

Evidence for reductionist or anti-reductionist approaches of mental processing

Edited by

Francesca Strappini, Mark Couch, Antonino Carcione
and Marialuisa Martelli

Published in

Frontiers in Psychology
Frontiers in Psychiatry
Frontiers in Neuroscience
Frontiers in Integrative Neuroscience



FRONTIERS EBOOK COPYRIGHT STATEMENT

The copyright in the text of individual articles in this ebook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this ebook is the property of Frontiers.

Each article within this ebook, and the ebook itself, are published under the most recent version of the Creative Commons CC-BY licence. The version current at the date of publication of this ebook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or ebook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714
ISBN 978-2-8325-3762-6
DOI 10.3389/978-2-8325-3762-6

About Frontiers

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

Frontiers journal series

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

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Evidence for reductionist or anti-reductionist approaches of mental processing

Topic editors

Francesca Strappini — University of Bologna, Italy

Mark Couch — Seton Hall University, United States

Antonino Carcione — Terzo Centro di Psicoterapia, Italy

Marialuisa Martelli — Sapienza University of Rome, Italy

Citation

Strappini, F., Couch, M., Carcione, A., Martelli, M., eds. (2024). *Evidence for reductionist or anti-reductionist approaches of mental processing*.

Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-3762-6

Table of contents

- 04 **Reductionistic Explanations of Cognitive Information Processing: Bottoming Out in Neurochemistry**
William Bechtel
- 16 **New research tools suggest a “levels-less” image of the behaving organism and dissolution of the reduction vs. anti-reduction dispute**
John Bickle, André F. De Sousa and Alcino J. Silva
- 28 **The multiple realization of human color vision revisited**
Ken Aizawa
- 38 **From analytic to synthetic-organizational pluralisms: A pluralistic enactive psychiatry**
Christophe Gauld, Kristopher Nielsen, Manon Job, Hugo Bottemanne and Guillaume Dumas
- 50 **Unhealthy mind in a healthy body: A criticism to eliminativism in psychopathology**
Francesco Mancini, Alessandra Mancini and Cristiano Castelfranchi
- 60 **The hard problem of consciousness—A perspective from holistic philosophy**
Jicheng Chen and Linlin Chen
- 79 **Mechanistic decomposition and reduction in complex, context-sensitive systems**
Daniel C. Burnston
- 91 **The concept of Narcissistic Personality Disorder—Three levels of analysis for interdisciplinary integration**
Kerrin A. Jacobs
- 103 **Neurobiological reduction: From cellular explanations of behavior to interventions**
David Parker
- 123 **Clarifying the relation between mechanistic explanations and reductionism**
Mark Couch



Reductionistic Explanations of Cognitive Information Processing: Bottoming Out in Neurochemistry

William Bechtel*

Department of Philosophy, University of California, San Diego, San Diego, CA, United States

A common motivation for engaging in reductionistic research is to ground explanations in the most basic processes operative in the mechanism responsible for the phenomenon to be explained. I argue for a different motivation—directing inquiry to the level of organization at which the components of a mechanism enable the work that results in the phenomenon. In the context of reductionistic accounts of cognitive information processing I argue that this requires going down to a level that is largely overlooked in these discussions, that of chemistry. In discussions of cognitive information processing, the brain is often viewed as essentially an electrical switching system and many theorists treat electrical switching as the level at which mechanistic explanations should bottom out. I argue, drawing on examples of peptidergic and monoaminergic neurons, that how information is processed is determined by the specific chemical reactions occurring in individual neurons. Accordingly, mechanistic explanations of cognitive information processing need to take into account the chemical reactions involved.

Keywords: mechanistic explanation, reduction, control mechanisms, neuropeptides, monoamines

INTRODUCTION

Where should reduction stop? Traditional philosophical accounts of reduction (Nagel, 1961) argue for stopping with the fundamental laws of nature. On these accounts, in a successful reduction, characterizations of higher-level phenomena such as cognitive information processing are derived from these basic laws. New mechanists in philosophy of science challenged the need to invoke laws in explanations in the life sciences and instead argue that explanations often take the form of characterizing the mechanism responsible for the phenomenon being explained (Machamer et al., 2000; Bechtel and Abrahamsen, 2005). These explanations are still reductionistic insofar as they decompose mechanisms into their component entities and activities and appeal to them to explain the phenomenon.¹ Given the compositional nature of mechanisms, these components can

OPEN ACCESS

Edited by:

Mark Couch,

Seton Hall University, United States

Reviewed by:

John Bickle,

Mississippi State University,

United States

David Parker,

University of Cambridge,

United Kingdom

*Correspondence:

William Bechtel

wbechtel@ucsd.edu

Received: 18 May 2022

Accepted: 13 June 2022

Published: 04 July 2022

Citation:

Bechtel W (2022) Reductionistic Explanations of Cognitive Information Processing: Bottoming Out in Neurochemistry. *Front. Integr. Neurosci.* 16:944303. doi: 10.3389/fnint.2022.944303

¹They also appeal to how these parts are organized and how the whole mechanism is situated. Different ways of organizing components, and different ways of embedding them in a larger context, can result in producing different phenomena from the same parts. Insofar as they recognize the importance of organization and the situatedness of mechanisms (Bechtel, 2009), mechanistic explanations are not solely reductionistic and do not fit the mode of Bickle's (2003, 2006) account of ruthless reduction. Nonetheless, appeal to components is a fundamentally reductionistic feature of mechanistic explanations.

be understood as at a lower level than the mechanism.² While not involving iterated derivations, mechanistic explanations often involve iterated decompositions of entities into components; accordingly, mechanistic explanations can involve multiple descents to lower levels.

On the mechanistic account, the question of where reduction should stop becomes: how many times should one iterate the process of decomposition? Machamer et al. (2000) speak of explanations bottoming out; according to them, the level at which mechanistic explanations bottom out depends on the interests and resources of the investigators. While not denying that explanatory interests are crucial in directing mechanistic inquiry, I argue that there is a principled basis for identifying the level at which mechanistic explanations should bottom out: they should bottom out at the level at which the specific kinds of work that are being performed account for the features of the phenomenon being explained. In the case of cognitive information processing mechanisms, the phenomenon involves the control or regulation of other mechanisms. As I develop in section “Control mechanisms: modifying constraints in controlled mechanisms,” the work that is required to perform control activities involves enabling relevant information to determine the internal constitution of the control mechanism so that it modifies the components of the controlled mechanisms, thereby determining how they operate.

Drawing upon this understanding of control mechanisms, I will argue for a conclusion that will be surprising to many researchers in cognitive science and cognitive and systems neurosciences (It will not, however, be surprising the researchers engaged in cell and molecular research in neuroscience whose research I have drawn on in what follows).³ The work that is performed in the nervous system when organisms process information so as to control their activities is not electrical but chemical in nature: it involves the chemical processes through which neurotransmitters (often several) are synthesized in one neuron, released from it, and responded to, often in multiple ways, by other neurons. It is, accordingly, with these various chemical reactions that neuroscientific explanations of cognitive information processing should bottom out.

The importance of the chemical work involved in cognitive information processing is often concealed by a perspective in which synapses are understood on the model of electrical switches. As I will develop in section “The war of the soups

and the sparks: who won?,” this perspective has deep roots in the history of neuroscience. Initially many neuroscientists resisted the contention that communication between neurons was chemical, insisting that it was a purely electrical process. Even when the “war of the soups and the sparks” (Valenstein, 2005) ended with the acceptance by the sparks that transmission was chemical, many neuroscientists continued to view neurons as much like electrical switches, with all neurons processing information in essentially the same way. This perspective is reflected in recent work in connectomics and in accounts of artificial neural networks. Connectome maps (Sporns, 2011, 2012, 2015) emphasize structural connections between neurons, and even when they appeal to functional connectivity, they do not address the chemistry through which neurons interact. Brezina (2010), Bargmann (2012), and Nusbaum et al. (2017), among others, have argued for the limitations of connectome maps that fail to take into account the richness of the chemical processes through which neurons interact. In artificial neural network research, the individual nodes are each viewed as summing incoming electrical activity and, based on the sum, initiating a response in the recipient neuron. To explain the processing when neural networks are differently trained, researchers appeal primarily to the weighted connections between neurons (Goodfellow et al., 2016; Aggarwal, 2018). If this reflected how information is processed in our nervous system, neuroscientific explanations could bottom-out with a characterization of how neurons are connected into networks. I will argue, however, that such a model of electrical switching mischaracterizes how the brain processes information. The critical work involved in processing information is performed through the chemical processes through which individual neurons alter their behavior, including their actions on other neurons, in response to specific chemical signals received on their receptors. Critically, these processes are of many different types. These processes provide for a much richer repertoire of ways of processing information than have figured in accounts that construe the brain as processing information through electrical switching.

To a first approximation, the electrical switch model applies to neurons insofar as they communicate through the release at synapses of amino-acid-based neurotransmitters, such as glutamate or GABA, which act on ionotropic receptors (receptors that modify ion channels) in the postsynaptic neuron, altering ion flow across the neuronal membrane and generating a current along it. But this is only one type of transmission between neurons. Even in this case, there are often multiple types of ionotropic receptors that are associated with different channels and produce different postsynaptic currents. Moreover, what current they generate depends not just on the receptor but also current electrical activity and electrochemical gradients in that neuron. I will not develop this, but it further supports the contention that attending to the specific chemical processes through which transmitters are processed in postsynaptic cells is important to understanding neural and cognitive activity. To demonstrate the need to ground cognitive information processing accounts in chemistry I will focus on information processing that involves the release and response to two other types of neurotransmitters, neuropeptides

²Mechanistic levels are defined relative to the mechanism that is being decomposed. A mechanism may be decomposed into components of vastly varying sizes, which nonetheless interact to produce the phenomenon and so are then denizens of the same level. Decomposing a different mechanism may combine in one level entities at a different level in another mechanism. Accordingly, as argued by Craver and Bechtel (2007), mechanistic accounts do not assume levels that span the natural world.

³The balkanization of science is often much lamented and the failure of many cognitive and systems-level neuroscientists to draw upon the results of decades of research on chemical signaling in the brain is another example. Often pursuing one's research in ignorance of potentially relevant contributions of others allows for important advances. This has certainly been the case in cognitive science and cognitive and systems neuroscience. Yet, as I will argue, what is overlooked is extremely important for the understanding one is trying to develop and has the potential to amend and enrich our understanding of how brains contribute to cognition and control behavior.

and monoamines, characterizing what is distinctive about the information processing activities in which these transmitters participate.

In section “Information processing with neuropeptides” I focus on information processing relying on peptidergic transmitters. Peptidergic transmitters are employed throughout the brain. One brain region in which they are especially important is the hypothalamus; accordingly, I focus on it as an example. Among other sources and targets, hypothalamic neurons receive inputs from and send outputs to the endocrine system and can be viewed as an extension of it. Unlike amino-acid-based neurotransmitters, neuropeptides are not restricted to the synapse but, like hormones, are disseminated widely and are responded to by whichever cells have appropriate receptors. In most cases, these receptors are metabotropic—they initiate a wide variety of metabolic activities, including gene expression, in the recipient neuron. As a result, the signal is not just an activator or inhibitor of electrical signaling in the recipient cell—what information is processed depends on the peptide synthesized, the receptors that respond to it, the chemical state of the postsynaptic neurons, and the metabolic activities initiated in response. This provides a much richer range of information processing activities than envisaged with the electric switch model.

In section “Information processing with monoamines” I turn to another group of neurotransmitters, monoamines such as dopamine and serotonin. These transmitters are synthesized only in neurons in a limited set of nuclei but are distributed very widely in the brain. In invertebrate research, they were characterized as *neuromodulators* as they were shown to modify how information is processed in local circuits whose pattern of connectivity was not altered. This demonstrated that connectivity alone does not determine how a circuit processes information; it depends on which modulators are bathing the circuit. Insofar as they are released in response to global information and determine the processing in circuits to which they project, they can be viewed as setting the agendas for information processing at classically characterized synapses.

In sections “Information processing with neuropeptides” and “Information processing with monoamines” I will, for the most part, focus on the action of individual neuropeptides and monoamines, but that itself is a serious oversimplification. Nearly fifty years ago some neuroscientists drew attention to the fact that some neurons release multiple transmitters (Burnstock, 1980). Co-transmission is now recognized as the rule, not the exception (Burnstock, 2004; van den Pol and Anthony, 2012; Nusbaum et al., 2017; Svensson et al., 2019). Drawing upon investigations of the feeding circuit in *Aplysia*, Brezina (2010) has shown that the interactions of multiple transmitters are often non-linear. As a result, when released together two or more transmitters may produce an effect that none of them alone produces. Even without developing these complications, the description of the information processing activities involving neuropeptides and monoamines presented in sections “Information processing with neuropeptides” and “Information processing with monoamines” reveals that the brain employs a wide variety of different modes of information processing. It is not limited to or even well characterized in terms of the activities exhibited by electrical

switches. Accordingly, as I further develop in the final section, chemical processing between neurons is the appropriate level to bottom out reductionist accounts of the mechanisms of neural information processing.

CONTROL MECHANISMS: MODIFYING CONSTRAINTS IN CONTROLLED MECHANISMS

The standard accounts of mechanisms advanced by the new mechanists in philosophy of science (Machamer et al., 2000; Bechtel and Abrahamsen, 2005) characterize them in terms of their parts, operations, and how these are organized inside mechanisms, not how mechanisms are controlled by external processes. Such control, however, is required if the mechanisms responsible for the core activities of an organism (e.g., contraction of muscles, secretion from glands, synthesis and repair of bodies parts) are to carry out this work⁴ when and only when those phenomena are needed. If these mechanisms are allowed to generate their phenomena (e.g., a muscle is allowed to contract) whenever resources are available, the result is, at best, wasted resources and, worse, generation of phenomena in circumstances in which they are actually harmful to the organism. What cognitive information processing mechanisms do is control other mechanisms.⁵ They do this by performing work on the components of these other mechanisms so that they operate as appropriate on different occasions. Just as with other mechanisms, the work control mechanisms perform depends on their own internal constitution. In virtue of this constitution, they process information that is procured either directly through the making of measurements or from other control mechanisms. In either case, the internal constitution of the control mechanism is altered, resulting in it acquiring information (it is literally, informed), which it then processes through the operations its parts perform.⁶

On this framing, mechanistic explanation of a given phenomenon should bottom out with the various work activities

⁴Traditionally, new mechanists have not characterized mechanisms as performing work. Focusing on work, however, highlights another omission in new mechanist accounts—that mechanisms require free energy to produce phenomena. Winning and Bechtel (2018; see also Bechtel and Bollhagen, 2021), drawing inspiration from Pattee's (1972a, b) have characterized mechanisms as performing work as a result of the configuration of their components constraining flows of free energy. In biological mechanisms, free energy usually takes the form of ATP. Even without specifically identifying where free energy is released and how the components constrain its effects, one can characterize the work that is thereby accomplished. That is the perspective adopted in this paper.

⁵Single cell organisms also require such control mechanisms; for discussion, see Bich and Bechtel (2022).

⁶The concept *information* has been characterized in different ways. Shannon (1948) developed a mathematical analysis of the quantity of information carried by a signal in terms of how much it reduced uncertainty. Dretske (1981) advanced a semantic characterization of information that emphasizes its content—what the signal is carrying information about. Dretske offers a basically Humean causal account of how a signal acquires content. An Aristotelian perspective in which a causal process alters the form of an object more fully captures Pattee's (1972a, b) understanding in which a control process performs and executes a process based on a measurement in which a state of the control system comes to correspond to the property being measured.

that together result in the phenomenon for which an explanation is sought. In the case of muscle contraction, it is the level at which myosin binds to an actin filament and, by hydrolyzing ATP, produces a force that pulls the actin filament along it. In the case of control mechanisms, this is the level at which they are altered by information and, based on that, act on and modify other mechanisms. In the case of muscle, control is achieved through chemical reactions which allow an influx of Ca^{2+} into the cytoplasm of the muscle cell, which serves to expose the binding sites at which myosin can bind actin (Bechtel, 2022). For both control and controlled mechanisms, explanation bottoms out in the characterization of the work that is done to produce the phenomenon.

In some cases, control can be carried out by a single control mechanism. But control can also be spread over multiple mechanisms as long as a signal is passed between them so that the action of the downstream mechanism is dependent on the processing of the upstream mechanisms and ultimately on the ones acquiring the information through making measurements. Such signaling radically expands the potential for information processing. A given control mechanism can be informed by measurements made by multiple mechanisms, process that information in a distinctive way, and send signals to different downstream control mechanisms that carry out further processing or act on controlled mechanisms. There need not be just one pathway through multiple control mechanisms; control mechanisms can form networks. This is exemplified by the integration of neurons into a nervous system in which information procured by some neurons is processed by numerous other neurons and those neurons that directly control muscle cells or secretory cells respond to inputs from many other neurons. The key point remains: Individual acts of information processing are carried out by processes within individual neurons that, in response to inputs, constrain the flow of free energy into the performance of work.

THE WAR OF THE SOUPS AND THE SPARKS: WHO WON?

As I indicated in section “Introduction,” the richness of how neurons process information is concealed in the conception of the brain as an electrical switching system. This focus on the nervous system as an electrical system has deep historical roots. Galvani (1791) not only showed that muscles respond to electrical stimulation, but inferred that muscles and nerves, like Leyden jars, contained their own source of electricity. Continuing this line of inquiry, du Bois-Reymond (1848–1884) both provided careful experimental demonstrations of currents in nerves and muscles and identified what he termed “the negative variation” through which nerves transmit signals when stimulated. His student, Bernstein (1868) established that the negative variation, which was later designated as the *action potential*, constituted the nerve pulse. Toward the end of his career Bernstein (1902) showed that, rather than a current, when not stimulated, nerves and muscles exhibit a potential due to ions being unequally distributed across the membrane of the neuron. (For further

discussion of this history, see Lenoir, 1987; Bechtel and Vagnino, 2022.)

Once Sherrington (in his contribution to Foster’s 1897, p. 929), named and characterized the synapse, the question arose as to how electrical transmission along one neuron could elicit a response in a post-synaptic neuron. Although Elliott (1904, 1905), Langley (1905), and Dixon (1907) all advanced evidence of chemical transmission, none of them pressed their claims and few researchers at the time accepted that transmission between neurons or neurons and muscles was chemical. Dale (1914), the researcher whose own detailed research on the effects of acetylcholine administration positioned him to embrace chemical transmission, did not [largely due to the lack of any “evidence that a substance resembling acetyl-choline exists in the body at all” (p. 188)]. By the time Dale and Dudley (1929) found acetylcholine as well as histamine in ox and horse spleens, Loewi (1921) had conducted an experiment (conceptually similar to one Dixon had conducted previously) that demonstrated that something he called *Vagusstoff* could be extracted from one heart muscle whose contractions were depressed and administered to another, depressing its contractions. Even when Loewi and Dale were awarded the Nobel Prize in 1936 for chemical transmission at the periphery of the autonomic nervous system, the dominant view was that in the brain and in peripheral nerves controlling skeletal muscles transmission must be electrical. Chemical mediation was deemed to be much too slow to account either for control of skeletal muscles or central processing—the electrical charge was simply understood to jump the gap between neurons. (For indepth historical discussion, see Davenport, 1991; Valenstein, 2005.)

This conflict, which Valenstein (2005) describes as the war between the soups and the sparks, only ceased after Eccles, who had been a chief proponent of electrical transmission, found evidence about inhibitory stimulation that he could not account for with a purely electrical hypothesis (Brock et al., 1952). This resulted in the general acceptance that transmission between neurons involves a chemical process. The issue of the slowness of chemical transmission was partly resolved by the discovery of fast chemical responses. Dale had identified both a fast and slow response to acetylcholine and much of the focus was on the fast response. It took a surprisingly long period to identify the amino acid derivatives glutamate and γ -aminobutyric acid (GABA) as the principal fast-acting neurotransmitters, in part because their presence in the brain was largely attributed to their potential role in metabolism. By the 1970’s they were regarded as “putative neurotransmitters” (Krnjević, 1970; Curtis and Johnston, 1974) and shortly after that glutamate was recognized as the principal excitatory transmitter and GABA as the chief inhibitory transmitter in the mammalian central nervous system.

Referring to a transmitter as excitatory or inhibitory is an oversimplification. Whether in a given case a transmitter generates excitation or inhibition depends on the receptor and conditions in the postsynaptic neuron. In prototypical cases, glutamate and GABA act on ionotropic receptors, opening or closing an ion channel, thereby determining whether an ion (of, e.g., sodium, potassium, calcium, or chloride) is transported through the membrane. This results in either reduced or increased polarization of the membrane and initiates a current

along the dendritic membrane. The postsynaptic neuron collects currents generated along its dendritic tree and, if these exceeded a threshold, initiates an action potential along its dendrite. Focusing on this role, chemical processing at synapses can be viewed as simply enabling conduction and switching of electrical signals, rendering the victory of the soups pyrrhic. Attention to the chemical processes may seem to add little to the understanding of neural information processing.

But this is a serious oversimplification. Even if one limits one's focus to actions on ionotropic receptors, the same transmitter can generate different currents in postsynaptic neurons depending on which ionotropic receptors are present and on the electrochemical gradient across the membrane of the post-synaptic neuron. In addition, though, amino acid transmitters such as glutamate and GABA often bind to not just on ionotropic receptors but also metabotropic receptors, through which they alter metabolic processes, including gene expression, in the postsynaptic cell.

An indication that the electrical transmission account is seriously incomplete is that, once the search for neurotransmitters began, the number of known neurotransmitters mushroomed (there are now more than twenty small molecule neurotransmitters and over a hundred peptidergic transmitters known to be operative in mammalian brains). If all neurotransmitters did were initiate movement of ions across the post-synaptic membrane, one might wonder why nature is so profligate with transmitters?⁷ An alternative perspective that makes sense of the diversity of chemicals acting as neurotransmitters is that the chemical interactions between neurons are not just transmitting information but, depending on the response elicited in the recipient neuron, processing it in different ways. Different receptors for different neurotransmitters result in the recipient neuron behaving differently. To illustrate the implications of focusing on the range of chemical interactions between neurons, I turn in the next two sections to two classes of transmitters that act principally on metabotropic receptors—neuropeptides and monoamines. Once we recognize the diversity of processing provided by chemical transmission between neurons, we can recognize the profound implications of the soups' victory: it is through a wide range of chemical responses to neural transmissions that information is processed in the mind-brain.

INFORMATION PROCESSING WITH NEUROPEPTIDES

I begin with one of the last class of chemicals to be recognized as neurotransmitters, neuropeptides. In his review of chemicals

involved in synaptic transmission, under the category “some other putative transmitters,” (Krnjević, 1974, p. 491) briefly discusses substance P and then, even more briefly, notes that polypeptides had been shown to excite neural activity. He comments, “Whether these are of significance for synaptic function remains to be established.” Substance P had been identified by von Euler and Gaddum (1931) after they found that an extract from whole equine brain depressed blood pressure even after they applied atropine, which was known to inhibit acetylcholine. They viewed it as a second transmitter in their preparation in addition to acetylcholine. Numerous other neuropeptides, such as vasopressin and oxytocin, were discovered in the early 20th century, but they were at first characterized as hormones and not as neurotransmitters.

Physiologists were investigating hormone signaling even as the sparks dominated discussions of neurotransmission. Drawing on his discovery of secretin, a peptide secreted by the intestines that initiates secretion of digestive fluids in the pancreas (Bayliss and Starling, 1902), Starling (1905) coined the term hormone (from the Greek ὁρμῶν, I excite or arouse) for “chemical messengers which, speeding from cell to cell along the blood stream,... coordinate the activities and growth of different parts of the body.” Subsequent research has resulted in identification of many peptides acting as hormones in coordinating the operation of the various organs responsible for physiological activities such as digestion, respiration, growth, reproduction, and sleep.⁸

Hormone signaling exhibits both features of control mechanisms identified in section “Control Mechanisms: Modifying Constraints in Controlled Mechanisms.” In synthesizing and secreting hormones, cells are responding to measurements of conditions registered in the release of the hormones (e.g., the presence of food). The cells that respond to hormones do so by altering their metabolic processes—catalyzing different reactions or expressing different genes. The differentiation of the processes of generating a signal and responding to it allows for the signal to be distributed to many responders that can respond differently and for responders to respond to different combinations of signals. The evolution of peptidergic neurons can be viewed as an extension of the information processing achieved with hormones.⁹ Essentially, a peptidergic neuron inserts an elongation between a receptor that responds to one or more peptides (and other transmitters) and the machinery for synthesizing new peptides and preparing them for secretion. In this elongation the signal can be propagated either by diffusion through the cytoplasm or electrical transmission along the membrane.

⁷Moroz et al. (2021, p. 7) argue that, on the electrical switching account, two, or even one, transmitter would suffice: “If a chemical messenger acts only as a pure transmitter (= messenger) at the synaptic cleft within a specific wiring diagram, only two neurotransmitters are needed (e.g., for excitation and inhibition, respectively). If there are two different receptors for the same transmitter (to induce excitation and inhibition)—then even one transmitter might be sufficient.” Considering circuit design, there may be good reasons to have more than one or two transmitters (e.g., to keep different messages segregated), but the number that have been found vastly exceeds what would be required to satisfy circuit requirements.

⁸Hormone signaling itself serves to process information. In animals without neurons, such as *Trichoplax adhaerens*, which has just six cell types, hormones coordinate activities in different cells and enable a variety of different behaviors (Senatore et al., 2017). Similar signaling occurs within plants and between single-cell organisms, including prokaryotes that act cooperatively in biofilms. Such signaling typically involves the secretion of peptides from one cell that are taken up by and metabolized in other cells. Such signaling, however, is not limited to peptide transmission: Prindle et al. (2015) have shown the bacteria also make use of electrical currents transmitted along cell membranes to coordinate activity in biofilms, even among bacteria of different species.

⁹Moroz et al. (2021) have argued that neurons first evolved to extend peptidergic signaling.

How information is processed by peptidergic neurons depends both on the process by which the peptide is disseminated and on the chemical responses to the peptide. Whereas amino acid transmitters are typically released at the synapse cleft and are restricted to that site, peptidergic transmitters are often volume transmitters—they may be released at various locations on one neuron (van den Pol and Anthony, 2012) and allowed to diffuse through the extracellular matrix as a volume transmitter to wherever they encounter an appropriate receptor, with the physical features of the matrix determining how much and where it diffuses (Agnati et al., 2010; van den Pol and Anthony, 2012; Fuxe et al., 2013). Volume transmitters can engage in multiple interactions—a peptide released by one neuron may act on transmitters of many neurons (this is referred to as *divergence*) and a given neuron may have receptors for transmitters released by many different neurons (*convergence*) (Brezina and Weiss, 1997; Swensen and Marder, 2000; Brezina, 2010). This potential for complexity is further extended when it is recognized that peptidergic neurons often release multiple different peptides as well as other neurotransmitters (Hökfelt et al., 1980; Hökfelt, 2009; Svensson et al., 2019).

In most cases, the response to neuropeptides begins with binding to a G-protein coupled receptor (GPCR) that crosses (seven times) the membrane (van den Pol and Anthony, 2012).¹⁰ Binding a peptide (or other ligand) on the outside of the cell alters the conformation of the protein, promoting reactions inside the cell. In particular, it activates a guanine nucleotide exchange factor (GEF) that causes the replacement of a GDP by a GTP in a heterotrimeric G-protein complex bound to the receptor on one of its passes inside the cell. The G-protein complex contains two subunits: G_α and $G\beta\gamma$. The G_α subunit is a GTPase that binds and eventually hydrolyzes the GTP. When GEF promotes the exchange of GTP for GDP in the G_α subunit (**Figure 1**), the subunits of the G-protein split apart, allowing each to catalyze reactions. This process is brought to a halt once the G_α subunit hydrolyzes GTP to GDP, enabling it to bind to a $G\beta\gamma$ subunit and becoming inactive. Regulator of G-protein signaling (RGS) proteins, in turn, modulate the rate of hydrolysis (for a detailed review, see McCudden et al., 2005).

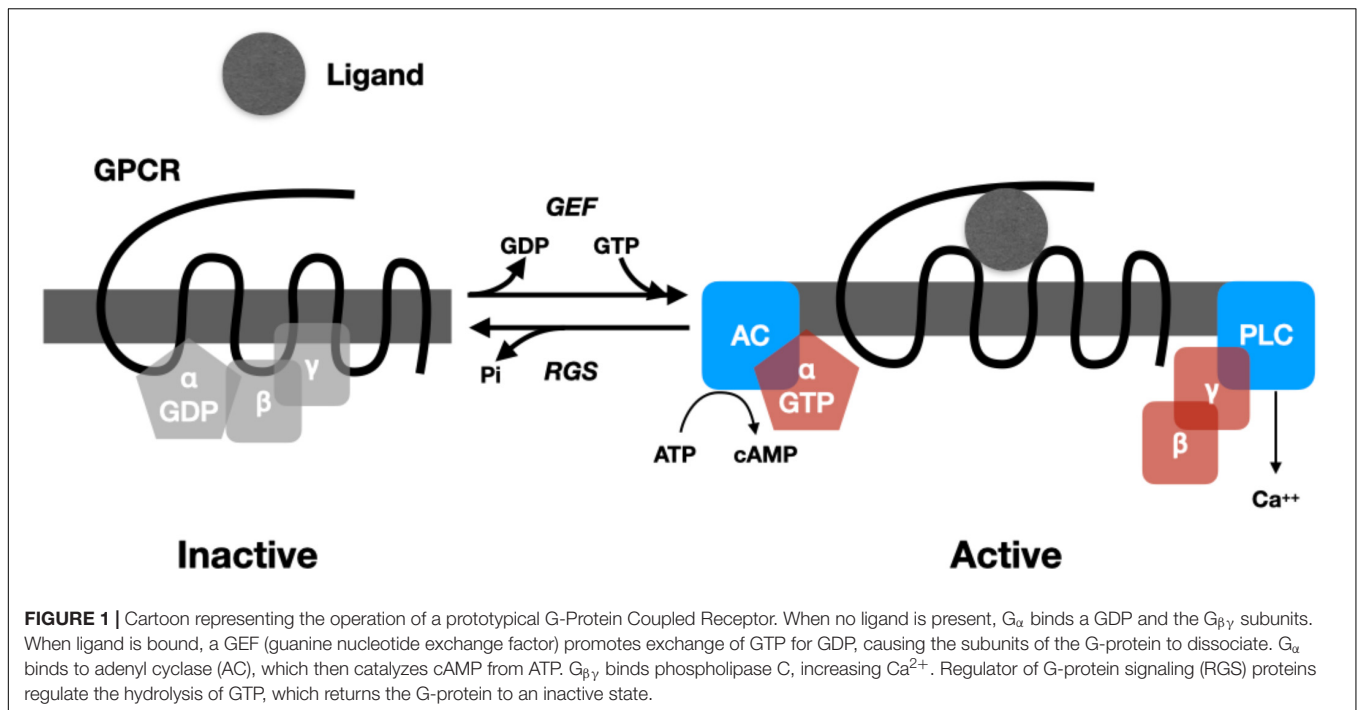
The splitting and activation of the subunits of the G protein can initiate a wide range of biochemical processes in the cell (for an accessible overview, see Marks et al., 2017). The subunits of G-proteins can activate enzymes such as adenylyl cyclase, which generates cyclic AMP (cAMP) from ATP, and phospholipase C, which, via the synthesis of inositol trisphosphate, generates an increase in Ca^{2+} . Both cAMP and Ca^{2+} are intracellular signals (second messengers) that initiate subsequent reactions depending on the constituents of the cell (cAMP through whatever protein kinase A is available and Ca^{2+} through

whatever protein kinase C is available). These diverse chemical reactions constitute the processing of the signal. Among the results of these reactions is altered gene expression, including the synthesis of new peptides. In addition to being synthesized in the endoplasmic reticulum, new peptides are subject to extensive post-translational modifications in the Golgi apparatus (different modifications resulting in different peptides) and then packaging into large dense core vesicles. One of the functions of a small but relatively long-lasting increase in Ca^{2+} concentration in the cytoplasm is the release of the contents of these vesicles into the extracellular matrix.

Neurons responding to and releasing neuropeptides play an especially important role in the hypothalamus. The hypothalamus consists of highly interconnected nuclei, each typically containing multiple cell types distinguished by their receptors and their machinery for synthesizing new peptides (Leng, 2018). Many of these nuclei are located adjacent to the median eminence at the base of the diencephalon, an ideal location for extending endocrine signaling since there is no blood-brain barrier at the median eminence. Instead, the fenestrated capillaries allow hormones to act on neurons in the hypothalamus and for peptides synthesized by hypothalamic neurons either to act as hormones by entering the bloodstream directly (oxytocin and vasopressin) or to initiate the synthesis of hormones in the pituitary. To illustrate the variety of information processing activities of peptidergic neurons in the hypothalamus, I will briefly describe the function of two peptides—orexin and vasopressin—released by populations of hypothalamic neurons.

When orexin-releasing neurons were discovered in the lateral hypothalamic area in the late 1990's (de Lecea et al., 1998; Sakurai et al., 1998), they were interpreted as promoting feeding behavior (the name *orexin* is derived from the Greek word for appetite). Researchers soon demonstrated that these neurons integrate signals from two populations of neurons in another hypothalamic nucleus, the arcuate nucleus. Neurons in one arcuate population respond to peptides such as leptin, which is released from adipose cells in proportion to fat mass, and release proopiomelanocortin (POMC). POMC can be viewed as signaling satiety (Yeo et al., 2021). Neurons in the other population respond to peptides such as ghrelin, which is synthesized in the gut and duodenum when no food is present. It generates neuropeptide Y and agouti-related peptide; high concentration of these peptides can be viewed as signaling hunger (Aponte et al., 2011; Al Massadi et al., 2017). (Although leptin and ghrelin are the best studied of these peptides, each population of arcuate nucleus neurons receives multiple peptidergic signals and integrates these to arrive at its input to the orexin neurons.) Orexin neurons have receptors for POMC, neuropeptide Y, and agouti-related peptide; these differentially effect their synthesis of orexin. Through their projections these neurons release orexin broadly through the brain; in many locations the presence of orexin initiates feeding behavior. Orexin-releasing neurons can thus be viewed as assessing information about the organism's nutritional state and reaching a decision as to whether to initiate feeding behavior. But the story is much more complex. Shortly after they were discovered, orexin neurons were also found to be especially active during sleep-to-wake transitions.

¹⁰There has been almost no discussion of GPCRs in the philosophical literature. A notable exception is Barwich and Bschir's (2017) analysis how GPCRs went from hypothetical posits to a class of entities occurring in nature. They argue that success in manipulating GPCRs played an important role, but only once researchers had "an adequate conceptual grasp of their potential structural and functional roles" (p. 1333). Barwich and Bschir emphasize the historical process of "epistemic iteration" (a concept they borrow from Chang, 2004)—an interplay between conceptual development and experimental interventions.



Stimulation of orexin neurons was found to promote waking.¹¹ Fittingly, these neurons also receive inputs from the reticular activating system in the brain stem. This revealed that orexin-releasing neurons integrate nutritional information and a variety of activating signals, initiating responses based on multiple sources of information. Subsequent research revealed that orexin-releasing neurons respond to an even wider range of peptides, signaling a variety of cell states, and contribute to initiating a broad range of cell responses (for a review, see Arrigoni et al., 2019).

Vasopressin-synthesizing neurons in the supraoptic and paraventricular nuclei of the hypothalamus reveal a similar pattern of integrating multiple sources of information and generating multiple responses (Stoop, 2012; Sternson, 2013; Leng, 2018; Watts et al., 2021). Neurons in these two nuclei receive excitatory inputs from the amygdala and inhibitory inputs from the hippocampus as well as noradrenergic and monoaminergic inputs from the brainstem. One result is that they register osmolarity and low-blood volume, as well as various stressors. Different cell populations in these nuclei synthesize and release vasopressin to different locations (Watts, 2010). Magnocellular neurons project into the posterior pituitary where they release vasopressin into the blood stream (Nestler et al., 2015, chapter 10). Vasopressin released into the blood has different downstream effects depending on which cells have either V1 or V2 receptors (each initiates a distinct metabolic cascade). (In addition to vasopressin, these neurons also release the opioid peptides

enkephalin and dynorphin, neuropeptide Y, cholecystokinin, and galanin, each of which acts on cells with appropriate receptors.) For example, the V2 receptor on the distal nephron of the kidney initiates the synthesis and insertion of water channels that result in the reabsorption of water into the circulation (in the process, rendering the urine more concentrated). Vasopressin in the blood also acts on V1a receptors in arterioles, causing them to contract and thereby raise blood pressure.

A second population of parvocellular neurons releases vasopressin, together with corticotropin releasing factor (CRF), into to the hypophyseal-portal circulation, which drains into the anterior pituitary (Nestler et al., 2015, chapter 10). There vasopressin and CRF together bind cells with V1b receptors and initiate a sequence of reactions beginning with the synthesis of POMC (discussed above as signaling satiety), which in turn stimulates synthesis and release of ACTH (Aguilera et al., 2008). ACTH both feeds back to inhibit POMC transcription and acts on receptors on different populations of cells in the adrenal cortex to initiate the synthesis and secretion from cholesterol of either glucocorticoids (cortisol) or other steroids including aldosterone and adrenal androgens. These in turn recruit energy for a flight-or-flight response. (Glucocorticoids also feeds back to repress both CRF and ACTH synthesis, thus stopping the action initiated by vasopressin.) In addition to these actions on the endocrine system, vasopressin synthesizing neurons also release vasopressin into other brain regions where it is implicated in reducing aggression and affiliative behaviors.

The two example neuropeptides I have discussed, orexin and vasopressin, reveal important features of how peptidergic neurons process information in the brain. The neurons that synthesize them do so in response to appropriate ligands (each

¹¹The output from orexin neurons is complex, often involving the corelease of orexin and glutamate, with the two transmitters operating on different timescales (Schöne et al., 2014).

of which carries information about one or more conditions in the organism). These ligands bind to GPCRs that, depending on the particular type of cell in which they occur, trigger the synthesis of specific peptides that can then be released into the extracellular matrix. The near endless variety of possible peptides, receptors, and intracellular signaling pathways (Brezina, 2010) allows peptidergic neurons to process information in a vast number of ways. Moreover, the ability of these neurons to incorporate receptors for multiple inputs and generate multiple outputs allows them to integrate information. In this way, they build upon the information processing capacities of the endocrine system, extending its capacity to process information. To understand how these neurons process information, one needs to take into account the different peptides, receptors, and intracellular signaling pathways involved, including the frequent release of multiple peptides by the same neuron and response to multiple peptides by a single neuron.

INFORMATION PROCESSING WITH MONOAMINES

To further develop the theme that one needs to ground accounts of neural information processing in the chemical activities in neurons, I turn to a second class of neurotransmitters, the monoamines norepinephrine, dopamine, and serotonin. They were among the earliest chemicals identified as neurotransmitters when they were found to meet the criteria of occurring naturally in brains and when administered, eliciting a detectable response. Only as researchers were able to localize their synthesis in the brain and investigate the receptors that responded to them did they come to recognize the distinctive type of information processing they support. On the one hand, while they are disseminated widely, in vertebrates they are only synthesized in a select set of nuclei—norepinephrine in the locus coeruleus and other nuclei in the pons and medulla, dopamine in the substantia nigra pars compactus (SNc) and the ventral tegmental area (VTA),¹² and serotonin in the raphe nuclei. Like neuropeptides, they mostly act through GPCRs. What is distinctive is how they act on other neurons—they alter the responses of recipient neurons to the main excitatory or inhibitory transmitters (glutamate and GABA). Accordingly, they are often referred to as neuromodulators.¹³ That, however, understates their role in determining how information is processed through these more traditional synapses. As they configure how circuits respond to

electrical signals, them might be viewed as setting the information processing agenda for these neural circuits.¹⁴

The fact that neuromodulators determine how neural circuits process information was first and most clearly demonstrated in invertebrate research. In many invertebrates the identity and connectivity of neurons is consistent organism to organism. This has made it possible to develop species-wide connectomes (maps of neural connectivity). Through serial electron microscopy, White et al. (1986) created a nearly complete map of neurons and their connections in the hermaphrodite nematode *C. elegans*. Based on this, researchers began to develop accounts of how specific circuits process information (Chalfie et al., 1985). However, other researchers discovered that the responses of neurons in these circuits can be modified by application of neuromodulators such as dopamine. Bargmann (2012) describes numerous cases in *C. elegans* in which application of neuromodulators changes the response properties of specific circuits without changing their physical connections. Marder has provided similar examples in a specific circuit, the stomatogastric ganglion network, in the lobster. This network of about 27 neurons regulates the foregut muscles that grind food and force it down to the gut. The network can be extracted and studied *in vitro*. Such investigations revealed that the circuit consists of two central pattern generators, one of which, the pyloric network, is constantly rhythmic while the other, the gastric mill network, generates rhythms only when it receives modulatory inputs produced by sensory inputs. Although there is a fixed pattern of physical connections between these neurons, the circuits exhibit different behavior when different monoamines and other neurotransmitters are added to the preparation (Marder and Bucher, 2007). The effects of dopamine are particularly dramatic as each neuron has dopamine receptors but responds differently to the addition of dopamine.¹⁵ The ability of neuromodulators to alter circuit behavior in invertebrates turns out to be the rule, not the exception (Marder, 2012).

Although it is more difficult to study the effects of monoamines on specific circuits in vertebrates, their effects in modulating neural activities are clear. The effects of dopamine on processing in the basal ganglia are illustrative. The basal ganglia are a network of nuclei implicated in selecting which other neural circuits process information. By default, the output regions of the basal ganglia send inhibitory GABAergic projections to regions throughout the brain (both those involved directly in action and those involved in central information processing). Only when activity in the basal ganglia inhibits these inhibitory outputs can these other brain regions carry out their activities. The basal ganglia are connected in loops to these other areas. In each loop, a brain region sends excitatory glutamatergic inputs to two sets of medium spiny neurons (MSNs) in the striatum, the input region of the basal ganglia, and the output regions of the basal ganglia act via the return loop either to maintain the inhibition or release it.

¹²A small collection of neurons in the arcuate nucleus of the hypothalamus also synthesizes dopamine. Its release suppresses synthesis of prolactin in the anterior pituitary. Dopamine is also synthesized in local circuit neurons in the retina.

¹³The term neuromodulator is used in different ways by different researchers. In an early review of neuromodulators, Kupfermann (1979, p. 448) asserts that they “do not simply excite or inhibit an electrically excitable cell, but rather are involved in altering the effects of other events occurring at the cell.” Katz (1999, p. 3) offers a far more inclusive definition: “Any communication between neurons, caused by release of a chemical, that is either not fast, or not point-to-point, or not simply excitation or inhibition will be classified as neuromodulatory.” Under either definition, the term includes the neuropeptides discussed in the previous section. In this section I single out the monoamines for their distinctive mode of neuromodulation.

¹⁴An important feature of neuromodulators, emphasized by Brezina (2010), is that commonly multiple neuromodulators are acting at the same time, resulting in multiple non-linear interactions that determine the output in any given case.

¹⁵These circuits exhibit yet further complexity as a result of co-transmission of neuropeptides and small molecules (Nusbaum et al., 2017).

The standard account of the operation of the basal ganglia function (originally advanced by Albin et al., 1989, to account for the features of different disorders associated with the basal ganglia) identifies two pathways originating with the MSNs in the striatum that receive the inputs. In what is referred to as the direct pathway, MSNs with D1 receptors send inhibitory projections directly to the output nuclei of the basal ganglia. By inhibiting these inhibitory outputs, active neurons with D1 receptors release the connected region from inhibition, allowing it to process information. Neurons with D2 receptors, in contrast, send inhibitory projects to intermediate regions of the basal ganglia, inhibiting their inhibitory effects on the output regions; the net effect is to enhance their inhibitory action on other brain regions. For contemporary presentations of this account, see Gerfen and Bolam (2016) and Clark et al. (2018). In developing a computational model of decision making based on this account, Hazy et al. (2007) characterize the direct pathway as generating a Go signal while the indirect pathway generates a NoGo signal.

On the standard account, it is the effect of dopamine in binding to the D1 and D2 receptors that determines the output of the basal ganglia.¹⁶ When dopamine binds the D1 receptor, it initiates a cascade involving cAMP and protein kinase A (PKA), with the PKA initiating responses that enhance the expression of both alpha-amino-3-hydroxy-5-methyl-4-isoxazolepropionate receptors (AMPA) and N-methyl-D-aspartate receptors (NMDARs), ionotropic receptors that respond to glutamate. The effect is to increase the responsiveness of these neurons to sustained glutamate inputs but to decrease it to transient inputs (Shen et al., 2016). When dopamine binds the DR2 receptor, it initiates responses including removal of AMPARs from the cell membrane and changes in Ca²⁺ and Na⁺ ion channels. In addition to these immediate effects, the combined action of dopamine and brain-derived neurotrophic factor released by cortical inputs on D1 receptors acts on a tyrosine receptor kinase B. This serves to initiate long-term activation (LTP), enhancing the likelihood that the neuron will respond to the same glutamatergic input in the future. In neurons with D2 receptors, when glutamate binds the mGluR5 receptor while Ca²⁺ is released into the neuron, it initiates long-term depression (LTD) (Shen et al., 2016).

Varying the amount of dopamine reaching the striatum from the SNc can thus alter how the basal ganglia processes inputs both immediately and in the longer term. When it is drastically reduced, as in Parkinson's, the response to inputs to D1 neurons is reduced while that to D2 neurons is enhanced, thereby reducing activity along the direct pathway and the release of other brain areas from inhibition. This explains the

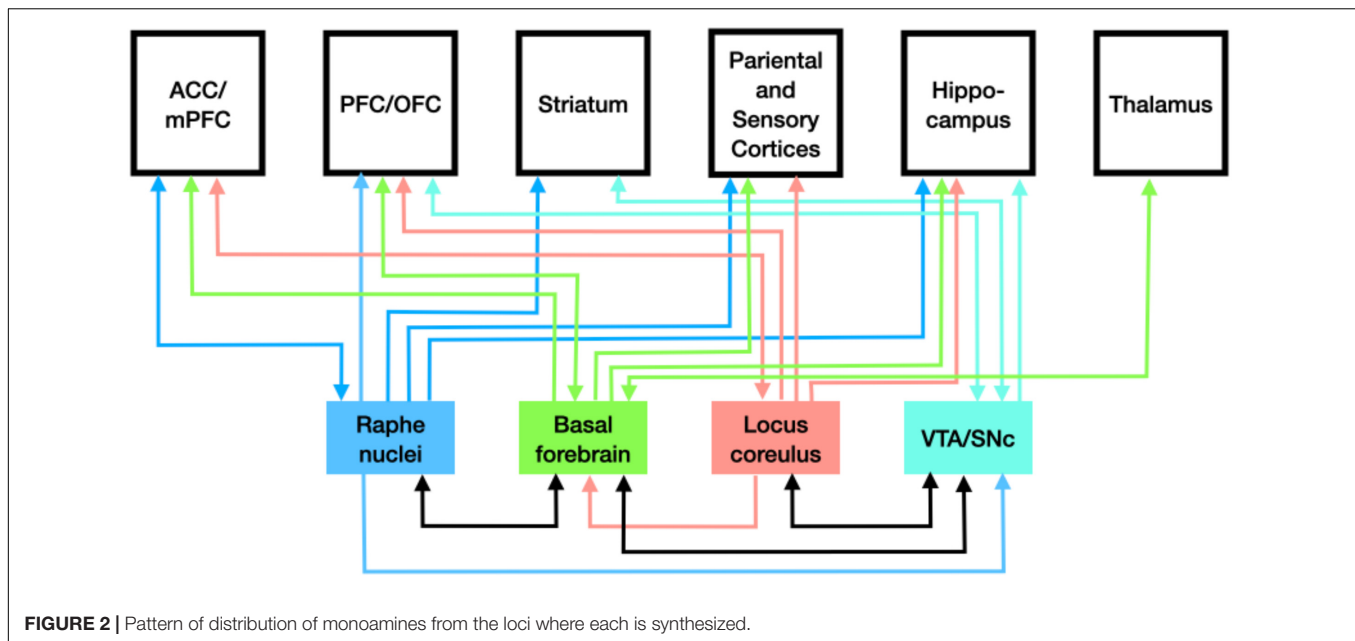
inability of Parkinson's patients to initiate voluntary actions. In most individuals, dopamine modulates how the basal ganglia process information. Starting with Schultz (Schultz et al., 1997; Schultz, 1998), a number of theorists have viewed dopamine as constituting a reward signal and have interpreted it as key to implementing reinforcement learning (as developed by AI theorists Sutton and Barto, 2018) by enabling neurons in the striatum to compare expected with actual reward and use that as a basis of learning. Others, such as Redgrave et al. (2016), contend instead that dopamine signaling enables striatal neurons to assess whether the organism is the agent of an outcome by detecting unexpected outcomes and relating them to efferent copies of motor commands. While there is disagreement of how to interpret the effects of dopamine on striatal neurons, all admit it plays a major role in structuring processing in the striatum and downstream in the basal ganglia.

