior in naturalistic settings: mutant mice lacking neurotrophin receptor TrkB in the forebrain show spatial learning but impaired behavioral flexibility. Hippocampus. 2002;12(1):27–38.