The action of dopamine produced in the SNc on the striatum of the basal ganglia is just one instance of neuromodulatory activity of the monoamines. As noted above, in vertebrates, each of the monoamines is synthesized only in select nuclei but neurons in these nuclei extend axons widely through the brain. These neurons release the monoamines as volume transmitters into the extracellular matrix where they are able to bind a variety of receptor types on different neurons in the region and alter how they respond to amino acid transmitters at synapses. When the whole set of monoamines is considered, they can be seen to have multiple effects on how other brain regions process information. As indicated in **Figure 2**, the pattern of distribution is complicated. Each nucleus in which monoamines are synthesized sends projections to many areas, including those where other monoamines are synthesized. Brain areas often receive inputs from multiple monoamines. Moreover, in many cases the projections are recurrent. The broad distribution from select nuclei suggests that these transmitters can determine how the whole brain processes information. This is supported by the range of neurological and psychiatric disorders associated with disrupted monoamine response in the various brain regions. Researchers face considerable challenges in determining how the monoamines individually and collectively modulate neural information processing, but it is apparent that they play different roles than just inputs to electrical switches.

CONCLUSION: GROUNDING INFORMATION PROCESSING EXPLANATIONS IN CHEMICAL ACTIVITY

Electrical switching has long been the model of how neurons process information. On such a perspective, it would seem sufficient for mechanistic explanations of cognitive information processing to bottom out with the connectivity of neurons. There would seem to be little reason to decompose further to the chemical processes of synthesizing and responding to neurotransmitters. One could acknowledge, as the sparks did in ending their war with the soups, that chemical transmitters intervened between neurons, but still insist that the details of chemical activity would not further enlighten our understanding

¹⁶The standard account is likely to be oversimplified. Graybiel (2005) identified six challenges based on then recent empirical evidence (e.g., evidence that there are many more collateral connections between the supposedly independent pathways, that the outputs of the basal ganglia are not solely inhibitory, and that the D1 and D2 neurons receive different inputs from cortex). Doubts about the standard account do not minimize the importance of dopaminergic signaling to the striatum. In her own research, Graybiel has investigated the role of dopamine projections to the striosomes of the striatum as relating mood and emotion to decision making (Graybiel and Grafton, 2015) and the role of dopamine in the acquisition of habits by modifying the behavior of striatal neurons (Graybiel and Grafton, 2015).



of how information is processed in the brain. Connectomic analyses and neural network models based on them could explain how information is processed.

The variety of neurotransmitters and the diversity of ways neurons respond to them suggests that, on the contrary, considerations of the chemical processes are pertinent to understanding how nervous systems process information. Neuropeptides and monoamines figure in quite different information processing than amino-acid-based transmitters acting on ionotropic receptors. The response to neuropeptides is not just the generation of an action potential in the recipient cell but a wide range of metabolic activities, including the synthesis of new peptides. The response to monoamines can significantly alter the processing in a circuit by amino-acid-based transmitters. The details of the chemical processing between neurons matters for how information is processed. As a foundation for understanding how brains process information, researchers need, in addition to a connectome detailing synaptic connections, a chemoconnectome: “an entire set of neurotransmitters, neuromodulators, neuropeptides, and receptors supporting chemical transmission in an animal (Moroz, 2021).

I have approached the issue of reduction from the perspective of developing mechanistic explanations of control mechanisms. A central feature of mechanistic approaches is decomposing systems. Until one reaches true atoms (indivisible components), further decomposition is always possible. Accordingly, one could argue that while chemical processes are important, mechanistic explanations of neural information processing should not bottom out there but, for example, continue on to the quantum processes at work in these chemical reactions. The account of mechanistic explanation in terms of the work performed by the mechanism, however, shows why further decomposition is not likely to be informative. In the case of control mechanisms, the relevant work is processing information. With the chemical processes between

neurons, one has reached a level at which one can account for the different ways in which information is processed. Further decomposition will not provide additional illumination about the ways information is processed in the brain.

In advocating for explanations of neural information processing bottoming out in chemical processing, I am not arguing that only the chemical level is required to understand such information processing. Organization at multiple levels helps determine how information is directed through an organism. Patterns of connections between neurons is important, as is the organization of neurons into nuclei and brain structures. Even higher levels of organization are also relevant, including levels that integrate neurons with different organs and connect activities in organisms to entities in their environment, including the social environment. In requiring both reductionistic and holistic research, mechanistic reduction differs from Bickle’s (2003, 2006) characterization of ruthless reduction. Insofar as one explanatory goal is to understand how circuits in the brain process information, though, the level of chemical processes that mediate between neurons is of critical importance as it is a level at which the work of processing information in particular ways is carried out.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article. Further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

REFERENCES

- Aggarwal, C. C. (2018). *Neural Networks and Deep Learning*. Cham: Springer.
- Agnati, L. F., Guidolin, D., Guescini, M., Genedani, S., and Fuxe, K. (2010). Understanding wiring and volume transmission. *Brain Res. Rev.* 64, 137–159. doi: 10.1016/j.brainresrev.2010.03.003
- Aguilera, G., Subburaju, S., Young, S., and Chen, J. (2008). “The parvocellular vasopressinergic system and responsiveness of the hypothalamic pituitary adrenal axis during chronic stress,” in *Progress in Brain Research*, Vol. 170, eds I. D. Neumann and R. Landgraf (Elsevier), 29–39.
- Al Massadi, O., López, M., Tschöp, M., Diéguez, C., and Nogueiras, R. (2017). Current understanding of the hypothalamic ghrelin pathways inducing appetite and adiposity. *Trends Neurosci.* 40, 167–180. doi: 10.1016/j.tins.2016.12.003
- Albin, R. L., Young, A. B., and Penney, J. B. (1989). The functional anatomy of basal ganglia disorders. *Trends Neurosci.* 12, 366–375. doi: 10.1016/0166-2236(89)90074-x
- Aponte, Y., Atasoy, D., and Sternson, S. M. (2011). AGRP neurons are sufficient to orchestrate feeding behavior rapidly and without training. *Nat. Neurosci.* 14, 351–355. doi: 10.1038/nn.2739
- Arrigoni, E., Chee, M. J. S., and Fuller, P. M. (2019). To eat or to sleep: that is a lateral hypothalamic question. *Neuropharmacology* 154, 34–49. doi: 10.1016/j.neuropharm.2018.11.017
- Bargmann, C. I. (2012). Beyond the connectome: how neuromodulators shape neural circuits. *Bioessays* 34, 458–465. doi: 10.1002/Bies.201100185
- Barwich, A.-S., and Bschor, K. (2017). The manipulability of what? The history of G-protein coupled receptors. *Biol. Philos.* 32, 1317–1339. doi: 10.1007/s10539-017-9608-9
- Bayliss, W. M., and Starling, E. H. (1902). The mechanism of pancreatic secretion. *J. Physiol.* 28, 325–353. doi: 10.1113/jphysiol.1902.sp000920
- Bechtel, W. (2009). Looking down, around, and up: mechanistic explanation in psychology. *Philos. Psychol.* 22, 543–564.
- Bechtel, W. (2022). Levels in biological organisms: hierarchy of production mechanisms, heterarchy of control mechanisms. *Monist* 105, 156–174. doi: 10.1093/monist/onab029
- Bechtel, W., and Abrahamsen, A. (2005). Explanation: a mechanist alternative. *Stud. History Philos. Biol. Biomed. Sci.* 36, 421–441.
- Bechtel, W., and Bollhagen, A. (2021). Active biological mechanisms: transforming energy into motion in molecular motors. *Synthese* 199, 12705–12729. doi: 10.1007/s11229-021-03350-x
- Bechtel, W., and Vagnino, R. (2022). Figuring out what is happening: the discovery of two electrophysiological phenomena. *History Philos. Life Sci.* 44:20. doi: 10.1007/s40656-022-00502-1
- Bernstein, J. (1868). Über den zeitlichen verlauf der negativen schwankung des nervenstroms. *Pflügers Archiv* 1, 173–207.
- Bernstein, J. (1902). Untersuchungen zur thermodynamik der bioelektrischen ströme. *Pflügers Archiv Für Die Gesamte Physiol. Des Menschen Und Der Tiere* 92, 521–562.
- Bich, L., and Bechtel, W. (2022). Control mechanisms: explaining the integration and versatility of biological organisms. *Adapt. Behav.* 2022:10597123221074429. doi: 10.1177/10597123221074429
- Bickle, J. (2003). *Philosophy and Neuroscience: A Ruthlessly Reductive Account*. Dordrecht: Kluwer.
- Bickle, J. (2006). Reducing mind to molecular pathways: explicating the reductionism implicit in current cellular and molecular neuroscience. *Synthese* 151, 411–434.
- Brezina, V. (2010). Beyond the wiring diagram: signalling through complex neuromodulator networks. *Philos. Trans. R. Soc. B Biol. Sci.* 365, 2363–2374. doi: 10.1098/rstb.2010.0105
- Brezina, V., and Weiss, K. R. (1997). Analyzing the functional consequences of transmitter complexity. *Trends Neurosci.* 20, 538–543. doi: 10.1016/S0166-2236(97)01120-X
- Brock, L. G., Coombs, J. S., and Eccles, J. C. (1952). The recording of potentials from motoneurons with an intracellular electrode. *J. Physiol.* 117, 431–460. doi: 10.1113/jphysiol.1952.sp004759
- Burnstock, G. (1980). “Do some nerve cells release more than one transmitter?” in *Commentaries in the Neurosciences*, eds A. D. Smith, R. Llinás, and P. G. Kostyuk (Pergamon), 151–160.
- Burnstock, G. (2004). Cotransmission. *Curr. Opin. Pharmacol.* 4, 47–52. doi: 10.1016/j.coph.2003.08.001
- Chalfie, M., Sulston, J. E., White, J. G., Southgate, E., Thomson, J. N., and Brenner, S. (1985). The neural circuit for touch sensitivity in *Caenorhabditis elegans*. *J. Neurosci.* 5, 956–964. doi: 10.1523/JNEUROSCI.05-04-00956.1985
- Chang, H. (2004). *Inventing Temperature: Measurement and Scientific Progress*. New York: Oxford University Press.
- Clark, D. L., Mendez, M. F., and Boutros, N. N. (2018). “Basal ganglia,” in *The Brain and Behavior: An Introduction to Behavioral Neuroanatomy*, 4 Edn, eds D. L. Clark, M. F. Mendez, and N. N. Boutros (Cambridge: Cambridge University Press), 103–123.
- Craver, C. F., and Bechtel, W. (2007). Top-down causation without top-down causes. *Biol. Philos.* 22, 547–563.
- Curtis, D. R., and Johnston, G. A. R. (1974). Amino acid transmitters in the mammalian central nervous system. *Ergebnisse der Physiol. Rev. Physiol.* 69, 97–188.
- Dale, H. H. (1914). The action of certain esters and ethers of choline, and their relation to muscarine. *J. Pharmacol. Exp. Ther.* 6, 147–190.
- Dale, H. H., and Dudley, H. W. (1929). The presence of histamine and acetylcholine in the spleen of the ox and the horse. *J. Physiol. London* 68, 97–123. doi: 10.1113/jphysiol.1929.sp002598
- Davenport, H. W. (1991). Early history of the concept of chemical transmission of the nerve impulse. *Physiologist* 129, 178–190.
- de Lecea, L., Kilduff, T. S., Peyron, C., Gao, X. B., Foye, P. E., Danielson, P. E., et al. (1998). The hypocretins: hypothalamus-specific peptides with neuroexcitatory activity. *Proc. Natl. Acad. Sci. U. S. A.* 95, 322–327. doi: 10.1073/pnas.95.1.322
- Dixon, W. E. (1907). On the mode of action of drugs. *Med. Magaz.* 16, 454–457.
- Dretske, F. I. (1981). *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press/Bradford Books.
- du Bois-Reymond (1848-1884). *Untersuchungen Über Thierische Elektrizität*. Berlin: Reimer.
- Elliott, T. R. (1904). On the action of adrenalin. *J. Physiol.* 31:1055. doi: 10.1113/jphysiol.1904.sp001055
- Elliott, T. R. (1905). The action of adrenalin. *J. Physiol.* 32, 401–467. doi: 10.1113/jphysiol.1905.sp001093
- Foster, M. (1897). *A Text Book of Physiology, Part Three: The Central Nervous System*, 7th Edn. London: Macmillan.
- Fuxe, K., Borroto-Escuela, D. O., Romero-Fernandez, W., Zhang, W., and Agnati, L. F. (2013). Volume transmission and its different forms in the central nervous system. *Chin. J. Integr. Med.* 19, 323–329. doi: 10.1007/s11655-013-1455-1
- Galvani, L. (1791). *De Viribus Electricitatis In Motu Musculari Commentarius*. Bologna: Ex typographia Instituti Scientiarum.
- Gerfen, C. R., and Bolam, J. P. (2016). “The neuroanatomical organization of the basal ganglia,” in *Handbook of Behavioral Neuroscience*, Vol. 24, eds H. Steiner and K. Y. Tseng (Elsevier), 3–32.
- Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. Cambridge, MA: The MIT Press.
- Graybiel, A. M. (2005). The basal ganglia: learning new tricks and loving it. *Curr. Opin. Neurobiol.* 15, 638–644. doi: 10.1016/j.conb.2005.10.006
- Graybiel, A. M., and Grafton, S. T. (2015). The striatum: where skills and habits meet. *Cold Spring Harbor Perspect. Biol.* 7:8. doi: 10.1101/cshperspect.a021691
- Hazy, T. E., Frank, M. J., and O'Reilly, R. C. (2007). Towards an executive without a homunculus: computational models of the prefrontal cortex/basal ganglia system. *Philos. Trans. R. Soc. B Biol. Sci.* 362, 1601–1613. doi: 10.1098/rstb.2007.2055
- Hökfelt, T. (2009). “Coexistence of neuromessenger molecules – a perspective,” in *Co-Existence and Co-Release of Classical Neurotransmitters: Ex uno plures*, ed. R. Gutierrez (Boston, MA: Springer), 1–13.
- Hökfelt, T., Johansson, O., Ljungdahl, Å., Lundberg, J. M., and Schultzberg, M. (1980). Peptidergic neurones. *Nature* 284, 515–521. doi: 10.1038/284515a0
- Katz, P. S. (1999). “What are we talking about? Modes of neuronal communication,” in *Beyond Neurotransmission: Neuromodulation and its Importance for Information Processing*, ed. P. S. Katz (New York: Oxford), 1–28.
- Krnjević, K. (1970). Glutamate and γ -aminobutyric acid in brain. *Nature* 228, 119–124. doi: 10.1038/228119a0
- Krnjević, K. (1974). Chemical nature of synaptic transmission in vertebrates. *Physiol. Rev.* 54, 418–540. doi: 10.1152/physrev.1974.54.2.418

- Kupfermann, I. (1979). Modulatory actions of neurotransmitters. *Ann. Rev. Neurosci.* 2, 447–465. doi: 10.1146/annurev.ne.02.030179.002311
- Langley, J. N. (1905). On the reaction of cells and of nerve endings to certain poisons, chiefly as regards the reaction of striated muscle to nicotine and to curari. *J. Physiol. London* 33, 374–413. doi: 10.1113/jphysiol.1905.sp001128
- Leng, G. (2018). *The Heart Of The Brain: The Hypothalamus And Its Hormones*. Cambridge: MIT Press.
- Lenoir, T. (1987). Models and instruments in the development of electrophysiology, 1845–1912. *Historical Stud. Phys. Sci.* 17, 1–54.
- Loewi, O. (1921). Über humorale übertragbarkeit der herznervenwirkung. *Pflüger's Archiv Für Die Gesamte Physiol. Des Menschen Und Der Tiere* 189, 239–242. doi: 10.1007/BF01738910
- Machamer, P., Darden, L., and Craver, C. F. (2000). Thinking about mechanisms. *Philos. Sci.* 67, 1–25. doi: 10.1086/392759
- Marder, E. (2012). Neuromodulation of neuronal circuits: back to the future. *Neuron* 76, 1–11. doi: 10.1016/j.neuron.2012.09.010
- Marder, E., and Bucher, D. (2007). Understanding circuit dynamics using the stomatogastric nervous system of lobsters and crabs. *Ann. Rev. Physiol.* 69, 291–316. doi: 10.1146/Annurev.Physiol.69.031905.161516
- Marks, F., Klingmüller, U., and Müller-Decker, K. (2017). *Cellular Signal Processing: An Introduction to the Molecular Mechanisms of Signal Transduction*. New York, NY: Garland Science.
- McCudden, C. R., Hains, M. D., Kimple, R. J., Siderovski, D. P., and Willard, F. S. (2005). G-protein signaling: back to the future. *Cell Mol. Life Sci.* 62, 551–577. doi: 10.1007/s00018-004-4462-3
- Moroz, L. L. (2021). Multiple origins of neurons from secretory cells. *Front. Cell Dev. Biol.* 9:669087. doi: 10.3389/fcell.2021.669087
- Moroz, L. L., Romanova, D. Y., and Kohn, A. B. (2021). Neural versus alternative integrative systems: molecular insights into origins of neurotransmitters. *Philos. Trans. R. Soc. B Biol. Sci.* 376:20190762. doi: 10.1098/rstb.2019.0762
- Nagel, E. (1961). *The Structure of Science*. New York: Harcourt, Brace.
- Nestler, E. J., Hyman, S. E., Holtzman, D. M., and Malenka, R. C. (2015). *Molecular Neuropsychopharmacology: A Foundation for Clinical Neuroscience*. New York, NY: McGraw Hill.
- Nusbaum, M. P., Blitz, D. M., and Marder, E. (2017). Functional consequences of neuropeptide and small-molecule co-transmission. *Nat. Rev. Neurosci.* 18, 389–403. doi: 10.1038/nrn.2017.56
- Pattee's, H. H. (1972a). "Laws and constraints, symbols and languages," in *Towards A Theoretical Biology*, ed. C. H. Waddington (Chicago: Adine-Atherton).
- Pattee, H. H. (1972b). "The nature of hierarchical controls in living matter," in *Foundations of Mathematical Biology*, ed. R. Rosen (New York: Academic Press), 1–22.
- Prindle, A., Liu, J., Asally, M., Ly, S., Garcia-Ojalvo, J., and Süel, G. M. (2015). Ion channels enable electrical communication in bacterial communities. *Nature* 527, 59–63. doi: 10.1038/nature15709
- Redgrave, P., Vautrelle, N., Overton, P. G., and Reynolds, J. (2016). "Phasic dopamine signaling in action selection and reinforcement learning," in *Handbook of Behavioral Neuroscience*, Vol. 24, eds H. Steiner and K. Y. Tseng (Elsevier), 707–723.
- Sakurai, T., Amemiya, A., Ishii, M., Matsuzaki, I., Chemelli, R. M., Tanaka, H., et al. (1998). Orexins and orexin receptors: a family of hypothalamic neuropeptides and G protein-coupled receptors that regulate feeding behavior. *Cell* 92, 573–585.
- Schöne, C., Apergis-Schoute, J., Sakurai, T., Adamantidis, A., and Burdakov, D. (2014). Coreleased orexin and glutamate evoke nonredundant spike outputs and computations in histamine neurons. *Cell Rep.* 7, 697–704. doi: 10.1016/j.celrep.2014.03.055
- Schultz, W. (1998). Predictive reward signal of dopamine neurons. *J. Neurophysiol.* 80, 1–27. doi: 10.1152/jn.1998.80.1.1
- Schultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. *Science* 275, 1593–1599. doi: 10.1126/science.275.5306.1593
- Senatore, A., Reese, T. S., and Smith, C. L. (2017). Neuropeptidergic integration of behavior in trichoplax adhaerens, an animal without synapses. *J. Exp. Biol.* 220:3381. doi: 10.1242/jeb.162396
- Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Techn. J.* 27, 623–656.
- Shen, W., Plokin, J. L., Zhai, S., and Surmeier, D. J. (2016). "Dopaminergic modulation of glutamatergic signaling in striatal spiny projection neurons," in *Handbook of Behavioral Neuroscience*, Vol. 24, eds H. Steiner and K. Y. Tseng (Elsevier), 179–196.
- Sporns, O. (2011). The human connectome: a complex network. *Ann. N.Y. Acad. Sci.* 1224, 109–125. doi: 10.1111/j.1749-6632.2010.05888.x
- Sporns, O. (2012). *Discovering the Human Connectome*. Cambridge, MA: MIT Press.
- Sporns, O. (2015). Cerebral cartography and connectomics. *Philos. Trans. R. Soc. B Biol. Sci.* 370:20140173. doi: 10.1098/rstb.2014.0173
- Starling, E. H. (1905). *Croonian Lecture: On the Chemical Correlation Of The Functions Of The Body*. London, Royal College of Physicians.
- Starling, E. H. (1905). *Croonian Lecture: On the Chemical Correlation of the Functions of the Body*. London: Royal College of Physicians.
- Stoop, R. (2012). Neuromodulation by oxytocin and vasopressin. *Neuron* 76, 142–159. doi: 10.1016/j.neuron.2012.09.025
- Sutton, R. S., and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. Cambridge, MA: The MIT Press.
- Svensson, E., Apergis-Schoute, J., Burnstock, G., Nusbaum, M. P., Parker, D., and Schiöth, H. B. (2019). General principles of neuronal co-transmission: insights from multiple model systems. *Front. Neural Circ.* 12:117. doi: 10.3389/fncir.2018.00117
- Swensen, A. M., and Marder, E. (2000). Multiple peptides converge to activate the same voltage-dependent current in a central pattern-generating circuit. *J. Neurosci.* 20, 6752. doi: 10.1523/JNEUROSCI.20-18-06752.2000
- Valenstein, E. S. (2005). *The War of the Soups and the Sparks: The Discovery Of Neurotransmitters And The Dispute Over How Nerves Communicate*. New York: Columbia University Press.
- van den Pol, P., and Anthony, N. (2012). Neuropeptide transmission in brain circuits. *Neuron* 76, 98–115. doi: 10.1016/j.neuron.2012.09.014
- von Euler, U. S., and Gaddum, J. H. (1931). An unidentified depressor substance in certain tissue extracts. *J. Physiol. London* 72, 74–87. doi: 10.1113/jphysiol.1931.sp002763
- Watts, A. G. (2010). "Neuroendocrine parvocellular neurons," in *eLS* (New York, NY: Wiley). doi: 10.1002/9780470015902.a0000047.pub2
- Watts, A. G., Kanoski, S. E., Sanchez-Watts, G., and Langhans, W. (2021). The physiological control of eating: signals, neurons, and networks. *Physiol. Rev.* 102, 689–813. doi: 10.1152/physrev.00028.2020
- White, J. G., Southgate, E., Thomson, J. N., and Brenner, S. (1986). The structure of the nervous system of the nematode *Caenorhabditis elegans*. *Philos. Trans. R. Soc. London B Biol. Sci.* 314, 1–340. doi: 10.1098/rstb.1986.0056
- Winning, J., and Bechtel, W. (2018). Rethinking causality in neural mechanisms: constraints and control. *Minds Mach.* 28, 287–310.
- Yeo, G. S. H., Chao, D. H. M., Siegert, A.-M., Koerperich, Z. M., Ericson, M. D., Simonds, S. E., et al. (2021). The melanocortin pathway and energy homeostasis: from discovery to obesity therapy. *Mol. Metab.* 48:101206. doi: 10.1016/j.molmet.2021.101206

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

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Bechtel. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



OPEN ACCESS

EDITED BY

Mark Couch,
Seton Hall University,
United States

REVIEWED BY

Daniel C. Burnston,
Tulane University,
United States
Glenn Hartelius,
Alef Trust,
United Kingdom

*CORRESPONDENCE

John Bickle
jbickle@philrel.msstate.edu

SPECIALTY SECTION

This article was submitted to
Consciousness Research,
a section of the journal
Frontiers in Psychology

RECEIVED 09 July 2022

ACCEPTED 05 August 2022

PUBLISHED 29 August 2022

CITATION

Bickle J, De Sousa AF and Silva AJ (2022)
New research tools suggest a “levels-less”
image of the behaving organism and
dissolution of the reduction vs. anti-
reduction dispute.
Front. Psychol. 13:990316.
doi: 10.3389/fpsyg.2022.990316

COPYRIGHT

© 2022 Bickle, De Sousa and Silva. This is
an open-access article distributed under
the terms of the [Creative Commons
Attribution License \(CC BY\)](#). The use,
distribution or reproduction in other
forums is permitted, provided the original
author(s) and the copyright owner(s) are
credited and that the original publication in
this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

New research tools suggest a “levels-less” image of the behaving organism and dissolution of the reduction vs. anti-reduction dispute

John Bickle^{1,2*}, André F. De Sousa³ and Alcino J. Silva^{3,4,5}

¹Department of Philosophy and Religion, Shackouls Honors College, Mississippi State University, Starkville, MS, United States, ²Department of Advanced Biomedical Education, University of Mississippi Medical Center, Jackson, MS, United States, ³Department of Neurobiology, University of California, Los Angeles, Los Angeles, CA, United States, ⁴Department of Psychiatry and Integrative Center for Learning and Memory, UCLA, Los Angeles, CA, United States, ⁵Department of Psychology and Integrative Center for Learning and Memory, UCLA, Los Angeles, CA, United States

A kind of “ruthless reductionism” characterized the experimental practices of the first two decades of molecular and cellular cognition (MCC). More recently, new research tools have expanded experimental practices in this field, enabling researchers to image and manipulate individual molecular mechanisms in behaving organisms with an unprecedented temporal, sub-cellular, cellular, and even circuit-wide specificity. These tools dramatically expand the range and reach of experiments in MCC, and in doing so they may help us transcend the worn-out and counterproductive debates about “reductionism” and “emergence” that divide neuroscientists and philosophers alike. We describe examples of these new tools and illustrate their practical power by presenting an exemplary recent case of MCC research using them. From these tools and results, we provide an initial sketch of a new image of the behaving organism in its full causal-interactive complexity, with its molecules, cells, and circuits combined within the single system that it is. This new image stands in opposition to the traditional “levels” image of the behaving organism, and even the initial sketch we provide of it here offers hope for avoiding the dreary metaphysical debates about “emergence” and “downward causation,” and even the reduction vs. anti-reduction dispute, all dependent upon the familiar “levels” image.

KEYWORDS

molecular and cellular cognition, ruthless reductionism, memory linking, C-C chemokine receptor type 5, head-mounted miniscopes, levels image, levels-less image

Ruthless reductionism guided the first 2 decades of molecular and cellular cognition

The neuroscience field of “molecular and cellular cognition” (MCC) began in the early 1990s. Gene targeting techniques, adapted into neuroscience from developmental biology, enabled experimenters to manipulate a single protein product that was part of some intra- or intercellular signaling pathway in cells in the brain, and to measure the effects of these manipulations on both cellular activities and organism behaviors. Silva et al. (1992a,b), in experiments, widely acknowledged to be the first ones published in this field, “knocked out” the gene for the α isoform of calmodulin kinase II at the embryonic stem cell stage of development in mice. They tracked the negative effects of this intervention on the induction of long-term potentiation (LTP) in mutant hippocampus tissue slices, and on learning in the intact mutants in the Morris water maze. Over the next decade, the spatial precision and temporal resolution of gene targeting in engineered mutants increased greatly; Silva et al. (2014) documents a number of these experiment tools and some key results in landmark MCC publications from its first 2 decades. Before the turn of the 21st century, Kandel and colleagues incorporated many of these early MCC findings into a “molecular model for the consolidation of the late phase of LTP and hippocampus-based long-term memory” (Abel et al., 1997, Figure 7, 623). Much of this original model is now current textbook neurobiology.

Reflecting metascientifically¹ on these early MCC practices and results, and paying close attention to the language these scientists used in their experimental publications to describe their results, Bickle (2003, 2006) proposed that a novel type of reduction, “ruthless reduction,” was at work. Contrasted with popular accounts of reduction from the philosophy of science, including varieties of intertheoretic reduction, functional reduction, and then-newly described mechanistic reduction, Bickle argued that the ruthless reductionism implicit in MCC practices was a matter of intervening experimentally into increasingly lower levels of biological organization, and then

tracking the effects of these interventions on behaviors *in vivo*, typically in rodents, using a variety of protocols widely accepted as operationalizing various cognitive functions. He argued that the most straightforward interpretation of what MCC scientists were doing in their experiments, and concluding in their discussion sections, was the direct reduction of cognitive functions to the intra- and inter-neuronal molecular pathways that the new gene targeting research tools were rendering experimentally manipulable. According to ruthless reductionism, the numerous other levels commonly thought to be interspersed between the molecular and the behavioral—especially the circuit and the “cognitive” (information-processing) levels—might be heuristically useful for finding new cellular and molecular mechanisms; but once these cellular/molecular mechanisms were found through rigorous MCC experimentation, the explanation of the behavior proceeded directly through those, embedded in the anatomical pathways that translated the activity of cells in the central nervous system out to muscle output and therefore measurable behavior.

Eventually Silva et al. (2014) provided a more extensive metascientific articulation of how molecular mechanisms account for behavior by defining the properties of “connection experiments” in MCC, which seek to establish causal relations between neurobiological kinds, including molecular mechanisms and single cognitive properties, for example. “Neurobiological kinds” had very broad scope in this context. Connection experiments in MCC could relate molecules, cellular physiology, circuit activities, or behavior. Within this more extensive account, “ruthless reduction” was modified into the typical MCC experimental practices to observe and manipulate hypothesized molecular mechanisms in “negative” and “positive manipulation” experiments that tracked the behavioral effects of molecular mechanisms in the well-known behavioral protocols for specific cognitive functions. Bickle and Kostko (2018) used this broader account to further modify ruthless reductionism, to include sets of experimental practices, well-illustrated by landmark MCC discoveries, by which multiple-experiment research programs are designed such that, if successful, the results of each component experiment will integrate directly with those of the other experiments. The goal of these multi-component research programs is to test activities throughout multiple-component causal pathways ultimately generating the behavior used to operationalize the cognitive function under investigation. Ruthless reductionism now incorporated the ongoing search for additional causal factors in the chains of causes revealed by multiple-experiment research programs, and the typical choice by MCC researchers to focus on components of the neurobiological kinds that connection experiments had already revealed to be parts of these causal chains. These later modifications of ruthless reductionism were also strictly metascientific hypotheses; they too were derived directly out of careful studies of landmark MCC publications, including the experiment designs, reported findings, and

¹ Bickle (2003) introduced this term and has used it since to denote a method for studying particular concepts or notions in science by carefully attending to scientific practices involving that concept or notion, and to the language scientists use to express and characterize it, especially in their experimental publications. The basic idea is to set aside all philosophical (metaphysical, epistemological, normative) assumptions about what that concept or notion is supposed to be or do, and try to describe its use by scientists as adequately as possible. The ‘-science’ component of ‘metascience’ is not intended to be honorific, but rather to distinguish this approach from more traditional normative ways of doing the philosophy of science, e.g., ‘metaphysics of science.’ See Bickle (2022) for his most recent attempt to characterize this method. We thank one of our reviewers for anonymously asking us to clarify this term (Bickle, 2022).

conclusions offered in discussion sections of landmark MCC experimental publications.

No one should deny that the reductionist approach at work over the first 2 decades of MCC has been extremely successful, and extremely valuable toward understanding how different aspects of brain biology contribute to a given cognitive process or behavior. The MCC track record from the early 1990s through the 2010s is a testament to those successes. However, it also seems apparent that even this sophisticated reductionist approach, when supplemented with increasingly sophisticated experiment tools built on the latest technological marvels, can miss important components and processes at work in systems as complex as the mammalian, much less the human brain. Nothing mystical or magical motivates this worry. Reductionist approaches of necessity divide complex phenomena, cognitive or otherwise, into their basic components, be those components molecules, cells, circuits, or behavior. They then carefully manipulate these components *individually*, in a tightly controlled fashion, to investigate that component's causal contribution to the system's behavior. This "decomposition" approach is necessary for exploring whether the factor in question is or is not a part of the causal nexus generating the system's target behaviors. But approaching a system's causal structure in this reductionist fashion inevitably leaves open the possibility that some key components in the complex chains of interacting, interconnected causes might still remain unnoticed or unspecified. This worry is exacerbated by the typically incomplete temporal, cellular, and circuit precision of early MCC tools, which made it difficult to integrate molecular, cellular, and circuit mechanisms in explanations of behavior.

There is also a deeper philosophical worry about reductionism, to which the ruthless reductionism implicit in early MCC experimental practices is not immune. This worry has been stressed in numerous recent publications by Michael Silberstein, Philippe Huneman, and Sara-Lee Green and Robert Betterman, and others. It holds that there are system-level properties that can only be understood by investigating the system "holistically," and which require their own type of nonreductive, emergentist explanations. These authors stress such properties investigated by different sciences, but neural system properties are common to many of their arguments.

If only MCC experimenters could manipulate a specific molecular component of the brain, e.g., some specific gene or molecule, in a specific collection of cells (neurons, glia, etc.) and then observe not only how that manipulation affects the system's overall behavior, but also somehow simultaneously how that manipulation affects other components throughout the system, e.g., activities in cellular networks or brain circuits. And if only these manipulations had the temporal and cellular resolution to permit for the first time, the investigation of the genetic or molecular component at time scales that are compatible with cellular and circuit operations that are relevant for behavioral output. Such an experimental tool might address the worries about even sophisticated reductionist approaches we just sketched. Results from using this new tool might even suggest an alternative

to the traditional categorization of behaving systems in terms of "levels of analysis"; it might offer us a novel "levels-less" image of the inevitable causal interdependence between an organism's behavior, circuit activities, and the cellular and molecular components that make up its brain.

As we will report here, new research tools have become available to do exactly the kinds of manipulation and observation experiments we just described. The increasing use of these tools has led one prominent MCC laboratory to assert the emergence of a new field, "molecular systems neuroscience" (Shen et al., 2022a). These new imaging and manipulation tools and the results they are generating carry considerable implications for the reduction vs. anti-reduction dispute. All discussions of reduction vs. anti-reduction in science occur against a backdrop of some account of distinct "levels": of analysis, of description, of experimental investigation, of mechanisms, or of reality. It is whether the reduction relation holds between components of these distinct "higher" and "lower levels" that reductionists and their opponents dispute. What is especially philosophically intriguing about these new studies in MCC is that they suggest an alternative image to the "levels" picture of the behaving, interacting system; and this alternative, "levels-less" image opens the possibility of sidestepping the reduction versus anti-reduction dispute. To a first approximation, what we find in these recent MCC studies is an image of the organism in its full causal-interactive complexity, with its molecules, cells, and circuits combined into the single system that it is. Our goal in the final section below is to provide a first sketch of this new image, and even to diagram it opposite to the traditional "levels" diagram. Even our initial sketch of this new image suggests that we can transcend the reduction vs. anti-reduction dispute, along with the dreary metaphysical arguments about how entities and processes across "levels" causally interact or relate. The "levels" metaphor ultimately traces back to medieval disputes about "levels of Being," rooted in neo-Platonism and theology. Surely, we can all welcome a 21st century scientific worldview freeing itself from that arcane, ancient scaffolding.

A case study from recent MCC that uses some of these emerging technologies

To present these new experiment tools in some detail, we next describe a recent example of research using them. Shen et al. (2022b) investigated how the mammalian brain links two individual memories acquired close in time to generate a novel mnemonic structure, "linked memories," that support adaptive behaviors. "Mnemonic structures" are theoretical constructs that the brain creates and uses to relate information linked by different dimensions of experience, including time, space, and perceptual or conceptual similarities (de Sousa et al., 2021). Recent studies in rodents had demonstrated that the linking of memories acquired close in time (e.g., 5 h apart) depends on the percentage of

overlapping neurons encoding each memory, with linked memories sharing more encoding neurons than non-linked ones (Cai et al., 2016; Rashid et al., 2016).

One possibility for why memories become linked when acquired in close temporal proximity is the “allocate-to-link” hypothesis, which is based on the observation that after a learning event, neurons involved in memory encoding have, for a period of time, increased activity of the cAMP response element-binding protein (CREB), a gene transcription modulator, and consequently a temporary increase in intrinsic excitability (Silva et al., 2009). Since more excitable neurons are more likely to be allocated for memory encoding (Han et al., 2007; Rogerson et al., 2014; Yiu et al., 2014; Josselyn and Frankland, 2018), subsequent and related events that occur close in time to the first event will have a higher likelihood of engaging neurons that were involved in encoding the first event. This way, two independent memories can become linked *via* their overlapping and shared neuronal ensembles. Future retrieval of one memory will increase the likelihood of retrieving the other due to the reactivation of the neuronal ensembles of both memories. However, while CREB expression and neuronal excitability have been thought to open the window for memory linking, it has not been known whether this window is closed by a passive process, or whether there is an active mechanism that closes the temporal window for memory linking. Shen et al. (2022b) addressed this question by investigating the critical role of C-C chemokine receptor type 5 (CCR5) as a negative regulator of CREB activity and neuronal excitability (Shepherd et al., 2013; Zhou et al., 2016). CCR5 is a major chemokine receptor that had been extensively studied in the context of HIV infection (Brelot and Chakrabarti, 2018). More recently, Zhou et al. (2016) demonstrated the role of this receptor in suppressing CREB signaling and affecting neuronal plasticity following learning, and Joy et al. (2019) then showed that the levels of this receptor can dynamically change in the brain. Given the role of CCR5 as a suppressor of CREB activity and neuronal excitability, and in turn, their involvement in memory linking, the results presented above raised the tantalizing possibility that changes in CCR5 levels or activity following learning may affect memory linking.

Shen et al. (2022b) first replicated the behavioral finding that mice link the memories of two different spatial contexts when the contexts are explored on the same day (e.g., 5 h apart), but not if they are explored on different days (e.g., 2 days apart; Cai et al., 2016). Mice were allowed to explore one novel context (context A) for 10 min, and then, 5 h or 1–7 days later, they explored a second novel context (context B) for 10 min. Two days following the second exploration, mice received a mild foot shock immediately after entering context B and their conditioned response, freezing level, was subsequently measured upon re-exposure to context A, context B, and a novel context. Freezing is an innate response that rodents display when presented with a threatening stimulus that may elicit fear. The animal crouches and remains motionless except for breathing. In this behavioral paradigm, freezing indicates that the mouse formed an association between a

particular context and the aversive foot shock. This paradigm is well established in the MCC field and mice usually develop a conditioned response (freezing) specifically to the context where they received the foot shock and not to other contexts (i.e., mice can discriminate between an aversive context and a neutral one). Interestingly, in the linking experiments described above, mice that had visited both contexts 5 h apart froze for the same amount of time when re-exposed to both contexts A and B, although they had never been shocked in context A. In contrast, mice that visited the two contexts 7 days apart only froze when re-exposed to the context where they received the shock. These results indicate that although mice in the 5-h group had never received a foot shock in context A, they were displaying the same conditioned response observed in context B, as if retrieval of the memory for context A induces the retrieval of the linked memory for context B, where mice were indeed shocked. Importantly, none of the groups displayed significant freezing when exposed to the novel context in the test phase, excluding the possibility of simple fear-to-context generalization.

Shen et al. (2022b) also characterized the expression of CCR5 messenger RNA (mRNA) in the mouse dorsal hippocampus, a brain region involved in memory linking (Cai et al., 2016). They observed that under baseline conditions, most CCR5 mRNA expression is found in microglia cells, with only some limited expression in neurons. However, 6–12 h following a learning event, expression of CCR5 mRNA dramatically increases in dorsal hippocampus neurons, especially in neurons involved in memory encoding, the so called “engram cells,” (Josselyn and Tonegawa, 2020). This increase in the expression of CCR5 mRNA was accompanied by a similar increase in the expression of CCL5 mRNA, one of the ligands of this receptor, suggesting a potential activation of the receptor during this time frame. However, these increases in mRNA levels do not necessarily translate into activity of the CCR5 receptor and classical MCC tools do not allow researchers to track the activity of CCR5 with the temporal and cellular resolution necessary for testing its involvement in memory linking. To address this problem, the authors created a new tool, CCR5-iTango2, to probe the activity of CCR5 receptor *in vivo* during a learning event. This system is based on the iTango tool, a sophisticated molecular system first reported in *Nature Methods* in 2017 (Lee et al., 2017). iTango is a ligand- and light-gated labeling system whereby neurons express a fluorescent reporter protein if the target molecular activity occurs within them while these cells are exposed to blue light. This optogenetic “light switch” insures that the activation observed actually took place within a precise time window marked by blue light activation. This enables the identification of the molecule’s activity in specific populations of cells during specific timepoints using immunohistochemistry techniques. Using this approach, Shen et al. (2022b) demonstrated that CCR5 is indeed highly activated for 6–24 h after the learning event, particularly in dorsal hippocampus neurons involved in memory encoding (“engram neurons”). This experiment is a good illustration of how novel tools in MCC are allowing researchers to probe the activity of molecules with unprecedented cellular and

temporal resolution that is essential to understand their activity in the integrated context of specific cell, circuit, and behavioral activity.

Given the role of the CCR5 receptor in suppressing CREB signaling, and the role of CREB in memory allocation, the authors hypothesized that activation of CCR5 during this period of time could be involved in closing the window for memory linking. To test this possibility, the authors exposed mice to context A and 4 h later infused CCL5 into dorsal hippocampus to activate CCR5 receptors. One hour after infusions, mice were exposed to context B and underwent the memory linking behavioral paradigm as described above. Remarkably, overactivation of CCR5 by CCL5 infusions prevented memory linking without disrupting fear memory for context B, suggesting that this signaling pathway is able to selectively modulate memory linking. Although still informative, this classical MCC approach suffers from low temporal and cellular resolution since ligand infusions can affect molecules for long periods of time and lack cellular or circuit specificity since the ligand can diffuse in the brain and affect multiple circuits in adjacent brain areas. To gain better cellular, circuit, and temporal resolution, the authors built a novel optogenetic tool, Opto-CCR5. With this tool, CCR5 can be activated by simply using blue light. Neuroanatomical analyses can also precisely confirm where CCR5 was activated. A key component of Opto-CCR5 is a receptor protein from the “Opto-XR” family, a group of opsin-receptor chimeric proteins developed through the fusion of a light sensitive receptor protein (rhodopsin) and different G protein-coupled receptors (GPCRs; in this case CCR5; Airan et al., 2009). Using the same memory linking behavioral paradigm, the authors showed that activation of Opto-CCR5 before exposure to context B, 5 h after exposure to context A, led to an impairment of memory linking. Thus, with two very different tools, the authors were able to demonstrate that CCR5 activation is *sufficient* to close the temporal window for memory linking. The new tools are giving much more than additional precision and specificity: they are allowing the design of experiments that explore and test the interactions between molecules, cells, circuits, and behavior in ways that were unthinkable even 10 years ago. This precision and specificity have freed MCC researchers from the previous reductionist logic that implicitly or explicitly dominated the field.

The authors then tested whether CCR5 activity is *necessary* for closing the memory linking window. To this end, the authors used mutant mice engineered to lack the CCR5 or CCL5 genes and they also expressed a short hairpin RNA (sh-RNA) to decrease CCR5 expression in dorsal hippocampal CA1 neurons of wild type mice. shRNA is a bioengineered artificial RNA molecule designed to inhibit the expression of a desired gene. In all three experiments, mice were exposed to context B (and the foot shock) 2 or 7 days after initial exposure to context A, a time frame over which mice do not show memory linking. Remarkably, all three manipulations not only decreased CCR5 or CCL5 expression, but they also dramatically expanded the temporal window for memory linking since the mice with these manipulations froze just as much in the

never-shocked context A as they froze in shocked context B that they saw either 2 or 7 days apart, times when normally mice fail to link memories. These results indicated a critical causal role for the CCR5/CCL5 system in memory linking by demonstrating that increasing or decreasing CCR5 activity directly impairs or extends (respectively) memory linking.

To understand how CCR5 could be affecting cellular and circuit properties relevant for memory linking, Shen et al. (2022b) used miniature head-mounted fluorescent microscopes (miniscopes; Ghosh et al., 2011; Cai et al., 2016), to image the activity of many individual neurons (>300 per animal) in the dorsal CA1 region of the hippocampus, in real time while mice were engaged in the memory linking behavioral paradigm. Specifically, Shen et al. (2022b) monitored the level of intracellular calcium ions (a proxy for neuronal activity) with a genetically-encoded fluorescent molecular reporter (GCaMP6f,) engineered to detect cytoplasmic free calcium ions in activated neurons. With this novel technology, the authors were literally observing neuronal activity (or lack thereof) throughout the dorsal CA1 region of the hippocampus in behaving mice.

Consistent with the hypothesis that the overlap between memory ensembles in the dorsal CA1 regions of the hippocampus determines memory linking (Cai et al., 2016), the authors showed that a manipulation that expanded the temporal window for memory linking (e.g., CCR5 knockout) also expanded the temporal window in which they saw higher overlap between the CA1 memory ensembles for each of the two contexts in the memory linking experiment. The use of miniscopes in these memory linking experiments was crucial since it allowed the authors to determine the active neurons that were present in both memory ensembles (the overlap neurons) with a precision and with time windows (e.g., 7 days) that were simply impossible with previously used MCC technologies, such as with intracranial recording electrodes. The authors further observed that when wild-type mice explore the two contexts 5 h apart, the overlapping activated cells had significantly less CCR5 expression than do non-overlapping cells, indicating that this receptor might be directly modulating the extent of ensemble overlap. To directly test this last hypothesis, the authors used the Opto-CCR5 system to selectively activate the CCR5 pathway with blue light in specific neurons before the mice explored a new context. Using immunohistochemistry techniques, they demonstrated that those neurons with optogenetic activation of CCR5 (neurons with Opto-CCR5) were excluded from memory encoding, a result consistent with the idea that delayed expression of CCR5 in neurons engaged by the first memory excluded these neurons from also participating in the encoding of the second memory, thus closing the temporal windows for memory ensemble overlap and memory linking.

Finally, the authors used neuronal recordings in brain slices to show that increases in CCR5 activity with CCL5 resulted in lower neuronal excitability, a finding that explains why higher activity levels of this receptor cause decreases in memory allocation, and consequently lower ensemble overlap and loss of memory linking.

Together, this impressive set of convergent and consistent findings involving molecular (CCL5/CCR5) cellular (neuronal excitability), circuit (memory allocation; memory ensemble overlap in CA1), and behavioral phenomena (the memory linking paradigm) provide a compelling example of how studies involving multiple entities typically defined at different levels of analyses not only provide a more complete explanation of behavioral phenomena such as memory linking, but they also help to strength the convergent and consistent findings that are at the basis of developing explanations of brain phenomena. Manipulations of CCR5 affected excitability, memory allocation, memory ensemble overlap, and memory linking in a consistent manner. For example, manipulations that increased CCR5 activity decreased neuron excitability, decreased memory allocation, reduced memory ensemble overlap (measured through the miniscopes), and prevented memory linking (measured behaviorally). The CCR5 manipulations not only tested the connections between this receptor and each of these other four phenomena, but they also tested predictions of the allocate-to-link hypothesis that these four phenomena are causally connected (Silva et al., 2009).

In the last section of the paper, Shen et al. (2022b) showed that increases in CCR5 also accounted for the loss of memory linking in middle-aged mice (Cai et al., 2016). Previous results had shown that aging can alter chemokine signaling in the brain (Felzien et al., 2001). The authors first measured the levels of CCR5 and CCL5 mRNA in middle-aged mice at baseline conditions and observed a significant increase in the expression of both genes compared to young mice. Moreover, 3h following learning, middle-aged mice showed a sharp increase in CCL5 expression, which was earlier than the peak observed in young adult mice (6–12h). This observation raised the possibility that an early increase in CCR5 signaling could be responsible for the impairment in memory linking in middle-aged mice. Remarkably, middle-aged mice with a CCR5 knockout were able to link memories encoded 5h apart, indicating that CCR5 signaling could indeed be responsible for the age-related deficits in memory linking. To further test this hypothesis, Shen and collaborators infused Maraviroc, an FDA approved CCR5 antagonist used in the treatment of HIV, into the hippocampus of middle-aged mice and showed that this treatment was sufficient to reverse the loss of memory linking in these mice. These results may have significant clinical implications since Maraviroc is an FDA approved drug and could be used in clinical trials to determine whether it is effective in treating deficits in memory linking associated with aging and psychiatric disorders.

The development of new technologies like miniscopes, iTango2, and Opto-CCR5 is allowing MCC and other neuroscience researchers to transcend constraints imposed by traditional reductionist experimental approaches that in part were imposed by technical limitations of previous approaches. With the increased temporal, cellular and even sub-cellular precision of the new measurement and manipulation techniques, it is now possible to not only test more precisely ideas about the role of molecules in behavior, but also to meaningfully test the impact of specific

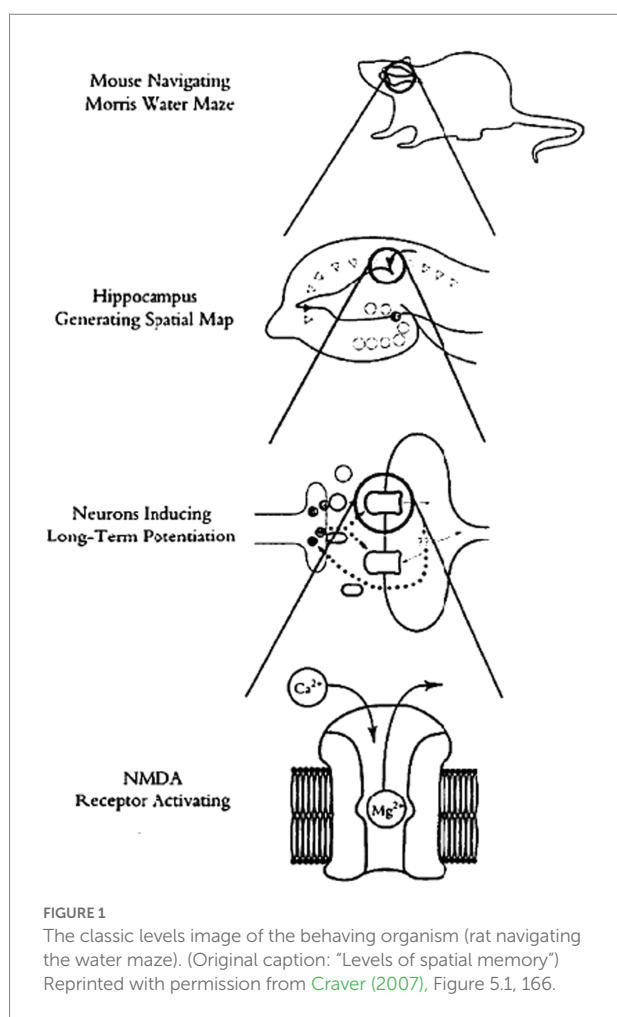
circuit changes caused by those molecular manipulations in behavior in ways that were unthinkable just a few years ago. Thus, these powerful new tools are helping neuroscientists to study how different biological components in the brain (e.g., a specific gene, group of cells, or targeted circuit) interact to generate brain states and behavior.

Here we have summarized the experiments in just a single recent publication from one lab, but they illustrate a number of these new imaging and manipulation research tools that are transforming research in neuroscience. The crucial next step is to both expand and shift the scope of studies, from what are traditionally assumed to be a single level of analyses to studies that are instead focused on how different phenomena classified traditionally at different levels of analyses interact to generate brain states and behavior. This collection of new tools makes such investigations possible. For example, using light-sheet microscopy and immediate early gene expression, researchers can now routinely image the entire brain of a mouse at cellular and even sub-cellular resolution, and thus map neurons across the brain that are active during specific behaviors (DeNardo et al., 2019). Likewise, identification and genetic profiling of neurons with specific roles in behavior (e.g., neurons forming overlapping ensembles in memory linking) open the door for a new type of understanding of the heterogeneity of neuronal populations working together across the brain to generate and modulate behavior. Finally, a large number of recent studies have consistently demonstrated that brain states and behavior not only depend on neurons, but also on a number of other cell types in the brain, including astrocytes, microglia, and oligodendrocytes. Ultimately, approaches that combine brain-wide imaging with single cell (even sub-cellular) resolution, with temporal and site-specific manipulations of molecules, different cell types, and circuits in the behaving animal will be key to understanding how all of these components interact causally to give rise to brain states and behavior.

A “levels-less” image of the behaving organism and a way around all of the philosophical conundrums that the levels image generates?

The new research tools we illustrated in the previous section promise scientific progress. But might the results that stem from their use also revamp our traditional image of the behaving organism and the organization of its interacting components; and thereby lead to philosophical progress as well?

Since the mid-20th century, the reduction vs. anti-reduction dispute has occurred explicitly against a backdrop of a “levels” account of reality, or of scientific inquiry, or of both. For Oppenheim and Putnam (1958, 9), the distinct levels were the universes of discourse of the various branches of science: social groups, multi-cellular living things, cells, molecules, atoms, and



elementary particles. According to Oppenheim and Putnam, the key relationship holding between elements across these levels was micro-reduction of the higher-level elements to those at the next level down. For Ernest Nagel, the levels at work in "heterogeneous" cases of intertheoretic reduction, where the "primary theory" reduces a "secondary theory" even though the latter contains descriptive terms not present in the former, are those of macroscopic phenomena dealt with by the secondary theory and "a microscopic constitution for those macroscopic processes" postulated by the primary theory (1961, 340). "Intertheoretic reduction" for Nagel was deduction, of the logical structure of the secondary (reduced) theory from that of the primary (reducing) theory, supplemented with whatever "conditions of connectability"² were necessary to link descriptive terms of the secondary theory not present in the primary theory to terms of the primary theory, plus whatever limiting assumptions or boundary conditions were necessary on the primary theory owing to its typically broader explanatory

scope. In relatively smooth intertheoretic reductions, the macroscopic entity denoted by some term from the secondary theory (e.g., heat) linked in some cross-theory condition of connectability were deemed identical to the microscopic entities denoted by the related terms from the primary theory (e.g., mean kinetic energy of the system's constituent molecules). Reductionists have always claimed to hold a *prima facie* metaphysical advantage over anti-reductionists, insisting that reduction implies (or at least provides evidence for) cross-level identities. Most anti-reductionists have sought to provide some cross-level relationship logically weaker than reduction, yet not committed to some kind of spooky dualist status for the non-reducible entities or processes.

Most recently, "new mechanists" have introduced an account of "levels" into these discussions that they claim to be less contentious than previous ones. In his comprehensive account of neuroscience from the "new mechanist" perspective, Craver (2007) spends an entire chapter (chapter 5) providing a "field guide" to levels in the philosophy of science. Ultimately, he elaborates and defends a novel "levels of mechanisms" account. The "next level down" from any given target system are that system's components, the individual dynamics of those components, and their organization that generates the system's input-output behavior. The entire system is thus a nested hierarchy of such mechanisms-within-mechanisms, as the mechanisms at the next level down that compose the higher-level system are themselves composed of components, their dynamics, and their organization at the next level down. Figure 1 is much discussed diagram of Craver (2007, Figure 5.1, 166) of the nested hierarchy of mechanisms-within-mechanisms of a rat navigating a water maze. Interestingly, mechanists who share this basic account of levels have differed about whether it is reductionist or not. Bechtel (2009) advocates it explicitly as an account of "mechanistic reductionism"; Craver (2007) remains ambivalent.

Is philosophy of science inevitably stuck with some vexing "levels" concept, in one form or another, and so with the inevitable and seemingly unresolvable disputes between those who insist that higher levels reduce to lower levels, and those who deny reduction and insist on some weaker cross-levels relationship? The recent directions in MCC research we illustrated in our case study in the previous section, guided by the new research tools, suggests a new image of the behaving organism, a "levels-less" one in which distinctions between "levels" need not sidetrack us into solutionless metaphysical disputes. Instead of picturing the behaving organism from its molecular level "up" to its cellular level, then "up" to its circuitry or network level, and finally "up" to the behaving organism level itself, as illustrated in Figure 1, and then wondering how components or ongoing activities at any one of these levels relate to those at others, picture instead scrunching the entire image *down*, "level" *within* "level," into the single interacting system that it is, replete with molecular pathways *in* cells, cells *in* networks, and networks *in* the behaving organism itself. Then start with some activated intracellular molecular

² This was term of Nagel (1961); numerous other names became popular for these conditions: 'bridge laws,' 'correspondence rules,'...

pathway in some central neurons, the usual target of positive and negative manipulations in MCC experiments. Then, move seamlessly from that intraneuronal molecular signaling pathway *outward*, first to the individual neurons whose bilipid membranes encase those manipulated molecular components and pathways. Continue to move seamlessly *outward* from activity in those individual neurons, into the wider cellular circuits they are part of; not only with other neurons, but also with glial, endocrine, immune, and muscle cells—with cells of all other types of tissues with which those neurons form active interacting circuits. Finally, move further seamlessly *outward* to the behaving organism itself, inside the skin of which all of those cellular circuits are active, of that single, marvelously interactive system. Some of the active molecular pathways encased within cellular membranes move molecules across these membranes selectively, into other cells. Those cells combine into circuits that actively communicate with one another, typically *via* molecular exchanges and interactions. Neural circuitries connect with sensory receptors of various sorts, which transform environmental energy of specific kinds into cellular activities, while motor neurons in these circuits communicate directly with muscle tissues to contract muscle fibers against the calcium frames (the bones) to which tendons connect these muscles. All of these components—the interacting molecules, the cells, the circuits, the sensory receptors, the muscles, and bones—are interacting elements of one and only one system, the behaving organism. [Figure 2](#) diagrams this alternate image to the classic “levels” image of [Figure 1](#).

The exquisite temporal and spatial specificities of the new technologies guiding recent MCC research allow researchers to transcend the ruthless reductionism of previous MCC work, since these technologies allow for meaningful integrated studies, simultaneously and in real time, across what tradition clumsily separates as various “levels.” For example, early MCC studies with alpha CaMKII were very specific at the molecular level. They deleted the alpha CaMKII gene without deleting others. But the widespread effects of this single deletion in multiple circuits at different developmental stages made cellular and circuit studies very difficult if not impossible to run. By contrast, the precision and specificity of the new tools used in the CCR5 studies described in section 2 (OptoCCR5, head-mounted miniscopes) complement the use of more traditional MCC tools such as the CCR5 knockout mutant mice. They make analyses and interpretation of circuit properties, such as CA1 neuronal ensemble overlap, measurable in real time, compelling, and meaningful. The use of miniscopes permitted meaningful and long-term imaging of the activation patterns of entire neuronal ensembles in freely behaving animals. “Long-term” is an important addendum here, because before miniscopes were developed, MCC researchers could only observe neuronal activity changes in freely behaving animals (e.g., mice) after specific molecular manipulations by using tools such as intracranial electrodes. Miniscopes permit experimenters to record from more neurons in key circuits and for longer periods of time. These advances are crucially important. With electrodes, one could never even be sure that the same neurons were being

studied as time went on. The molecules of intracellular signaling pathways in specific neurons, the effects on circuit activities to single neuron resolution, and the behaving animal can now be manipulated and monitored simultaneously, in individual experiments. Hence our first attempt to sketch the levels-less image of the behaving organism that these new tools are revealing.

Thinking upwardly in terms of different “levels,” even in terms of seemingly innocuous “levels of mechanisms,” only clouds our emerging capacities to manipulate and track interacting components of the entire system, simultaneously and in real time. Thinking outwardly instead, from the molecular pathways in specific cells, to the cell assemblies those cells are part of, to the brain networks those cell assemblies connect up, and finally to the behaving system itself, which these new precision molecular intervention and circuit-activity imaging technologies now permit, all at once, seems to absolve us of any need to separate the organism into distinct “levels.” The power of this intriguing alternative image of the behaving organism is that all of the philosophical conundrums that the levels-image generates no longer demand answers. Start with the long-standing reduction vs. anti-reduction dispute. Armed with this new image of the behaving organism and its myriad interacting components, we now face no mysteries about how intervening into a specific molecular component in selected neurons can directly affect specific network or circuitry activities in a specific brain region, or even the entire organism’s behavior. Because with the new MCC experiment tools, we can now *observe* these network or circuitry effects directly, and at the same time, we are observing the effects of our molecular manipulations on the organism’s behavior. No multitude of “levels” is being “leaped in a single bound,” from molecules to behaving organism. There is just the single system that is the behaving organism and its myriad interacting components. No elaborate cross-levels metaphysics is needed to “bridge multiple levels,” since the new image suggested by results garnered using these new research tools does not relegate these components into distinct levels.

In the section “Ruthless reductionism guided the first 2 decades of molecular and cellular cognition,” we mentioned two worries that the requisite reductionist focus on single components of the behaving system generates. One was that this focus inevitably leaves out too many possible causes contributing to a complex system’s behavior that have yet to be investigated, especially those of broader systems that the single manipulated component is a part of. The new MCC experiment tools we illustrated in the section “A case study from recent MCC that uses some of these emerging technologies” are tailor-made to investigate contributions of exactly these kinds of circuitry components, simultaneous with the molecular interventions and the standard behavioral measures. The second worry was the challenge that systems level properties can only be explored by investigating the system “holistically,” and requires a special kind of explanation. While it is true that the new network-level imaging technologies like head-mounted miniscopes are a novel addition to MCC research, their use in MCC experiments is just one part

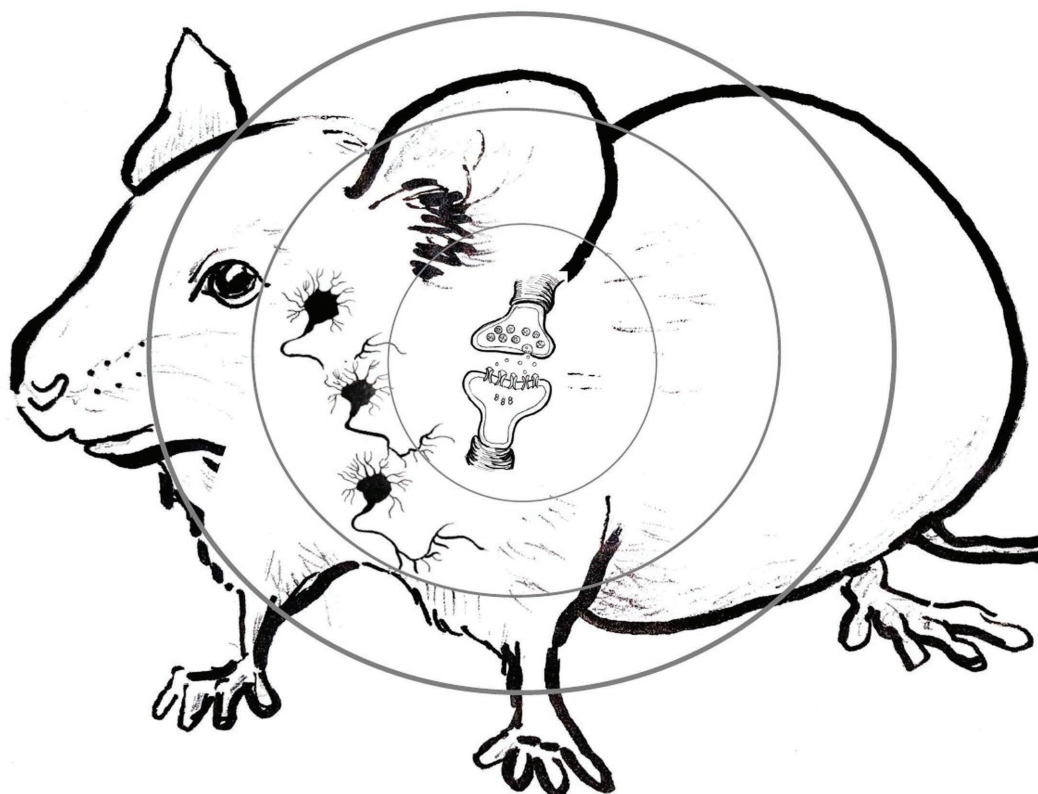


FIGURE 2

The “outwards” levels-less image of the behaving organism suggested by results using the new research tools of molecular and cellular cognition (MCC), with myriad causal relations obtaining between intra- and intercellular molecular pathways (innermost concentric circle), cells and networks of cells (middle concentric circle), and the behaving organism (outermost concentric circle; Original artwork by Caroline Cooper.).

of complex experimental designs that combine these measures in real time with precise molecular manipulations and behavioral measures of cognitive functions. All of these molecular, cellular, network, and behavioral techniques merge together in studies like the one we described in the section “A case study from recent MCC that uses some of these emerging technologies,” in a fashion that no reductionists or anti-reductionists ever previously considered. And now no special explanations are required for the network or system properties, when they are investigated in the fashion illustrated by our example.

As philosophical discussions can be, our initial attempt here to sketch an alternative to the standard levels-account of the behaving organism might seem annoyingly abstract; some might worry that we have construed this new image mostly negatively so far, as not the familiar levels image.³ So let us try to flesh it out further and state its philosophical advantages in more concrete, positive terms. Consider a hypothetical experiment. Suppose we engineer CCR5 knock-out mice, which would normally increase their memory linking capacities. But suppose we couple these mutant mice with

some artificial method for decreasing activity in specific neurons in the dorsal hippocampus to prevent neuronal ensemble overlap. Presumably, we would thereby inhibit memory linking in the manipulated mutants.⁴ Within any levels-framework, even within the innocuous nested-hierarchy-of-mechanisms-within-mechanisms framework (Figure 1), we would seem here to have generated a cellular-level mechanism, namely, our artificial method for decreasing cellular activity and ensemble overlap. That mechanism seems to override our molecular pathways-level mechanism for increasing memory linking, *via* our CCR5 gene knock-out mutation. On the traditional “levels” picture, this combination of experimental manipulations would seem to be an instance

³ As one of our reviewers anonymously worried.

⁴ Something like this experiment has been done. Using amygdala-dependent tone-fear conditioning and optogenetic manipulation of neuronal activation in mouse lateral amygdala, Rashid et al. (2016) suggested that GABA-releasing (inhibitory) parvalbumin interneurons in lateral and basal amygdala inhibit lateral amygdala neurons from encoding tone-shock memories to control ensemble overlap during memory linking. That study seems to be an actual example of a “circuit-level” control of memory linking mechanisms of the sort we are envisioning here.

of a higher-level mechanism causally overriding a lower-level one. And we thereby generate the logical and metaphysical conundrums tied up with “downward causation,” with the higher-level components and activities in their capacities as (or *qua*) higher level components, causally affecting the lower-level processes.⁵ Or we face articulating the problematic cross-level identities that reductionists champion, but which they rarely articulate in any detail (Exactly which activities in which intercellular molecular pathways are our artificial cell-level interventions identical to?). But when we replace the mechanists’ nested hierarchy of mechanisms-within-mechanisms image with the single interacting causal system we are sketching here (Figure 2), these logical and metaphysical mysteries vanish. There are no levels that need to be crossed. The dorsal hippocampus neurons whose activity we envision causally decreasing experimentally contain those CCR5 molecular pathways, so do the circuits of interacting neurons and the behaving organism itself. All components of the system are in ongoing causal flux. No special “downward” causes from higher to lower “levels” of the system’s components are needed; just ordinary causation of the sorts that connection experiments in science routinely provide evidence for.

Finally, what about the anti-reductionist worry that mapping neuron activity across the brain at single-cell resolution provides a uselessly complex data set that cannot be understood without appealing to organizational principles, and with these principles necessarily cashed in terms of distinct levels?⁶ Investigating biological systems using experimental tools that measure or produce narrow changes in specific biological kinds at any given time is a necessary evil in our endeavors to understand the behaving organism in all of its causal complexity. There simply is no absolute way, currently or for the foreseeable future, to capture completely all changes occurring at once inside an organism as it behaves. However, a new capacity to map brain-wide changes of experimentally-induced neuronal activities, against a background image that assumes no intrinsic levels of analysis and no hierarchal streams of causation, has the benefit of not compartmentalizing the changes we can affect, and now observe and measure, as separate from all the other biological kinds that make up the behaving organism. In this sense, the mapping of active neurons within brain circuits while manipulating a step in a molecular cascade in some of those neurons is an excellent example to support our sketch of a level-less image of the behaving organism.

As an example, to map neuron activity across the brain at single-cell resolution, it is now routine to observe the expression of specific immediate early genes as a proxy for individual neuronal activity, because we now know that these genes are only expressed once neurons fire above a certain rate threshold; we know that firing rate induces the changes in those genes’ expression. Using the levels image, this would suggest that a higher-level kind—a neuronal or circuit feature—directly affected changes at a lower level—the mechanisms of immediate early gene expression; thus implying a direction of causation from higher to lower levels. However, one can readily see that in order for a neuron to fire in the first place, it needs activities in the genetic and molecular components that constitute it. Upon a synaptic input, synaptic receptors are activated by neurotransmitters at the postsynaptic terminal, ions flow through ion channels in the cell membrane leading to changes in membrane potential and the neuron may fire an action potential that propagates down its axon. What caused the neuron to fire? The synaptic input? The neurotransmitter receptors? The ion channels? Each of these “lower level” components is reasonably considered to be part of the causal chain that makes the neuron fire. In turn, neuronal firing will change immediate early gene expression, which can change synaptic responses *via* changes in synaptic plasticity and membrane excitability, which in turn might change the way the neuron fires the next time. So, on a levels view, there seems to be a loop of causality between higher and lower levels. These loops pose challenges to reductionist views, but they also saddle anti-reductionists with explaining all of these multi-level causal interactions in a scientifically legitimate way.

Instead, if we understand the entire organism as a single system in which all components interact in a single plane of causal interactions, we can start to appreciate the constant flux of interdependencies that make up all the material components of the brain. Ions, molecules, cells, circuits, and the entire brain are parts of a single construct that we call a living organism. To make reductionist claims about any of these components is to ignore all the magnificent complexity that exists between all these biological kinds. To make anti-reductionist claims invites imputing activities that cannot be cashed out scientifically. The new MCC experiment tools presented here, although far from perfect, allow us to take the first steps toward a new levels-less image of the behaving organism. An ideal future scenario would be able to observe and manipulate each of the different components at timescales relevant for each process, and to understand their joint impact on all the other components. Obviously, we are not there yet. But the new tools MCC researchers now have and the results they are starting to generate suggest the levels-less image of the behaving organism that we provide a first sketch of here. It obviates the need for philosophical accounts of how biological kinds at different levels interact. The behaving animal is a single, complex system of muscles contracting against skeletal frames, receptors being activated, neurons firing, genes being activated or repressed, molecules interacting, even atoms moving from one place to another, with all of these in ongoing causal interactions. But there

⁵ These logical and metaphysical conundrums have not been lost on new mechanists. See Craver and Bechtel (2007), and the logical knots they confront trying to make sense of asserting “top down causation without top down causes.”

⁶ Again, we thank one of our reviewers for anonymously raising this worry for our level-less image.

is nothing mysterious about these components or interactions calling out for special philosophical theorizing. They are just nature's building blocks, interacting in ways optimized by natural selection to promote the evolutionary fitness of the organism. And our levels-less picture better reflects our increasing capacities to both intervene and image activities simultaneously in many of these components than does the traditional levels image.

The levels metaphor entered into Western intellectual discourse from speculations about "levels of being" in medieval theology. Results obtained using the new research tools that are revolutionizing recent MCC research, such as the ones we described in the section "A case study from recent MCC that uses some of these emerging technologies," suggest a new image of the behaving organism and its myriad interacting components that may finally let us lay that antiquated "levels" notion thankfully to rest. With the levels notion also go the many philosophical puzzles it has generated. The reductionism vs. anti-reductionism dispute is one of those puzzles. Does more need to be said to further flesh out this alternate image, and about how results using these new MCC experiment tools contribute to that fleshing out? Absolutely! Ours is only a first attempt to draw out this alternative image from ongoing science. But its promises seem well worth the effort. A useful next step could be an account of how this new image might generate different kinds of experiments that the traditional levels-image obscures. Our hypothesized experiment just above could be a first step toward providing that.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

References

- Abel, T., Nguyen, P. V., Barad, M., Dueul, T. A., Kandel, E. R., and Bourchouladze, R. (1997). Genetic demonstration of a role for PKA in the late phase of LTP and in hippocampus-based long term memory. *Cell* 88, 615–626. doi: 10.1016/S0092-8674(00)81904-2
- Airan, R. D., Thompson, K. R., Fenno, L. E., Bernstein, H., and Deisseroth, K. (2009). Temporally precise in vivo control of intracellular signaling. *Nature* 458, 1025–1029. doi: 10.1038/nature07926
- Bechtel, W. (2009). "Molecules, systems and behavior: another view of memory consolidation," in *The Oxford Handbook of Philosophy and Neuroscience*. ed. J. Bickle (New York: Oxford University Press), 13–39.
- Bickle, J. (2003). *Philosophy and Neuroscience: A Ruthlessly Reductive Account*. Dordrecht: Springer
- Bickle, J. (2006). Reducing mind to molecular pathways: explicating the reductionism implicit in current cellular and molecular neuroscience. *Synthese* 151, 411–434. doi: 10.1007/s11229-006-9015-2
- Bickle, J. (2022). Metascience, not metaphysics of neuroscience. *J. Conscious. Stud.* 29, 175–184. doi: 10.53765/20512201.29.7.175
- Bickle, J., and Kostko, A. (2018). Connection experiments in neurobiology. *Synthese* 195, 5271–5295. doi: 10.1007/s11229-018-1838-0
- Brelot, A., and Chakrabarti, L. A. (2018). CCR5 revisited: how mechanisms of HIV entry govern AIDS pathogenesis. *J. Mol. Biol.* 430, 2557–2589. doi: 10.1016/j.jmb.2018.06.027
- Cai, D. J., Aharoni, D., Shuman, T., Shobe, J., Biane, J., Lou, J., et al. (2016). A shared neural ensemble links distinct contextual memories encoded close in time. *Nature* 534, 1–16. doi: 10.1038/nature17955
- Craver, C.F. (2007). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. New York: Oxford University Press
- Craver, C. F., and Bechtel, W. (2007). Top down causation without top down causes. *Biol. Philos.* 23, 547–563. doi: 10.1007/s10539-006-9028-8
- de Sousa, A. F., Chowdhury, A., and Silva, A. J. (2021). Dimensions and mechanisms of memory organization. *Neuron* 109, 1–14. doi: 10.1016/j.neuron.2021.06.014
- DeNardo, L. A., Liu, C. D., Allen, W. E., Adams, E. L., Friedmann, D., Fu, L., et al. (2019). Temporal evolution of cortical ensembles promoting remote memory retrieval. *Nat. Neurosci.* 22, 460–469. doi: 10.1038/s41593-018-0318-7
- Felzien, L. K., McDonald, J. T., Gleason, S. M., Berman, N. E. J., and Klein, R. M. (2001). Increased chemokine gene expression during aging in the murine brain. *Brain Res.* 890, 137–146. doi: 10.1016/S0006-8993(00)03090-0
- Ghosh, K. K., Burns, L. D., Cocker, E. D., Nimmerjahn, A., Ziv, Y., Gamal, A. E., et al. (2011). Miniaturized integration of a fluorescence microscope. *Nat. Methods* 8, 871–878. doi: 10.1038/nmeth.1694
- Han, J.-H., Kushner, S. A., Yiu, A. P., Cole, C. J., Matynia, A., Brown, R. A., et al. (2007). Neuronal competition and selection during memory formation. *Science* 316, 457–460. doi: 10.1126/science.1139438
- Josselyn, S. A., and Frankland, P. W. (2018). Memory allocation: mechanisms and function. *Annu. Rev. Neurosci.* 41, 389–413. doi: 10.1146/annurev-neuro-080317-061956
- Josselyn, S. A., and Tonegawa, S. (2020). Memory engrams: recalling the past and imagining the future. *Science* 80, 4325. doi: 10.1126/science.aaw4325

Author contributions

ASi and ASo contributed to the research and publication of the scientific results described in the section "A case study from recent MCC that uses some of these emerging technologies." JB, ASo, and ASi contributed equally to the writing of this manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by grants from the NIMH (R01 MH113071), NIA (R01 AG013622), NINDS (R01 NS106969), and from the Miriam and Sheldon G. Adelson Medical Research Foundation to ASi.

Conflict of interest

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

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

- Joy, M. T., Assayag, E. B., Shabashov-Stone, D., Liraz-Zaltsman, S., Mazzitelli, J., Arenas, M., et al. (2019). CCR5 Is a Therapeutic Target for Recovery after Stroke and Traumatic Brain Injury. *Cell* 176, 1143–1157.e13. doi: 10.1016/j.cell.2019.01.044
- Lee, D., Creed, M., Jung, K., Stefanelli, T., Wendler, D. J., Oh, W. C., et al. (2017). Temporally precise labeling and control of neuromodulatory circuits in the mammalian brain. *Nat. Methods* 14, 495–503. doi: 10.1038/nmeth.4234
- Nagel, E. (1961). *The Structure of Science*. New York: Harcourt, Brace and World.
- Oppenheim, P., and Putnam, H. (1958). Unity of Science as a Working Hypothesis, Minnesota Studies in Philosophy of. *Science* 2, 3–36.
- Rashid, A. J., Yan, C., Mercaldo, V., Hsiang, H. L., Park, S., Cole, C. J., et al. (2016). Competition between engrams influences fear memory formation and recall. *Science* 80, 383–388. doi: 10.1126/science.aaf0594
- Rogerson, T., Cai, D. J., Frank, A., Sano, Y., Shobe, J., Lopez-Aranda, M. F., et al. (2014). Synaptic tagging during memory allocation. *Nat. Rev. Neurosci.* 15, 157–169. doi: 10.1038/nrn3667
- Shen, Y., Luchetti, A., Fernandes, G., Do Heo, W., and Silva, A. J. (2022a). The emergence of molecular systems neuroscience. *Mol. Brain* 15, 1–19. doi: 10.1186/s13041-021-00885-5
- Shen, Y., Zhou, M., Cai, D. J., Filho, D. A., Fernandes, G., Cai, Y., et al. (2022b). CCR5 closes the temporal window for memory linking. *Nature*. 606, 1–7. doi: 10.1038/s41586-022-04783-1
- Shepherd, A. J., Loo, L., and Mohapatra, D. P. (2013). Chemokine Co-Receptor CCR5/CXCR4-Dependent Modulation of Kv2.1 Channel Confers Acute Neuroprotection to HIV-1 Glycoprotein gp120 Exposure. *PLoS One* 8:e76698. doi: 10.1371/journal.pone.0076698
- Silva, A. J., Landreth, A., and Bickle, J. (2014). *Engineering the Next Revolution in Neuroscience*. New York: Oxford University Press.
- Silva, A. J., Paylor, R., Wehner, J. M., and Tonegawa, S. (1992b). Impaired spatial learning in alpha-calcium-calmodulin kinase mutant mice. *Science* 257, 206–211. doi: 10.1126/science.1321493
- Silva, A. J., Stevens, C. F., Tonegawa, S., and Wang, Y. (1992a). Deficient hippocampal long-term potentiation in alpha-calcium-calmodulin kinase II mutant mice. *Science* 257, 201–206. doi: 10.1126/science.1378648
- Silva, A. J., Zhou, Y., Rogerson, T., Shobe, J., and Balaji, J. (2009). Molecular and cellular approaches to memory allocation in neural circuits. *Science* 80, 391–395. doi: 10.1126/science.1174519
- Yiu, A. P., Mercaldo, V., Yan, C., Richards, B., Rashid, A. J., Hsiang, H. L. L., et al. (2014). Neurons are recruited to a memory trace based on relative neuronal excitability immediately before training. *Neuron* 83, 722–735. doi: 10.1016/j.neuron.2014.07.017
- Zhou, M., Greenhill, S., Huang, S., Silva, T. K., Sano, Y., Sano, Y., et al. (2016). CCR5 is a suppressor for cortical plasticity and hippocampal learning and memory. *elife* 5, 1–30. doi: 10.7554/eLife.20985



OPEN ACCESS

EDITED BY

Massimo Marraffa,
Roma Tre University, Italy

REVIEWED BY

Błażej Skrzypulec,
Jagiellonian University, Poland
Corey Maley,
University of Kansas, United States

*CORRESPONDENCE

Ken Aizawa
ken.aizawa@gmail.com

SPECIALTY SECTION

This article was submitted to
Cognitive Science,
a section of the journal
Frontiers in Psychology

RECEIVED 03 July 2022

ACCEPTED 30 August 2022

PUBLISHED 20 September 2022

CITATION

Aizawa K (2022) The multiple
realization of human color vision
revisited.
Front. Psychol. 13:985267.
doi: 10.3389/fpsyg.2022.985267

COPYRIGHT

© 2022 Aizawa. This is an open-access
article distributed under the terms of
the [Creative Commons Attribution
License \(CC BY\)](#). The use, distribution
or reproduction in other forums is
permitted, provided the original
author(s) and the copyright owner(s)
are credited and that the original
publication in this journal is cited, in
accordance with accepted academic
practice. No use, distribution or
reproduction is permitted which does
not comply with these terms.

The multiple realization of human color vision revisited

Ken Aizawa*

Department of Philosophy, Rutgers University, Newark, NJ, United States

Over the last 25 years, there has been a concerted effort to settle questions about multiple realization by bringing detailed scientific evidence to bear. Ken Aizawa and Carl Gillett have pursued this scientific approach to multiple realization with a precise theory and applications. This paper reviews the application of the Dimensioned approach to human color vision, addressing objections that have appeared in the literature.

KEYWORDS

realization, multiple realization, Dimensioned realization, human color vision, trichromacy

Over the last 25 years, there has been a concerted effort to settle questions about multiple realization by bringing detailed scientific evidence to bear. [Bechtel and Mundale \(1999\)](#), proposed that scientific work on human brain mapping presented a challenge to multiple realization.¹ [Bickle \(2003\)](#) proposed that the biochemistry of memory consolidation presented a challenge to multiple realization.² [Weiskopf \(2011\)](#) proposed that the visual systems of *Limulus polyphemus* illustrate multiple realization. Many others have investigated multiple realization in the context of evolution by natural selection.³ And even these examples do not exhaust this approach.⁴

Some philosophers of science have pursued this scientific approach to multiple realization with a precise theory and applications. On the theoretical side, [Gillett \(2002, 2003\)](#) proposed a Dimensioned view of realization. [Aizawa and Gillett, 2009a,b](#), added a complementary theory of Dimensioned multiple realization. On the applied side, Aizawa and Gillett have considered a number of examples from neurobiology, the most detailed of which concerns human color vision. [Aizawa and Gillett \(2009a, 2011\)](#) and [Aizawa \(2013, 2020\)](#), proposed that normal human color vision is multiply realized by distinct sets of property instances of the absorption spectra of retinal cone opsins. [Aizawa and Gillett \(2011\)](#) also proposed that normal color vision is multiply realized by distinct sets of property instances of proteins, such as transducin, in the phototransduction biochemical cascade.

1 For a contrary assessment, see [Aizawa \(2009\)](#).

2 For a contrary assessment, see [Aizawa \(2007\)](#).

3 See, for example, [Rosenberg \(2001\)](#) and [Balari and Lorenzo \(2015, 2019\)](#).

4 See, for example, [Batterman \(2000\)](#), [Fang \(2018, 2020\)](#), and [Koskinen \(2019\)](#).

Corresponding to the two-fold character of the Aizawa-Gillett project, one can object to both the theory and its applications. One might argue that the theoretical account does not correctly characterize the compositional relations in science that it is meant to characterize, or one can argue that the examples do not fit the theory.

Polger and Shapiro (2016), presses both kinds of objection. It rejects the Aizawa-Gillett theories of realization and multiple realization and the Aizawa-Gillett conclusion that human trichromatic vision is multiply realized. Against the Aizawa-Gillett account of multiple realization, they claim that “philosophers like Ken Aizawa and Carl Gillett . . . who allow variation of any sort to distinguish between realizations—as little as a difference of a single molecule—are heading down the wrong path” [Polger and Shapiro (2016), p. 62]. Concerning color vision, they write, “the example of variations in human cone opsins does not make for concrete direct evidence of actual multiple realization of a psychological capacity” [Polger and Shapiro (2016), p. 110].

Balari and Lorenzo (2019) offer a bold criticism of Aizawa and Gillett’s applied work. In a section of their paper labeled “The dismissal of scientific practice,” Balari and Lorenzo claim that Aizawa and Gillett “actually ignore scientific practice.” Although they aim their fire at Aizawa and Gillett’s handling of long-term potentiation (LTP), one can easily see how their concern would extend to the discussion of human color vision.

Strappini et al. (2020) make a case for the multiple realization of visual crowding in humans. This is an instance of what is sometimes described as “intraspecific multiple realization,” by which they mean a property that is multiply realized by members of a single biological species. For this case, they embrace the Aizawa-Gillett theory of multiple realization. Moreover, they noted the significance of human color vision as a potential case of intraspecific multiple realization. Despite their sympathies with the Aizawa-Gillett approach to multiple realization, however, they expressed various reservations one might have about concluding that human color vision as treated by scientists is multiply realized.⁵

Given the interest in the Aizawa-Gillett approach to multiple realization and the possible multiple realization of human color vision, it is important to address both theoretical and applied objections that have appeared in the literature. To this end, some guidance is needed about the sometimes complicated features of the Dimensioned framework for realization and multiple realization and their application to some of the relevant science of human color vision. The term “guidance” should be noted. The goal here is not to work once again through all the scientific and theoretical details that have been presented in earlier works (Gillett, 2002, 2003, 2013a, 2016, Aizawa, 2007, 2018a,b, 2020,

Aizawa and Gillett, 2009a, 2011, 2019—perhaps there is no need for *that*—but instead to highlight the principal points that would help interested readers to navigate those details.

Section “Realization and multiple realization: The theories” reviews the Aizawa-Gillett theory of Dimensioned realization and the complementary account of Dimensioned multiple realization. Section “An application of the theories: Human color vision” reviews the application of the theory to human color vision. Section “Critiques of the Aizawa-Gillett theories” will address Polger and Shapiro’s objections to the theories of realization and multiple realization. Section “Critiques of the Aizawa-Gillett application” addresses the multiple critiques of the application of the theory to human color vision.

Realization and multiple realization: The theories

Dimensioned realization and multiple realization have been given extensive and detailed exposition in other works.⁶ The goals of the presentation here, therefore, are more focused. One goal is to provide a simple and accessible presentation of the view, setting aside various details. The second goal is to highlight features of the theory that address objections.

The core idea of Dimensioned Realization is that the relations between properties in scientific explanations are often a species of many-one compositional determination relation—one type of ontological determination relation.⁷ Consider an extremely simple example, the dipole moment of a molecule of hydrogen fluoride (HF). The dipole moment of a molecule is its “charge imbalance.” A HF molecule is more negative on the fluorine side of the molecule than it is on the hydrogen side. The standard scientific explanation of this charge asymmetry is that fluorine is more electronegative than is hydrogen, so that electrons tend to cluster closer to it than to the hydrogen. This makes the fluorine side more negative. Thus, we have a scientific explanation of a property instance of a whole—the dipole moment of a molecule—in terms of property instances of its constituent parts—the electronegativities of its constituent atoms. This example is about as theoretically simple as the Dimensioned view allows.

A more precise account of the matter requires certain complications. First, one wants a theory of the property instances involved. To this end, Aizawa and Gillett rely upon a version of the causal theory of properties according to which properties confer powers upon individuals. Second, property instances stand in this relation only under certain background conditions, the most familiar of which are temperature and

⁶ Gillett (2002, 2003) and Aizawa and Gillett, 2009a,b, 2011

⁷ For a discussion that places this species of explanation alongside others, see Aizawa and Gillett (2019).

⁵ Strappini et al. (2020, p. 8).

pressure. To give an example, a given cone opsin molecule will have a specific absorption spectrum only under a limited range of temperatures. Above a certain temperature, the protein changes its conformational structure, thereby changing its sensitivity to different frequencies of light.

Setting aside many important features, we get this precise schema for a “Dimensioned” account of realization:

Property/relation instance(s) F_1 – F_n realize an instance of a property G , in an individual s under conditions $\$$, *if and only if*, under $\$$, F_1 – F_n together contribute powers, to s or s ’s part(s)/constituent(s), in virtue of which s has powers that are individuating of an instance of G , but not vice versa (Aizawa and Gillett, 2011, p. 202).

In this schema, the inclusion of the contribution of powers reflects the commitment to the causal theory of properties.⁸ Further, the reference to conditions $\$$ reflects the acknowledgment of the role of background conditions.⁹

The core idea of Dimensioned multiple realization is that one must have one set of property instances F_1 – F_n in certain parts that realizes an instance of G in some whole and another non-identical set of property instances F^*_1 – F^*_m that realizes an instance of G , where the parts and whole may be different. Take an example selected for simplicity. Both H_2 and O_2 have no dipole moments. This is because the electronegativity of one atom in the molecule is that same as the electronegativity of the other atom in the molecule. The electronegativity of one atom balances the electronegativity of another. It is easy to see how one might generalize this. Dipole moments are vector quantities. They are directional magnitudes. Any two molecules that have the same vector sum of dipoles among their constituent chemical bonds will have the same dipole moment. Benzene, with its planar symmetric structure, also lacks a dipole moment. The explanans for the dipole moment of benzene will be different than the explanans for the dipole moment of O_2 .

Commentators have sometimes objected to the example of dipole moments. Why, one might ask, do we need the complicated schema Aizawa and Gillett offer (see below) in order to understand such a simple bit of science as the dipole

moment? The answer is that the theory is not meant to illuminate the dipole moment; instead, the dipole moment is supposed to illuminate the schema. Indeed, there is one realized property of having no dipole moment that molecules of both H_2 and O_2 both have. Further, the electronegativities of H and O are different. There is, thus, one realized property and two different realizer properties. This is about as simple as it can get. Examples from psychology are likely to be much more complicated as many more properties will be involved. Further, the example is far from contested territory in cognitive science. Further, the example is scientific and one can, in fact, fit the example into the proposed schema.¹⁰

The core idea of Dimensioned multiple realization is that two sets of property instances F_1 – F_n and F^*_1 – F^*_m realize instances of G . Matters are not, however, that simple. One does not count the realization of, say, pain at the neuronal level and at the biochemical level as multiple realizations of pain. One wants the distinct realizers of G to be at the same scientific level. All of this is captured in the following schema:

A property G is multiply realized if and only if (i) under condition $\$$, an individual s has an instance of property G in virtue of the powers contributed by instances of properties/relations F_1 – F_n to s , or s ’s constituents, but not vice versa; (ii) under condition $\* (which may or may not be identical to $\$$), an individual s^* (which may or may not be identical to s) has an instance of property G in virtue of the powers contributed by instances of properties/relations F^*_1 – F^*_m of s^* or s^* ’s constituents, but not vice versa; (iii) F_1 – $F_n \neq F^*_1$ – F^*_m and (iv), under conditions $\$$ and $\* , F_1 – F_n of s and F^*_1 – F^*_m of s^* are at the same scientific level of properties.

Notice that, since realization is a many-one relation, multiple realization obtains when one set of property instances is distinct from another set of property instances.

An application of the theories: Human color vision

Aizawa and Gillett aspire to providing an account of the compositional relations among properties that scientists postulate. It is an account of relations that scientists implicitly rely upon in providing compositional explanations. Given this

⁸ One important feature this schema understates is how the powers of realized and realizer properties may be qualitatively distinct (see Gillett, 2002).

⁹ Balari and Lorenzo (2019), make a cryptic claim about the schema for Dimensioned realization. They write, Dimensioned realization “does not really go beyond telling us when and that (multiple) realization occurs, but falls short when it comes to explaining why or how some asymmetric dependence exists between entities at different levels” (Balari and Lorenzo, 2019, p. 5). The schema clearly states that asymmetric dependency between the property instances at one level, F_1 – F_n , and the property instance at another level G arises when F_1 – F_n confers properties that are individuating of G . It is the conferring of properties that explains what Balari and Lorenzo ask for. (See also the discussion of Gillett, 2003 in section 2.0 below.) What is missing?

¹⁰ As an aside, the dipole moment example fares well in comparison to Polger and Shapiro’s favorite example of corkscrews. Corkscrews are artifacts. Moreover, the example does not fit Polger and Shapiro’s “Official Recipe”. See Aizawa (2020), but also Polger and Shapiro (2016, p. 67). Polger and Shapiro do not try to fit the corkscrew example into the schema. Instead, they merely gesture at what the fourth condition is supposed to do.

goal, it is important to show that it applies to actual cases. The dipole moment cases show that. But we can show that it also applies to much more complicated cases in psychology, as in vision science.

Aizawa and Gillett choose an example from the science of human color vision as it is *prima facie* a case of Dimensioned realization and Dimensioned multiple realization. They suppose that individual humans have normal color vision and that individual humans have normal color vision in virtue of, among other things, the spectral sensitivities of some of their parts, namely, the opsins contained in retinal cones. The structure of the case is relatively simple. Scientists screen individuals for normal color vision using simple tests, such as the Ishihara test. Individuals making the correct identifications of numerals in the set of plates are deemed to have normal color vision. It is also possible to use genetic tests to determine, for example, whether an individual male has, say, Red(Ser¹⁸⁰) versus Red(Ala¹⁸⁰). (Since color vision is a sex-linked trait, we presuppose a male so that there is only one red cone opsin.) Biophysical measurements reveal that these distinct cone opsins—Red(Ser¹⁸⁰) and Red(Ala¹⁸⁰)—have distinct absorption spectra.¹¹ Their peak sensitivities are somewhat different. Thus, we have the same property realized in the two males, but the two males have different realizers at the biochemical level. This is, in essence, Dimensioned multiple realization.

The preceding part of the story focuses on the cone opsins, but there is another part that focuses on subsequent steps in the biochemical phototransduction pathway.¹² The discussion in Aizawa and Gillett (2011) is complicated, so a simpler presentation is in order. For present purposes, it suffices to focus on the second protein in the phototransduction pathway, a G-protein sometimes called “transducin.” Upon absorption of a photon, a single photopigment molecule will change conformation. After this conformational change, the molecule breaks into two components, a retinal chromophore and an opsin protein. The opsin component binds to a single transducin molecule. This transducin molecule, in turn, activates a molecule of an enzyme, cGMP phosphodiesterase. There are known genetic mutations to transducin.¹³ Such mutations are likely to give rise to differences in property instances in distinct transducin molecules that realize normal color vision, hence give rise to multiple realization. What is important, and underappreciated, about this example is that the differences in transducin properties do not induce individual differences in color discrimination. Differences in transducin property instances are causally downstream from the cone opsins that are differentially sensitive to the frequency of captured photons. If

this analysis is correct, then the putative multiple realization of normal color vision cannot be dismissed on the grounds that differences among realizer property instances induce individual differences among those individuals bearing a multiply realized property.¹⁴

Critiques of the Aizawa-Gillett theories

As noted in the introduction, critics have raised objections to both the theoretical component of the work and to its application. In this section, we begin with objections Polger and Shapiro have raised to the theoretical component of the project.

Aizawa and Gillett assess their account by how well it captures the relation implicit in certain compositional explanations in the sciences. Polger and Shapiro, however, have objected to Dimensioned Realization on a different basis:

An account of realization should discriminate between realization and other dependence relations—other ways that things can be made up.. Moreover, and again in contrast to Gillett, it is informative because it does not posit realization everywhere. Some things are realized, and some are not [Polger and Shapiro (2016), pp. 29–30, cf. p. 28, fn. 14.]

It is easy to see how one might have this objection, as there is no one place in Aizawa and Gillett’s works that specifically addresses it. One must, instead, survey a number of their works for the response to come into focus. The first piece of what has become the Aizawa-Gillett picture—the Dimensioned view of realization—was first broached in Gillett (2002, 2003). Gillett (2013a), adds to this a theory of scientific constitution as another dependence relation alongside realization. This is a theory of the dependence relation between an individual and its parts, as for example the relation between a cell and its organelles. Further, Gillett (2013b, 2016) outlines a theory of implementation that characterizes the dependence relation between the activity of an individual, such as the contraction of a muscle, and the activities of its constituent parts, such as the binding of myosin to actin filaments, the hydrolysis of ATP, and the conformational change of myosin. Aizawa and Gillett (2019) propose that these distinct species of dependency relations figure into distinct species of compositional explanations. Thus, Polger and Shapiro’s contentions notwithstanding, Aizawa and Gillett do discriminate between realization and other dependence relations.

Moreover, Aizawa and Gillett do not posit realization everywhere. For one thing, some things (property instances)

¹¹ See, for example Merbs and Nathans (1992).

¹² For further details, see Aizawa and Gillett (2011), section 10.3.3.

¹³ For a review and entry into this literature (see, e.g., Weinstein et al., 2006). For some reason, Aizawa and Gillett (2011), did not reference this literature.

¹⁴ Cf., e.g., Polger and Shapiro (2016, pp. 66f) and Strappini et al. (2020, p. 8).

are realized, and other things (individuals and activities) are not. These other things are constituted or implemented. For another, Aizawa and Gillett do not think the property instances of microphysics have been shown to be realized (Gillett, 2016). They think that some property instances are realized (i.e., property instances of non-basic individuals of the special sciences) and some are not (i.e., property instances of basic individuals of basic physics).

Turn now to Polger and Shapiro's criticism of the theory of multiple realization. They claim that "On some views, variation of any sort suffices for multiple realization" (Polger and Shapiro, 2016, p. 38) and suggest that Aizawa and Gillett have one of these views. This is not correct. Consider two individual cone opsins, Red(ala¹⁸⁰), a "red" cone opsin with an alanine amino acid at position 180, and Red(ser¹⁸⁰), a "red" cone opsin with a serine amino acid at position 180. These molecules differ in their absorption spectra. They also differ in their polarity, since serine has a hydroxyl group where alanine has only a proton. The differences in absorption spectra are relevant to the multiple realization of normal human color vision because the absorption spectra contribute powers that are individuating of the property of normal human color vision. By contrast, the differences in the polarity are not relevant to the multiple realization of normal human color vision, because the polarity does not contribute powers that are individuating of normal human color vision. Many other properties of Red(ser¹⁸⁰) and Red(ala¹⁸⁰), such as their shape, size, etc., would serve to make the same point.

Gillett (2003) made essentially this point informally even before the formulation of the schema for multiple realization. Gillett, first, proposes that "only properties/relations that result in the powers of the realized property are taken to be relevant to (multiple realization)" (Gillett, 2003, p. 598). As an example, Gillett proposes that the properties/relations of aluminum atoms in one corkscrew, label them F_1-F_n , provide one realization of the property of being a corkscrew, whereas the different properties/relations of steel atoms, label them $F^*_1-F^*_m$, provide a distinct realization of the property of being a corkscrew. Why? Because F_1-F_n and $F^*_1-F^*_m$ both contribute to the same the property or capacity of removing corks, G. It should be emphasized that, consistent with his Dimensioned approach to realization, Gillett does not say that the two corkscrews are multiple realizations of the property of being a corkscrew. Instead, he says that the distinct property instances, $F_1-F_n \neq F^*_1-F^*_m$, of the aluminum and steel are distinct realizations.

Gillett further illustrates the view with a case in which other properties/relations of some of the constituent atoms does *not* lead to multiple realization. He writes,

Do proponents of the dimensioned metaphysics . . . take all differences of composition to be instances of multiple realization? To see that they do not, consider two aluminum

corkscrews that are similar in all other respects except that one is made of aluminum containing a trace element. This element does not chemically bond with the aluminum, or change the metallic structure of aluminum atoms, but it does absorb a certain wavelength of light giving this corkscrew a yellow tinge. The same structure of aluminum atoms is therefore responsible for rigidity in both corkscrews, but there is a trace element in one of them (Gillett, 2003, pp. 598–599).

Gillett's point is that the properties/relations of the atoms of the trace element do not contribute to the second corkscrew's property of/capacity for removing corks, hence that the properties/relations of the atoms of the trace element do not realize the second corkscrew's property of/capacity for removing corks. Thus, the properties/relations of the atoms of these two corkscrews represent only one realization of corkscrew. It should again be emphasized that, consistent with his Dimensioned approach to realization, Gillett does not say that the two corkscrews are a single realization of the property of being a corkscrew. Instead, he says that the numerically distinct properties/relations of the constituent aluminum atoms provide for a single realization of the property of being a corkscrew.

As a separate objection, Polger and Shapiro comment that "It would be odd indeed if the autonomy of psychology from neuroscience could be secured in virtue of tiny differences in potassium atoms" [Polger and Shapiro (2016), p. 39]. This, however, is not the Aizawa-Gillett view. Aizawa and Gillett (2009a) proposed that claims of realization and multiple realization are always indexed to particular levels and specific properties at these levels.

We can quickly see the importance of this point. Suppose that some higher level property G is multiply realized by microphysical properties of fundamental particles and hence multiply realized at the microphysical level. This does not, of course, mean that G is multiply realized in, say, distinct physiological properties (Aizawa and Gillett, 2009a, p. 550).

Applying what Aizawa and Gillett write, one does not get the multiple realization (or autonomy) of psychology from neuroscience by appealing to chemical properties of potassium.

Polger and Shapiro further object that the Aizawa-Gillett approach "entails an undesirable profligacy of distinct realizations for every kind, and undermines the significance of realization within debates over the autonomy of the special sciences" [Polger and Shapiro (2016), p. 39].¹⁵ Aizawa and Gillett (2009a) propose to develop a theory of realization and

¹⁵ Cf., Balari and Lorenzo (2015, p. 883).

multiple realization that is meant to characterize compositional relations in the sciences. Once one has this theory, it is to a first approximation an empirical matter just how much multiple realization there is in the world. One should not judge *a priori* that a form of multiple realization is, or is not, pervasive.

We should perhaps go beyond what Aizawa and Gillett have already written to consider a confusion that seems to underlie Polger and Shapiro's reasoning. Polger and Shapiro do not distinguish two claims. On the one hand, there is the claim that Dimensioned multiple realization is pervasive and, on the other, there is the claim that Dimensioned multiple realization is in some sense unimportant or trivial. Aizawa and Gillett believe that multiple realization is a pervasive feature of the biological world. They explicitly endorse, for example, the massive multiple realization of psychological properties.¹⁶ But, what is the connection between Dimensioned multiple realization being pervasive and Dimensioned multiple realization being trivial or undermining the significance of realization? Polger and Shapiro do not say. They appear not to see the difference between these claims, so perceive no need for an argument.¹⁷ Take an analogy to illustrate the point. Sexual reproduction is a pervasive feature of the biological world but is it not a trivial feature of the world. It is a kind of serious fact about life on earth that evolutionary biologists are very much concerned to understand and explain. Similarly, Aizawa and Gillett take pervasive Dimensioned multiple realization to be a serious fact about the world that merits philosophical attention.

Critiques of the Aizawa-Gillett application

The Polger and Shapiro critique

What reasons do Polger and Shapiro give to challenge the application of Aizawa and Gillett's theory to the science of human color vision? They begin by switching from the property of having normal color vision to the property of trichromacy. They, then, caution that trichromacy might be a behavior or a behavioral capacity. They write,

To say that human beings are trichromats or that normal human color vision is trichromatic is to say that normal human beings exhibit a certain behavioral pattern.

¹⁶ Aizawa and Gillett (2009a, p. 540).

¹⁷ Balari and Lorenzo seem to be making essentially the same mistake in this passage: "assuming a criterion of identity that imposes strict equivalence of form or shape of biological structures, then, given the fact that inter- and intraspecific variation are the norm rather than the exception, multiple realization of any property will be trivially (and vacuously) true." (Balari and Lorenzo, 2015, p. 883).

Trichromacy is the capacity to do a certain task—to match a sample using three primary lights. "Being trichromatic" is more like "being graceful" than it is like "being a vertebrate"; it is a behavior or effect that might have many causes. This makes us hesitant about whether the example of "normal human color vision" is an example of an internal or cognitive process at all, rather than the output of such a process [Polger and Shapiro (2016), p. 107].

This "cautionary note" is entirely misplaced. There are two distinct claims here. First, that trichromacy is a behavior or a behavioral pattern and, second, that it is a behavioral capacity. Let us consider these in order.

It is unclear why they think trichromacy is a behavior or a behavioral pattern. An individual might be a trichromat even in the dark or while sleeping. An individual is not a trichromat at just the time that individual is performing a matching test. Many, perhaps most, individuals who are trichromats never take such tests. Surely almost all the non-human primates that are trichromats never take such tests. As for the idea of a behavioral capacity, one can fathom how they got this idea. Earlier in their discussion they comment, "Normal human color vision is trichromatic, meaning that normally sighted human beings can match almost any color sample by mixing three different "primary" lights (Surridge et al., 2003)." So, let us concede for the sake of argument that there is a trichromatic behavioral capacity which is a capacity to successfully match. We might then ask how an agent has this behavioral capacity. Presumably the agent has the behavioral capacity in virtue, in part, of some visual perceptual capacity. In the typical case, if the agent did not have the visual perceptual capacity, the agent would not have the behavioral capacity. The picture here is the quite familiar one in cognitive science in which behavioral capacities depend on a lot of other capacities, many perceptual and cognitive, acting together. One of the core contentions of the cognitive revolution was that a behavioral capacity to speak a natural language involves a psychological linguistic capacity. The point is that even if there is a trichromatic behavioral capacity that does not show that there is not also a trichromatic visual perceptual capacity. Indeed, in typical cases, the latter would seem to be required for the former. Polger and Shapiro say nothing to undermine this familiar picture.¹⁸

¹⁸ Polger and Shapiro argue that, on their theory of multiple realization, trichromacy is not multiply realized. Strictly speaking, this does not amount to a criticism of the Aizawa-Gillett approach. Readers might, however, want to have some idea of what to make of this. The simplest point is that the Polger-Shapiro view, being a flat view of realization, simply does not fit a Dimensioned case, like that of cone opsins. Consider their claim "(ii): There is a taxonomic system that distinguishes among trichromats—it sorts them by their peak sensitivities, say" [Polger and Shapiro (2016), p. 109]. (ii) is incorrect, as vision science does not distinguish among human trichromats in terms of their peak sensitivities. It distinguishes between cone opsins in terms of their peak sensitivities.

The Balari and Lorenzo critique

Consider, next, what Balari and Lorenzo have to say about the science on which Aizawa and Gillett rely. The drift of their critique is that Aizawa and Gillett ignore scientific practice, because Aizawa and Gillett do not use homology as a standard for similarity and difference. Balari and Lorenzo focus on Aizawa and Gillett's discussion of the biochemistry of memory consolidation, but their discussion could apply just as easily to the biochemistry of human color vision. Here is the crucial passage

“[e]ven homologous proteins will differ to a greater or lesser degree in their amino acid sequences, so that they will differ to a greater or lesser degree in their physico-chemical properties” (Aizawa and Gillett, 2009b, p. 200). These words are illustrative, because they suggest that Aizawa and Gillett are here disregarding what counts as the same or different in the sciences, in molecular biology in this case. They appear not to be at all impressed by the fact that biologists and biochemists consider these proteins homologous and have developed their methods to determine homologies at this (and other) levels (Balari and Lorenzo, 2019, p. 18).

Notice the two dramatically different claims in this passage. The first is the claim that Aizawa and Gillett disregard what counts as the same or different in the sciences. The second is, in essence, the claim that Aizawa and Gillett disregard what counts as the same or different in terms of homology. The first would be problematic, if true. But it is false. The second is true but is unproblematic. There is a reason that Aizawa and Gillett do not adopt the criterion of homology, namely, there are other scientific standards of similarity and difference, those standards are the ones that are used in the portion of vision science under examination, and that is the science that is relevant for understanding the compositional relations in science.

Consider, first, the biochemistry of memory consolidation.¹⁹ In outline, the Aizawa-Gillett claim is that memory consolidation, G, is *probably* multiply realized by one set of property instances, F_1 – F_n , in mice, another set of property instances, F^*_1 – F^*_m , in *Drosophila*, and another set of property instances, F^{**}_1 – F^{**}_l , in *Aplysia*. The argument for this begins with the observation that biochemists have identified distinct proteins, i.e., distinct chains of amino acids, in each of these species. Aizawa (2007), cites scientific work by Bartsch et al. (1998), Bergold et al. (1992), Beushausen et al. (1988), Kalderon and Rubin (1998), and Yin et al. (1994), in support of this view. They next proposed that differences in amino acid sequences are likely to generate differences in the properties of the proteins, thus, *probably* yielding multiple realization.

Clearly, scientists distinguish proteins in terms of their amino acid sequences and distinguish them in terms of the properties, such as their binding constants, that they contribute to memory consolidation. So, it is clearly false to say that Aizawa and Gillett disregard what counts as the same or different in the sciences. The science was previously set out in Aizawa (2007).

Return now to the science of human color vision. In the memory consolidation case, there was an inference from differences in amino acid sequence to a difference in property instances that realize memory consolidation. The experimental work cited did not include direct measurements of, for example, the binding constants of the different proteins involved in LTP.²⁰ Thus, there was, in point of logic, some room for empirical doubt. That was the basis of the italicized qualifier *probably*. The human color vision case addresses that source of empirical doubt. In the human color vision case, vision scientists know both the amino acid sequences and the absorption spectra of the cone opsins. Further, vision scientists know that two individuals with normal color vision can differ in the absorption spectra of their cone opsins. In support of this, Aizawa (2018b, cites Winderickx et al., 1992; Neitz and Neitz, 1998; Sjöberg et al., 1998; Sharpe et al., 1999). Balari and Lorenzo do nothing to square Aizawa and Gillett's use of these scientific facts with the idea that they fail to respect scientific practice. Surely the charges of ignoring scientific facts must be dismissed.

Balari and Lorenzo are correct in noting that Aizawa and Gillett focus on, for example, whether two cone opsins have the same or different absorption spectra, but not on whether two cone opsins are homologous. One reason is that even if one accepts the need for identity criteria based on homology, one also needs identity criteria that are not so based. Clearly scientists recognize that distinct amino acid sequences have distinct properties, such as their absorption spectra.

What explains the connection between distinct amino acid sequences and distinct absorption spectra? Aizawa and Gillett (2019) propose that scientists give Standing Compositional explanations of such things. The absorption spectrum of a given amino acid chain is scientifically explained in terms of the individual amino acids of that chain, their primary sequence, and individual property instances. Scientists have this basic picture—though the complexity of the case makes it typically practically impossible—and Aizawa and Gillett offer a theory of this scientific picture.²¹

The Strappini et al., critique

Like Polger and Shapiro and Balari and Lorenzo, Strappini et al., have doubts about the extent to which the human color

¹⁹ Here we aim for brevity. For more details, consult (Aizawa, 2007; Aizawa and Gillett, 2009a).

²⁰ Recall the discussion of transducin above.

²¹ For further explanation of the work the Aizawa-Gillett theory might do for philosophy on this score, see (Aizawa, 2020).

vision case illustrates multiple realization. Here is the bulk of their critique,

Somehow in line with Polger and Shapiro (2016), we think that in the example provided by Aizawa and Gillett [*sic*], the cognitive property is missing. We do not exclude *a priori* that color perception (or being trichromat) can be considered a psychological property; however, we think that its phenomenology, its behavioral outcome, is missing from the proposal. We further conjecture that this example could provide concrete evidence of multiple realization if the psychological level was added by showing that there are no differences in color perception among trichromats that have those polymorphisms. Indeed, even slight differences among these normal trichromats would exclude that color vision is multiply realized (Strappini et al., 2020, p. 8).

To begin with, there are unclarity in what Strappini et al., are saying in the first part of this passage. What is this “cognitive property” they have in mind. And, “phenomenology” is often understood to be a kind of subjective feel, rather than a behavioral outcome. That, however, does not seem to be the core of their objection. Instead, their substantive claim is that there must be no differences in color perception between individuals with, say, Red(ala¹⁸⁰) and Red(Ser¹⁸⁰).

There is a long-standing idea that multiple realization requires, at the least, that the realizers be distinct and that the realized must be the same. Aizawa and Gillett’s application is meant to respect this. Indeed, it does so in three ways. First, Aizawa and Gillett propose that “normal color vision” as scientists use it in this context focuses on one property, but excludes certain other properties that one might lump under a pedestrian concept of normal color vision or of other scientific conceptions that might be labeled “normal color vision.” It focuses on the ability to make certain visual color discriminations. It excludes, for example, rapidity of response, luminance sensitivity, etc.²² So, there are some differences in color perception that are not included in the concept of “normal color vision” that are in play in this example.

Second, Aizawa and Gillett note that normal color vision, as used in the context, is a property that individual humans may have, even though there are individual differences in color discrimination among those who have this property. The Ishihara test, for example, is widely accepted as screening for normal human color vision, but a more sensitive Rayleigh color matching test is able to detect color matching differences among individuals with distinct cone opsins.²³ Thus, there is a constant property that persists in the face of individual diversity.

Third, there is Aizawa and Gillett’s example of transducin. The crucial points of the example are that (1) property instances of individual transducin molecules realize normal color vision and (2) differences in these property instances do not induce individual differences in normal human color discriminations. See section “Critiques of the Aizawa-Gillett theories” above.²⁴ Thus, the case cannot be dismissed on the grounds that the differences in transducin property instances merely induce individual differences.²⁵

One reviewer has proposed that some further clarifications are required to address Strappini et al.’s objections:

When presenting the Aizawa-Gillett approach, the author adopts a theory of properties according to which properties are individualized by their causal powers.

It seems plausible that the property of normal color vision is, *inter alia*, individualized by powers related to abilities for discriminating between colors. However, as suggested by Strappini et al., people with different absorption spectra of retinal cone opsins differ in abilities for color discrimination. It suggests that normal color vision is not, in fact, a single property, but rather a set of similar properties such that people with different absorption spectra possess different properties from this set. In this case, normal color vision is not a good example of multiply realized property.

The core of the “suggestion” here is that there is no property of normal human color vision, so no property to be multiple realized.

There is a lot to be said to address this, but much of it takes us far afield of the science of human color vision. To begin with, Strappini et al., do not give the foregoing argument. Although they do mention that Aizawa and Gillett are committed to the causal theory of properties, they do not use it as part of an objection to the proposal that there is a property of normal color vision. The closest they come is saying that “We do not exclude *a priori* that color perception (or being trichromat) can be considered a psychological property” (Strappini et al., 2020, p. 8). Second, what reason is there to think that there is no property of human color vision detected by, for example, the Ishihara test, but that there are other properties that are detected by, for example, Rayleigh matching?

The reviewer’s proposal invites emphasizing the importance of the role of transducin G-protein in human color vision. Suppose, simply for the sake of argument, that there is no property of normal color vision just as the reviewer proposes.

²⁴ See also Aizawa and Gillett (2011, section 10.3.3).

²⁵ For the record, Polger and Shapiro (2016) and Strappini et al., seem to think that it is some sort of theory-neutral philosophical “datum” that individual differences cannot or should not give rise to multiple realization. In truth, this assumption is a consequence of Polger and Shapiro’s take on realization as a matter of an individual being a member of a kind. In other words, it is “theory-laden” presupposition.

²² For a more detailed exposition and defense of this point, see (Aizawa and Gillett, 2011, p. 211).

²³ Cf., Aizawa and Gillett (2011, pp. 213–214).

Instead, there are only the “fine grained” properties of individual color discriminations as might be detected through Rayleigh matching. Even those very fine color discriminations will remain the same in the face of differences in the binding properties of transducin. One would not have multiple realization of normal human color vision by different cone opsins, but one would have multiple realization of “fine color discriminations” by instances of the binding constants of transducin. This story bears a lot more attention than it has so far received.

Conclusion

For the last 25 years or so, a group of philosophers of science have tried to resolve questions of realization and multiple realization by closer attention to scientific practice. Aizawa and Gillett have long been a part of this. Over many years, they have developed a detailed theory of realization and multiple realization that is part of a broader account of compositional relations and compositional explanations in the special sciences. Further, they have provided numerous detailed case studies intended to illustrate its ability to account for actual scientific theorizing. The goal of this paper has been to draw together some of principal features of their work to show how the Aizawa-Gillett package of ideas addresses some of the objections that have appeared in the literature.

References

- Aizawa, K. (2007). The biochemistry of memory consolidation: A model system for the philosophy of mind. *Synthese* 155, 65–98. doi: 10.1007/s11229-005-2566-9
- Aizawa, K. (2009). Neuroscience and multiple realization: a reply to Bechtel and Mundale. *Synthese* 167, 493–510. doi: 10.1007/s11229-008-9388-5
- Aizawa, K. (2013). Multiple realization by compensatory differences. *Eur. J. Philos. Sci.* 3, 69–86. doi: 10.1007/s13194-012-0058-6
- Aizawa, K. (2018a). Multiple realization and multiple “ways” of realization: A progress report. *Stud. History Philos. Sci. Part A* 68, 3–9. doi: 10.1016/j.shpsa.2017.11.005
- Aizawa, K. (2018b). “Multiple realization, autonomy, and integration,” in *Explanation and Integration in the Mind and Brain Sciences*, ed. D. M. Kaplan 215–235. doi: 10.1093/oso/9780199685509.003.0010
- Aizawa, K. (2020). The many problems of multiple realization. *Am. Philos. Quart.* 57, 3–16. doi: 10.2307/48570642
- Aizawa, K., and Gillett, C. (2009a). “Levels, individual variation and massive multiple realization in neurobiology,” in *The Oxford Handbook of Philosophy and Neuroscience*, ed. J. Bickle (Oxford: Oxford University Press), 539–581. doi: 10.1093/oxfordhb/9780195304787.003.0023
- Aizawa, K., and Gillett, C. (2009b). The (multiple) realization of psychological and other properties in the sciences. *Mind Lang.* 24, 181–208. doi: 10.1111/j.1468-0017.2008.01359.x
- Aizawa, K., and Gillett, C. (2011). “The autonomy of psychology in the age of neuroscience,” in *Causality in the Sciences*, eds P. M. Illari, F. Russo, and J. Williamson (Oxford: Oxford University Press), 202–223. doi: 10.1093/acprof:oso/9780199574131.003.0010
- Aizawa, K., and Gillett, C. (2019). Defending pluralism about compositional explanations. studies in history and philosophy of science part C. *Stud. History Philos. Biol. Biomed. Sci.* 2019:78. doi: 10.1016/j.shpsc.2019.101202
- Balari, S., and Lorenzo, G. (2015). Ahistorical homology and multiple realizability. *Philos. Psychol.* 28, 881–902. doi: 10.1080/09515089.2014.949004
- Balari, S., and Lorenzo, G. (2019). Realization in biology? *History Philos. Life Sci.* 41, 1–27. doi: 10.1007/s40656-019-0243-4
- Bartsch, D., Casadio, A., Karl, K. A., Serodio, P., and Kandel, E. R. (1998). CREB1 encodes a nuclear activator, a repressor, and a cytoplasmic modulator that form a regulatory unit critical for long-term facilitation. *Cell* 95, 211–223. doi: 10.1016/S0092-8674(00)81752-3
- Batterman, R. W. (2000). Multiple realizability and universality. *Br. J. Philos. Sci.* 51, 115–145. doi: 10.1093/bjps/51.1.115
- Bechtel, W., and Mundale, J. (1999). Multiple realizability revisited: Linking cognitive and neural states. *Philos. Sci.* 66:175. doi: 10.1086/392683
- Bergold, P. J., Beushausen, S. A., Sacktor, T. C., Cheley, S., Bayley, H., and Schwartz, J. H. (1992). A regulatory subunit of the cAMP-dependent protein kinase down-regulated in aplysia sensory neurons during long-term sensitization. *Neuron* 8, 387–397. doi: 10.1016/0896-6273(92)90304-V
- Beushausen, S., Bergold, P., Stürner, S., Elste, A., Roytenberg, V., Schwartz, J. H., et al. (1988). Two catalytic subunits of cAMP-dependent protein kinase generated by alternative RNA splicing are expressed in aplysia neurons. *Neuron* 1, 853–864. doi: 10.1016/0896-6273(88)90133-X
- Bickle, J. (2003). *Philosophy and Neuroscience: A Ruthlessly Reductive Account*. Boston: Kluwer Academic Publishers. doi: 10.1007/978-94-010-0237-0
- Fang, W. (2018). The case for multiple realization in biology. *Biol. Philos.* 33, 1–24. doi: 10.1007/s10539-018-9613-7
- Fang, W. (2020). Multiple realization in systems biology. *Philos. Sci.* 87, 663–684. doi: 10.1086/709733
- Gillett, C. (2002). The dimensions of realization: A critique of the standard view. *Analysis* 62, 316–323. doi: 10.1093/analysis/62.4.316

Author contributions

The author confirms being the sole contributor of this work and has approved it for publication.

Acknowledgments

The author thank to Carl Gillett for the comments on a draft of this manuscript.

Conflict of interest

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

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

- Gillett, C. (2003). The metaphysics of realization, multiple realizability, and the special sciences. *J. Philos.* 100, 591–603.
- Gillett, C. (2013b). “Understanding the sciences through the fog of “functionalism(s)”,” in *Functions: Selection and Mechanisms*, ed. P. Huneman (New York: Springer), 159–181. doi: 10.1007/978-94-007-5304-4_9
- Gillett, C. (2013a). Constitution, and multiple constitution, in the sciences: Using the neuron to construct a starting framework. *Minds Mach.* 23, 1–29. doi: 10.1007/s11023-013-9311-9
- Gillett, C. (2016). *Reduction and Emergence in Science and Philosophy*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9781139871716
- Kalderon, D., and Rubin, G. M. (1998). Isolation and characterization of drosophila cAMP-dependent protein kinase genes. *Genes Dev.* 2, 1539–1556. doi: 10.1101/gad.2.12a.1539
- Koskinen, R. (2019). Multiple realizability as a design heuristic in biological engineering. *Eur. J. Philos. Sci.* 9, 1–15. doi: 10.1007/s13194-018-0243-3
- Merbs, S. L., and Nathans, J. (1992). Absorption spectra of the hybrid pigments responsible for anomalous color vision. *Sci.* 258, 464–466. doi: 10.1126/science.1411542
- Neitz, M., and Neitz, J. (1998). “Molecular genetics and the biological basis of color vision,” in *Color Vision: Perspectives from Different Disciplines*, eds W. G. K. Backhaus, R. Kliegel, and J. S. Werner (Berlin: Walter de Gruyter & Co), 101–119. doi: 10.1515/9783110806984.101
- Polger, T. W., and Shapiro, L. (2016). *The Multiple Realization Book*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780198732891.001.0001
- Rosenberg, A. (2001). On multiple realization and the special sciences. *J. Philos.* 2001, 365–373. doi: 10.2307/2678441
- Sharpe, L. T., Stockman, A., Jägle, H., and Nathans, J. (1999). “Opsin genes, cone photopigments, color vision, and color blindness,” in *Color Vision: From Genes to Perception*, eds K. R. Gegenfurtner and L. T. Sharpe (New York: Walter de Gruyter), 3–51.
- Sjoberg, S. A., Neitz, M., Balding, S. D., and Neitz, J. (1998). L-cone pigment genes expressed in normal colour vision. *Vision Res.* 38, 3213–3219. doi: 10.1016/S0042-6989(97)00367-2
- Strappini, F., Martelli, M., Cozzo, C., and di Pace, E. (2020). Empirical evidence for intraspecific multiple realization? *Front. Psychol.* 11:1676. doi: 10.3389/fpsyg.2020.01676
- Surridge, A., Osorio, D., and Mundy, N. (2003). Evolution and selection of trichromatic vision in primates. *Trends Ecol. Evol.* 18, 198–205. doi: 10.1016/S0169-5347(03)00012-0
- Weinstein, L. S., Chen, M., Xie, T., and Liu, J. (2006). Genetic diseases associated with heterotrimeric G proteins. *Trends Pharmacol. Sci.* 27, 260–266. doi: 10.1016/j.tips.2006.03.005
- Weiskopf, D. A. (2011). The functional unity of special science kinds. *Br. J. Philos. Sci.* 62, 233–258. doi: 10.1093/bjps/axq026
- Winderickx, J., Lindsey, D. T., Sanocki, E., Teller, D. Y., Motulsky, A. G., and Deeb, S. S. (1992). Polymorphism in red photopigment underlies variation in colour matching. *Nature* 356, 431–433. doi: 10.1038/356431a0
- Yin, J. C., Wallach, J., Del Vecchio, M., Wilder, E., Zhou, H., Quinn, W., et al. (1994). Induction of a dominant negative CREB transgene specifically blocks long-term memory in *Drosophila*. *Cell* 79, 49–58. doi: 10.1016/0092-8674(94)90399-9



OPEN ACCESS

EDITED BY

Marialuisa Martelli,
Sapienza University of Rome, Italy

REVIEWED BY

William Sulis,
McMaster University, Canada
Jasmina Mallet,
Assistance Publique Hôpitaux de
Paris, France
Enara García,
University of the Basque
Country, Spain

*CORRESPONDENCE

Christophe Gauld
christophe.gauld@chu-lyon.fr

SPECIALTY SECTION

This article was submitted to
Social Neuroscience,
a section of the journal
Frontiers in Psychiatry

RECEIVED 29 June 2022

ACCEPTED 08 September 2022

PUBLISHED 27 September 2022

CITATION

Gauld C, Nielsen K, Job M,
Bottemanne H and Dumas G (2022)
From analytic to
synthetic-organizational pluralisms: A
pluralistic enactive psychiatry.
Front. Psychiatry 13:981787.
doi: 10.3389/fpsy.2022.981787

COPYRIGHT

© 2022 Gauld, Nielsen, Job,
Bottemanne and Dumas. This is an
open-access article distributed under
the terms of the [Creative Commons
Attribution License \(CC BY\)](#). The use,
distribution or reproduction in other
forums is permitted, provided the
original author(s) and the copyright
owner(s) are credited and that the
original publication in this journal is
cited, in accordance with accepted
academic practice. No use, distribution
or reproduction is permitted which
does not comply with these terms.

From analytic to synthetic-organizational pluralisms: A pluralistic enactive psychiatry

Christophe Gauld^{1,2*}, Kristopher Nielsen³, Manon Job⁴,
Hugo Bottemanne^{5,6,7} and Guillaume Dumas^{8,9}

¹Department of Child Psychiatry, Hospices Civils de Lyon, Grenoble, France, ²Institut des Sciences Cognitives Marc Jeannerod, UMR 5229 CNRS and Université Claude Bernard Lyon 1, Paris, France, ³School of Psychology, Te Herenga Waka - Victoria University of Wellington, Wellington, New Zealand, ⁴Institut Jean Nicod, École Normale Supérieure-EHESS, Paris, France, ⁵Paris Brain Institute - Institut du Cerveau (ICM), Institut National de la Santé et de la Recherche Médicale (INSERM), Center for the National Scientific Research (CNRS), APHP, Pitié-Salpêtrière Hospital, DMU Neuroscience, Sorbonne University, Paris, France, ⁶Department of Psychiatry, Pitié-Salpêtrière Hospital, DMU Neuroscience, Sorbonne University, Assistance Publique-Hôpitaux de Paris (AP-HP), Paris, France, ⁷Department of Philosophy, Sorbonne University, SND Research Unit, Center for the National Scientific Research (CNRS), UMR 8011, Paris, France, ⁸CHU Sainte-Justine Research Center, Department of Psychiatry, Université de Montréal, Montréal, QC, Canada, ⁹Mila - Québec Artificial Intelligence Institute, Université de Montréal, Montréal, QC, Canada

Introduction: Reliance on sole reductionism, whether explanatory, methodological or ontological, is difficult to support in clinical psychiatry. Rather, psychiatry is challenged by a plurality of approaches. There exist multiple legitimate ways of understanding human functionality and disorder, i.e., different systems of representation, different tools, different methodologies and objectives. Pluralistic frameworks have been presented through which the multiplicity of approaches in psychiatry can be understood. In parallel of these frameworks, an enactive approach for psychiatry has been proposed. In this paper, we consider the relationships between the different kinds of pluralistic frameworks and this enactive approach for psychiatry.

Methods: We compare the enactive approach in psychiatry with wider analytical forms of pluralism.

Results: On one side, the enactive framework anchored both in cognitive sciences, theory of dynamic systems, systems biology, and phenomenology, has recently been proposed as an answer to the challenge of an integrative psychiatry. On the other side, two forms of explanatory pluralisms can be described: a non-integrative pluralism and an integrative pluralism. The first is tolerant, it examines the coexistence of different potentially incompatible or untranslatable systems in the scientific or clinical landscape. The second is integrative and proposes to bring together the different levels of understanding and systems of representations. We propose that enactivism is inherently a form of integrative pluralism, but it is at the same time a component of the general framework of explanatory pluralism, composed of a set of so-called analytical approaches.

Conclusions: A significant number of mental health professionals are already accepting the variety of clinical and scientific approaches. In this

way, a rigorous understanding of the theoretical positioning of psychiatric actors seems necessary to promote quality clinical practice. The study of entanglements between an analytical pluralism and a synthetic-organizational enactivist pluralism could prove fruitful.

KEYWORDS

psychiatry, enaction, cognitive science, enaction and embodied cognition, pluralism, neuroscience

Introduction

The field of clinical and scientific psychiatry deals with a vast spectrum of phenomena, from subjective experiences and social dynamics to brain activity and computational models, or psychotherapies and pharmacological treatments. By nature then, psychiatry is a discipline requiring a plurality of explanations and perspectives. Indeed, a patient seen in consultation can be understood in genetic, neurological, cognitive, psychoanalytical, developmental, and/or socio-cultural terms (1). As a result, reliance on sole *reductionism*, whether explanatory, methodological or ontological, is difficult to support in clinical psychiatry. Faced with the multiplicity of perspectives and kinds of explanations observed in clinical practice, conceiving that a single explanation or perspective can summarize the patient and her/his subjectivity is a view largely abandoned in the literature and in the field (2, 3). Anti-reductionist approaches have played an interesting role in deconstructing this conception (4). The notion of emergence in particular allows for a move away from a reductionist perspective. Consider for example the notion of supervenience, developed especially by Jaegwon Kim, which conceives that phenomena at higher scales depend those beneath them, without being reducible to it (5), e.g., a beautiful painting depends on the physical qualities of the paint in that it reflects certain wave-lengths of light, but this dependence says almost nothing about why the painting is beautiful. Such strongly *anti-reductionistic* approaches however, provide few practical answers to understand and support heterogeneous and multi-perspectivist daily clinical practice.

In contrast, the acceptance of all explanations and perspectives without weighing the value of each seems irrelevant. Such a *radical holism* would consider each perspective with the same value as another (e.g., the psychoanalytic explanation explains acoustico-verbal hallucinations as well as the neurobiological explanation) and would require that all these perspectives be systematically considered (e.g., the psychoanalytical approach and the neurobiological approach must be necessarily considered for any psychiatric phenomenon). For these reasons, this kind of holism seems also untenable in clinical psychiatry.

It is in this context that explanatory pluralism has come to be seen as relevant for psychiatry. Explanatory pluralism constitutes a general epistemological framework, an *umbrella term*, under which it is accepted that there are multiple ways to explain the world, and in the case of psychiatry, multiple ways to explain our patients presenting life problems and our ways of treating them. In other words, what is plural under explanatory pluralism are the explanations themselves – we are left free to utilize multiple simultaneous explanations of any type (e.g., mechanistic, causal, dynamic, etc.), drawing from any perspective, system of representation, or scientific/clinical method, so long as it adds sufficient explanatory or pragmatic value. Explanatory pluralism thus allows us to hybridize various clinical practices – e.g., pharmacological treatment and psychotherapy – despite the fact that such practices are often grounded in different kinds of explanations of what is happening to the patient (e.g., neurochemical vs. cognitive-behavioral). As a general epistemological framework, explanatory pluralism allows multiple explanations to co-exist, facilitating a flexible use of various evidence-based/well-reasoned clinical practices even when the underlying explanations may not completely align (while also considering other constraints of practice such as client preferences, ethics, policy, etc.). Such an explanatory pluralism, in which several kinds of explanations are mobilized, e.g., neuroscientific and psychoanalytical, goes hand in hand with: (i) an ontological pluralism, positioned on the metaphysical level and for which there are several objects in the world according to the explanations, e.g., brains and *Dasein* both exist, (ii) and methodological pluralism, in which a variety of tools and treatments are used, e.g., genetic testing and interviewing, psychotherapy, and pharmacotherapy. Explanatory pluralism at the explanatory level can thus be seen as entailing an ontological pluralism and conditions a pluralism at the level of methods (both clinically and in research).

Psychiatry can thus be understood as engaging with explanatory pluralism. Many attempts claiming (more or less explicitly) such explanatory pluralism have been conducted in the history of psychiatry, such as the so-called clinical “integrative” approaches (6). Scientific disciplines as diverse as molecular genetics, biochemistry and neurobiology have been integrated into this explanatory pluralism, e.g., through the Research Domain Criteria (RDoC) project (7). More

recently, computational sciences have grafted themselves onto this dynamic of explanatory pluralism in psychiatry (8). They propose to model a variety of levels of understanding of living organisms (7) within statistical models (such as symptom network models) (9, 10), for instance with Bayesian models (11).

Recognition of this pluralism in psychiatry is sometimes related to the Engel's biopsychosocial model (12), although this relationship was maybe not intended initially (13). The biopsychosocial model was described by Pascal Engel in 1977 as a representation of the human being in which biological, psychological, and social factors are considered to participate simultaneously in the maintenance of health or the development of disease (12). Although adopted by a large number of clinicians, this model has been the target of numerous criticisms since its development, criticisms coming jointly from the philosophy of science and clinical psychiatry. The biopsychosocial model, at least in its initial version (14), constitutes a juxtaposition of three levels of analysis (biological, psychological, and social), randomly chosen and vaguely described according to a systems theory transposed from physics (13, 15). Moreover, in clinical practice and in research, this model is disappointing because it does not give equal weight to these three levels – the American psychiatrist Steven Sharfstein, in his inaugural speech of his presidency of the APA, thus argued that it was in practice a “bio-bio-bio” approach (16). The biopsychosocial model in its initial form has artificially clear boundaries, without any real attempts at integration, causality, or communication between these levels (17, 18). In particular, due to the absence of mutual causality between the levels (19), this model does not consider the first-person experience of the psychiatric phenomena, nor the meaning that individuals give to their existence (or to others and to the world).

A more recent approach for considering the integrated or interwoven nature of causes in psychiatry is the enactive approach (13, 20). Enactivism, not reducible to psychiatry, is a philosophy of mind approach of human functioning rooted in systems biology, dynamical systems theory, cognitive sciences, and phenomenology (21). This approach is based on the idea that cognition is an embodied activity that is *enacted* through the interaction of an organism with its environment. Instead of the generally received hierarchical and brain-centric view whereby chemical structures are organized into neurons and neurons are organized into neural circuits, and these structures, in turn, are seen to ‘process’ the world *via* sensory input, enactivism represents a much more *loopy view* (22). Indeed, under enactivism it is understood that biological objects such as neural circuits do not simply “cause” behaviors, rather they are one part of a wider network of causal chains that simultaneously cause and are caused by behavior. These causal chains are themselves constrained and influenced by other parts of the organism as well as its wider dynamics and intentions. Enactivism focuses on the biologically constituted organism standing in relation to

the world. Thus, to understand behavior, we need to consider the wider brain-body-environment system evolving over time (22–26). Another important facet of enactivism refers to the central role given to meaning and experience. Regarding the first, cognition is the act of making sense of the world (including oneself), often referred to as *sense-making*. Regarding the second, phenomenological experience is not understood as a product of the brain, but as the world from the concerned point of view of a self-maintaining and adaptive organism (26).

Given the complexity that the enactive approach demands to be reckoned with, it has been argued that enactivism entails a kind of pluralism – that under an enactive understanding of psychiatry there should be many different legitimate ways to explain mental disorders (24, 27, 28). However, the kind of explanatory pluralism entailed by enactivism is different to the general explanatory pluralism discussed above, as we will develop. In this article, we will show that the general pluralist framework is a much broader epistemological construction than enactivism. We will argue that they are of different statuses since the first is a general epistemological framework while the second is an approach to conceptualizing human functioning.

Therefore, as a theoretical approach that does not *itself* provide tools for exploring all relevant mechanisms (e.g., neuropsychological), enactivism would be more restricted than explanatory pluralism for the clinical practice of psychiatry. Enactivism would be only one of the approaches contained within a general framework of pluralism, albeit a very useful and integrative one.

In this article, we compare the general framework of explanatory pluralism and the enactivist approach. Although psychiatry can be understood as both a clinical activity and a research activity, in this article, we are focused on psychiatry as a clinical activity. Indeed, we seek to identify a perspective in which the clinician in psychiatry places himself, and more generally any individual who is interested in psychiatry. We question the methodological and pragmatic gain that each of these approaches brings to clinical psychiatry. First, we show the advantages of the enactive approach for clinical psychiatry, by analyzing how it can be conceptually and methodologically used in pedagogy and clinical practice. Secondly, we consider the different kinds of explanatory pluralism applied to psychiatry, detailing its clinical, pedagogical, and theoretical implications. Finally, in the third part, we discuss the issues of the relations between the general and philosophical framework of explanatory pluralism and the enactive approach, in clinical psychiatry. This third part aims to explore the challenges and benefits of crossing an explanatory pluralist framework and the enactive approach. The paper is neutral on the point of whether enactivism should be seen as part of a pluralistic approach or whether pluralistic methods can be understood beneath a wider enactive frame.

The enactive stance

The enactive approach, an embodied cognition

Enactivism is based jointly on phenomenology (a philosophical discipline centered on the analysis of the experience lived by a subject), theory of dynamic systems (a mathematical discipline studying the laws applied to the evolution of a system), and systems biology (a biological discipline seeking to integrate different levels of biological information to understand the functioning of an organism). It seeks to provide an approach for understanding human behavior and subjective phenomena, such as belief or perception, based on a set of principles which we will review in the following section.

The enactivist approach is initially based on the idea of *autopoiesis* (21, 29, 30), an observation within cellular biology that cells represent self-maintaining and adaptive, operationally-closed systems (31), capable of coupling and changing through the interaction with their environment. In *The Embodied Mind* published in 1991, Francisco Varela, Evan Thompson, and Eleanor Rosch (21) hypothesized that this concept of autopoiesis was a fitting analogy for the mind and could be used to ground a new approach to philosophy of mind and the mind sciences.

These authors sought to move away from an understanding of the mind grounded in a metaphor of computation and representation, and instead understand the “mind” through analogy to life forms, especially notions of biological autonomy and coupling. Under this approach, they proposed cognition is a relational process that is *enacted* through the *embodied* interaction of an organism *embedded* in the world. This formulation comes in response to the questions provoked by the growing explanatory gap between cognitive sciences and phenomenology, the former often finding themselves unable to transcribe, explain, or represent the subjective reality experienced in the first person by an individual. A branch of phenomenology known as neurophenomenology is related to but differs from the enactive approach and seeks above all to address the hard problem of consciousness at the crossroads of neuropsychology, neuro-anthropology and behavioral neuroscience. The enactive approach recognizes that cognitive activities are carried out by organisms in constant interaction with their environment. This assumption ensures that individuals and environments continually co-construct each other, the action of the former drastically influencing the nature of the latter, and vice versa. This formulation contrasts with the dated traditional cognitivist view according to which the brain forms a fixed and immutable representative cartography, i.e., an internal model which would replicate the world, as a mirror of sensory reality (32). For instance, enactivism sees perception as a (potentially predictive) activity in its own right, generating meaning through interaction with

the environment – rather than a matter of static internal representation of the external world (33).

The 4Es approach

Enactivism has led to four important principles concerning the nature of human functioning and the mind. These are that cognition is: (1) embodied, (2) embedded, (3) enacted, (4) and emotive (21, 32, 34). This “4E” approach essentially constitutes a modern iteration on enactivism (35, 36). The landscape and the philosophical and scientific communities around enactivism and 4E approaches are complex, in particular because the “4E” approach is not synonymous with enactivism despite much crosstalk, and because proponents often incorporate or exclude different ‘e’ principles when using the overarching banner [e.g., Clark and Chalmers (37), proponents of ‘extension’ of mind – an alternative ‘e’ principle – claim to be part of the 4E approach but not to the enactive community]. It is beyond the scope of this paper to disentangle these various approaches fully. For now, it is important to note that the theses of all 4E positions overlap significantly, as do the ‘e’ principles themselves, and that as we will describe later enactivism is the most integrative of the 4Es. In psychiatry, this “4E” approach has proven valuable in understanding the mechanistic and phenomenological processes involved in psychiatric disorders. We now will briefly detail this “4E” approach in regard to psychiatric disorders (34, 38, 39).

First, the *embodied* dimension of psychiatric disorders recognizes that physical, temporal, and social embodiment in one’s environment is what makes experience possible. For Gallagher (32) and Thompson (35), there is an inseparable relationship between sensation, action and environment: cognitive systems embody a dynamic sensorimotor loop. For instance, individuals move depending on their feelings, and their feelings depend on how they move (35). The physical body (e.g., sensations or sensitivity to negative events), and the subjectively experienced body are co-components. They should be considered simultaneously in the exploration of psychiatric disorders. This enactive principle could also be one of the foundations of contemporary computational theories, in particular in active inference (40).

The *embedded* dimension of psychiatric disorders means that individuals are contextually situated in their environment. An embedded approach to psychiatry means that each clinical situation should be based on the patient’s context and how the client makes sense of this context (41). A patient’s life experience cannot be dissociated from the environment in which her/his experience takes place. In this perspective, the expression of the paranoid delusion of a patient about his next-door neighbor can only be understood through the understanding and analysis of his home and his daily living conditions. In other words, manifestations of the disorders depend on the meaning given by the patient to her/his experience, and they can never be

sufficiently described out of their cultural and social context (38, 42).

An *enactive* understanding of psychiatric disorders means that cognition is not understood to occur solely ‘inside the head’ or to involve representing a pre-given world as accurately as possible. Rather, cognition – or *sense-making* – is understood as an active process, constantly unfolding as someone explores and makes sense of their environment. Under enactivism, all living systems are sense-making systems. They are autonomous, adaptive, and they regulate their own activity and exchanges with the environment, in accordance with their needs – be this a basic metabolic need for a food source or a deeply held socio-cultural value. For instance, pathological social anxiety typically represents complex feedback between attentional and behavioral engagement with social situations, heightened autonomic response, and the subsequent over-estimation of the threat of negative social evaluation, to the point that the individual struggles to source important needs from their social environment. In other words the way that someone is making sense of social situations is not helping them meet their important needs. Mental disorders are so often about something going wrong in the way we make sense of the world (42). We will therefore return extensively to this notion of sense-making.

The *emotional* dimension of psychiatric disorders under a 4E or enactive approach considers affective states as an embodied and enactive mode of evaluation by which the patient engages with and gives meaning to the world (including his/her disorder) (43). Emotions as a felt sense are seen to facilitate actions that have been adaptive or otherwise rewarded in the evolutionary or developmental past. This is congruous with but runs deeper than talk of emotions as ‘tools’ for engaging with and making sense of the environment through emotional states (44, 45). For example, within evolutionary psychology, emotions are often considered as processes allowing the survival of an organism in front of a threat (46). It also roughly accords with the various different theories bearing on the definition of emotions, understood either as physiological changes for authors like William James (47), or according to the cognitive appraisals of a situation for authors like Walter Cannon (48), or as functions for processing information from the environment, for authors like Stanley Schachter (49). Enactive approaches, however, see emotions/affectivity as more than just ‘tools’ that are added on top of our cognition, instead viewing cognition as thoroughly affective in nature. Giovanna Colombetti describes a *primordial affectivity*, an essential dimension of our embodied existence and not a contingent phenomenon of the mind. This affectivity would be the condition of the possibility of more specific affective states such as emotions and moods, and it is through the enactive approach that a meaning is conferred on this affectivity (50). Enactive versions of emotion are always intertwined and inseparable from experience: during an episode of paranoid-themed delirium, the person *feels* constantly threatened and emotions facilitate responses to this threat.

The enactive approach has been applied to many scientific fields in recent decades [e.g., (51–54)]. Only very recently has it been applied as a comprehensive approach to understanding clinical psychiatry (23, 42, 55). We will then detail an aspect of the enactive approach important for clinical practice: the notion of sense-making.

Sense-making

Enaction is indeed totally applicable to clinical practice with respect to sense-making in the patient-clinician dyad (42). Psychiatric disorders are deeply entangled with values and norms (39). In this context, one of the central concepts of the enactive approach corresponds to the notion of sense-making (35, 56, 57). Sense-making corresponds to the fact that the patient, embedded in their environment, gives meaning to this environment in order to maintain their life and identity, and the alternation or loss of this sense-making is one of the keys to understanding psychiatric disorders under the enactive frame (58).

The notion of sense-making corresponds to the diversity of understandings and engagements with the world across organisms, and that how a particular organism makes sense of and engages with the world depends upon on the structure, capacities, needs, and values of the organism, as well the environment itself. For example, sense-making accounts for the fact that a substance attracts the consumer thanks to the addictive characteristics of a substance and thanks to the individual characteristics of the consumer and their history (e.g., genetics and behavioral reinforcement) (59). Similarly, postpartum blues (non-pathological) constitutes a reaction deemed normal due to a set of biological, physiological, environmental, and cultural characteristics related to a particular context, i.e., the demanding reorganization of meaning and experience in response to the appearance of new concerns related to the newborn and navigating this reorganization in light of ones culturally informed expectations regarding motherhood. In other words, experiencing some emotional turmoil or flattened mood after going through pregnancy, birth, and the sudden demand to reorganize your life around a dependent other is fairly understandable and a normal part of the process. Conversely, postpartum depression (pathological) hinders the patient’s relationship to the world and to their new child: the meaning that the patient gives to the world no longer corresponds to her/his needs and values, but to a system of meaning characteristic of depression to the point that this becomes a problem for mother and baby (60). The agent is always situated within a world of meanings. However, psychiatric disorders demonstrate by contrast the loss or significant alternation of this meaning, resulting in a dysfunctional engagement with the world.

This notion of sense-making highlights also well how the enactive approach does not simply take a third person perspective where by people and the psychiatric challenges they face are simply objects of study. Rather enactivism and the notion of sense-making explicitly invites first and second person perspectives (61, 62). These intersubjective or second-person perspectives necessitate that clinical decision making should be informed not only by clinical and scientific standards, but also if reference to the cultural background, habits, beliefs and preferences of the patient. As we will discuss in the later section dealing with the limits of pluralism, such first and second person perspectives are missed within many approaches to explanatory pluralism.

Limits of the enactive approach

One of the postulates of enactivism is that behavior is the product of complex and irreducible causal interactions across multiple scales of enquiry. This does not mean that neurobiological, behavioral or social explanations are confused or claimed not to happen. In the enactive approach, a distinction is still made, and labeling is always possible between objects and processes at different scales (e.g., genes, proteins, dendritic spine density, political parties, and cultural processes all exist). However, enactivism by itself as a perspective from philosophy of mind, does not contain the conceptual tools to analyze such processes and objects. In this way, it does not itself sufficiently account for the mechanisms and material relationships that constitute multi-scale autonomous systems to provide a pragmatic framework for psychiatry (63). Indeed, instead of explaining psychiatric phenomena in terms of mechanism, many enactive approaches seek to explain these phenomena in terms of closed networks of self-sustaining constraints (37). Such holism would seem to make it difficult to provide causal explanations of phenomena fit for psychiatry's purposes. In other words, the enactive approach to psychiatry is predominantly theoretical/conceptual in nature. To use Varela's term, enactivism constitutes a research ethics (64). As such, the 'nuts and bolts' required for the modeling of many important aspects of psychiatric disorders are missing from enactivism itself.

In sum, enactivism sits primarily as a theory of cognition/mind (65) that does not itself provide the tools to study the mechanisms of distress/dysfunction at different levels of analysis relevant to living organisms. Such tools are necessary parts of explanatory research and clinical practice in psychiatry. Simply put, a strictly enactive or 4E approach is not enough by itself. An enactive approach to psychiatry should therefore be open to other perspectives or systems of representation. It should in other words be either itself pluralistic or be used as one way of understanding within a wider pluralistic approach.

Explanatory pluralisms in psychiatry

Definition of explanatory pluralism

When we speak about explanatory pluralism in psychiatry, we are referring to the simultaneous acceptance of multiple different perspectives and ways to explain mental disorders and their constituent phenomena. These perspectives may be targeted at or across any level/scale of enquiry and represent and conceptualize disorders in different ways. In psychiatry, the existence of multiple representations leads to considering different levels of understanding of life and functioning, ranging from a biological perspective to a social perspective. In the view of scientific democracy, explanatory pluralism encourages considering a set of intersecting perspectives to understand the patient. This consideration of a variety of perspectives raises the question of their integration (66–68). In other words, can we (or should we) integrate different perspectives (e.g., neurobiology, psychoanalysis, behaviorist, computational, systemic, phenomenological, sociological or anthropological approaches)? In order to answer this question, different kinds of pluralisms have been developed. Such a typology of pluralisms distinguishes non-integrative from integrative pluralisms. This will now be discussed. For clarity, we are interested here in explanatory (or "epistemological") pluralism, which differs from an ontological pluralism that we will not discuss further.

Non-integrative pluralisms

Non-integrative pluralism seeks to understand how different potentially incompatible or untranslatable levels of understanding, perspectives or systems of representations can coexist in a scientific or clinical landscape. It does not seek to bring together or link the different perspectives of psychiatry. For instance, it aims to question how several perspectives or levels of understanding can coexist in clinical practice, without being translated one vis-à-vis the other. At least two types of non-integrative pluralisms have been proposed: tolerant non-integrative pluralism and dappled non-integrative pluralism.

First, tolerant non-integrative pluralism has been defended by authors such as Longino (69) or Mitchell (70), with a view to promoting a division of labor between disciplines. This division would allow avoiding any form of scientific imperialism, i.e., the predominance of one perspective over the others. Tolerant pluralism considers that the choice of one perspective over another depends on the question asked and the answer expected (71–73). The choice of a neurobiological perspective can be relevant to guiding the initiation of a pharmacological treatment; the choice of a psychodynamic perspective can be relevant to understanding family dynamics in an adolescent. The relevant perspectives or level of explanation would thus depend on

the epistemic and pragmatic interests of the researcher and the clinician.

The second type of non-integrative pluralism is a dappled one (74). Under dappled non-integrative pluralism, each explanation is seen to explain different aspects of the wider reality, like paint dappled on different parts of a canvas gradually revealing a wider picture. One way to explain this is in regard to the 'laws' of scientific disciplines (for accuracies sake we should specify that given human behavior and dysfunction is rarely if ever law-like, disciplines such as psychology and psychiatry generally utilize general rules or generalized models rather than postulating laws). Laws/rules/models generally belong to certain scientific fields and apply only to these fields but looking across multiple scientific domains gives us the richest view of reality. In this way, any particular disciplines' set of laws/rules/models describes one spot of the dappled landscape of reality. Applied to psychiatry, some neurobiological rules may explain certain psychiatric phenomena, and some behaviorist rules or cognitive models help explain others. This patchwork of rules and models certainly leads to apparent disunity in the discipline, but also makes it a strength in the consideration of such a pluralism. Indeed, the scientist or the clinician can then choose the rules which best correspond to her/his objectives, in an opportunistic way. She/he can use a set of rules according to her/his will, her/his medical culture, his/her intuitions, her/his expertise, her/his relationship to risk and uncertainty, or even the institutional and social pressures exerted on him/her (75, 76).

How many different groups of rules are there? Some authors claim that this number is limited, in particular, because of the limited number of "styles" for doing science (77). Thus, only seven styles could sum up all of the past and present sciences: a mathematical style, including the geometric style and the combinatorial style, a laboratory-style (of instruments, of the creation of phenomena, of measurement), a Galilean style (of hypothetical modeling), a taxonomic style, a style of probabilities and statistical style, a "historico-genetic" style (as in geology, philology or psychoanalysis), and an experimental style (77).

In short, non-integrative pluralism recognizes that choosing one perspective on the world does not reduce the possibility of choosing others. Rather, the choice of one perspective is secondary to the consideration of all perspectives, and one is free to utilize multiple perspectives or system of representation, so long as doing so adds epistemic and/or practical value.

Integrative pluralisms

In order to deal with the variety of representations in psychiatry, another form of pluralism has been proposed: integrative pluralism. This pluralism proposes the development of a general framework bringing together the different levels of understanding, perspectives, systems of representations,

their tools and their objectives (78). Therefore, integrative pluralism aims to study how one of these levels or system can be translated into another. Unlike non-integrative pluralism, integrative pluralism does not deal with the question relative to the researcher or clinician (tolerant non-integrative pluralism) and does not consider the existence of different groups of laws (dappled non-integrative pluralism). Within integrative pluralism, for a given psychiatric disorder, there is a concentration of certain perspectives or levels of understanding (e.g., neurobiology or social influences) that can best explain the production of given clinical manifestations. Thus, the understanding of psychiatric disorders is disseminated over several levels of understanding or perspectives (74, 79).

For instance, the levels of understanding that explain the manifestations of the spectrum of schizophrenia (or even more in the case of a genetic syndrome with psychiatric expression, such as Williams syndrome) rather belong to the biological domain. Conversely, the levels of understanding that explain major depressive disorder tend to belong to the psychological (such as ruminations that maintain mood sadness) or environmental (such as detrimental social factors) domains. Finally, the manifestations of substance use disorder are better explained by all of the interacting levels: for example, in alcohol use disorder, we find levels of explanation ranging from biology (genetic variants influence ethanol metabolism), cultural factors (norms regarding alcohol consumption), psychological (certain personality traits), and social (peer availability and use) explanations (79). These levels are neither necessary nor sufficient: they influence the statistical probability of the presence of the disorder in a given individual. Environment influences gene expression and biological manifestations, and vice versa. Because of these mutual influences, such a pluralistic framework can be modeled in the form of *patterns* testifying to the conditional independence between heterogeneous variables, within symptom networks (10, 80, 81).

Among these characteristics, four factors characterize integrative pluralism: (1) the need for interdisciplinary practice in order to conceive and analyze the levels of explanation and perspectives; (2) the implication of synchronicity of the different explanations (they occur in one or more time intervals); (3) the non-exclusivity of these levels and perspectives; (4) a degree of cumulativeness (82). This last factor is particularly important because it refers to the fact that the perspectives and levels of explanation tend to accumulate over the development of psychiatry: there is no replacement of one by another, but a widening of the palette of perspectives available to clinicians and scientists (67).

Limits of explanatory pluralism

In clinical practice and in research, it is possible to adopt a non-integrative pluralism to answer different questions,

according to the needs of clinicians and researchers. An integrative pluralism, considering the entanglement of different levels of explanations, could also be interesting. However, the general framework of pluralism has limitations.

First, in its application in clinical practice, the pluralist framework is only used in a fragmented way. Such fragmentation could be partly related to the complexity and heaviness of the use of pluralism. Indeed, it involves the knowledge and manipulation of a huge corpus, almost impossible to acquire, and absolutely impossible to manage on a daily basis. In clinical practice, the use of a plurality of practices, selected according to contexts, questions and patients, requires mastering each of these practices and to know how to apply them precisely. Being loosely mastered and defined, clinicians find it difficult to apply and teach such perspectives in their entirety. Thus, clinicians can hide a certain wooliness behind their “pluralist” label, which could be just a banner made up of the perspectives it incorporates. Without the study of these perspectives, pluralism is weak. For example, often “pluralists” are not specialists in enactivism *and* phenomenology *and* biology *and* social psychiatry, etc. They are *philosophically* or *conceptually* pluralists and there are practical limitations on the breadth and depth of any clinician’s knowledge and skills. Thus, clinicians cannot be ‘perfectly pluralistic’ in practice, in the sense of grasping all possible scales and ways of understanding. This impossibility is sometimes managed by a simplification of the complexity at hand, which can ultimately lead to a form of managed reductionism (83). An enactive approach, meanwhile, is (or at least should be) open to multiple scales and ways of understanding yet demands that the resulting explanations be placed in the context of the embodied and meaning-experiencing organism standing in relation to its environment. Hence our interest in advocating both pluralism *and* enactivism, as we will do in the third section.

Secondly, contrary in particular to the enactive approach, first person experiences are often not directly considered in pluralistic frameworks. When conceiving of a pluralistic approach to psychiatry, it is common (but not necessary) to do so using the structure of traditional levels of enquiry (i.e., chemical, genetic, cellular, organs, and so on). However, such a conceptualization often side steps first person experience. Similarly, pluralistic frameworks often struggle to, or otherwise miss, consideration of what de Haan (42) refers to as the existential dimension of mental disorders (84, 85). For instance, when applying pluralism to major depressive disorder, there is a tendency to separate patients’ sadness or anhedonia into two domains (biological or psychosocial), three domains (biological, psychological or social), or four domains (biological, psychological, social and phenomenological). However, even when phenomenological analysis is incorporated into a pluralistic approach, it is often seen as adjunct and purely descriptive, artificially separated from the other ‘domains’ rather than intimately related with them (20, 42). In other words,

even when it is addressed, a patient’s personal experience of hearing and feeling his/her life is often seen as only one level of description in this framework, and one with little causative power. In sum, pluralistic frameworks often do not do justice to the subjective experiences of patients (4).

Similarly, the clinical application of pluralism is often deeply dualistic (86). This duality leads to a separation between the pluralist model of the clinician and the experiential model of the patient. The clinician’s pluralist model can break down and localize the prejudices experienced by the patient (87), ultimately providing the patient with overly naturalistic (i.e., referring to possible cerebral dysfunctions) or overly normativist (i.e., referring to the failure of the patient’s values) explanations. Value-Based Psychiatry provides a recent example of this duality (88).

The relationship between enactivism and explanatory pluralism

Analytical and synthetic-organizational pluralisms

Based on the discussion so far, we argue that one can simultaneously take a pluralist and enactive stance on psychiatry. This can be true in the sense that an enactive approach can be one component of a wider pluralist framework, and in the sense that (as argued by co-author KN’s wider work) an enactive approach to the *conceptualization* of mental disorder can demand and incorporate a plurality of explanatory approaches (55). Enactivism can be seen as one perspective within a wider pluralistic framework, or pluralistic methods and ways of understanding can be understood beneath a wider enactive frame. Whichever way we conceive of it, nothing appears to impede this integration. Part of the originality of this article is to go further than a merely *anti-reductionist* proposal. By incorporating explanatory pluralism *and* an enactive approach, we suggest that reductionist explanations or ways of understanding can be resituated within an enactive understanding of human functioning as complex, dynamic, thoroughly affective/meaning-involving, and self-determined/operationally-closed. In this way we suggest that the utilization of both enactivism *and* pluralism, may allow for clinicians to maintain an awareness of a wider holistic/meaningful/experiential reality, without throwing away the practical knowledge that reductionist explanations sometimes have to offer.

Comparing a strict enactivism with explanatory pluralism reinforces the practical weakness of the first. A strictly enactive approach, such as described by de Haan (42), considers that integration is necessary for a relevant and fruitful understanding of psychiatry. It is an integrative pluralism (and, as we will discuss, a synthetic one) in that it demands consideration of

how the different understandings relate to the dynamic whole – a person standing in relation to their environment. Enactivism is thereby in tension with a purely non-integrative pluralism as it is constantly asking us to consider how the parts and ways of understanding them come together to dynamically constitute human functioning and experience.

It is also important to note that integrativity for different perspectives or systems of representation of psychiatry does not necessarily require integrativity for levels/scales of explanation, and vice versa. Thus, an enactive approach may integrate different levels/scales of explanation while constituting only a part of integrative pluralism. However, for psychiatry, such a non-integrative pluralism (and in particular a tolerant non-integrative pluralism) seems particularly relevant. In addition to avoiding any form of scientific imperialism (the predominance of one system of representations over the others), non-integrative pluralism allows the clinician and the researcher to be flexibly free to choose relevant perspectives according to their medical interest (e.g., diagnostic, prognostic or therapeutic), interests of the patient, or non-medical interests (e.g., administrative, social).

It strikes us that there does not seem to be a language available to describe this difference. We therefore propose that the broader framework of explanatory pluralism should be described as an *analytical* pluralism, since, at first, it tends to break targets down across levels of understanding (e.g., biological from social), before it is considered whether these different understandings can be integrated or happily *co-exist*. An enactive pluralism, meanwhile, can be described as a *synthetic-organizational* pluralism, since it demands a constant return to consideration of all levels of understanding in relation to each other, in a synthetic and organizational way. We use the term “organizational” to avoid the confusion of “synthetic” being commonly used to refer to an artificial, synthetic product. In logic, the *synthesis* allows verifying that an object (e.g., an explanation) does indeed possess the properties

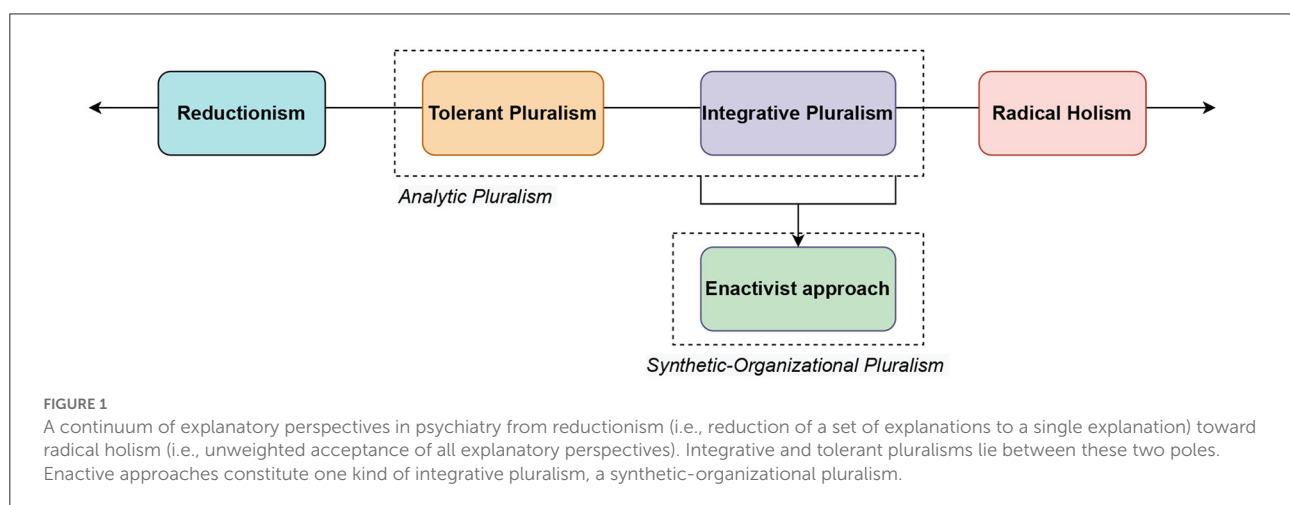
of the set in which it is located. In other words, enactivism is a synthetic-organizational integrative pluralism because each mode of explanation (e.g., experiential, physiological, etc.) can only have meaning by virtue of the other modes/levels and of the organism in its globality. Similar ideas can be seen in a discussion by Thompson and Varela (89) regarding the possibility for the enactive approach to try to capture “concrete wholes [body] in all their complexity,” without falling into the trap of unifying such complex phenomena under a single explanatory framework (90).

To summarize, the general framework of pluralism is typically *analytical*, while an enactive pluralism can be understood as a synthetic-organizational type of integrative pluralism. Such an enactive *and* pluralistic approach to psychiatry constitutes a subtype of explanatory pluralist frameworks (here named synthetic-organizational). Figure 1 allows considering the range of explanatory perspectives of psychiatry on a continuum from reductionism to radical holism *via* integrative and tolerant pluralisms. We propose that an enactive approach constitutes one kind of integrative pluralism which can be labeled synthetic-organizational pluralism.

Conclusion

Psychiatric practice requires understanding a variety of questions, tools, systems of representations, levels of explanation and perspectives. Reductionist approaches for clinical psychiatry can no longer be sustained. An opposing radical holism seems also untenable in practice. Psychiatry, therefore, demands to be understood pluralistically.

An enactive approach to psychiatry is beginning to emerge. It proposes that the different dimensions of understanding life and psychiatric disorders (and especially experiential, physiological, socio-cultural, and existential) are linked and integrated with each other. This stance provides an



integrative conception to explain psychiatric disorders – considering their embodied, embedded, enacted and emotive (4E) dimensions. This pluralist approach is integrative and synthetic (in organizational terms) because it allows integration of different explanations and perspectives within the same theory of cognition.

A general framework of explanatory pluralism allows the simultaneous conception and the possible integration of multiple levels and perspectives within our understanding of mental disorders and psychiatry. This general epistemological approach is a broader one than enactivism and makes fewer conceptual commitments regarding mental disorder and human functioning. This potentially makes it more encompassing and flexible than enactivism as an epistemological framework, however in practice, it can often result in the glossing over of first-person experience and can allow for the importation of dualism and unhelpful eclecticism.

Subsequently, a number of perspectives should be developed, including the need to consider the second-person approach to enactive psychiatry in relation to the pluralistic framework, the issue of pragmatic choices and epistemic gains of the clinician in enactive integrative pluralism, and the intertwined understanding of enactivism as a form of pluralism or as an approach that should add pluralism.

We have here considered the relationship between enactive and explanatory pluralism. We have argued that explanatory pluralism and enactivism are mutually compatible in their application to psychiatry. We have suggested that an enactive approach to psychiatry can itself be understood as a synthetic-organizational form of an integrative pluralistic approach. In sum, an enactive approach to psychiatry has great potential as an integrative framework, but we should not give up a wider commitment to explanatory pluralism.

Data availability statement

The original contributions presented in the study are included in the article/supplementary

material, further inquiries can be directed to the corresponding author/s.

Author contributions

CG wrote the first draft of the manuscript. CG and KN contributed to conception and design of the study. MJ and HB wrote sections of the manuscript and reviewed this work. GD wrote sections of the manuscript, supervised, reviewed, edited, and validated the work. All authors contributed to the article and approved the submitted version.

Funding

GD received funding from the Institute for Data Valorization (IVADO) and the Fonds de recherche du Québec—Santé (FRQS).

Conflict of interest

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

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Gijsbers V. Explanatory pluralism and the (dis)unity of science: the argument from incompatible counterfactual consequences. *Front Psychiatry*. (2016) 7:32. doi: 10.3389/fpsy.2016.00032
2. Putnam H. Reductionism and the nature of psychology. *Cognition*. (1973) 2:131–46. doi: 10.1016/0010-0277(72)90033-9
3. Murphy D. *Psychiatry in the Scientific Image*. Cambridge, MA: MIT Press (2006). Xi:410 p.
4. Brendel DH. Beyond engel: clinical pragmatism as the foundation of psychiatric practice. *Philos Psychiatr Psychol*. (2007) 14:311–3. doi: 10.1353/ppp.0.0145
5. McLaughlin B, Bennett K. Supervenience. In: Zalta EN, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab: Stanford University (2021). Available online at: <https://plato.stanford.edu/archives/sum2021/entries/supervenience/> (accessed July 25, 2022).
6. Cloninger R, Svrakic D. Integrative psychobiological approach to psychiatric assessment and treatment. *Psychiatry*. (1997) 60:120–41. doi: 10.1080/00332747.1997.11024793
7. Cuthbert. The RDoC framework: facilitating transition from ICD/DSM to dimensional approaches that integrate neuroscience and psychopathology. *World Psychiatry*. (2014) 13:28–35. doi: 10.1002/wps.20087
8. Ferrante M, Redish AD, Oquendo MA, Averbach BB, Kinnane ME, Gordon JA. Computational psychiatry: a report from the 2017 NIMH workshop on opportunities and challenges. *Mol Psychiatry*. (2019) 24:479–83. doi: 10.1038/s41380-018-0063-z

9. Gauld C. *Les Réseaux de Symptômes en Psychopathologie*. Grenoble: UGA Editions. (2021). 196 p.
10. Borsboom D. A network theory of mental disorders. *World Psychiatry*. (2017) 16:5–13. doi: 10.1002/wps.20375
11. Friston KJ, Stephan KE, Montague R, Dolan RJ. Computational psychiatry: the brain as a phantastic organ. *Lancet Psychiatry*. (2014) 1:148–58. doi: 10.1016/S2215-0366(14)70275-5
12. Engel GL. The need for a new medical model: a challenge for biomedicine. *Science*. (1977) 196:129–36. doi: 10.1126/science.847460
13. Aftab A, Nielsen K. From engel to enactivism: contextualizing the biopsychosocial model. *Eur J Philos*. (2021) 17:2–5. doi: 10.31820/ejap.17.2.3
14. Bolton D, Gillett G. *The Biopsychosocial Model of Health and Disease: New Philosophical and Scientific Developments*. Cham, CH: Palgrave Pivot (2019). Available online at: <http://www.ncbi.nlm.nih.gov/books/NBK552029/> [Accessed February 20, 2022] doi: 10.1007/978-3-030-11899-0
15. Von Bertalanffy L. An outline of general system theory. *Br J Philos Sci*. (1950) 1:134–65. doi: 10.1093/bjps/1.2.134
16. John R. The bio-bio-bio model of madness. *Psychologist*. (2005) 18:596–7. Available online at: <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.466.6413&rep=rep1&type=pdf>
17. Ghaemi SN. The rise and fall of the biopsychosocial model. *Br J Psychiatry*. (2009) 195:3–4. doi: 10.1192/bjp.bp.109.063859
18. Drayson Z. Embodied cognitive science and its implications for psychopathology. *Philos Psychiatr Psychol*. (2009) 16:329–40. doi: 10.1353/ppp.0.0261
19. Sadler J, Hulgus Y. Clinical problem solving and the biopsychosocial model. *Am J Psychiatry*. (1992) 149:1315–23. doi: 10.1176/ajp.149.10.1315
20. de Haan S. The existential dimension in psychiatry: an enactive framework. *Ment Health Relig Cult*. (2017) 20:528–35. doi: 10.1080/13674676.2017.1378326
21. Varela FJ, Rosch E, Thompson E. *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge, MA: MIT Press (1991). 328 p. doi: 10.7551/mitpress/6730.001.0001
22. de Haan S. Bio-psycho-social interaction: an enactive perspective. *Int Rev Psychiatry*. (2021) 33:471–7. doi: 10.1080/09540261.2020.1830753
23. Nielsen K, Ward T. Towards a new conceptual framework for psychopathology: embodiment, enactivism, and embedment. *Theory Psychol*. (2018) 28:800–22. doi: 10.1177/0959354318808394
24. Nielsen K. Comparing two enactive perspectives on mental disorder. *Philos Psychiatr Psychol*. (2021) 28:175–85. doi: 10.1353/ppp.2021.0028
25. Fuchs T. The circularity of the embodied mind. *Front Psychol*. (2020) 11:1707. doi: 10.3389/fpsyg.2020.01707
26. Fuchs T. *Ecology of the Brain: The Phenomenology and Biology of the Embodied Mind*. Oxford: OUP Oxford (2017). 368 p. doi: 10.1093/med/9780199646883.001.0001
27. Donovan C, Murphy D. De Haan on sense-making and psychopathology. *Philos Psychiatr Psychol*. (2020) 27:29–30. doi: 10.1353/ppp.2020.0003
28. Nielsen K. Comparing two enactive perspectives. *Philos Psychiatr Psychol*. (2021) 28:197–200. doi: 10.1353/ppp.2021.0031
29. Maturana HR, Varela FJ. *Autopoiesis and Cognition: The Realization of the Living*. Berlin: Springer Science & Business Media (2012). 172 p.
30. Varela FG, Maturana HR, Uribe R. Autopoiesis: The organization of living systems, its characterization and a model. *Biosystems*. (1974) 5:187–96. doi: 10.1016/0303-2647(74)90031-8
31. Di Paolo E. Extended life. *Topoi*. (2008) 28:9. doi: 10.1007/s11245-008-9042-3
32. Gallagher S. *Enactivist Interventions: Rethinking the Mind*. Oxford: Oxford University Press (2017). 262 p. doi: 10.1093/oso/9780198794325.001.0001
33. Thompson E. *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*. Cambridge, Mass: Belknap Press of Harvard University Press (2007). 543 p.
34. Newen A, Bruin LD, Gallagher S. *The Oxford Handbook of 4E Cognition*. Oxford: Oxford University Press (2018). 961 p. doi: 10.1093/oxfordhb/9780198735410.001.0001
35. Thompson E. Sensorimotor subjectivity and the enactive approach to experience. *Phenom Cogn Sci*. (2005) 4:407–27. doi: 10.1007/s11097-005-9003-x
36. Fuchs T, De Jaegher H. Enactive intersubjectivity: participatory sense-making and mutual incorporation. *Phenom Cogn Sci*. (2009) 8:465–86. doi: 10.1007/s11097-009-9136-4
37. Clark A, Chalmers D. The Extended Mind. *Analysis*. (1998) 58:7–19. doi: 10.1093/analys/58.1.7
38. Stilwell P, Harman K. An enactive approach to pain: beyond the biopsychosocial model. *Phenom Cogn Sci*. (2019) 18:637–65. doi: 10.1007/s11097-019-09624-7
39. Nielsen K, Ward T. Mental disorder as both natural and normative: Developing the normative dimension of the 3e conceptual framework for psychopathology. *J Theor Philos Psychol*. (2020) 40:107–23. doi: 10.1037/teo0000118
40. Ramstead MJ, Kirchhoff MD, Friston KJ. A tale of two densities: active inference is enactive inference. *Adapt Behav*. (2020) 28:225–39. doi: 10.1177/1059712319862774
41. Gieseck JJ, Mangold W, Katz C, Low S, Saegert S. *The People, Place, and Space Reader*. Oxfordshire: Routledge (2014). 481 p. doi: 10.4324/9781315816852
42. de Haan S. *Enactive psychiatry*. Cambridge: Cambridge University Press. (2020). doi: 10.1017/9781108685214
43. Bottemanne H. Bayesian brain: can we model emotion? *L'encephale*. (2020) 47:58–63. doi: 10.1016/j.encep.2020.04.022
44. Maiese M. An enactivist approach to treating depression: cultivating online intelligence through dance and music. *Phenom Cogn Sci*. (2020) 19:523–47. doi: 10.1007/s11097-018-9594-7
45. Bottemanne H, Barberousse A, Fossati P. Multidimensional and computational theory of mood. *L'Encéphale*. (2021). doi: 10.1016/j.encep.2022.02.002. [Epub ahead of print].
46. Ekman P, Davidson RJ. *The Nature of Emotion: Fundamental Questions*. Oxford: Oxford University Press. (1994).
47. James W. Discussion: The physical basis of emotion. *Psychol Rev*. (1894) 1:516–29. doi: 10.1037/h0065078
48. Cannon WB. Again the James-Lange and the thalamic theories of emotion. *Psychol Rev*. (1931) 38:281–95. doi: 10.1037/h0072957
49. Schachter S, Singer J. Cognitive, social, and physiological determinants of emotional state. *Psychol Rev*. (1962) 69:379–99. doi: 10.1037/h0046234
50. Colombetti G. *The Feeling Body: Affective Science Meets the Enactive Mind*. Cambridge, MA: MIT Press (2014). 288 p. doi: 10.7551/mitpress/9780262019958.001.0001
51. Soto-Andrade J. Enactive Metaphorising in the Learning of Mathematics. In: Kaiser G, Forgasz H, Graven M, Kuzniak A, Simmt E, Xu B, editors. *Invited Lectures from the 13th International Congress on Mathematical Education*. ICME-13 Monographs. Cham: Springer International Publishing (2018). p. 619–637. doi: 10.1007/978-3-319-72170-5_34
52. Kyselo M. The enactive approach and disorders of the self—The case of schizophrenia. *Phenomenol Cogn Sci*. (2016) 15:591–616. doi: 10.1007/s11097-015-9441-z
53. Ongaro G, Ward D. An enactive account of placebo effects. *Biol Philos*. (2017) 32:507–33. doi: 10.1007/s10539-017-9572-4
54. De Jaegher H. Embodiment and sense-making in autism. *Front Integr Neurosci*. (2013) 7:15. doi: 10.3389/fnint.2013.00015
55. Nielsen K. What is Mental Disorder? *Developing an Embodied, Embedded, and Enactive Psychopathology*. Victoria University of Wellington. (2020).
56. De Jaegher H, Di Paolo E. Participatory sense-making. *Phenom Cogn Sci*. (2007) 6:485–507. doi: 10.1007/s11097-007-9076-9
57. Varela F. Organism: A Meshwork of Selfless Selves. In: *Organism: The Origins of Self*. Boston, Studies in the Philosophy of Science. (1991). p. 79–107. doi: 10.1007/978-94-011-3406-4_5
58. de Haan S. An enactive approach to psychiatry. *Philos Psychiatr Psychol*. (2020) 27:3–25. doi: 10.1353/ppp.2020.0001
59. Uexküll J. von. *Theoretische Biologie*. Berlin: Paetel. (1920).
60. James. Lecture II: what pragmatism means. In: *Pragmatism: A new name for some old ways of thinking*. New York, NY: Longmans, Green and Co (1907). p. 43–81. doi: 10.1037/10851-000
61. Di Paolo EA, De Jaegher H. Enactive ethics: difference becoming participation. *Topoi*. (2022) 41:241–56. doi: 10.1007/s11245-021-09766-x
62. Galbusera L, Fellin L. The intersubjective endeavor of psychopathology research: methodological reflections on a second-person perspective approach. *Front Psychol*. (2014) 5: 1150. doi: 10.3389/fpsyg.2014.01150
63. Moreno A, Mossio M. *Biological Autonomy: A Philosophical and Theoretical Enquiry*, 2015th edition. Dordrecht: Springer (2015). 255 p.

64. Varela FJ. *Ethical Know-How: Action, Wisdom, and Cognition*. Stanford: Stanford University Press (1999). 96 p.
65. Adams F, Aizawa K. Why the Mind is Still in the Head. In: Aydede M, Robbins P, editors. *The Cambridge Handbook of Situated Cognition*. Cambridge: Cambridge University Press (2009). p. 78–95. doi: 10.1017/CBO9780511816826.005
66. Hacking I. Historical Ontology. In: Gärdenfors P, Woleński J, Kijania-Placek K, editors. *In: The Scope of Logic, Methodology and Philosophy of Science: Volume Two of the 11th International Congress of Logic, Methodology and Philosophy of Science, Cracow, August 1999*. Synthese Library. Dordrecht: Springer Netherlands (2002). p. 583–600
67. Ruphy S. *Scientific Pluralism Reconsidered: A New Approach to the (Dis)Unity of Science*. University of Pittsburgh Press (2017). 179 p. doi: 10.2307/j.ctt1mtz6n9
68. Maxwell JC. *The Scientific Papers of James Clerk Maxwell*. Cambridge: Cambridge University Press (1890). 838 p.
69. Kellert SH, Longino HE, Waters CK. *Scientific Pluralism*. Minneapolis, MI: University of Minnesota Press (2006). 288 p.
70. Mitchell. *Unsimple Truths: Science, Complexity, and Policy*. Chicago, IL: University of Chicago Press (2009). 161 p. doi: 10.7208/chicago/9780226532653.001.0001
71. Van Bouwel J. What can democratic theory teach us about scientific pluralism, objectivity and consensus? *Three rivers philosophy conference 2011 : science, knowledge, and democracy, Abstracts*. (2011). Available online at: <http://hdl.handle.net/1854/LU-1853227> (accessed December 25, 2020).
72. De Vreese L, Weber E, Van Bouwel J. Explanatory pluralism in the medical sciences: theory and practice. *Theor Med Bioeth*. (2010) 31:371–90. doi: 10.1007/s11017-010-9156-7
73. Dupré J. Scientific Classification. *Theory Cult Soc*. (2006) 23:30–2. doi: 10.1177/026327640602300201
74. Cartwright N. *The Dappled World: A Study of the Boundaries of Science*. Cambridge: Cambridge University Press. (1999). doi: 10.1017/CBO9781139167093
75. Bhugra D, Easter A, Mallaris Y, Gupta S. Clinical decision making in psychiatry by psychiatrists. *Acta Psychiatr Scand*. (2011) 124:403–11. doi: 10.1111/j.1600-0447.2011.01737.x
76. Sober E. The multiple realizability argument against reductionism. *Philos Sci*. (1999) 66:542–64. doi: 10.1086/392754
77. Hacking I. 'Style' for historians and philosophers. *Stud Hist Philos Sci*. (1992) 23:1–20. doi: 10.1016/0039-3681(92)90024-Z
78. Kendler. Toward a philosophical structure for psychiatry. *AJP*. (2005) 162:433–40. doi: 10.1176/appi.ajp.162.3.433
79. Kendler. The dappled nature of causes of psychiatric illness: replacing the organic–functional/hardware–software dichotomy with empirically based pluralism. *Mol Psychiatry*. (2012) 17:377–88. doi: 10.1038/mp.2011.182
80. McNally RJ. Can network analysis transform psychopathology? *Behav Res Ther*. (2016) 86:95–104. doi: 10.1016/j.brat.2016.06.006
81. Borsboom D. Psychometric perspectives on diagnostic systems. *J Clin Psychol*. (2008) 64:1089–108. doi: 10.1002/jclp.20503
82. Goyer S. Pour un modèle de l'explication pluraliste et mécaniste en psychiatrie. (2012) 94 p.
83. Bottemanne H, Chevance A, Joly L. Psychiatry without Mind? *L'Encéphale*. (2021) 47:605–12. doi: 10.1016/j.encep.2021.05.006
84. Potochnik A. *Idealization and the Aims of Science*. University of Chicago Press (2017). 263 p. doi: 10.7208/chicago/9780226507194.001.0001
85. Korterink JJ, Diederik K, Benninga MA, Tabbers MM. Epidemiology of Pediatric Functional Abdominal Pain Disorders: A Meta-Analysis. *PLoS ONE*. (2015) 10:e0126982. doi: 10.1371/journal.pone.0126982
86. Adam Carter J. Epistemic Pluralism, Epistemic Relativism and 'Hinge' Epistemology. In: Coliva A, Jang Lee Linding Pedersen N, editors. *Epistemic Pluralism. Palgrave Innovations in Philosophy*. Cham: Springer International Publishing (2017). p. 229–249. doi: 10.1007/978-3-319-65460-7_9
87. Bechtel W, Richardson RC. *Discovering complexity: Decomposition and localization as strategies in scientific research*. Princeton, NJ: Princeton University Press (1993). XIV:286 p.
88. Fulford. Values-based practice: a new partner to evidence-based practice and a first for psychiatry? *Mens Sana Monogr*. (2008) 6:10–21. doi: 10.4103/0973-1229.40565
89. Thompson E, Varela FJ. Radical embodiment: neural dynamics and consciousness. *Trends Cogn Sci*. (2001) 5:418–25. doi: 10.1016/S1364-6613(00)01750-2
90. Di Paolo E. Process and individuation: the development of sensorimotor agency. *Hum Dev*. (2019) 63:1–25. doi: 10.1159/000503827



OPEN ACCESS

EDITED BY

Marialisa Martelli,
Sapienza University of Rome, Italy

REVIEWED BY

Simone Pollo,
Sapienza University of Rome, Italy
Marco Del Giudice,
University of New Mexico,
United States

*CORRESPONDENCE

Francesco Mancini
mancini@apc.it

SPECIALTY SECTION

This article was submitted to
Psychopathology,
a section of the journal
Frontiers in Psychiatry

RECEIVED 10 March 2022

ACCEPTED 08 September 2022

PUBLISHED 30 September 2022

CITATION

Mancini F, Mancini A and
Castelfranchi C (2022) Unhealthy mind
in a healthy body: A criticism to
eliminativism in psychopathology.
Front. Psychiatry 13:889698.
doi: 10.3389/fpsy.2022.889698

COPYRIGHT

© 2022 Mancini, Mancini and
Castelfranchi. This is an open-access
article distributed under the terms of
the [Creative Commons Attribution
License \(CC BY\)](#). The use, distribution
or reproduction in other forums is
permitted, provided the original
author(s) and the copyright owner(s)
are credited and that the original
publication in this journal is cited, in
accordance with accepted academic
practice. No use, distribution or
reproduction is permitted which does
not comply with these terms.

Unhealthy mind in a healthy body: A criticism to eliminativism in psychopathology

Francesco Mancini^{1,2*}, Alessandra Mancini¹ and
Cristiano Castelfranchi³

¹Schools of Cognitive Psychotherapy (APC-SPC), Rome, Italy, ²Department of Psychology, Telematic University of Rome "Guglielmo Marconi", Rome, Italy, ³Istituto di Scienze e Tecnologie della Cognizione - CNR, Rome, Italy

In this article we criticize the thesis "The diseases we treat are diseases of the brain". A first criticism is against the eliminativist perspective and in favor of a perspective that is still reductionist but emergentist and functionalist. In a second part, we try to answer the question "under which conditions can we consider this statement legitimate?". We argue that only those mental disorders whose neural substrate has clearly neuropathological characteristics, i.e., anomalies with respect to the laws of good neural functioning, can be considered "brain diseases." We propose that it is not sufficient to observe a simple difference between the brains of people with psychopathology, that is, with anomalies with respect to the laws of good psychological functioning, and that of people without psychopathology. Indeed, we believe it is a categorical error to postulate a neuropathology starting from a psychopathology. Finally, we summarize some research that shows how purely psychological interventions can reduce or eliminate the differences between the brains of people with or psychopathology and those of people without.

KEYWORDS

reductionism, emergence, intentions, psychopathology, mind-body problem, mental processing

Introduction

In 2015, an editorial entitled *The Future of Psychiatry as Clinical Neuroscience*, was published on the important JAMA Psychiatry (1). The main argument concerned the increase in knowledge on the differences between the brains of people affected by a given psychopathology (e.g., depression) and those of people not affected or affected by a different psychopathology. "Technologic advances have enhanced our ability to study the brain, and new findings have reshaped the fundamental way in which we understand psychiatric illness. For example, although depression was once characterized as simply a monoaminergic deficit, new research is expanding our understanding of depression across multiple levels of analysis—from circuits, to neuro transmitters, to synaptic plasticity, to second messenger systems" (1). The conclusion was somewhat apodictic: "The diseases we treat are diseases of the brain" (1).

This sentence, beyond the real intentions of those who wrote it, lends itself well to discuss two theses, connected to each other, concerning psychopathology, and which, albeit with different nuances, can be glimpsed behind the so-called “biological psychiatry” (2). The first thesis is that psychopathology, like any other psychic event, is substantially reducible to neural events, *therefore* it should be described and explained *only* with neuroscientific concepts (i.e., eliminating the mental level of description). Following this approach, with neuroscientific advances, mentalistic concepts, still used today for describing, explaining and treating psychopathology, run the risk to be considered like phlogiston in chemistry (i.e., according to the “eliminative materialist”, the explanation in terms of *beliefs*, *desires*, and *fears*, will become obsolete, thanks to advancements in neuroscience) (3). In that, following this eliminativist thesis, psychotherapy and the psychological models of psychopathology on which it is based, would have no value.

The second thesis, connected with the previous one, argues that psychopathology, that is, mental illnesses, are actually “diseases of the brain.” These two theses deserve to be discussed in order to restore the right dignity to mentalist descriptions and explanations, and therefore also to strictly psychological interventions, without however falling back into the mind-brain dualism.

Therefore, in the first part of this article, we argue against eliminativism, and in favor of an emergentist approach which, while on the one hand considers the brain as the foundation of the mind and the mind as an expression of brain activity, on the other hand believes it necessary to adopt a multilayer, complex, self – emergent view of reality, that allows to give mental concepts the right role in the explanation of psychopathology.

Having ascertained the scientific legitimacy of using the mental level to explain psychopathology, in the second part of the article, we interrogate ourselves on what are the conditions for affirming or denying that “The diseases we treat are diseases of the brain”.

Eliminativism and the emergentist criticism

Eliminativism is defined by the Dictionary of Philosophy of Mind: “*The view that, because mental states and properties are items posited by a protoscientific theory (called folk psychology), the science of the future is likely to conclude that entities such as beliefs, desires, and sensations do not exist. The alternate most often offered is physicalist and the position is thus often called “eliminative materialism”* (4–7).

One of the most convincing criticisms to eliminativism comes from the so-called emergentism, which rejects the mind-brain dualism and accepts that the mind is a product of the brain, that any mental phenomenon corresponds to a neural phenomenon and that it cannot exist a mind without

a brain or without its implementation in a material support. Emergentism, however, is different from eliminativism, because it assumes that the mind is an emergent phenomenon, or rather that mental phenomena are emergent features of complex brains (8) and thus they are not entirely reducible to it. They should be described and modeled at the macro-function layer implemented in- and emergent from the underlying micro-function of the material substrate. “A property of a system is said to be emergent if it is a new outcome of some other properties of the system and their interaction, while it is itself different from them” (9). As Chalmers put it “We can say that a high-level phenomenon is strongly emergent with respect to a low-level domain when the high-level phenomenon arises (in some sense) from the low-level domain, but truths concerning that phenomenon are not deducible even in principle from truths in the low-level domain (...). I think there is exactly one clear case of a strongly emergent phenomenon, and that is the phenomenon of consciousness” (10). However, strong epistemological emergence is different from ontological emergence, which rejects the layered model of reality as divided into a discrete hierarchy of levels (11).

At the basis of emergence there is the idea, ancient and shared by many, that the whole is greater than the sum of its parts. Nature (and, in nature, society) has different levels of structure, organization, dynamics, and “*functions*,” each macro-layer is grounded on the entities, properties and mechanisms of the lower layer (micro) but implies the emergence of macro-layer properties (12). For example, consider the concept of “Information” (that of Information Theory, Computer Science, etc.). It could be argued that “Information” is merely “energy.” Yes, but it is a specific level of energy dynamics and a *function* that energy assumes at a certain level of organization of the matter and of its processes. We could not eliminate the concept of “information,” because in nature there is only “energy.” However, is “information” nothing but energy? No, it is something more; it’s energy with new characteristics, processes, laws; a new level of functions and effects, requiring their own “laws” and “concepts,” not meta-physical, but physical at a different level.

Nature organizes itself into emerging levels of complexity, with new structures, *which require their own scientific concepts and laws*, not existing at the micro level. Indeed, “Emergence occurs in *complex systems* in which novel properties emerge through the *aggregate functions* of the parts of that system” (8). As said, this holds even within the neural level: human experience and behavior are due to the brain and to bodily processes. These are due to micro-biological (cellular) processes, which in turn are due to biochemical processes and so on. But biochemistry or underlying physics are not enough and concepts - and the physical objects captured by them - such as “neurons,” “neural networks,” “activations” are essential. Indeed, they are a level of organization of a physical reality that possess new properties and dynamics.

An example of the very reductive outcome obtained by several attempts to establish the neural foundation of psychological (and social) notions is about the concept of “trust.”

As Fehr writes: “the rationale for the experiment originates in evidence indicating that oxytocin plays a key role in certain pro-social approach behaviors in non-human mammals. (...) Based on the animal literature, Kosfeld et al. (13), hypothesized that oxytocin might cause humans to exhibit more behavioral trust as measured in the trust game” (13, 14). In these experiments they also show how oxytocin has a specific effect on social behavior because it differently impacts on the trustor and the trustee (only in the first case there is a positive influence). In addition, it is also shown that the trustor’s sensitivity to risk is not reduced as a general behavior but it depends on the partner nature (human vs. non-human). These are without any doubts very interesting data. However, the multidimensional and very articulated notion of trust (so crucial for individual feelings and conduct and for social relations) (15), should not be reduced to a generic pro-social attitude and to a particular chemical response or the mere activation of a given brain area. Trust is not a simple, vague, and unitary notion and disposition; it is made of rather complex evaluations, expectations, attributions, decisions to rely, sentiments. It should be a componential and analytical psychological model of trust to *drive* neural research rather than searching for a simplistic and direct solution, just localist and correlational (16).

Indeed, even the most accurate and complete knowledge of the micro-level does not allow us to infer structure, organization, dynamics, and “functions” of the macro-level. For instance, the explanation of cellular roles and activities and their laws cannot be reduced to the micro-description of their underlying chemical processes without losing necessary information. Cells are indeed implemented, founded on their chemical substrate and laws but we need the other layer of notions/concepts, their new functions, their laws [see also (17)]. Reduction is micro-foundation, material grounding, but not necessarily elimination.

Let us consider, for example, the following case: we want to evaluate whether a dancer correctly performs a certain dance step. Suppose it is possible to detect all changes in all of the dancer’s muscles as she dances. Even if we have a computer with an enormous computational capacity, could we entrust the computer with the evaluation of her dance? That is, does the complete and accurate recording of the activation of the dancer’s muscles allow for an aesthetic evaluation? No. For at least two reasons. The first is that we should also codify the parameters describing the muscle activation patterns relevant for the evaluation; an information which could otherwise not be inferred just by the sum of the data concerning the movements of the different muscles. The second is that we will also have to translate in a computational form the aesthetic criteria discriminating the activation patterns that characterize good executions; an information which is also not inferable just by muscle registration. In other words, we should enter into the

computer information concerning the macro level and which cannot be inferred from the data coming from the micro level. It would be non-sense to pretend to understand if the movement of the dancer corresponds to aesthetic criteria, only by studying the movements of her muscles and without knowing the aesthetic criteria. And for those involved in dance, for example a choreographer, aesthetic criteria are indispensable.

It seems plausible that a (very large) machine learning model fed with enough labeled examples could be trained to reproduce a fair aesthetic assessment of a dance from a stream of pixels in a video. But on the condition of providing labeled examples of correct and incorrect movements, that is, examples of the application of aesthetic criteria that nevertheless belong to a different level from that of muscle movements. Aesthetic criteria can be reduced to movements but they are not necessarily deduced on the basis of movements. In other words, aesthetic criteria *supervene* on movements [for a definition of Supervenience and its distinction to emergence see (18)].

We do not think that the problem raised by eliminativism is just “practical” and one destined to be overcome as the knowledge about the brain advances.

Rather, we believe that *at the epistemological level* (i.e., in order to understand reality) another level of description of reality is needed and more specifically, the level of emerging macro-functions which define and model processes and mechanisms. Science should be modeling, conceptualization, description and explanation not just at the micro-micro level but also at the different functional levels of complexity. This does not involve a dualism of reality but a dualism of theory and concepts (as also in the physical and natural sciences: material vs. functional concepts, and not on two levels but on layers). Indeed, we assume that reality is one and material but we believe that, in order to understand it, we need to consider different levels of emergent properties that can be grasped with conceptual categories appropriate to that specific level and cannot be grasped otherwise (i.e., with categories belonging to a lower or upper level). For example, given that viruses are ultimately made of atoms and atoms of electrons, using just lower-level atomic-physics conceptual categories to understand how viruses work, does not appear substantially appropriate, because aspects that are crucial for the understanding of viruses, such as for example their architecture and methods of reproduction, are not captured by the lower level concepts of atomic physics. To answer these questions, the knowledge of the virologist is necessary, that is, a body of knowledge that grasps reality at a different level than that of atomic physics. Indeed, other conceptual categories are needed, and these are not only pragmatically more useful than those of the atomic physics; they are irreplaceable for understanding and explaining viruses as well as for acting on them. Those of the atomic physicists can contribute to enrich the knowledge of the virologist, but not replace them, as well as those of the epidemiologist and sociologist, who look at the phenomenon at even more macro levels, can complement those

of the virologist but not replace them. Importantly, since we assume that reality is one, even if it can be described at different levels and from different points of view, it follows that we cannot strictly speak about “causality” between different levels. As suggested by Kim (19) psychophysical causal relations should be viewed as epiphenomenal supervenient causal relations (20). To understand this concept Kim (19) proposes the following example: “Thus, if a pain causes the sensation of fear an instant later, this account tells the following story: the pain is supervenient on a brain state, this brain state causes another appropriate brain state, and given this second brain state, the fear sensation must occur, for it is supervenient upon that brain state” (19). A mental event is not caused by a neural event since they are the same thing, described at two different levels, with different categories that are able to grasp the characteristic properties of one level but not the other. In this article, of the many possible levels, we are interested in two, i.e., the neural and the mental (e.g., not the molecular and not the social), it seems interesting to observe an asymmetry between the two.

While it is true that the characteristics of the macro-level cannot necessarily be inferred from the characteristics of the micro-level, the opposite is true. Inferences from the macro to the micro level are possible, and therefore the study of the micro-level could not only be used, but it should be used as a bench test for psychological hypotheses. It *should* because, if it is true that the mind is implemented in the brain, then any mental hypothesis must be compatible with the structure or functioning of the brain.

A research (21) tested the hypothesis, strictly psychological, that there are two types of guilt feelings, one altruistic and one deontological. Deontological guilt was induced in one group and altruistic guilt in another group of non-clinical participants. During the induction, brain activity was detected *via* fMRI. The results showed that the two guilt feelings have a different neural substrate. Therefore, the hypothesis has been corroborated. It should be noted that no CNS analysis, however accurate and exhaustive, could have made sense of the neural activation patterns detected in this study, had it not been accompanied by psychological assumptions. Furthermore, it cannot be overlooked that renouncing to the psychological construct of guilt would imply renouncing to explain and predict many behaviors and interactions between people. It is interesting to observe that from the aforementioned study, it emerged that deontological guilt, but not altruistic guilt, shares part of the neural substrate with disgust, specifically the insular cortex. These results might also explain another psychological problem which concerns the relationship between guilt and disgust in the so-called Macbeth effect, in which the induction of guilt increases disgust sensitivity and washing the body reduces guilt (22). The Macbeth effect has been found inconsistently in some studies but not in others. However, it becomes clear only if the type of guilt induced is deontological and not if it is altruistic (23). Taken together these studies well represent an example

of the use of neural data to assess psychological hypotheses. Specifically, here H_1 was that guilt can be conceptualized in two distinct emotional patterns and that these differences are also reflected in brain activity. Furthermore, the results also helped to clarify why the Macbeth effect was observed only in some studies but not in others. Indeed, previous research did not consider separately the effects of deontological and altruistic guilt.

In keeping, two behavioral studies have shown that induction of deontological guilt implies more thorough and prolonged washings than induction of altruistic guilt (24, 25), and two other studies, using transcranial direct current stimulation (tDCS), showed that a stimulation of the insular cortex implies an enhancement in disgust and orient moral judgments in a deontological sense, while the inhibition of the insula has the opposite effect. On the other hand, there is no effect on altruistic moral judgments (26, 27).

In a similar vein, some researchers observed that the dysfunction of the social brain in schizophrenia is modulated by intention type. Specifically, patients showed significantly less activation in three regions typically activated in ToM tasks, i.e., paracingulate cortex and bilateral temporo-parietal junctions. However, this dysfunction was present only for social but not for non-social intentions (28). In this case, neuroscientific findings helped to determine that also the psychological concept of “intention” can be differentiated on the basis of the object of the intention and that only certain types of intention are abnormal in schizophrenic patients.

An anonymous reviewer suggested that one could collect a large number of guilt instances and corresponding brain activation patterns, then run some kind of clustering to see if distinct grouping emerges; it is possible that such a micro \rightarrow macro approach would reveal partially differentiated clusters of brain activity, which could then reveal corresponding differences in the corresponding guilt episodes. However, to carry out this operation of searching for differences between guilt feelings starting from the neural data collection it is necessary to have psychological categories, such as “guilt feelings,” and to define corresponding differences in the corresponding guilt episodes, such as the absence or presence of an affective relation between the guilty and the victim. Moreover, without the knowledge contained at the mental level, the neuroscientist might incur in the multiple realization problem (i.e., the thesis that the same mental state can be realized by different physical states), (29–31).

Mental representations, functions, and processing are just material, informational entities; emergent *functions*, described in informational/functional terms, but if they are brought back to their underlying micro-processes, they will not be redundant and eliminable. The psychological notions should be preserved for understanding and explaining “what the brain is *doing*”: perceiving, memorizing, retrieving, deciding, pursuing, and so on; at its emergent, macro-functional level of activity.

Neural correlates cannot be the right *vocabulary* for explaining human behaviors, just because they refer to concepts

pertaining to the micro-level and do not represent and discriminate the complex “patterns” and their properties and functions (not of their sub-components) at the cognitive and motivational macro-level. Once we will have the real neural representation of a complex object like a “motivating goal,” or an “altruistic intention,” or of a real “trust attitude,” or a “complex emotion with its appraisal components” like envy, we will have a quasi-complete explanation of it, but we could not renounce to that psychological vocabulary¹; since it holds at the functional/informational macro level (12).

More in general: there are no alternatives to the need for *reading* and *understanding* body in terms of functions, not just in terms of “simple” matter and its physico-chemical processes description. We look at the kidney as a “filter,” at glands in terms of “secretion.” Otherwise we do not understand what they do, that is, what they are; which is the sense of the physico-chemical processes that we are describing. Indeed, we know the world through its functions. Even the most basic categories (e.g., fruit, apples) are organized to give information on the functions of a certain element. In this way we also know biology or economics and so on. The same obviously holds for our brain. Neuroscientists shouldn’t try to “skip” psychology and its information-processing models of structures and manipulations, for directly connecting brain with behavior (neuro-economics, neuro-aesthetics, neuro-ethics, neuro-politics,...). On the contrary they should take the procedural (possibly computational) models of the cognitive sciences and find their neural grounding or - if this proves unfeasible - change them. In fact, a cognitive model that is not grounded in our brain and somatic processes is just wrong, unacceptable. And - on the other side - psychology should provide models of proximate processes; not just correlational “theories” (7, 12, 32, 34).

Are the diseases we treat diseases of the brain?

Under what conditions can we consider this statement legitimate?

As is well known, the problem regarding the definition of psychopathology is still debated and concerns the possibility of basing the diagnosis on objective and non-evaluative criteria.

For instance, according to Christopher Boorse’s biostatistical account, to define a (mental) disease value-laden judgements are not necessary: “if diseases are deviations from the species

biological design, their recognition is a matter of natural science, not evaluative decision” (35). This definition holds for mental disorders on the condition that a definition of mental disorder is informed by our knowledge of biological design. Differently, Jerome’s Wakefield *hybrid naturalism*² accepts a value component (harm), while still embracing an objective, evolutionary account of natural functions (36).

Here, we do not enter into the merits of this still unresolved debate on the definition of psychopathology, (i.e., on the criteria that differentiate psychopathology from normality). We simply base our definition of psychopathology on the DSM 5 or ICD 11. Indeed, rather than drawing a final conclusion about what psychopathology is or not, here we discuss the differences between psychopathology and neuropathology at the brain level.

Secondly, from neuroscience, for the moment, no criterion has emerged that allows a reliable psychiatric diagnosis, that is, without an exaggerated number of false positives and negatives, but, even if a neural marker is found as a valid diagnostic tool, would this justify such a conclusion?

The answer is necessarily articulated.

Let’s consider an example of a psychopathological disorder underlying a brain disease: progressive paralysis. It is a serious neuropathological form caused by the treponema of syphilis which manifests itself, among other things, with mood changes and delusions. The symptoms are predominantly psychiatric and the cause is exclusively neurological and, specifically, infectious. Similarly, important psychopathological, emotional and behavioral alterations, up to real personality disorders, can be caused by traumatic, neoplastic, infectious or degenerative lesions of the frontal lobe. In these cases, the brains of patients are different from that of non-patients for their neuropathological characteristics. Here the mental disorder is underpinned by a true brain disease a true brain disease, in fact, there are characteristics of the CNS that are compatible with the anatomy and physiology typical of neurological diseases. In these cases, the statement of Ross and colleagues is justified.

There are other cases in which psychopathological disorders are accompanied by brain damage but which nevertheless do not justify Ross’s conclusion. It is well known that the incidence of psychopathology in people with intellectual development disorder is higher than usual (37, 38).

It is plausible that at the basis of some forms of intellectual disability there is a brain damage due to infectious, neoplastic, metabolic, degenerative, autoimmune, traumatic or genetic causes. It is equally evident that the cognitive outcomes of these damages interact with psychological variables, for example with greater difficulty in regulating emotions, and with social variables, for example with social exclusion, which in turn interacts with self-esteem, producing psychiatric symptoms. Also, in this case there is a neurological damage, but the brain injury and its cognitive consequences are just a vulnerability factor to psychopathology and not the necessary and sufficient cause, as it happens in progressive paralysis.

¹ Let’s also remark that the criticism of Elimitativism to Psychology, i.e., that psychology would just use common-sense words, would just be “folk psychology” without scientific notions, is false/wrong: consider for example the notion of “goal,” which is very contrary to common sense (with its notions of feedback, circular causality, exc.) and it is directly derived from Cybernetics (32, 33).

There are differences in the brain due to neuropathological alterations but these are not the cause of psychopathology, rather, their consequences represent a vulnerability factor for psychopathology. Let us now consider, for instance, the brain of a person suffering from OCD. With a certain approximation it can be said that his brain is anatomically and functionally different from that of other people (39), but not in the same way as in patients with progressive paralysis or with frontal injuries. In fact, the brains of patients with OCD do not show the typical signs of neurological diseases, in which neurons are abnormal with respect to the laws of neuroanatomy and neurophysiology, for instance, the electrical activity of an epileptic brain, the presence of beta amyloid plaques, demyelinated plaques or gliotic infiltrates. A similar consideration can be extended to synaptic mediators. For instance, some results suggest that the density of serotonin (5-HT) transporter ^3H -Par binding sites was significantly lower in OCD patients than in controls. Could we infer from these data a damage in serotonin metabolism in OCD patients? Not necessarily, because the same alteration has been observed in people who are in love (40). Thus, the fact that the density of ^3H -Par binding sites is significantly lower in OCD patients than in controls is not necessarily an expression of a brain disease unless we also claim that love is a brain disease. It would seem more correct to state that we are in the presence of normal variations of serotonin metabolism which are connected to different mental states.

Certainly, it cannot be excluded that, in the future, the knowledge of pathological anatomy and pathophysiology will increase, enabling us to recognize signs of actual neuropathologies in the brain of obsessive patients, but at the moment it does not seem to be so, without prejudice to that nosographic entity (i.e., the Pediatric Autoimmune Neuropsychiatric Disorders, PANDAS) (41), whose existence is still debated and scarcely accepted by most and which in any case would concern a small subset of people with obsessive compulsive disorder. The differences that the brains of patients with OCD have to those of healthy controls is more similar to the differences found in the brains of “experts” (42). For instance, the brains of professional pianists are structurally different from that of other people but the neurons are not pathological, rather they are well functioning with respect to the laws of neuroanatomy and neurophysiology (43, 44).

Similarly, we can assume that a football fan has a different brain functioning than a person who is completely disinterested in football or a fan of an opposing team (45, 46). Even in this case we can speak of differences in terms of behaviors, assessments, and emotions, but we cannot say that the fan’s brain is abnormal with respect to the laws of neurology. Let’s now consider the case of a person that is moved not by the passion for the piano or for a football team but by the passion for cleaning, and they are an expert not in pianos and not even in playing schemes but in the prevention and neutralization of contamination. We can observe that her brain is different from

that of other people. Now suppose a psychiatrist tells us that this person is suffering from Obsessive Compulsive Disorder, that is, from a psychopathology. Would this diagnosis be sufficient to affirm that the observed cerebral diversity is similar to that of the patient suffering from progressive paralysis or from lesions of the prefrontal lobe? No, unless we observe anatomo-functional abnormalities with respect to the neuropathological criteria that discriminate a healthy nervous system from a sick one, for example degenerative or neoplastic lesions, outcomes of trauma, signs of infection or autoimmune reactions. If these conditions are not met, then we are in the presence of the many individual differences that characterize every organ of the human body. It does not appear legitimate, therefore, to infer a disease of the brain just because a diversity is observed, even if the diversity observed in the brain corresponds to a psychopathology. If this limit is not admitted, there is a risk of a paradox. Let’s see it. We can imagine, for the benefit of our argument, that the brain of a homosexual person is different from that of a heterosexual [extensive findings indeed suggest that human sexual orientation is associated with brain morphology, e.g., (47)].

Nowadays, no one would say that homosexuality is a form of psychopathology, therefore the observed diversity appears similar to that found in pianists: different interests, different ways of being that correspond to different brains.

Now, suppose we go back in time, to 70 years ago. Homosexuality was considered a form of psychopathology. Would this have implied that the diversity of the brains of homosexuals was analogous to that of the patient with progressive paralysis? That is, can brain diversity be neuropathological or cease to be so, only as a consequence of conventional decisions about what is or is not psychopathological? Here it seems very pertinent what Protopapas and Parrila (49) write about the dyslexia: “... differences in brains are certain to exist whenever differences in behavior exist, including differences in ability and performance. Therefore, findings of brain differences do not constitute evidence for abnormality; rather, they simply document the neural substrate of the behavioral differences. We suggest that dyslexia is best viewed as one of many expressions of ordinary ubiquitous individual differences in normal developmental outcomes. Thus, terms such as “dysfunctional” or “abnormal” are not justified when referring to the brains of persons with dyslexia” (49).

A mental pathology does not necessarily imply a malfunction, an anomaly in the neural mechanisms in

2 It should be noted that also in agreement with Wakefield and Conrad (48), in order to define psychopathology, an evaluation criterion is indispensable. “The HDA maintains that a disorder is a harmful condition—judged by social values, thus value laden—caused by a dysfunction, where “dysfunction” is a factual concept that refers to a failure of some feature of the organism to perform a natural function...”.

which it is implemented. A psychic malfunction does not imply a neural malfunction. To use the computer analogy, a software may not work, if it is poorly made or damaged, without any damage or problem to the hardware in which it is implemented. Similarly, a complex algorithm may not work well, even if the basic software in which it is written is perfect and works smoothly; it is that very algorithm to be faulty.

Of course, if a brain disease is there, there can be psychopathological repercussions. Similarly, if the hardware is damaged, the software and the algorithm might also not work properly. These examples portray well how there is a *non sequitur* between the (obvious) idea that dysfunctional / psychopathological processes are *brain processes* and the assumption that *therefore* their cause *must* be a brain damage, a neural or biochemical dysfunction, a neural disease (12).

Psychotherapy and the brain

Similarly, there is a *non sequitur* between the (obvious) idea that dysfunctional/psychopathological (and recovery) processes are *brain processes* and the assumption that *therefore* the intervention must necessarily and *directly* be on the brain and its functioning [see also (50)].

To think something is a new state of our brain; to learn something is to modify our brain; to relearn, adjust previous learning, is to modify our brain again (12). There might have been, for several concurrent factors, a *dysfunctional* learning, dysfunctional thoughts, and the challenge is, restructuring the learned representations and processes, through new cognitive and affective experiences and mental elaborations. Any change in our conduct or attitudes is a change in our minds; any change in our minds is a change in our brains. Our brain has also been materially “written” by our conduct, experience, and environment. In psychotherapeutic, educational or rehabilitation interventions the challenge is to preserve this route, and this view. For changing our brain, we do not need to directly act on our brain. Similarly, for producing water we do not need (and it is even worst) to join oxygen and hydrogenous; or for changing genes regulation not necessarily we manipulate genes (epigenetics).

According to Karlsson (51): “Psychotherapy outcomes and the mechanisms of change that are related to its effects have traditionally been investigated on the psychological and social levels, by measuring changes in symptoms, psychological abilities, personality, or social functioning. Many psychiatrists have also held the unfortunate dichotomized position that psychotherapy is a treatment for “psychologically based” disorders, while medication is for “biologically based” disorders. During the past several decades, it

has become clear that all mental processes derive from mechanisms of the brain. This means that any change in our psychological processes is reflected by changes in the functions or structures of the brain. Straightforward reductionistic stances, however, are unfounded because there is clear evidence that our subjective experiences affect the brain”.

Empirical and meta-analytical data have shown that:

Several types of psychotherapies modify the brain structure and its functioning. “...cognitive-behavioral therapy (CBT), dialectic behavior therapy (DBT), psychodynamic psychotherapy, and interpersonal psychotherapy alter brain function in patients suffering from major depressive disorder (MDD), obsessive-compulsive disorder, panic disorder, social anxiety disorder, specific phobias, posttraumatic stress disorder, and borderline personality disorder (BPD)” (51); these changes sometimes appear similar to those obtained with drugs and sometimes different. “The majority of these studies have reported similar brain changes after psychotherapy and medication. However, some recent studies have also shown clear differences among these treatment modalities” (51); sometimes psychotherapy modifies precisely the brain characteristics that are considered specific to a disorder. e.g., in depression, “Behavioral therapy for anxiety disorders was consistently associated with attenuation of brain-imaging abnormalities in regions linked to the pathophysiology of anxiety, and with activation in regions related to positive reappraisal of anxiogenic stimuli,” and in OCD: “The symptoms of obsessive-compulsive disorder (OCD) include intrusive thoughts, compulsive behavior, anxiety, and cognitive inflexibility, which are associated with dysfunction in dorsal and ventral corticostriato-thalamocortical (CSTC) circuits” (52). Psychotherapy involving exposure and response prevention has been established as an effective treatment for the affective symptoms, 16 studies measuring neural changes after therapy were included in the review. Post-treatment decreases of symptoms and activity in the ventral circuits during symptom provocation, as well as mainly increased activity in dorsal circuits during cognitive processing. These effects appear to be common to both psychotherapy and medication approaches” (53). It could be argued that these changes are functional and not structural and that the latter may not be affected by psychotherapeutic interventions. However, some data suggest that prolonged psychological interventions can modify those structural aspects that are considered distinctive of a given psychopathological disorder. Some examples: “Research in recent decades has (...) *provided* compelling evidence that learning new behavior can alter the structure of the adult human brain” (42). This learning-dependent structural plasticity has been shown for psychotherapy. Two years of cognitive remediation therapy increased gray matter volume in the fusiform gyrus, hippocampus and amygdala (54) as well as fractional anisotropy in the genu of the corpus callosum in

patients with schizophrenia (55). Ten weeks of cognitive behavioral group therapy reduced gray matter volume in parieto-occipital and prefrontal regions and increased fractional anisotropy in the uncinate and inferior longitudinal fasciculus and structural connectivity in a frontolimbic network in patients with social anxiety disorder (56). “We found that DBT increased gray matter volume of brain regions that are critically implicated in emotion regulation and higher-order functions, such as mentalizing. The role of the angular gyrus for treatment response may reside in its cross-modal integrative function. These findings enhance our understanding of psychotherapy mechanisms of change and may foster the development of neurobiologically informed therapeutic interventions” (57). Hoexter et al. (58) found that abnormalities in gray matter volume in the left putamen were no longer detectable after CBT. Finally, Zhong and colleagues, (59) found that white matter alterations in some regions (i.e., left orbital frontal cortex, right cerebellum, right putamen nucleus, which play an important role in the neural mechanisms of OCD) can be reversible following an effective course of CBT (58, 59).

These data lend themselves to two considerations. The first is that psychotherapy changes the brain. It is worth noting, that affirming this does not necessarily imply mental causation (a very complex and still debated problem) (60). Indeed, as pointed out by Davidson “each individual mental event is in fact a physical event in the following sense: any event that has a mental description has also a physical description. Further, it is only under its physical description that a mental event can be seen to enter into a causal relation with a physical event (or any other event) by being subsumed under a causal law” (61). Psychotherapy consists of an exchange of information that takes place through verbal and non-verbal channels, and since information is nothing more than energy, organized in different ways, but still energy, psychotherapy must have an impact on the brain, and ultimately on the atoms that compose it.

The second consideration is that the influence of psychotherapy on the brain is not non-specific but, as at least suggested by some research, it modifies aspects of the brain that are specifically involved in the psychological disorder which is being treated. It is important to note that this is different for instance from what happens through rehabilitation after a brain injury. For instance, a thrombosis in a cerebral artery is likely to cause the death of a group of neurons which will be substituted by glial cells. Let’s imagine that this causes a functional damage, e.g., aphasia. The function of language can be restored through speech therapy, which thanks to neural plasticity, can modify the micro-anatomic organization of the brain, but it cannot repair the specific area of the brain that was damaged (i.e., its specific substratum), that is, it cannot turn glial cells back into neurons again. The difference with psychotherapy here consists in the

observation that psychotherapy is able to change those same neural characteristics that are considered as proof of the putative neuropathological origin of those mental disorders. For instance, glucose metabolic rates in the right head of the caudate change when OCD is successfully treated with either fluoxetine or behavior therapy (62). This means that psychotherapy is able to restore the specific substratum of a psychopathological disorder, precisely because this substratum was never “damaged.”

If psychotherapy is able to change the specific substratum of a psychopathological disorder, then it is difficult to argue that “The diseases we treat are diseases of the brain,” only on the basis of the discovery of specific cerebral, functional and structural characteristics. If psychopathologies were true neurological diseases, such as Alzheimer’s or multiple sclerosis or Huntington’s chorea, their specific neural substrate would not be modifiable by psychotherapy. Indeed, it is not plausible that a psychotherapy can reduce the beta amyloid plaques in Alzheimer’s disease, even if psychotherapy could reduce anxiety and depression reactive to the awareness of being affected by this serious disease.

Conclusions

Interpreting the statement “The diseases we treat are diseases of the brain” in a literal way implies, in our opinion, two critical points. The first is the assumption of an eliminativist perspective, at least in the domain of psychopathology. Psychopathological manifestations would be devoid of intrinsic meaning and therefore would need an explanation at the neural level, a level that Dennett would define as “sub-personal” (63). Moreover, according to this perspective it would be useless, or even misleading, to try to explain psychopathology by resorting to the contents of the patient’s mind, (i.e., his mental representations, his beliefs and his own goals); in other words, to use the explanation level which, according to Dennett, we could define as “personal.” In short, the statement “The diseases we treat are diseases of the brain” appears underpinned by an eliminativist reductionism that we here challenged by presenting arguments in favor of emergent reductionism.

The second point is the following. The differences found in the brains of people with psychopathology would be neuropathological differences, that is, abnormal with respect to the anatomical and physiological criteria that define the healthy brain. Here, we contested the idea that it is enough to find a difference between the brains of people suffering from psychopathology and that of people who are not affected or affected by different psychopathologies. We therefore disentangled between psychopathological disorders underlying a true pathology

of the brain from those underlying simple anatomical or functional differences. Differences that are similar to those that are normally found between individuals, even among those who are not affected by psychopathologies. Finally, we considered some studies which show how purely psychological interventions can reduce or eliminate the differences between the brains of people with psychopathology and those of people without.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work, and approved it for publication.

Funding

The publication of this work has been funded by the School of Cognitive Psychotherapy, SPC, Rome, Italy.

References

- Ross DA, Travis MJ, Arbuckle MR. The future of psychiatry as clinical neuroscience: why not now? *JAMA Psychiatry*. (2015) 72:413–4. doi: 10.1001/jamapsychiatry.2014.3199
- Johnson B, Panksepp J. *Textbook of Biological Psychiatry*. Hoboken, NJ: Wiley (2002).
- Churchland PM. *Matter and Consciousness (revised ed.)*. Cambridge, MA: MIT Press (1988).
- Feyerabend P. Materialism and the mind-body problem. *Rev Metaphys*. (1963) 17:49–66.
- Rorty R. In Defense of eliminative materialism. *Rev Metaphys*. (1970) 24:112–21.
- Rosenthal D.. *Materialism and the Mind-body Problem*. Prentice-Hall, NJ: Englewood cliffs (1971).
- Churchland PM. *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science*. Cambridge, MA: MIT Press (1992). doi: 10.7551/mitpress/4940.001.0001
- Feinberg TE, Mallatt J. Phenomenal consciousness and emergence: eliminating the explanatory gap. *Front Psychol*. (2020) 11:1041. doi: 10.3389/fpsyg.2020.01041
- O'Connor T, Wong HY. *Emergent Properties*. Stanford, CA: The metaphysics Research Lab. (2015).
- Chalmers D. Strong and weak emergence. In: *The Re-Emergence of Emergence: The Emergentist Hypothesis from Science to Religion*, Clayton P, Davis P, eds. 244–256 (2006).
- Silberstein M. In defence of ontological emergence and mental causation. In: *The Re-Emergence of Emergence: The Emergentist Hypothesis from Science to Religion*, Clayton P, Davis P, eds. (2006). p. 203–226
- Castelfranchi C. For a science of layered mechanisms: beyond laws, statistics, and correlations. *Front Psychol*. (2014) 5:536. doi: 10.3389/fpsyg.2014.00536
- Kosfeld M, Heinrichs M, Zak PJ, Fischbacher U, Fehr E. Oxytocin increases trust in humans. *Nature*. (2005) 435:673–6. doi: 10.1038/nature03701
- Fehr E. On the Economics and biology of trust. *J Eur Econ Assoc*. (2009) 7:235–66. doi: 10.1162/JEEA.2009.7.2-3.235
- Castelfranchi C, Falcone R. *Trust Theory. A Socio-Cognitive and Computational Model*. Hoboken, NJ: Wiley (2010).
- Castelfranchi C. *Review of Neuroeconomics: Decision Making and the Brain* (2009).
- Morowitz H. *The Emergence of Everything*. Oxford: Oxford University Press. (2002).
- Kim, J. *Being realistic about emergence*. In: *The Re-Emergence of Emergence: The Emergentist Hypothesis from Science to Religion*. (2006). p. 189–203.
- Kim, J. *Supervenience and the Mind. Selected Philosophical Essays*. Cambridge, MA: Cambridge studies in Philosophy (1993).
- Borsboom D, Cramer AOJ, Kalis A. Brain disorders? Not really: Why network structures block reductionism in psychopathology research. *Behav Brain Sci*. (2018) 42:e2. doi: 10.1017/S0140525X17002266
- Basile B, Mancini F, Macaluso E, Caltagirone C, Frackowiak RSJ, Bozzali M, et al. Deontological and altruistic guilt: evidence for distinct neurobiological substrates. *Hum Brain Mapp*. (2011) 32:229–39. doi: 10.1002/hbm.21009
- Zhong CB, Liljenquist K. Washing away your sins: threatened morality and physical cleansing. *Science*. (2006) 313:1451–1452. doi: 10.1126/science.1130726
- D'Olimpio F, Mancini F. Role of deontological guilt in obsessive-compulsive disorder-like checking and washing behaviors. *Clin Psychol Sci*. (2014) 2:727–39. doi: 10.1177/2167702614529549
- Ottaviani C, Mancini F, Petrocchi N, Medea B, Couyoumdjian A. Autonomic correlates of physical and moral disgust. *Int J Psychophysiol*. (2013) 89:57–62. doi: 10.1016/j.ijpsycho.2013.05.003
- Ottaviani C, Collazzoni A, D'Olimpio F, Moretta T, Mancini F. I obsessively clean because deontological guilt makes me feel physiologically disgusted! *J Obsessive Compuls Relat Disord*. (2019) 20:21–9. doi: 10.1016/j.jocrd.2018.01.004
- Ottaviani C, Mancini F, Provenzano S, Collazzoni A, D'Olimpio F. Deontological morality can be experimentally enhanced by increasing disgust: a transcranial direct current stimulation study. *Neuropsychologia*. (2018) 119:474–81. doi: 10.1016/j.neuropsychologia.2018.09.009
- Salvo G, Provenzano S, Di Bello M, D'Olimpio F, Ottaviani C, Mancini F, et al. Filthiness of immorality: manipulating disgust and moral rigidity through non-invasive brain stimulation as a promising therapeutic tool for obsessive compulsive disorder. *Clin Psychol Sci*. (2021) 10:127–140. doi: 10.1177/21677026211009508
- Walter H, Ciaramidaro A, Adenzato M, Vasic N, Ardito RB, Erk S, et al. Dysfunction of the social brain in schizophrenia is modulated by intention type: an fMRI study. *Soc Cogn Affect Neurosci*. (2009) 4:166–76. doi: 10.1093/scan/nsn047

Acknowledgments

We wish to thank Prof. Mauro Giacomantonio for his helpful comments.

Conflict of interest

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

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

29. Putnam H. Minds and machines. In: *Dimensions of Mind*, ed. Hook S. New York: New York University Press (1960). p. 138–164.
30. Block NJ, Fodor JA. What psychological states are not. *Philos Rev.* (1972) 81:159–81. doi: 10.2307/2183991
31. Bechtel W, Mundale J. Multiple realizability revisited: linking cognitive and neural states. *Philos Sci.* (1999) 66:175–207. doi: 10.1086/392683
32. Rosenblueth A, Wiener N, Bigelow J. Behavior, purpose and teleology. *Philos Sci.* (1943) 10:18–24. doi: 10.1086/286788
33. Miller GA, Galanter E, Pribram KH. Plans and the Structure of Behavior. New York, NY: Holt (1960). doi: 10.1037/10039-000
34. Castelfranchi C, Devescovi A, Burani C. Understanding causal relations. In: *Proceedings of the Conference On Language, Reasoning and Inference: An Interdisciplinary Conference in Cognitive Science*. Edinburgh: University of Edinburgh, School of Epistemics (1982). pp. 2–6.
35. Boorse C. Health as a theoretical concept. *Philos Sci.* (1977) 44:542–73. doi: 10.1086/288768
36. Wakefield JC. The biostatistical theory vs. the harmful dysfunction analysis, part 1: is part-dysfunction a sufficient condition for medical disorder? *J Med Philos A Forum Bioeth Philos Med.* (2014) 39:648–82. doi: 10.1093/jmp/jhu038
37. Dekker M, Ende J, Verhulst F. Emotional and behavioral problems in children and adolescents with and without intellectual disability. *J Child Psychol Psychiatry.* (2002). 43:1087–98. doi: 10.1111/1469-7610.00235
38. Einfeld SL, Ellis LA, Emerson E. Comorbidity of intellectual disability and mental disorder in children and adolescents: a systematic review. *J Intellect Dev Disabil.* (2011) 36:137–43. doi: 10.1080/13668250.2011.572548
39. Whiteside SP, Port JD, Abramowitz JS. A meta-analysis of functional neuroimaging in obsessive-compulsive disorder. *Psychiatry Res Neuroimaging.* (2004) 132:69–79. doi: 10.1016/j.psychres.2004.07.001
40. Marazziti D, Akiskal HS, Rossi A, Cassano GB. Alteration of the platelet serotonin transporter in romantic love. *Psychol Med.* (1999) 29:741–5. doi: 10.1017/S0033291798007946
41. Swedo EA, Leckman JF, Rose NR. From research subgroup to clinical syndrome: modifying the PANDAS criteria to describe PANS (Pediatric Acute-onset Neuropsychiatric). *Pediatr/ and Ther.* (2012) 2:1–8. doi: 10.4172/2161-0665.1000113
42. May A. Experience-dependent structural plasticity in the adult human brain. *Trends Cogn Sci.* (2011) 15:475–82. doi: 10.1016/j.tics.2011.08.002
43. Münte TF, Altenmüller E, Jäncke L. The musician's brain as a model of neuroplasticity. *Nat Rev Neurosci.* (2002) 3:473–8. doi: 10.1038/nrn843
44. Gärtnert H, Minnerop M, Pieperhoff P, Schleicher A, Zilles K, Altenmüller E, et al. Brain morphometry shows effects of long-term musical practice in middle-aged keyboard players. *Front Psychol.* (2013) 4:636. doi: 10.3389/fpsyg.2013.00636
45. Molenberghs P, Halász V, Mattingley JB, Vanman EJ, Cunnington R. Seeing is believing: neural mechanisms of action-perception are biased by team membership. *Hum Brain Mapp.* (2013) 34:2055–68. doi: 10.1002/hbm.22044
46. Andrews TJ, Smith RK, Hoggart RL, Ulrich PIN, Gouws AD. Neural correlates of group bias during natural viewing. *Cereb Cortex.* (2019) 29:3380–9. doi: 10.1093/cercor/bhy206
47. Votinov M, Goerlich KS, Puiu AA, Smith E, Nickl-Jockschat T, Derntl B, et al. Brain structure changes associated with sexual orientation. *Sci Rep.* (2021) 11:5078. doi: 10.1038/s41598-021-84496-z
48. Wakefield JC, Conrad JA. Harm as a necessary component of the concept of medical disorder: Reply to Muckler and Taylor. *J Med Philos.* (2020) 45:350–70. doi: 10.1093/jmp/jhaa008
49. Protopapas A, Parrila R. Is dyslexia a brain disorder? *Brain Sci.* (2018) 8:61. doi: 10.3390/brainsci8040061
50. Carcione A. Psychotherapy and psychiatric drugs in psychiatry: what relationship, and what hierarchy? (2011).
51. Karlsson H. How psychotherapy changes the brain: understanding the mechanisms. *Psychiatr Times.* (2011) 28:21–3. doi: 10.1136/bmj.e1188
52. Roffman J, Marci C, Glick D, Dougherty D, Rauch S. Neuroimaging and the functional neuroanatomy of psychotherapy. *Psychol Med.* (2005) 35:1385–98. doi: 10.1017/S0033291705005064
53. Thorsen AL, van den Heuvel OA, Hansen B, Kvale G. Neuroimaging of psychotherapy for obsessive-compulsive disorder: a systematic review. *Psychiatry Res Neuroimaging.* (2015) 233:306–13. doi: 10.1016/j.pscychres.2015.05.004
54. Eack SM, Hogarty GE, Cho RY, Prasad KMR, Greenwald DP, Hogarty SS, et al. Neuroprotective effects of cognitive enhancement therapy against gray matter loss in early schizophrenia: results from a 2-year randomized controlled trial. *Arch Gen Psychiatry.* (2010) 67:674–82. doi: 10.1001/archgenpsychiatry.2010.63
55. Penadés R, Pujol N, Catalán R, Massana G, Rametti G, García-Rizo C, et al. Brain effects of cognitive remediation therapy in schizophrenia: a structural and functional neuroimaging study. *Biol Psychiatry.* (2013) 73:1015–23. doi: 10.1016/j.biopsych.2013.01.017
56. Steiger VR, Brühl AB, Weidt S, Delsignore A, Rufer M, Jäncke L, et al. Pattern of structural brain changes in social anxiety disorder after cognitive behavioral group therapy: a longitudinal multimodal MRI study. *Mol Psychiatry.* (2017) 22:1164–71. doi: 10.1038/mp.2016.217
57. Mancke F, Schmitt R, Winter D, Niedtfeld I, Herpertz SC, Schmahl C, et al. Assessing the marks of change: how psychotherapy alters the brain structure in women with borderline personality disorder. *J Psychiatry Neurosci.* (2018) 43:171–81. doi: 10.1503/jpn.170132
58. Hoexter MQ, de Souza Duran FL, D'alcante CC, Dougherty DD, Shavitt RG, Lopes AC, et al. Gray matter volumes in obsessive-compulsive disorder before and after fluoxetine or cognitive-behavior therapy: a randomized clinical trial. *Neuropsychopharmacology.* (2012) 37:734–45. doi: 10.1038/npp.2011.250
59. Zhong Z, Yang X, Cao R, Li P, Li Z, Lv L, et al. Abnormalities of white matter microstructure in unmedicated patients with obsessive-compulsive disorder: changes after cognitive behavioral therapy. *Brain Behav.* (2019) 9:e01201. doi: 10.1002/brb3.1201
60. Murphy N. Emergence and mental causation. In: *The Re-Emergence of Emergence: The Emergentist Hypothesis from Science to Religion*. Oxford: Oxford University Press. (2006). p. 227–243.
61. Davidson D. Mental events. In: *Contemporary Materialism*, Foster L, Swanson JW, eds. Amherst: University of Massachusetts Press. (1970).
62. Baxter Jr LR, Schwartz JM, Bergman KS, Szuba MP, Guze BH, Mazziotta JC, et al. Caudate glucose metabolic rate changes with both drug and behavior therapy for obsessive-compulsive disorder. *Arch Gen Psychiatry.* (1992) 49:681–9. doi: 10.1001/archpsyc.1992.01820090009002
63. Dennett DC. *Content and Consciousness*. London: Routledge & Kegan Paul. (1968).



OPEN ACCESS

EDITED BY

Antonino Carcione,
Terzo Centro di Psicoterapia, Italy

REVIEWED BY

J. Shashi Kiran Reddy,
ThoughtSeed Labs, India
Zhen-Dong Wang,
Shanghai University of Traditional
Chinese Medicine, China
Daniela Flores Mosri,
Universidad Intercontinental, Mexico

*CORRESPONDENCE

Jicheng Chen
chenjicheng2008@163.com
Linlin Chen
linlinchen0707@163.com

SPECIALTY SECTION

This article was submitted to
Perception Science,
a section of the journal
Frontiers in Neuroscience

RECEIVED 22 June 2022

ACCEPTED 30 September 2022

PUBLISHED 25 October 2022

CITATION

Chen J and Chen L (2022) The hard
problem of consciousness—A
perspective from holistic philosophy.
Front. Neurosci. 16:975281.
doi: 10.3389/fnins.2022.975281

COPYRIGHT

© 2022 Chen and Chen. This is an
open-access article distributed under
the terms of the [Creative Commons
Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use,
distribution or reproduction in other
forums is permitted, provided the
original author(s) and the copyright
owner(s) are credited and that the
original publication in this journal is
cited, in accordance with accepted
academic practice. No use, distribution
or reproduction is permitted which
does not comply with these terms.

The hard problem of consciousness—A perspective from holistic philosophy

Jicheng Chen* and Linlin Chen*

Department of Vasculocardiology, Shenzhen Longhua District Central Hospital, Guangdong
Medical University, Shenzhen, China

Based on a material view and reductionism, science has achieved great success. These cognitive paradigms treat the external as an objective existence and ignore internal consciousness. However, this cognitive paradigm, which we take for granted, has also led to some dilemmas related to consciousness in biology and physics. Together, these phenomena reveal the interaction and inseparable side of matter and consciousness (or body and mind) rather than the absolute opposition. However, a material view that describes matter and consciousness in opposition cannot explain the underlying principle, which causes a gap in interpretation. For example, consciousness is believed to be the key to influencing wave function collapse (reality), but there is a lack of a scientific model to study how this happens. In this study, we reveal that the theory of scientific cognition exhibits a paradigm shift in terms of perception. This tendency implies that reconciling the relationship between matter and consciousness requires an abstract theoretical model that is not based on physical forms. We propose that the holistic cognitive paradigm offers a potential solution to reconcile the dilemmas and can be scientifically proven. In contrast to the material view, the holistic cognitive paradigm is based on the objective contradictory nature of perception rather than the external physical characteristics. This cognitive paradigm relies on perception and experience (not observation) and summarizes all existence into two abstract contradictory perceptual states (Yin-Yang). Matter and consciousness can be seen as two different states of perception, unified in perception rather than in opposition. This abstract perspective offers a distinction from the material view, which is also the key to falsification, and the occurrence of an event is inseparable from the irrational state of the observer's conscious perception. Alternatively, from the material view, the event is random and has nothing to do with perception. We hope that this study can provide some new enlightenment for the scientific coordination of the opposing relationship between matter and consciousness.

KEYWORDS

hard problem of consciousness, reductionism, holistic philosophy, perception, contradiction, free energy principle, quantum mechanics

Introduction

In the past few 100 years, biology and physics have achieved remarkable success. On the basis of material view and reductionism, we regarded the external as an objective being and ignored the inner conscious experience. The natural phenomena and laws are described by observation and statistics and the macroscopic phenomena are explained by microscopic quantum. For example, the phenomenon of life is explained by cells and the origin of the universe is explained by microscopic quantum. We have long been accustomed to deploying the cognitive paradigm of reductionism. Its underlying assumptions and methods are taken for granted. However, this cognitive paradigm has brought about a series of puzzles about consciousness in both biology and physics, reflecting its limitations.

Over the past few decades, neural and cognitive scientists have made remarkable progress in studying consciousness from a physical level (Dehaene and Changeux, 2011; Boly et al., 2013; Owen, 2019). Koch et al. (2016) argue that we are now at a point where we can understand consciousness in a scientific way, such as neuronal correlates of consciousness (NCC), and not as a philosophical question, especially in the field of visual consciousness (Crick and Koch, 1998; Koch et al., 2016), and these represent the functional side of consciousness research. However, subjective experiences cannot be explained from an objective standpoint. Relatedly, how do organisms produce the meaning of life that we experience, and how it relates to the brain (the mind–body problem) (Reddy, 2016; Levin, 2020)? This represents the “hard problem of consciousness” (Chalmers, 1998; Solms, 2014, 2021; Solms and Friston, 2018).

From another point of view, similar to the above problem, there is a contradiction between free will and causality based on time and space, which cannot be currently explained by reductionism (Heisenberg, 2009; Rappaport, 2011; Hillman, 2018). For humans, if our brain produces certain thoughts, we can detect the electrical activity in the corresponding regions of the brain with instruments, but we do not have an idea what causes nerve cells to become excited. We do not get excited by an external electrode stimulation, which is perceptually called free will.

We establish causality based on time and space but, in an experiment like this, the electrical excitation of the brain’s nerves is triggered by invisible thoughts or motivation that we think of as autonomous without any physical cause. But we do not know exactly how invisible thoughts lead to physical changes in the brain. This feature of consciousness undoubtedly challenges the idea of causality, dependent on space and time. Is the sense of freedom we perceive not subject to the laws of the physical world? If we attribute the neuroelectric excitation to the external physical environment, it means that we are like a robot, free will is just a mechanical reflection of the environment, a kind of illusion. Although there is some neurobiological evidence

against the nature of free will, the evidence is not convincing. More importantly, if free will is an illusion, how do we explain the meaning of life? (Brass et al., 2019; Lavazza, 2019).

The cognitive paradigm of material view and reductionism also leads to the puzzle of consciousness in quantum physics. Matter and consciousness, which used to be philosophical issues, have become concrete scientific problems (Frank, 2015). Quantum mechanics has revealed some puzzling microscopic phenomena, such as wave-particle duality and quantum entanglement. These phenomena have challenged classical thinking regarding the objective physical reality and suggest an inseparable aspect of matter and consciousness, in which we cannot treat consciousness as an illusion. To solve the core problem of how quantum random collapse produces a well-ordered world, scientists have focused on consciousness as the key. John von Neumann argued that only consciousness could eventually collapse the wave function to produce a definite reality (Neumann, 2020). Eugene Paul Wigner argued that the role of conscious creatures in quantum mechanics must be different from that of inanimate measuring devices (Wigner et al., 1992). In 2007, Robert Lanza and Bob Berman came up with a new concept termed biocentrism (Lanza, 2012). They proposed that order or reality requires the presence of a conscious observer. However, how consciousness causes wave function collapse (or affects reality) remains unclear.

In conclusion, we think that although these consciousness-related puzzles take different forms in different disciplines, what they have in common is that they jointly reveal that matter and consciousness (body-mind) interact and cannot be separated, but they lack a scientific explanation of the underlying principle and mechanism. For example, how can abstract subjective experiences lead to physical neural excitation (we cannot observe any medium)? How does consciousness affect wave function collapse? The cognitive paradigm of the material view, which puts matter and consciousness in opposition, will lead to such gaps in interpretation. We propose there is another cognitive paradigm that can reconcile the antagonistic relationship between matter and consciousness and reconcile these dilemmas.

Currently, scientists are trying to build models to understand the nature of consciousness (Seth and Bayne, 2022). The free energy principle proposed by Friston and Stephan (2007) is applied to explain this puzzle and it has become a compelling solution (Solms, 2018, 2021). The precursor to the free-energy principle was a way of describing how the brain works. At every level, the brain’s prediction of what the most likely experience will be in a given environment is compared with the actual information received from the senses. If the prediction is not correct, then higher levels of the nervous system are required (Friston and Stephan, 2007; Ramstead et al., 2018). The free energy principle describes the mind–brain system as any other adaptive biological system, connecting psychological sciences, neuroscience, and

related fields in confluence and synergy with psychoanalytic concepts (Cieri and Esposito, 2019). In addition, there are some other well-known theories. Integrated information theory (IIT), developed by Tononi and collaborators, focuses on the objectivity of subjective experience itself (Koch et al., 2016). The orchestrated objective reduction (or “Orch OR”) model, developed by Hameroff and Penrose, has suggested that consciousness is the result of the collapse of wave functions caused by quantum gravity in microtubules (Hameroff, 2012; Hameroff and Penrose, 2014). These hypotheses offer a deeper insight into the understanding of the phenomenal aspect of consciousness (Rees et al., 2002).

Philosophical perspective may offer inspiration for scientific studies and provide theoretical foundations for understanding the relationship between matter and consciousness (or the nature of consciousness; Churchland and Churchland, 1997; Sturm, 2012). The confusion afflicting physics today has led scientists to understand the universe from a more holistic perspective. Niels Bohr believed that the Taiji diagram (the logo of holistic philosophy) contained the principle of wave-particle duality (Capra, 2000), and quantum physicist Bohm (2004) tried to explain the origins of order from the perspective of wholeness in his ontological picture of the universe. However, we still need to build a more detailed theoretical model of consciousness that can be described scientifically on the basis of a deeper understanding of holistic philosophy.

There are significant cultural and cognitive differences between the East and West (Wang et al., 2021). The material view is not the only cognitive paradigm in which we describe the movement and development of the universe, the *Book of Changes* and *Tao Te Ching* tend to understand the world from a holistic perspective (Yutang, 1948).

This holistic philosophy has profoundly affected different cultural forms of the East and provided a series of effective social applications (Liu, 2008; Kafatos and Yang, 2016). As an important work from the perspective of holistic philosophy, this study discusses our understanding of *Tao Te Ching*. We propose that the theory of scientific cognition exhibits a paradigm shift in terms of perception. With a tendency implying that reconciling the dilemma of consciousness requires an abstract theoretical model that is not based on physical forms, the Taiji diagram in the philosophy of the holistic view is a candidate. We propose that the holistic perspective provides a potential solution and new inspirations to solve current reductionism-based scientific dilemmas.

Objectivity is the foundation for establishing a theoretical system in both the material view and holistic view. The cognitive paradigm of holistic philosophy is based on the basic objectivity of perception, which shows the objective nature of contradiction beyond the control of the individual, but intuitively, we think of it as subjective or as belonging to an individual. Although objectivity is abstract, it is the basis and key to establishing a holistic description system, just

like our description of the objectivity of different physical quantities. The holistic view relies on conscious experience (rather than observation) and reduces everything to two abstract perceptual states: Yin-Yang. We regard matter and consciousness as two contradictory perceptual states that are unified in perception. Their unity implies that the inner and the outer are not absolute opposites, but that there is an interconvert relationship between the two. This perspective avoids the dilemma of consciousness caused by the emphasis of the material view on the external objective description. We will elaborate on the holistic philosophy in the following paragraphs.

Holistic philosophy

Different explanations of the origin and evolution of the universe from *Tao Te Ching*

Tao Te Ching, written by Lao Zi, has had a profound influence worldwide. It offers a representative interpretation of holistic philosophy, although there is no unified interpretation of the book. In this part, we first introduce the core ideas of the holistic philosophy in the *Tao Te Ching*. We will discuss this in detail with some examples in the following chapters. Its first and most important chapter includes a brief exposition of the origin of *everything*:

The Tao that can be told of, Is not the Absolute Tao;/The Names that can be given, Are not Absolute Names./The Nameless is the origin of Heaven and Earth, The Named is the Mother of All Things./Therefore: Oftentimes, one strips oneself of passion./In order to see the Secret of Life; Oftentimes, one regards life with passion./In order to see its manifest forms./These two (the Secret and its manifestations) Are (in their nature) the same;/They are given different names, When they become manifest./They may both be called the Cosmic Mystery:/Reaching from the Mystery into the Deeper Mystery, Is the Gate to the Secret of all life (Yutang, 1948).

We think this chapter has the following three meanings:

Existence is relative and an objective material entity independent of perception is essentially indescribable

The creation of everything (reality or a phenomenon to be described, such as the state of a particle rather than a particle entity) and perception occur simultaneously and irreplaceably; they are two sides of the same coin. This differs greatly from the objective observations that we assume in intuition and basic scientific assumptions. According to our understanding,

a state of being always requires symmetric physical descriptors, such as up and down, large and small, light and heavy, and more and less. An alphabet without letters cannot be described, and, in the same way, being or matter cannot exist without these contradictory descriptors (difference), which cannot exist independently from perception. In other words, an objective material entity or physical descriptor (state) cannot exist independent of perception (Lanza, 2012). For example, the statement “our bodies are made up of cells,” is usually thought of as an objective phenomenon, and cells are an objective existence that is the same for everyone, whether you know or observe them or not. However, according to our understanding of *Tao Te Ching*, the observed cells (“The Named” in *Tao Te Ching*) and the feeling of the person describing the cells (“The Nameless” in *Tao Te Ching*, cannot be concretized) are two objective states of existence that appear simultaneously and cannot be replaced by one another.

On one hand, cells exhibit external physical characteristics, such as mass and size, which can be observed, and the objectivity of cells is thus recognized. On the other, a doctor who knows a lot about cells and a patient who is being treated have different perceptions of the “cell” (a material entity regarded as objective, which forms different perceptual states for doctor and patient) that result in different realities or state of feelings caused by their different roles (the doctor and the patient are, respectively, active or passive reality). This has important meaning beyond physical form; although it cannot be described concretely, it is also an objective aspect. Therefore, if the cell is defined as a material entity that is objective or the same to everyone, this sort of cognition is one-sided. Although we can explain the body in terms of the laws of cells based on reductionism, we also know that a group of cells does not equal a person. Subjective experience cannot be explained (Figure 1). A holistic view focuses on the objectivity of these abstract perceived states (the meaning of feeling and subjective experience) rather than external physical forms. It can be used to build a different descriptive system depending on the perception that completely differs from material views. Many social applications of the East, such as traditional Chinese medicine, acupuncture, architectural design styles, and culture overall, derive from this philosophical view.

An objective law describing everything independent of perception does not exist too

While we may ponder whether the world is deterministic or non-deterministic, which remains unclear, holistic philosophy may provide a reference point for this question. When we undertake scientific explorations, we see ourselves as observers

based on the distinction between our physical form and external objects, which we take for granted. According to reductionism, we have succeeded in explaining macroscopic phenomena in terms of microscopic quanta. This leads us to believe that we can construct a theory explaining *everything* based on reductionism. Following this assumption, external objective laws and the movement of the universe have nothing to do with perception. However, according to our understanding of holistic philosophy, a theory of *everything* cannot exist independently of perception. An objective law describing the evolution of everything (reality) independent of perception does not exist, and we cannot make objective remarks as independent observers or separate perception from the laws of nature. Conscious experiences and preferences in feelings participate in the creation of order/reality. We will discuss this assumption in detail in the next chapter.

The way order and reality occur depends on the state of perception

In Chapter 42 of *Tao Te Ching*, Lao Zi defined the development trend of *everything* as follows:

“Out of Tao, One is born;/Out of One, Two;/Out of Two, Three;/Out of Three, the created universe./The created universe carries the yin at its back, and the yang in front;/Through the union of the pervading principles it reaches harmony” (Yutang, 1948).

As per our understanding, based on the perceived differences in the positions of the two sides of a conflict, the development of reality always tends toward the “good” side of feelings (it is a relative concept that depends on perception), such as reasonableness, balance, equality, unity, and fairness (the created universe carries the Yin at its back and the Yang in front). Otherwise, it will increasingly encounter resistance, making this form untenable and leading to either a collapse or a shift to the opposition.

In general, the movement of everything is always from opposition to unity. The frame of reference that influences reality is an internal rather than an external concrete frame of reference. In other words, causality is not external but inseparable from perception, being a relative concept depending on the state of perception.

If body and mind are two appearances (aspects) of the same underlying thing, then what stuff is the underlying thing made of? In other words, using the analogy of thunder and lightning, what is the metapsychological equivalent of “electricity” i.e., the thing that gives rise to thunder and lightning, both? (Solms, 2018).

According to holistic philosophy, matter and consciousness (body-mind, external or internal) are not opposites, but two

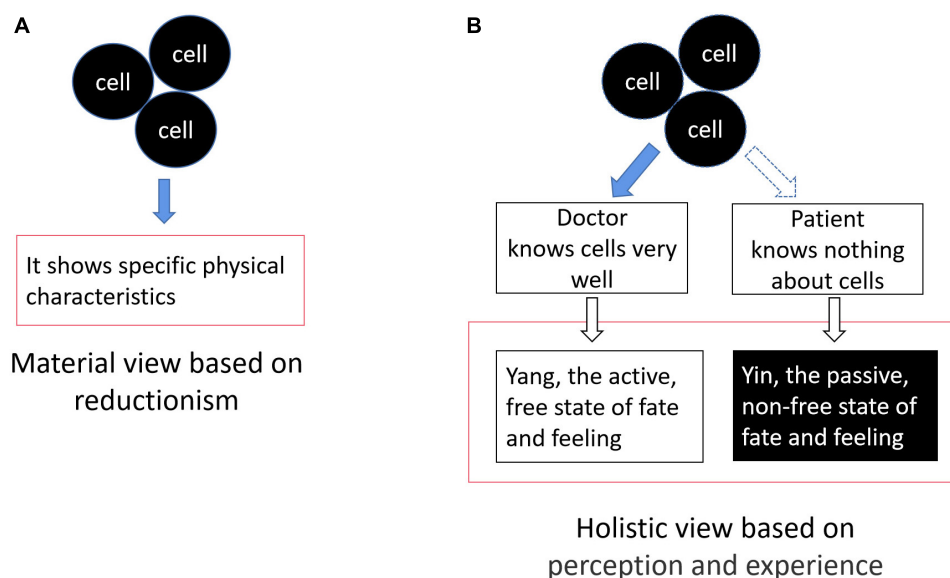


FIGURE 1

The material view and the holistic view put different emphasis on objectivity. (A) Cells are considered objective material entities that are the same for everyone whether you know/observe it or not. This paradigm focuses on external objectivity based on observation. (B) Understanding the rules of the cell can cure disease, although this effect is universally applicable, however, the different experience on the “same and objective” cells results in different fates. Different “fates” are also another aspect of objectivity. This paradigm emphasizes the objectivity of feelings and experiences rather than external physical forms.

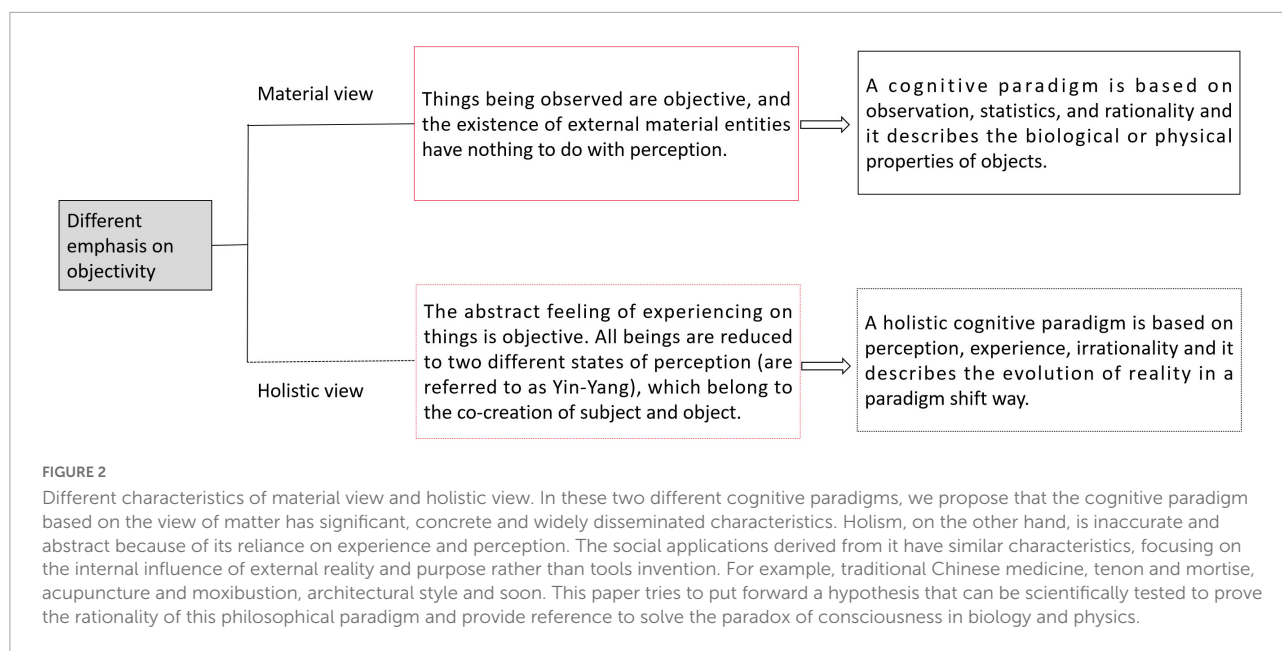
sides of the same coin. They are seen as two states that can be separated by perception (represented by Yin-Yang), which can be used to describe the evolution of reality in a paradigm shift way. Rather than relying on observations and statistics, this kind of description relies on conscious experience. In contrast to the material view, the unity of matter and consciousness at the level of perception also implies that the occurrence of reality is an inseparable process from conscious experience.

For example, over the process of evolution, the nests that ants build to adapt to their environment must have changed dramatically, and we are using the ant nest as a metaphor for reality (the result of an event or phenomenon being observed). If we observe a nest of ants and replace any individual ant, we will find that the construction of the ant nest will not be affected. We think of the individual as unimportant; ants build nests by instinct, which is unconscious behavior. This is just a description of a static phenomenon, the individual conscious experience of ants is ignored. However, we know (through perception and empathy, rather than observation) that the structure of ant nests also evolves and this process is not random; it depends on the constant adjustment of the invisible individual's perceptual state to adapt to the external environment.

In other words, reality comes from the interaction of the individual (inner feelings and experiences) with the external environment and depends on the preferences or tendencies in perception. The irrational sense (to seek a more harmonious state of feeling) is a factor in the creation of observed reality, although this is not visible to the observer. It is an abstract

and relative concept. This is an important difference between holism and material view. As an observer, reductionism tries to find causes from the outside and descriptions of concrete objects' motion. Following a holistic view, perception is an objective being, the occurrence of reality is inseparable from the state of conscious experience (inner). Although we use ants as an example, according to holism, this description applies to *everything*, even if the object is a microparticle or abiotic (the created universe carries the “Yin” at its back and the “Yang” in front).

This is not to say that inanimate objects or particles have perception or consciousness, but that the state of perception itself is objective. It is an objective existence born together from subject and object. Unlike material views, which treat external material entities as objective beings, in a holistic view, the state of perception (Yin-Yang; is the individual's reality passive or active, free or not free, purpose achieved or not achieved) and the preference of feeling is the most objective existence. It does not depend on external physical forms to describe objects, but different realities (reality is passive/Yin or active/Yang) experienced through perception. Therefore, the object described through a holistic view is abstract reality (reality is relative, and the “same” phenomenon means different things to different people) inseparable from perception, strongly differing from material views (Figure 2). Why can reality be described? In later chapters, we will discuss the objectivity of perception, its character beyond individual control determines the basis on which reality can be described.



Discussion on the development of characteristics of scientific cognition

Scientific cognition shows a significant paradigm shift trend from opposition to unity in terms of perception

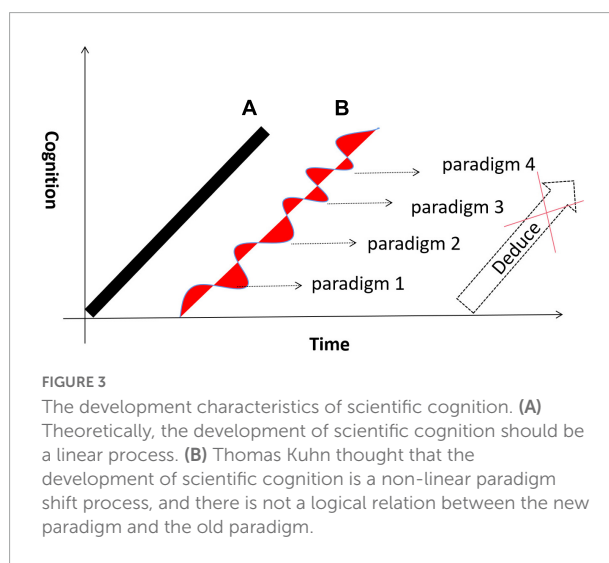
In (section “The paradigm-shifting trends of reality depend on irrational feelings”) we propose that an objective law describing everything independent of perception does not exist. Here, we will discuss in detail why the process of scientific exploration is not just a process of objective observation.

The term “paradigm shift” was first coined by the twentieth-century philosopher Thomas Kuhn. In Kuhn’s view, a paradigm is a specific knowledge system formed by a series of results obtained from scientific research (Kuhn, 2004). The widespread acceptance of a paradigm indicates the maturation of a scientific field. When existing paradigms fail to explain certain natural phenomena, new paradigms that can explain them emerge. Kuhn also believed that there is nological relation between the new paradigm and the old paradigm. In other words, new theoretical paradigms cannot be deduced from old paradigms by relying on logic (Figure 3). However, if we analyze key scientific theories of physics and biology from the perspective of their historical development, we will find that the cognitive pattern (we are concerned not with the mathematical or physical form of the theories but with their abstract meaning in perception) abstracted by these revealed phenomena show surprising similarities and development trends.

Physics

Before Copernicus, the Earth was thought to be the center of the universe. In 1543, Copernicus formally proposed the heliocentric theory, placing the sun at the center of the universe. In terms of spatial arrangement, this theory overturned the self-centered (human) cognitive model.

In 1687, Newton proposed the law of universal gravitation, which became the cornerstone of classical physics. This allows us to accurately describe the motion of objects based on the principle of force interactions. A simple linear causal cognitive model is created based on the opposition between time and space. Newton’s view of space and time dominated physics for over 200 years until Einstein’s theory of relativity deepened our



understanding of Time and Space (quality and energy), which do not exist independently and are naturally linked. This discovery overturned the opposing relationship of time and space and indicated the universality of connections on a material level. Nonetheless, the observer and the object remained in a state of opposition (more basic forms of opposition).

In the nineteenth century, scientists came together to develop quantum theory, represented by the phenomenon of wave-particle duality and quantum entanglement. At the microscopic quantum level, the description of a quantum state requires a conscious observer, and the observer and the object (subject and object) are inseparable. It further deepened the scope of a universal connection based on the indivisibility of relativistic space-time, thereby challenging the most basic scientific assumptions about the distinction between subject and object based on physical form opposites.

Biology

In biology, scientific cognition developed similarly. In the seventeenth century, species were believed to be created by God and human beings had a core status in nature. However, in 1858, Charles Darwin and Alfred Russel Wallace proposed the theory of natural selection at the Linnean Society in London, explaining the orderly evolution of biological species, including humans. This theory subverted the self/human-centered cognitive paradigm for positioning biological species in nature.

Before the work of the Austrian biologist Gregor Johann Mendel, people's understanding of biological traits was vague and abstract. Biological traits showed the dual characteristics of heredity and variation, similar to the cognition of the motion of stars and objects before Newton's theory of universal gravitation. In the middle of the nineteenth century, findings in molecular biology revealed that interactions between ligands and receptors produce information transfers that form the basis of microscopic activities. In 1865, Mendel revealed the laws of segregation and independent assortment in genetics (dominant and recessive genes) following 8 years of experiments with hybrid peas. Based on the principle of interaction, biology has moved away from the abstract perception of phenomena to a concrete description, to a more unified understanding of biological traits at the microlevel.

Microbiological studies in recent decades have revealed a very complex network of molecular interactions. Although the importance of molecules varies, in essence, there is no simple linear cause and effect in the determination of biological phenotypes, and compensatory effects are common among molecules. At the microlevel, the transition has moved from simple linear causal cognition to non-linear universal molecular interactions.

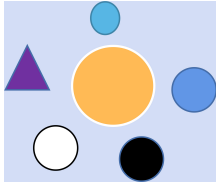
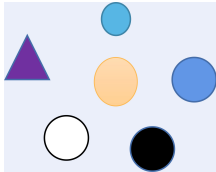
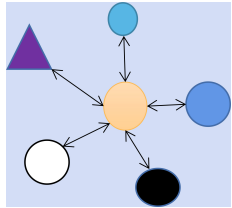
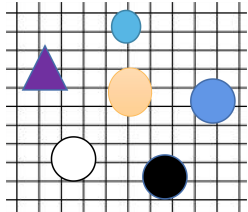
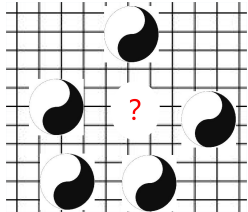
In 1992, the discovery of mirror neurons further demonstrated that we learn about the world not just by independent observation but through perception and imitation (Dapretto et al., 2006; Falck-Ytter et al., 2006;

Rizzolatti and Sinigaglia, 2016). For example, monkeys watching one another eat have neural activity in the same regions of the brain. This reflects the non-absoluteness and indivisibility of the role of the observer and the observed object, which is a relationship of inclusion, analogous to quantum phenomena in physics from an abstract meaning. By describing the developmental process from the core scientific theories mentioned above, we can roughly conclude that scientific cognition has the following characteristics.

First, from self/human-centered cognition to self/humans are not special cognitive conclusions/trends and from the antagonistic relationship based on physical forms to the indivisibility of the subject and object (internal and external, consciousness and matter). We are aware of our existence because of the differences in external physical characteristics with the outside world, and we define ourselves/human beings in terms of these physical or biological characteristics. In the development of science, our cognition of these descriptive quantities is also a synchronous process of redefining ourselves/human beings. This process shows a trend from the perception that the self/human is special to the perception that the self is not special (or from self-centeredness to the cognitive conclusion of the self/human is not special). For example, we originally defined ourselves in terms of unique biological traits, distinguishing ourselves from other living or non-living beings through biological trait differences. Then, the theory of natural selection (the form of contradiction is the relationship between the individual and environment) broke down this notion of the special status of humans in nature. Subsequent scientific exploration revealed that these traits are non-special and can be explained uniformly at the microlevel by genes and proteins (proteins are extrinsic exhibitors and genes are arbiters).

In other words, in the development of cognition, the contradictory form is always from the most intuitive and obvious form (in terms of perception) to the most subtle and hidden form of contradiction (this relates to the relationship between two basic descriptors of a theoretical paradigm. For example, natural selection explains evolution by describing the relationship between individuals and the environment, and the theory of relativity describes the relationship between mass and energy). This blurs the boundary between the subject and objects and is beyond the distinction of external physical forms (Table 1). Some experiments reveal inclusive relations between the two sides of the contradiction, and the contradictory form is ultimately the manifestation of the most basic relationship between the subject (consciousness) and the object (matter). This is from the antagonistic relationship established in the earliest view of the matter to the phenomenon that they contain one another, but a theoretical framework that can be proved scientifically describing the unified relationship between matter and consciousness (or object and subject) is still lacking (Figure 4). For example, our most conspicuous perception is

TABLE 1 The paradigms in different stages of scientific cognition show a regular tendency in terms of perception.

	Core theories listed represent phenomena or reality revealed by different stages of paradigm		The shared meaning abstracted from the revealed phenomenon or reality (in terms of perception rather than specific mathematical or logical descriptors)	Diagrams used to represent the state of perception abstracted from the phenomena and its evolution tendency
	Biology	Physics		
Paradigm 1	Humans are unique in nature	Geocentrism; The earth is the center of the universe	Depend on the distinction of external physical forms, the relationship between subject and object is antagonistic. It was believed that the self/human has a special status or location in the universe. The spheres of different shapes and colors represent the difference in physical form (mass, volume, speed, anything that can be described) between the self/human and external objects. The big yellow circle represents subject, and the surrounding spheres are around it, indicating that self/human has a strong feeling that self/human is very special in universe.	
Paradigm 2	Theory of natural selection (individuals need to adapt to environment)	Heliocentric theory	Human beings have no special or core status in the universe and are common species in nature. The heliocentric theory suggests that the earth revolves around the sun. The circle in the center of the diagram became significantly smaller and lighter, indicating that the feeling that human is very special is weakened compared with the former paradigm.	
Paradigm 3	Ligand and receptor interaction. Laws of inheritance based on dominant and recessive genes	The theory of gravity	A linear cognitive model is established based on the principle of interaction and determinism is gradually formed. Physics takes force as the basic concept; while biology forms information transmission based on the interaction between ligand and receptor. Make a mathematically descriptive connections between descriptive quantities/characteristics, but the relationship between time (energy) and space (mass) is antagonistic.	
Paradigm 4	A universal network of interactions between molecules	The theory of relativity	This paradigm further expands the scope of universal connection, time and space are inseparable and not antagonistic relation. On the physical level, there was a universal connection between objects, and the abstract concept of “field or network” was more suitable to describe the real connection pattern between objects. Meanwhile, determinism reached its peak, but subject and object are opposites. The use of grids to describe the interaction between objects in the diagram is used to represent ubiquity and abstraction.	
Paradigm 5	Mirror neurons were found	Wave-particle duality and quantum entanglement	At the micro level, the cognition of the objective world is challenged; the scope of connection is further expanded, not only is there a universal connection between the physical level, the contradictory nature (wave or particle characteristics) of objects is inseparable from the subject. We are more confused about the nature of consciousness than ever before. For example, in mirror neuron experiments, relying solely on brainwaves cannot distinguish between object and subject, showing inclusion relation of object and subject beyond external physical form. In the diagram, the observer is inseparable from the state of objects being observed. The different features of exterior objects are represented by contradictory nature (black or white). A grid describes an indivisible abstraction connection between subject and object, external and internal. But a scientific theoretical framework to describe the relationship between inner and outer is still lacking.	

We think it contains the following three features: (1) In terms of perception, the contradictory form of different paradigm (for example, the theory of relativity describes the relationship between mass and energy) goes from obvious to basic, from kinds of distinction on physical forms to subject and object (external and internal) inseparable. (2) The development of scientific cognition is also a synchronous process of breaking the feeling state that the self/human is very special in terms of external physical form distinctions. (3) The trend of reality paradigm shift dependent on the state of perception, in other words, state of consciousness acts as an abstract frame of reference to determine the occurrence of reality, reality is not a purely objective process observed by the observer. There are no objective laws independent of perception.

that the self is a unique and conscious species of animal in the universe. The revealed reality or phenomenon of scientific activities always tends to take place in the opposite direction, which is a natural tendency for unifying the differences and unique characteristics in an external physical form that one feels naturally at the beginning. This trend and regular development process cannot be separated from the state of perception as reference frames, and they move in a formalized non-linear way. From this point of view, scientific cognition itself (the reality or phenomenon revealed) based on perception can be deduced in a paradigmatic way.

Second, referring to external physical form, different disciplines and theories are independent. However, there are obvious similarities in the abstract meanings of phenomena revealed by paradigms in different disciplines. For example, the concept of mirror neurons in biology shares similar abstract implications to the phenomenon of quantum wave-function collapse in physics. They both revealed the inclusive relations and inseparable relationship between two aspects in a contradiction (subject and object).

For a long time, science has been defined as an objective description of the laws of nature. We are only objective observers exploring objective laws. Biology and physics, based on reductionism, seek to explain the universe through the motions of microscopic cells and quanta, but why do scientific theories (the reality or phenomenon revealed) show regular paradigm shifts and significant similarities in perception between different disciplines (Figure 4)? We argue that this implies that scientific cognition itself may not just be an objective and independent process of observation but a synchronous process that constantly breaks down self-particularity (based on external physical forms) perception states contained in a more unified framework of laws, reflecting the natural trend for *everything* to move from opposition to unity. Hegel did not separate nature from history. For the first time, he described the natural, historical, and spiritual world as a unified process and tried to reveal the regularity and objectivity of its movement and believed it was the contradiction that led to the change and development of movement (Hegel, 1994, 2004). This view is also supported by our discussion of trends in scientific cognition in this section, as follows: the laws of nature and the laws of society are an inseparable process, which can be seen as the paradigm shift process, the reality or phenomena revealed by scientific theories is inseparable from perception state, and the occurrence of reality is not just governed by an external objective law.

It is worth noting that the dominant view of our current “human-special” concept is that only humans or higher animals have consciousness. Thus, the distinction between the inanimate and the living, matter and consciousness, and external and internal may also be a problem we need to address.

The next possible paradigm: the unity relations of matter and consciousness at the perception level

To reconcile the gap between body and mind, Solms (1997) argues that the solution to this problem must reduce its psychological and physiological to a single physical abstraction (Solms, 1997, 2014). From the perspective of holistic philosophy, the development trend of scientific cognition also supports this view. Depending on perception, scientific cognitive theory itself presents a regular paradigm shift trend. Based on the discussion in the previous chapter, we try to deduce the next possible paradigm or framework and propose scientific hypotheses that can be verified through experiments.

The external (object) and the internal (subject) are indivisible (showing that both sides of the contradiction are inclusive). For example, the wave and particle properties exhibited by a quantum are inseparable from the observer, and the observing subject and object in the mirror neuron cannot be distinguished by external forms. The paradigmatic shift trend of reality is from self/human-centered cognition to the conclusion that the self is not special, from the external physical characteristics of antagonistic relationships to unity (inseparable from perception), and it shows a tendency to distinguish objects beyond external physical forms. A paradigm that can be mathematically described must be represented in contradictory forms (e.g., mass and energy, dominant and recessive genes), and the contradictory forms that construct the new theory increasingly tend to be the most basic and the most subtle forms distinguished by perception.

Based on the above three features, we propose the next possible cognitive paradigm: a more abstract presentation form of contradiction that transcends physical form, which is distinguished by perception to describe the shift of reality in a paradigmatic manner. This possible next paradigm is the basis of holistic philosophy, marked by the Taiji Diagram (depending on conscious experience, all existence is summarized as contradictory perception states of Yin-Yang, replacing concrete physical features). The two sides of contradiction are interdependent, interlaced, and inter-transformed. This interpretation is based on our understanding of the *Tao Te Ching*, and its rationality needs to be supported scientifically (Table 2).

We propose that matter and consciousness are not two opposite existences but two completely different contradictory states of perception. They can be represented by the abstract Yin-Yang (Yin: has characteristics in perception like abstract, internal, hidden, defensive, and passive; Yang: has characteristics in perception like concrete, external, prominent, aggressive, and active). They can be distinguished and are unified at the perception level. Since they are two sides of the same coin, this can be demonstrated by the influence of “internal” state of consciousness on the “external” occurrence of reality.

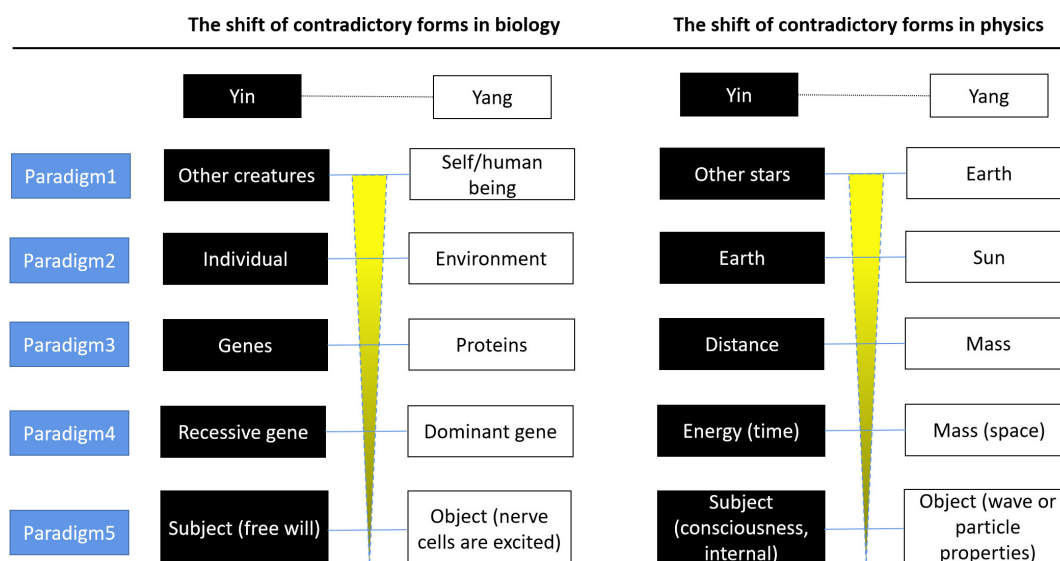


FIGURE 4

Contradiction forms of paradigm in different stages are from obvious to hidden, from intuitive to basic in terms of perception. Objects in completely different physical forms (For example, genes and the earth) can be reduced to two contradictory (Yin-Yang) perceptual states. Yin: shows characteristics in perception like abstract, internal, hidden, defensive, passive etc.; and Yang: On the other hand, shows characteristics in perception like concrete, external, prominent, aggressive, positive etc. On the contrary, in order of precedence, Yin is in front and Yang is behind (Yin-Yang instead of Yang-Yin). For example, the exterior of a tree comes from hidden roots. Genes are the determinants of heredity, not proteins. Therefore, the causal view of holistic philosophy is internal, and the material view seeks external cause and effect. **Paradigm 1:** In the most intuitive state of perception, according to external physical form, human beings are the most advanced creatures in nature and are very special beings. Thus, the self/human has core status (exhibiting abstract features of "Yang") relative to the other creatures, while other creatures are relatively insignificant (exhibiting abstract features of "Yin"); the other stars revolve around the earth and are therefore passive and active relationship (Yin and Yang, respectively). **Paradigm 2:** The theory of natural selection suggests that the individual needs to adapt to the environment, so the individual is in a passive status (Yin) and the environment is in an active status (Yang); the heliocentric theory proposes that the earth revolves around the sun, so the sun is active (Yang) and the earth is passive (Yin). **Paradigm 3:** Genes are the internal determinants of biological traits (Yin) with hidden characteristics, while external traits are mainly presented by proteins (Yang); two descriptors in the equation of universal gravitation, $F = Gm_1m_2/r^2$. Mass (m) is perceived first and distance (r) second, so distance is Yin and mass is Yang (Just like a cup, the surrounding outer wall is focused by perception firstly, and the empty part inside is focused secondly, but the two together can hold water as a cup). **Paradigm 4:** Mendel proposed the most basic genetic law was based on dominant gene (A) and recessive gene (a). The recessive gene (Yin) was in a hidden status compared with the dominant gene (Yang); in the theory of relativity ($E = mc^2$), the relationship between two fundamental descriptors, mass exhibits concrete characteristics (Yang), while energy is an abstract state (Yin). **Paradigm 5:** Matter exhibits remarkable and concrete characteristics (Yang), consciousness is abstract and hidden (Yin), you can't see it but you can feel it. Both physics and biology reveal an inseparable relationship between matter and consciousness or subject and object. It should be emphasized that this empirical division of Yin and Yang according to experience and perception is not absolute. For example, in rare cases, proteins can also serve as genetic material, reflecting a mutually inclusive relationship between contradictions.

Modern biology and physics are based on material views and reductionism, whereby scientists seek to explain the universe through an understanding of the laws of microscopic cells and material "entities" such as quanta. However, the most fundamental assumptions of this cognitive paradigm have been challenged by some of the phenomena and most fundamental problems revealed by recent science (Figure 5A). For example, biology is confused about issues like free will and causality based on space-time, the phenomenon of hyperspace-time quantum entanglement discovered by physicists, and how consciousness causes wave function collapse (affecting reality). These phenomena reveal the inclusive relations between two sides of contradictions (subject and object, consciousness and matter, internal and external), but the descriptive theoretical framework for reconciling (unifying) the opposite relationship of matter and consciousness has not been established. We

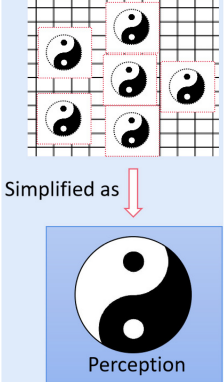
propose that the cognitive paradigm of a holistic view based on perception provides the possibility to reconcile these paradoxes. From this point of view, the confusion of biology and physics about consciousness can be unified into one same problem: how reality is created depends on the state of perception (not subjective intent and purpose; Figure 5B). We will elaborate on this hypothesis in the following sections.

The paradigm-shifting trends of reality depend on irrational feelings

The occurrence of reality is inseparable from irrational feelings

Panksepp's work led to the recognition of the importance of emotion in the study of consciousness and he coined the

TABLE 2 Describing the evolution of reality in terms of perceived abstract contradictions (Yin-Yang) may be the next paradigm reconciling the antagonistic relationship between matter and consciousness.

	Core theories listed represent phenomena or reality revealed Y different stages of paradigm		The shared meaning abstracted from the revealed phenomenon or reality (in terms of perception rather than specific mathematical or logical descriptors)	Diagrams used to represent the state of perception abstracted from the phenomena and its shift tendency
	Biology	Physics		
The supposed paradigm 6	What is the function of subjective experience? The free will is at odds with causality based on material view.	How consciousness causes wave function collapse (affecting the occurrence of reality/order)	Matter and consciousness are not opposites but are unified in perception, they are two different states of perception (it is represented by the abstract Yin-Yang). Both the subject and the object are represented by the abstract Taiji diagram, and there is no external physical form distinction between subject and objects (self/human is not special in terms of external physical form). The dotted box in the diagram indicates that the essential state of things is relative and inseparable from perception. Describe fate or reality in terms of perception (experience) is the subject of this paradigm. Fate or reality can be described in a paradigm shift way and inseparable from perception (irrational side)? This paradigm also happens to be the cognitive basis of holistic philosophy.	

phrase Affective Neuroscience in 1991 (Panksepp, 1992; Davis and Montag, 2018). To distinguish it from rationality more obviously, we used the word irrationality rather than sensibility (in this manuscript these two words have the same meaning) to refer to the sense of reasonableness as a result of experience, the experience involving emotion with no thinking. According to Solms, if we want to coordinate psychological and physiological aspects to explain consciousness, we must focus on the feeling and experience.

If the internal experience of having a memory and the neuronal assemblage embodying that same memory (pictured externally, through optogenetics, for example) are two realizations of a single underlying thing, then what is “memory” itself made of? The answer is that it is abstracted from both manifestations. Memory is not a stuff; it is a function. If we want to identify a mechanism that explains the phenomena of consciousness (in both its psychological and physiological aspects) we must focus on the function of feeling, the technical term for which is “affect.” That is why it is easy to agree that consciousness is not just another cognitive function (Solms, 2018).

In the first paragraph of this section, we mentioned Thomas Kuhn’s suggestion that scientific cognitive processes are paradigm-shifting processes that cannot be logically inferred. We believe that the shifting of scientific paradigms depends on the irrational side of feelings. It is an irrational tendency of movement, irreversible and relative, and is also an objective movement form that exists in contrast to logical/rational characteristics. For example, in the ancient days of human

civilization, people advocated for “an eye for an eye, a tooth for a tooth,” which they deemed reasonable. Now, if someone hurts another person with an axe, instead of punishing them in the same manner, we imprison them. This reality evolves depending on the abstract and irrational “sense of reasonableness.” From simple linear causal to non-linear compensation, the development trend of reality is formed. This trend cannot be independent of perception and experience, as there is no logical or physical connection between imprisonment and the axe. It cannot be strictly quantified, but it shows a clear trend at the perceptual level. If we turn this phenomenon upside down, we will find that it is extremely disharmonious at the feeling level and will cause chaos and collapse in reality, which does not conform to the developmental trend of things, and thus the resistance encountered will grow.

Cognitive trends abstracted from phenomena revealed by different paradigms share similar characteristics. New paradigms can be more consistent with the coordination of irrational feelings than previous ones. These trends make sense at the level of perception (a relative concept that can only be defined concerning previous paradigms) but not the other way around. At the level of physical forms, some are even as far apart as an axe and imprisonment. For example, waves and particles in physics are completely unrelated in a way similar to stem cells and differentiated cells in biology, and yet, they are unified/similar in irrational feelings and represent a potential state (which can become any specific state) and a specific state of beings (abstract and concrete; Yin and Yang, respectively, not the other way around). The irrational feelings

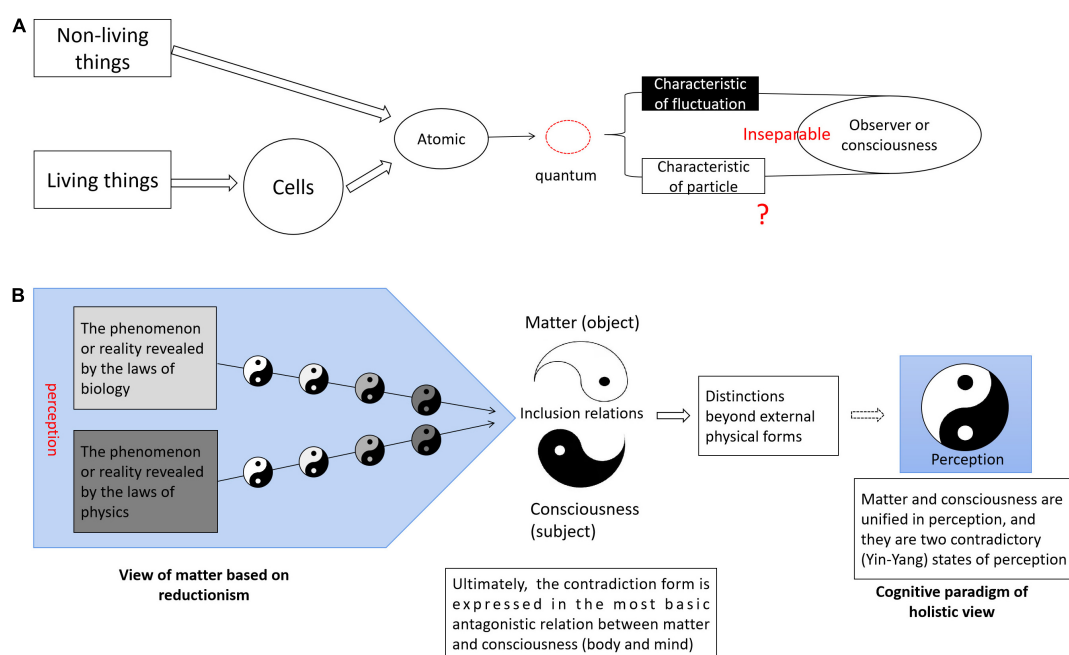


FIGURE 5

Dilemmas of consciousness in biology and physics may be reconciled by the holistic cognitive paradigm. **(A)** Reductionism seeks to explain everything (the theory of everything has nothing to do with perception) by understanding the motion of microscopic atoms (quanta). Phenomena revealed by quantum mechanics, such as wave-particle duality, challenge this basic cognitive paradigm. The nature of a quantum (wave or particle) is inseparable from the observer (or consciousness). However, further mathematical descriptive theoretical framework based on this paradigm has not been established. **(B)** The reality revealed by biology and physics presents a regular paradigm shift trend and has similarities in terms of perception. The contradiction form (paradigm) in terms of perception gradually changes from obvious to the most basic, from concrete to abstract. Both disciplines ultimately reveal, respectively, the limitations of cognitive paradigms based on material view, reflecting the inclusive relationship between matter and consciousness (It was first believed that the objective material world had nothing to do with perception, and the two were opposites, using different ways of describing human society and nature). Following the previous paradigm, we conjecture that the next paradigm will not be described in concrete physical form, but in the paradigm evolution of reality based on perception and experience. In this paradigm, individuals are defined in terms of different fates and realistic outcomes rather than physical forms.

influence the occurrence of an external reality. According to holistic philosophy, the description of the world depends on perception rather than observation, which means that the self and *everything* are connected and indivisible from perception (the distinction between subject and object is independent of physical forms). Therefore, the irrational feelings we experience are not subjective or individually owned, but one of the objective tools for creating reality.

The reference frames for the occurrence of reality are perceptual states rather than external physical entities

In the process of scientific exploration, we are used to the material view of cognition that sees the self/human as an observer, and we separate social activity from scientific exploration. Whereas human social activities focus on inner feelings and experience, scientific activities are considered to explore objective laws of nature as observers. However, we find that scientific laws that are regarded as objectives show a regular paradigm shift trend that cannot be separated from irrational perception, and finally find that subject and object (or

consciousness and matter) cannot be separated. This shows that there is no external objective law or material world separating “inner” perception, and the absolute assumption of the self as a pure observer is also limited. There is a more holistic framework that can unify the antagonistic relationship between consciousness and matter (subject and object), and we propose that holistic philosophy offers such a possibility.

Empiricists believe that human knowledge of the world comes from human experiences, while rationalists believe that human knowledge comes from human reasons. Kant, on the other hand, reconciled the two views to some extent. Kant believed that knowledge is obtained by human beings through sense and reason at the same time. Experience is necessary for the generation of knowledge, but it is not the only factor. Rationality is needed to convert experience into knowledge. We further extend this definition in this study (Kant, 1949).

Let us illustrate this abstract meaning with an example. If we watch a basketball game, we might think that spectators have no impact on the results of the game. In fact, from a holistic perspective, the preferences of perception (irrational “sense of reasonableness”) can influence the paradigm/style of the game in

discontinuous ways (reality occurrence in paradigm-shift ways). For example, a National Basketball Association (NBA) offense might first shift from advocating a near-the-basket offense to a mid-range jump shot and finally shift to a focus on shooting from outside the three-point line (the evolution of reality was initially the most obvious pattern of rationality, the closer the players were to the basket, the easier it was to shoot, based on observation or rationality). This process is the change in which the spectator can empathize through feeling with the player (individual), feel/seek the path of least resistance to attack, and not just as an objective spectator that has nothing to do with the game, the reality tends to the path formed by the game of two contradictory “forces or tools” (the state of feeling as a spectator or as a player). Without empathy (relying on perception rather than observation), the spectator will not be able to understand or predict the paradigm shift trend of this game (reality), and reality will always be subject to this contradictory state and evolve regularly and periodically. This process is a gradual movement from opposition in rationality to unity in irrationality, from the outside to the inside, from the most intuitive form in perception to a covert one (the dominant offense evolved from near the basket to the three-point line, just as the cognitive paradigm of science is gradually shifting from the outside world to the inner world), which describes a holistic framework for the evolution of the reality. This is the basic descriptive feature of holistic philosophy, which relies on irrational perception and is abstract, rather different from reductionism.

Some successful old-school coaches may not understand the development trend of this phenomenon and still yearn for the game’s original shoot-under-the-basket style. This is because they are used to the perception of preferences as spectators (the spectator always tends to like an intense game) while ignoring the feeling of the optimal offensive choice as a player in the game (lack of empathy based on perception). As a result, they have no way of predicting where the game paradigm is going. Just as the scientific system based on reductionism *ignores the objectivity of perception* and regards the self/human as an independent observer (or treats experiences and feelings as subjective beings, independent of the physical world), it cannot continue to rely on the basic assumptions of observation and rationality (the view of matter as opposed to consciousness) to explain the origin of macroscopic order. This cognitive paradigm develops so habitually that we take it for granted, until it is challenged by paradox phenomena revealed by biology and physics. Therefore, we think that the evolution of reality paradigms is always driven by these two abstract contradictory tools or two contradictory statuses of feeling.

It must be emphasized that this phenomenon does not apply only to the activities of human social activity but to the scope of the evolution of *everything* (the object described by the holistic view is reality itself). These two tools (Yin and Yang, rational and irrational perception) are two objective tools for

creating reality, and not the subjective form that belongs to the individual. In a holistic view, the individual is regarded as non-special, and the subject and objects are connected, indivisible, and unified through perception. Self/human are also a contradictory existence connected to perception, so they cannot be independent of natural laws as observers, and their perception state as the reference system participates in the occurrence of reality.

The objectivity of perception

Contradictory nature that can be experienced by perception is the objective form of being

In this chapter, we will discuss the theoretical basis on which reality can be described objectively by perception. The split-brain experiments shed some light on the contradictory nature of consciousness. In the 1940s, scientists cut the corpus callosum of epilepsy patients who did not respond to medical treatment. However, split-brain individuals whose corpus callosum has been incised have a distinct feeling of division or the act of division. For example, when Sperry injected a command to raise one’s hand or bend one’s knee into the left side of the split-brain, the patient’s right side obeys the command but the left side does not; there are many other similar contradictory behaviors (Gazzaniga, 2005; Volz and Gazzaniga, 2017; de Haan et al., 2020).

According to Hegel, everything contains a contradiction, which is the root of all movements. At the same time, he also regarded the development of contradiction as a process from in-itself to self-action. Opposition, distinction, and unity are different stages of contradiction development. Only after experiencing the contradiction of opposites will the unity of new contradictions be realized; opposite contains unity, and unity also contains the opposite. This dialectical thought is similar to the connotation of the Taiji diagram (Hegel, 1983, 2004).

Intuitively, though, internal perception is seen as subjective, it also has an objective nature of contradiction beyond the control of any individual. It does not belong to anyone/existence, but it is dynamically connected with any existence on the level of perception. Therefore, the contradictory nature of being experienced by perception is actually the source of creation (it divides existence into two contradictory perceptual states, Yin-Yang, which are connected with perception), which exists dynamically in an abstract but in a perceivable way and is not a subordinate feature that we can induce from material world phenomena, and this subjective level of objectivity is one of the bases for the holistic philosophy to describe the development of the evolution of reality.

Perception, which is perceived as subjective and is considered unique to human beings or higher animals, is

remarkably objective. For example, if told not to think about what a tall, red spruce looks like, a person cannot help but imagine it. If we flip a coin 10 times and every time it comes out heads, we will become increasingly curious and try to find the cause and effect, and if the results become increasingly balanced, we will experience the same kind of confusion. Once science finds and defines a rule, for example, that a random mutation of a base is selected by nature to lead to the evolution of a species, it also means that this statement will no longer be objective. We will continue to master targeted base-editing techniques to modify living species and reality will always move in the opposite direction, since objective observation and feeling preference cannot coexist. As science advances, we acquire even more sophisticated and high-resolution photography, but by contrast, forms of artistic expression begin to seek more abstract ways of expressing the uniqueness and meaning of their existence, we worship the spotlight of the stage, but once it reaches its peak, the noise makes feelings pursue being ordinary and vice versa. This perceptual contradiction is the most basic objective feature of the inner. The holism cognitive paradigm reduces everything to an abstract state of Yin and Yang, which is not merely the nature of external self-expression, but inseparable from internal perception or conscious experience. Therefore, Yin-Yang is regarded as the co-creation of subject and object and is inseparable from perception. On the other hand, the material view focuses on describing the objectivity of external physical properties (describing the object from the relationship of opposites), ignoring the basic unity relationship between internal perception and external.

This state of contradiction is a fundamental pattern beyond the control of any individual and, at this level, everything is connected in perception rather than demonstrating an external space split, and this suggests that, contrary to our intuition, we are not masters of perception; a new self-model based on the principle of Taiji expounded on the contradictory nature of the self (Wang et al., 2019; Wang and Wang, 2020).

Of note, when we interpret the relationship between matter and consciousness, we define consciousness as one of the abstract states of perception, as opposed to the states of concrete perception that form. For the most part, we use the word perception alone as the creator of all things, similar to the concept of Tao, but this description is not appropriate. The creator cannot be described by language or any concrete measure. If it is defined as one of these, then it means that it is no longer the other, and it loses its totipotency. We still use this word, in part to highlight the indivisibility of creation and perception, but habit is what keeps us from finding more appropriate words to replace this word. In other words, perception is not owned by the individual. What defines us as individuals are the different realities (active or passive, goals achieved or not achieved, the different outcomes of events experienced by individuals) that we perceive. Since they can only be experienced and not observed, we intuitively think

of them as belonging to the individual/human, not to other animals or non-living things, forming a worldview of matter as opposed to consciousness. The view of matter defines the individual in terms of biological or physical characteristics, while what determines reality is the contradictory nature of inner perception, which is beyond the control of the individual.

We believe that this interdependence of internal and external at the perception level is the most fundamental form of movement. At this level, everything is naturally connected and in perpetual motion, resulting in reality never occurring in a static, linear, or causal manner, detached from the perception of the subject. At any time, in the form of potential indivisible contradictory non-linear fluctuations, the dynamics alternating between the law and the irregular are the basic characteristics of the universe, nature, society, and other evolution of reality.

Explanation of quantum entanglement based on holistic philosophy

Our explanation for why the hyperspace-time entanglement occurs between two particles born in the same system is as follows. (1) We will ignore the objectivity of the inner and simply seek the connection between “two particles,” from the material view. It is not that two-particle entities are mysteriously interacting in hyperspace. As two ends of a contradiction, internal and external are born at the same time and are inseparable. The quantum state (topspin or backspin) observed by the observer is inseparable from perception, so they are not constrained by time and space. From a holistic view, the “particle” of a material entity independent from perception neither exists nor can be described, and beings can only exist in terms of contradictory physical properties (up and down, right and left, black and white, etc.) that cannot be separated from perception. Different observers are essentially different states of a conscious experience.

(2) In the previous section, we argued that the contradictory nature of inner perception is the most objective form of existence and that the material entity, which although gives us a very real objective sense, is not the most essential existence, objectivity is relative. It is only the side that is conspicuously perceptible to perception (Yang: a conspicuous, specific, prominent perceptual state). The occurrence of quantum entanglement depends on the rational “sense of reasonableness” (it is reasonable for external phenomena to remain symmetrical if they are based on observation or rationality).

(3) In other words, if we regard that the symmetric descriptors topspin and backspin are born at the same time as an occurrence of the result of an event (reality), the reference frame for quantum entanglement occurrence is the rational “sense of reasonableness” of perception. What we see is only the conspicuous aspect of conscious perception (Yang). But according to the material view, we only mistakenly regard it

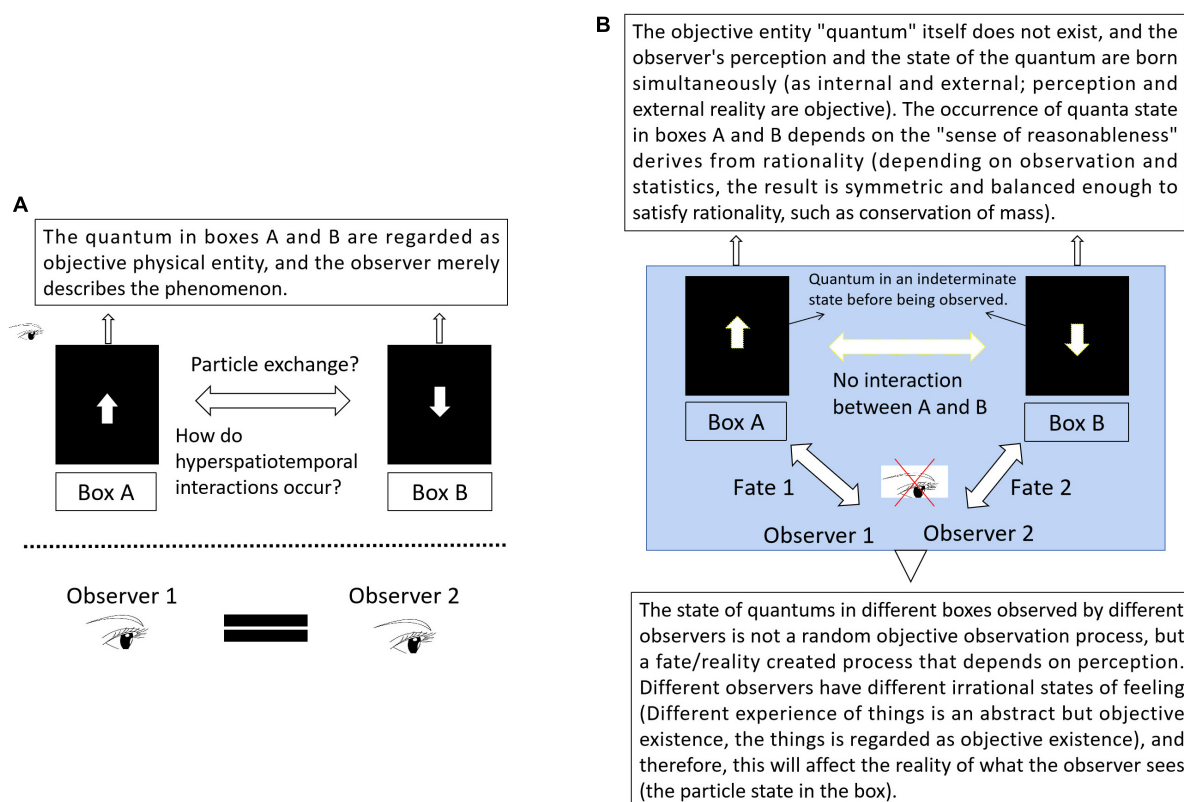


FIGURE 6

Interpretation of quantum entanglement based on holistic view. **(A)** Material view. Based on the difference in external physical form, observer and objects are opposites. External phenomena are irrelevant to the observer (or perception/consciousness). It cannot explain the hypertemporal effects of quantum entanglement. **(B)** Holistic view. Depending on perception, observer and object are unified. Perception is co-created by subject and object. In other words, what we intuitively regard as subjective, individual and abstract being (experience and feeling) can be described scientifically (Yin-Yang) and used as frame of reference to describe the occurrence of fate or reality of the observers, from the material view, the state of the quantum in boxes A and B is an objective random process.

as an observed process rather than a happening process that is inseparable from perception.

(4) The other irrational side (a state of conscious feeling that depends on the experience of the observer) is hidden and dynamic and is difficult to describe using explicit linear equations (Yin: a hidden, abstract state of perception). For example, just as one plus one equals two is generally accepted from rationality and logic, we find that in social activities and specific situations, one plus one can have any possible outcome in different realistic situations. However, this conscious experience is dynamic and needs to be in a relative situation to form a generally recognized (objective) state of reasonable feeling. It is in a hidden and unstable status that cannot be absolutized. However, that does not mean that it is not one objective side creating the order of reality.

Therefore, if the above explanation is correct, what we need to demonstrate is that the quantum state observed in different places is not random (the reality of the observer; which is regarded as random according to material views). It can be affected by the state of perception of the observer and, more

specifically, the state of conscious feeling formed by experience on the irrational side (Figure 6). Shaping the specific state of perception (irrational side) of the observer can affect the reality of the observer (the observed probability of upspin or downspin).

A preliminary discussion of hypothesis proof

From a material view, the connection between objects requires force as a medium, and the force originates in the interchange of microscopic particles. Therefore, we are puzzled by the spatio-temporal nature of quantum entanglement. In a holistic view, the nature of connections between objects depends on irrational perception (for example, the fact that trees and grass, as well as the sun and moon, can produce similar Yang and Yin states of perception, rather than the other way around, is a fundamental property of perception, despite their differences in physical level). Unity in perception

TABLE 3 A contrast between the material view and the holistic view cognitive paradigm.

	Material view	Holistic view	Additional remarks
Description method	Rely on observation, rationality (it can be referred to by an asterisk ■■).	Rely on conscious experience, irrationality. (it can be referred to by an asterisk ■).	The materialistic view is the cognitive paradigm to which we are most accustomed, and the underlying assumption of human beings as observers is not even questionable, while the holistic view, which relies on perception, is easily ignored or regarded as the ownership of individuals.
The relationship between matter and consciousness	Matter and consciousness (or object and subject) are opposites.	As two objective states of being that can be distinguished by perception, unified in perception.	In terms of the objectivity of perception, the abstract perception of Yin-Yang is objective existence, just as the objective features of external objects is observed, but it is an abstract form.
The object being described	It describes external physical characteristics, and has nothing to do with perception.	It contains both subject and object; describes the happen probability of an event (reality, object) of an observer (subject).	In the holistic view, the internal state of conscious perception affects reality, so a precise description must include subject and object.
Different emphasis on objectivity	Focus on the objectivity of external physical form. Mass, size, momentum, etc.,	Focus on the objectivities of perception. (1) Contradictory nature; (2) Irrational conscious perception; (3) Always seek a more harmonious state of being (a relative concept that depends on perception).	Objectivity is the basis on which a cognitive paradigm is constructed. We argue that objectivity has two different way of description, focusing on the external and the internal respectively, satisfying rational and irrational “sense of reasonableness”.
The contact between objects	Particles exchange (physical medium that can be observed) can form different forces, which mediates the connection between objects.	Depending on conscious experience (it's an abstract being that can be experienced but cannot be observed), the things are summed up as Yin-Yang perceptual state, which is not arbitrary and can form a common irrational feeling.	In the interpretation of the origin of all things (the habitual description of the material view) or the evolution of reality (the habitual description of the holistic view); they try to explain through microscopic quantum and abstract Yin- Yang respectively, inseparable from conscious experience.

means that connection does not need any material medium. Taking the double slit interference experiment as an example, the particle's choice to take slit A or slit B as the path is regarded as random under a material view. Researchers shaping the specific state of consciousness of the observer (irrational side, depending on individual experience) will not influence the result of the observer. Otherwise, doing so would affect the result without any material medium participation. Due to space limitations, we do not discuss further details of the experimental design in this study.

Many scientists believe that consciousness is the key to the collapse of the wave function, but we still need a model that can be scientifically validated. According to our previous analyses, existence is manifest in a contradictory way that is inseparable from perception. It is a comprehensive feeling state of the connection between objects generated based on experience and a dynamic superposition of irrationality and rationality (for the sake of the statement and later experimental verification, we replace Yin-Yang with rationality–irrationality), rather than objective material entities.

What we want to emphasize is that in the perception of the relationships between objects, which is also the essential/objective state of connection between objects in a holistic view, the abstract state of perception is a state of superposition of rationality and irrationality (the mutual inclusion relationship in Taiji diagram), and the rational conclusion is only a one-sided illusion, a state in the conspicuous

feeling state (Yang). Just as feelings of self tend to focus on physical features, body shape, and needs, rather than on the other side, empathy (or we can say both selfishness and selflessness) is the natural contradiction of the self, but selflessness is often in the hidden status of paradoxical side (Yin) as it forms a scientific cognitive system based on rationality/physical forms. However, the development trend of reality/paradigms leans toward the direction pointed out by this abstract irrational perception side, as we discussed in the previous chapter, on the development of scientific cognition. Thus, by shaping and influencing the irrational side of the observer's state of perception, the reality of the observer can be described and predicted by the researcher. This process is considered random or indescribable based on a material view.

Conclusion

Physics and biology together reveal some puzzling phenomena related to consciousness. Although they are manifested in different forms, they both imply the interaction and indivisible features between matter and consciousness, which cannot be reconciled by the cognitive paradigm of materialism and reductionism which describes matter and consciousness in opposition. In this study, we propose that a perception-based holistic view cognitive paradigm is a potential solution to reconcile this dilemma.

The material view explains phenomena based on forces, and the connection between objects depends on the exchange of microscopic particles. We can explain some natural phenomena from the level of physics and form causality based on time and space. Therefore, the definition of causality in the material view requires a physical medium and is confined to space and time, which we take for granted. However, this cognitive paradigm fails to explain some consciousness-related phenomena such as, how does the mind lead to physical changes in the body, how does consciousness lead to the collapse of wave functions, and why does quantum entanglement appear to be independent of space and time? In these phenomena and effects, we cannot observe any medium involved, so this violates the causality based on the material view, leading to a paradox or gap in interpretation.

On the contrary, the holistic view relies on the conscious experience to distinguish the thing into a Yin–Yang perceptual state, subject and object can not independently described at the perceptual level. Yin and Yang, in other words, do not simply refer to the characteristics of the object itself. It is inseparable from the subject's descriptive means (different conscious experiences), but intuitively, we think of conscious experience as subjective and belonging to the individual. The cognitive paradigm of holism does not describe matter and consciousness in opposition, or ignore either of them, but regards them as two objective states of being that can be distinguished by perception (e.g., concrete or abstract; rational or irrational), unified in perception. This has two meanings: (1) The inner is as objective as the outer, (2) because they are two sides of the same coin unified in perception, the reality (outer) and the inner can interact, which can be seen as a state switch manifested by the objective nature of consciousness. To be more precise, the occurrence of an event (external) from with the sense of reasonableness that come from irrationality as a reference frame, which does not require any physical medium and is not constrained by space and time.

This leads to a significant difference. The material view opposes the subject (internal) to the object, which is described as an objective physical being that is same to all observers. The description of the holistic view describes the subject (conscious experience) and the object (external) together on the basis of perception. Therefore, it describes the results of events rather than physical features, such as the reality of an observer. The two cannot be independent or described in opposition because the reality varies with the observer's state of consciousness (Table 3).

For example, in the experiment of quantum entanglement, from the material view, the observer only describes the state of the particle objectively, which has nothing to do with the conscious experience of the observer. Therefore, the state of the particle observed by the observer is considered to be random. According to holism, the observer's irrational conscious state will affect the probability of the observed particle state, and the conscious state is closely related to life experience. In other

words, the perceived asymmetry of what is regarded as an objective being is the root cause of order (reality does not happen randomly), and this makes it possible to scientifically prove the cognitive paradigm of holistic view philosophy. Due to space constraints, we do not discuss the details of the experimental design here. We elaborate it in detail in another article published in the preprint (doi: 10.31219/osf.io/c3neq).

Objectivity is the foundation on which a discipline or theory is built. The holistic cognitive paradigm is based on the objectivity of perception. Although it is abstract, it can be described scientifically (do not rely on reason and logic). According to our understanding of the *Tao Te Ching*, in this study, we introduce three basic objective properties of perception. (1) Contradictory nature. Conscious experience, though intuitively viewed as subjective, belongs to the individual. On the contrary, perception presents a contradictory and objective nature beyond individual control. What the individual experiences is only the sensory state (Yin or Yang; positive or negative) formed by the different reality determined by this objective nature. This means that the observer's reality is descriptive, not random and unpredictable, but indescribable from the viewpoint of force and reductionist. Let's use the analogy of the relationship between DNA and living things. On the surface, we believe that organisms possess DNA, but in fact, this relationship is inverted. DNA determines biological traits according to objective genetic laws, and individuals only show different biological traits that have been determined.

(2) Irrational conscious experience. The irrational conscious experience can summarize things into Yin–Yang states of perception. This induction is not arbitrary but has an objective standard, which is not essentially different from the description of external physical objectivity that we rely on reductionism, except that one is abstract and the other is concrete. (3) Always seek a more harmonious state of being (a relative concept that depends on perception). It determines that the development of reality has a relative direction and trend, for example, scientific cognition itself shows a regular inertial development trend.

The holistic cognitive paradigm also provides the possibility to coordinate the contradiction between determinism and non-determinism. Due to the objectivity of perception, the occurrence of reality is descriptive and regular. However, this does not mean that the observer's reality is completely determined (determinism), it is probabilistic and independent of perception. Since it is inseparable from perception, which always pursues a more harmonious state of being (the third objectivity), reality will change due to the switch of the observer's state of consciousness, reflecting the subjective initiative of consciousness.

For example, according to holism, we can describe a person's reality in terms of conscious experience. If we just observe (in

fact, this assumption is not entirely accurate because observation and feeling contain each other and can only be said to have a minimal probability of influence), then this probabilistic description can be verified. However, if we tell the observer what is going to happen to him, the state of consciousness of the person who is told will change because perception is always pursuing a more harmonious state, and then the probability of his reality will naturally change. Due to the contradictory nature of consciousness, it refuses to be completely determined and it also rejects absolute disorder. For example, if we are told not to think about a big mangrove, we cannot help but imagine it. We think that there is no right or wrong distinction between the material view and the holistic view. The construction of these two cognitive paradigms originates from two different ways of looking at things (observation or perception) and forms two different sense states of rationality. The different cognitive paradigms and ways of solving problems (it is also a reality-creation process) have limitations and should be complementary (Capra, 2000).

In conclusion, based on the understanding of *Tao Te Ching*, a representative work of holistic philosophy, we (1) deduced the next possible cognitive paradigm from a holistic view through trends of scientific cognitive development and proposed a preliminary scientific hypothesis; (2) summarized the confusion around consciousness in biology and physics as the same problem (of how to describe the evolution of reality depending on perception) and highlighted that holistic philosophy can solve this problem; and (3) we provided a new interpretation of quantum entanglement according to holistic philosophy, which is falsifiable. As interdisciplinary propositions, different disciplines are trying to describe consciousness from different perspectives (Friston and Stephan, 2007; Arsiwalla and Verschure, 2018; Marchetti, 2018). We believe that combinations of approaches from these different disciplines in the future will help us uncover the puzzles related to consciousness.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

References

- Arsiwalla, X. D., and Verschure, P. (2018). Measuring the Complexity of Consciousness. *Front. Neurosci.* 12:424. doi: 10.3389/fnins.2018.00424
- Bohm, D. J. (2004). *Wholeness and the Implicate Order*. Shanghai: Shanghai Science and Technology Education Press.
- Boly, M., Seth, A. K., Wilke, M., Ingmundson, P., Baars, B., Laureys, S., et al. (2013). Consciousness in humans and non-human animals: Recent advances and future directions. *Front. Psychol.* 4:625. doi: 10.3389/fpsyg.2013.00625
- Brass, M., Furstenberg, A., and Mele, A. R. (2019). Why neuroscience does not disprove free will. *Neurosci. Biobehav. Rev.* 102, 251–263. doi: 10.1016/j.neubiorev.2019.04.024
- Capra, F. (2000). *The Tao of Physics: An Exploration of the Parallels between Modern Physics and Eastern Mysticism*. Boston, MA: Shambhala Publications, Inc.
- Chalmers, D. J. (1998). The problems of consciousness. *Adv. Neurol.* 77, 7–16.
- Capra, F. (2000). *The Tao of Physics: An Exploration of the Parallels between Modern Physics and Eastern Mysticism*. Boston, MA: Shambhala Publications, Inc.

Author contributions

JC constructed the theory and wrote the manuscript. LC revised and edited the manuscript. Both authors approved the submitted version.

Funding

This work was supported by the Shenzhen Outbound Postdoctoral Research Fund (2020) and the Guangdong Regional Joint Fund (Grant no. 2021A1515111224).

Acknowledgments

We are grateful to the editor and the reviewers for their constructive suggestions, which have greatly improved this article. Mengmeng Han provided valuable suggestions during manuscript preparation. Some literature that inspired the author includes the following: Erwin Schrodinger's *What is Life?*; Stephen Hawking's *A Brief History of Time*; Peter Russell's *From Science to God: A Physicist's Journey into the Myths of Consciousness*; Jared Diamond's *The Third Orangutan*; and Joe DiSpine's *A Course of Change*.

Conflict of interest

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

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

- Churchland, P. M., and Churchland, P. S. (1997). Recent work on consciousness: Philosophical, theoretical, and empirical. *Semin Neurol.* 17, 179–186. doi: 10.1055/s-2008-1040928
- Cieri, F., and Esposito, R. (2019). Psychoanalysis and neuroscience: The bridge between mind and brain. *Front. Psychol.* 10:1790. doi: 10.3389/fpsyg.2019.01983
- Crick, F., and Koch, C. (1998). Consciousness and neuroscience. *Cereb. Cortex* 8, 97–107. doi: 10.1093/cercor/8.2.97
- Dapretto, M., Davies, M. S., Pfeifer, J. H., Scott, A. A., Sigman, M., Bookheimer, S. Y., et al. (2006). Understanding emotions in others: Mirror neuron dysfunction in children with autism spectrum disorders. *Nat. Neurosci.* 9, 28–30. doi: 10.1038/nn1611
- Davis, K. L., and Montag, C. (2018). Selected Principles of Pankseppian Affective Neuroscience. *Front. Neurosci.* 12:1025. doi: 10.3389/fnins.2018.01025
- de Haan, E. H. F., Corballis, P. M., Hillyard, S. A., Marzi, C. A., Seth, A., Lamme, V. A. F., et al. (2020). Split-brain: What we know now and why this is important for understanding consciousness. *Neuropsychol. Rev.* 30, 224–233. doi: 10.1007/s11065-020-09439-3
- Dehaene, S., and Changeux, J. P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron* 70, 200–227. doi: 10.1016/j.neuron.2011.03.018
- Falck-Ytter, T., Gredeback, G., and von Hofsten, C. (2006). Infants predict other people's action goals. *Nat. Neurosci.* 9, 878–879. doi: 10.1038/nn1729
- Frank, A. (2015). Uncertain for a century: Quantum mechanics and the dilemma of interpretation. *Ann. N.Y. Acad. Sci.* 1361, 69–73. doi: 10.1111/nyas.12972
- Friston, K. J., and Stephan, K. E. (2007). Free-energy and the brain. *Synthese* 159, 417–458. doi: 10.1007/s11229-007-9237-y
- Gazzaniga, M. S. (2005). Forty-five years of split-brain research and still going strong. *Nat. Rev. Neurosci.* 6, 653–659. doi: 10.1038/nrn1723
- Hameroff, S. (2012). How quantum brain biology can rescue conscious free will. *Front. Integr. Neurosci.* 6:93. doi: 10.3389/fnint.2012.00093
- Hameroff, S., and Penrose, R. (2014). Consciousness in the universe: A review of the 'Orch OR' theory. *Phys. Life Rev.* 11, 39–78. doi: 10.1016/j.plrev.2013.08.002
- Hegel, G. W. F. (1983). *Hegel and the Human Spirit*. Detroit, MI: Wayne State University Press.
- Hegel, G. W. F. (1994). *Lectures on the Philosophy of Spirit 1827–8 (R. R. Williams, Trans.)*. Oxford: Oxford University Press.
- Hegel, G. W. F. (2004). "Hegel's Philosophy of Mind," In *Hegel's Philosophy of Nature*. Eds A. V. Miller and J. N. Findlay (Oxford: Oxford University Press).
- Heisenberg, M. (2009). Is free will an illusion? *Nature* 459, 164–165. doi: 10.1038/459164a
- Hillman, B. J. (2018). Free Will. *J. Am. Coll. Radiol.* 15:1355. doi: 10.1016/j.jacr.2018.07.012
- Kafatos, M. C., and Yang, K. H. (2016). The quantum universe: Philosophical foundations and oriental medicine. *Integr. Med. Res.* 5, 237–243. doi: 10.1016/j.imr.2016.08.003
- Kant, I. (1949). *The Philosophy of Kant*. New York, NY: Modern Library Press.
- Koch, C., Massimini, M., Boly, M., and Tononi, G. (2016). Neural correlates of consciousness: Progress and problems. *Nat. Rev. Neurosci.* 17, 307–321. doi: 10.1038/nrn.2016.22
- Kuhn, T. S. (2004). *The Structure of Scientific Revolutions*. Beijing: Beijing University Press.
- Lanza, B. B. A. R. (2012). *Biocentrism*. Chongqing: Chongqing Press.
- Lavazza, A. (2019). Why Cognitive sciences do not prove that free will is an epiphenomenon. *Front. Psychol.* 10:326. doi: 10.3389/fpsyg.2019.00326
- Levin, C. F. A. M. (2020). How do Living Systems Create Meaning? *Philosophies* 5:36. doi: 10.3390/philosophies5040036
- Liu, J. (2008). *An Introduction to Chinese Philosophy*. Cambridge: Cambridge University.
- Marchetti, G. (2018). Consciousness: A unique way of processing information. *Cogn. Process* 19, 435–464. doi: 10.1007/s10339-018-0855-8
- Neumann, J. V. (2020). *Mathematical foundations of quantum mechanics*. Beijing: Beijing Science Press.
- Owen, A. M. (2019). The search for consciousness. *Neuron* 102, 526–528. doi: 10.1016/j.neuron.2019.03.024
- Panksepp, J. (1992). A critical role for "affective neuroscience" in resolving what is basic about basic emotions. *Psychol. Rev.* 99, 554–560. doi: 10.1037/0033-295X.99.3.554
- Ramstead, M. J. D., Badcock, P. B., and Friston, K. J. (2018). Answering Schrodinger's question: A free-energy formulation. *Phys. Life Rev.* 24, 1–16. doi: 10.1016/j.plrev.2017.09.001
- Rappaport, Z. H. (2011). The neuroscientific foundations of free will. *Adv. Tech. Stand Neurosurg.* 59, 3–23. doi: 10.1007/978-3-7091-0673-0_1
- Reddy, C. P. J. S. K. (2016). Science, subjectivity & reality. *J. Conscious. Explor. Res.* 7, 333–336.
- Rees, G., Kreiman, G., and Koch, C. (2002). Neural correlates of consciousness in humans. *Nat. Rev. Neurosci.* 3, 261–270. doi: 10.1038/nrn783
- Rizzolatti, G., and Sinigaglia, C. (2016). The mirror mechanism: A basic principle of brain function. *Nat. Rev. Neurosci.* 17, 757–765. doi: 10.1038/nrn.2016.135
- Seth, A. K., and Bayne, T. (2022). Theories of consciousness. *Nat. Rev. Neurosci.* 23, 439–452. doi: 10.1038/s41583-022-00587-4
- Solms, M. (1997). What is consciousness? *J. Am. Psychoanal. Assoc.* 45, 681–703. doi: 10.1177/00030651970450031201
- Solms, M. (2014). A neuropsychanalytical approach to the hard problem of consciousness. *J. Integr. Neurosci.* 13, 173–185. doi: 10.1142/S0219635214400032
- Solms, M. (2018). The Hard Problem of Consciousness and the Free Energy Principle. *Front. Psychol.* 9:2714. doi: 10.3389/fpsyg.2018.02714
- Solms, M. (2021). *The Hidden Spring: A Journey to the Source of Consciousness*. New York, NY: W. W. Norton & Company Press.
- Solms, M., and Friston, K. (2018). How and why consciousness arises: Some considerations from physics and physiology. *J. Conscious. Stud.* 25, 202–238.
- Sturm, T. (2012). Consciousness regained? Philosophical arguments for and against reductive physicalism. *Dialogues Clin. Neurosci.* 14, 55–63. doi: 10.31887/DCNS.2012.14.1/tsturm
- Volz, L. J., and Gazzaniga, M. S. (2017). Interaction in isolation: 50 years of insights from split-brain research. *Brain* 140, 2051–2060. doi: 10.1093/brain/awx139
- Wang, F. Y., Wang, Z. D., and Wang, R. J. (2019). The Taiji Model of Self. *Front. Psychol.* 10:1443. doi: 10.3389/fpsyg.2019.01443
- Wang, Z. D., and Wang, F. Y. (2020). The taiji model of self II: Developing self models and self-cultivation theories based on the chinese cultural traditions of taoism and Buddhism. *Front. Psychol.* 11:540074. doi: 10.3389/fpsyg.2020.540074
- Wang, Z. D., Wang, Y. M., Li, K., Shi, J., and Wang, F. Y. (2021). The comparison of the wisdom view in Chinese and Western cultures. *Curr. Psychol.* 6, 1–12. doi: 10.1007/s12144-020-01226-w
- Wigner, E. P., Wightman, A. S., and Mehra, J. (1992). *The Collected Works of Eugene Paul Wigner*. Berlin: Springer-Verlag.
- Yutang, L. (1948). *The Wisdom of Lao-tse*. New York, NY: Random House.



OPEN ACCESS

EDITED BY

Marialisa Martelli,
Sapienza University of Rome, Italy

REVIEWED BY

John Bickle,
Mississippi State University,
United States
Matthieu M. de Wit,
Muhlenberg College, United States

*CORRESPONDENCE

Daniel C. Burnston
dburnsto@tulane.edu

SPECIALTY SECTION

This article was submitted to
Consciousness Research,
a section of the
Frontiers in Psychology

RECEIVED 12 July 2022

ACCEPTED 13 October 2022

PUBLISHED 07 November 2022

CITATION

Burnston DC (2022) Mechanistic
decomposition and reduction in complex,
context-sensitive systems.
Front. Psychol. 13:992347.
doi: 10.3389/fpsyg.2022.992347

COPYRIGHT

© 2022 Burnston. This is an open-access
article distributed under the terms of the
[Creative Commons Attribution License \(CC
BY\)](#). The use, distribution or reproduction in
other forums is permitted, provided the
original author(s) and the copyright
owner(s) are credited and that the original
publication in this journal is cited, in
accordance with accepted academic
practice. No use, distribution or
reproduction is permitted which does not
comply with these terms.

Mechanistic decomposition and reduction in complex, context-sensitive systems

Daniel C. Burnston*

Philosophy Department, Tulane University, Tulane Brain Institute, New Orleans, LA, United States

Standard arguments in philosophy of science infer from the complexity of biological and neural systems to the presence of emergence and failure of mechanistic/reductionist explanation for those systems. I argue against this kind of argument, specifically focusing on the notion of context-sensitivity. Context-sensitivity is standardly taken to be incompatible with reductionistic explanation, because it shows that larger-scale factors influence the functioning of lower-level parts. I argue that this argument can be overcome if there are mechanisms underlying those context-specific reorganizations. I argue that such mechanisms are frequently discovered in neuroscience.

KEYWORDS

mechanism, emergence, context-sensitivity, network organization, complex systems

Introduction

Biological systems are complex. They are multi-scale, heavily interactive, and context dependent. In this paper, I will assess the ramifications of these facts for reductive and mechanistic explanation. One common reaction to the recognition of complexity is to deny that mechanistic and reductive explanations are possible, or, more weakly, to suggest that their scope is extremely limited. Instead, it is often argued, we should embrace an emergence thesis, and concomitantly a commitment to using distinct forms of explanation for emergent properties in complex systems.

I will question this line of thinking. In particular, I will question the idea that widespread context sensitivity across scales is tantamount to emergence. I will focus on the brain. Neural systems have recently been recognized to involve complex interactions between their parts, multi-functionality of individual parts, and context sensitive forms of organization (Anderson, 2014; Burnston, 2016a,b, 2021; de Wit and Matheson, 2022). As such, the brain, and the cognitive phenomena to which it gives rise, provide a good test case for assessing emergentist claims.

I will endorse, with others in the literature (Silberstein, 2021), the idea that functional decomposition and localization are the *sine qua non* of mechanistic explanation. The question is then best phrased as: do widespread context sensitivity and multi-scale relations in neural systems require us to embrace emergence and abandon localization and decomposition as explanatory strategies? I will argue that they do not, so long as

mechanisms by which context is recognized and used to implement functional reorganization are discoverable. If so, then the system is mechanistically explicable despite context sensitivity.

I begin (section 2) by laying out some of the intersecting dialectical dimensions that comprise the current debate. I endorse a pragmatic construal of the debate along the lines above, and offer a version of reductionism that is based on what I call the pragmatic downward pull of research – the idea that it is normatively better to seek and discover mechanisms at lower levels that comprise one's phenomenon of interest. In section 3 I argue, using some toy examples, that there is nothing *inherently* emergentist about context-sensitivity and multi-scale structure. I then go on (section 4) to illustrate a variety of mechanisms for context-recognition and functional reorganization in the brain, and I suggest that seeking these mechanisms is required for understanding how the brain produces cognitive phenomena. I then (section 5) give my general interpretation of the cases and consider some possible objections. Section 6 concludes.

Dimensions of the reduction and emergence debate

Something of an anti-reductionist consensus has arisen in philosophy of biology (Huttemann and Love, 2011; Kaiser, 2015; Brigandt et al., 2018). There are many reasons for this. For one, many have recognized that traditional reductionist approaches fare badly in accounting for the multi-scale organization involved in biological systems (Wimsatt, 2006). Another is the increased importance, within the last few decades, of dynamical systems and network-based approaches in understanding, e.g., genetic and neural systems (Green et al., 2018; Huneman, 2018). These approaches paint biological systems as inherently interactive and multi-scale, and as falling into classes of topological organization.

The best response a reductionist can make in these circumstances, in my view, is to admit that traditional reductive forms of explanation are indeed hopeless in light of these developments, but to suggest that traditional forms are not the only possible ones. For instance, traditional reductionist approaches have tended strongly towards “atomism” (Burnston, 2021), a view on which explanation proceeds by first discovering the intrinsic functional properties of the relevant lower-level parts, and then (and only then) explaining the properties of the system as interactions between those intrinsic properties. This style of explanation has indeed characterized some eras of investigation in biology and neuroscience, but it is not obvious (and, I will argue, not true) that this is the only way a reductionist thesis might be phrased. Why not come up with more complicated reductive schemas in an attempt to account for complexity in these systems?

I will assume that reductionist and mechanistic approaches are closely allied (although there are non-reductionist accounts of mechanisms; see Couch, this issue), in the sense of “explanatory

reductions” (Sarkar, 1992).¹ This is because mechanistic approaches are committed to decomposition and localization of system properties at lower levels. The question, on this view, is whether a sophisticated enough, but still genuinely mechanistic, account can be given that integrates with dynamical and network descriptions in a productive way. In the remainder of this section, I will lay out some of the extant dialectic surrounding the issue, and then give my preferred reading of reductionism in light of that extant discussion.

Some extant dialectical dimensions

While I make no claim to exhaustiveness, the following are some of the important dimensions surrounding debates about emergence. Note that these are related in numerous ways, and intuitions along one may correlate with intuitions along others. I do not plan to explore the details of this space in full, but instead to lay out some relevant issues so as to better express my version of the reductionism thesis in the next subsection.

Strong vs. weak emergence

Strong emergence is a view of emergence on which there is a discontinuity in nature between lower-level and higher-level phenomena. On traditional views, this has been expressed as the idea that new laws of nature apply to higher-level phenomena, that are not determined by basic physical laws. This has come to be viewed, with some exceptions (Boogerd et al., 2005), as too strong of a position. Views that posit weaker kinds of emergence, on the other hand, posit that there is no discontinuity in nature, but instead that certain organizational features at higher levels are emergent, even if they are ultimately the outcome of basic physical processes. These views have to be careful not to devolve into being too weak – i.e., they should not take basic aggregative and relational properties to be emergent. Take the property of being *five stones in a box*. This property is, trivially, not a property of any of the individual stones or of the box. But no reference to anything beyond the basic physical objects and their arrangement is required to account for the existence and causal powers of this property. Hence, views of emergence must situate themselves with regards to what kind of distinctions they posit between levels, and when those differences are robust enough to justify positing of emergence.

Ontological vs. epistemic emergence

If emergence occurs, is it a feature of the world or a feature of human descriptions of the world? On the former view, certain natural systems are organized such that novel higher-level

¹ While the distinction between explanatory and “theoretical” reduction has become entrenched, I do not actually think it is all that deep. Mechanistic models, on my view, are parts of theories about how the system is organized. I will not pursue this further here, though.

properties are generated in those systems, and hence emergence is a feature of the world. On the latter view, emergence is an epistemic phenomenon – that is, perceived differences between levels are the result of limitations of human classification, imagination, computational resources, etc., rather than any independent feature of the world. Emergence, on this kind of position, is the outcome of epistemic limitation and/or convenience. We posit emergence when we find it convenient or necessary to move away from descriptions of a given type at a given level, to descriptions of a different type at a higher level. A view on emergence must make clear whether it is positing an ontological or epistemic version.

Definitional vs. pragmatic argument

This distinction is not so much a distinction between *types* of emergence, but instead a distinction between kinds of arguments given for emergence. On the definitional approach, one posits that emergence is co-extensional with complexity of certain kinds. One then attempts to define emergence in terms of the relevant kinds of complexity in natural systems. For instance, consider Deacon's (2006) claim that one should define “a technical sense of emergence that explicitly describes a specific class of causal topologies.” On this kind of view, emergence is taken to just be the way that complexity is to be understood, and hence complexity is, definitionally, evidence for emergence. The pragmatic approach is much different.

Pragmatic arguments involve abductions over scientific practice and explanation. The necessity of different descriptions at different levels in science, pragmatic arguers suggest, is evidence that emergence is present – otherwise we would be able to close the explanatory gaps between different types of explanation at distinct levels. Note that a pragmatist need not be an epistemicist (although they may be). It is perfectly compatible with pragmatism to suggest that the best explanation for the presence and necessity of distinct modeling practices in the sciences is that emergence occurs in the world.

Emergence vs. mechanism

Is emergence incompatible with mechanistic and reductive explanation? Most views suggest that there is at least a strong tension between these positions. But this is not obviously the case. Bechtel (2016) has asserted that “reductionists must be holists too!,” arguing that any worthwhile explanation of a system at a lower level must make reference to systemic properties and organization – otherwise, one would never know *which* kinds of organization must be implemented at the lower level. Moreover, it has always been a part of Bechtel's program that mechanistic explanations must go hand-in-hand with dynamical explanations in order to account for phenomena (Bechtel and Abrahamsen, 2010). And he has recently applied this further to network explanation (Bechtel, 2019). Similarly, mechanists such as Kaplan and Craver (2011) have argued that dynamical models, to attain explanatory status, must be “mapped” to mechanistic descriptions.

Other recent mechanist proposals have embraced the ideas that some take to hallmarks of emergentist positions. For instance, in previous work I have argued extensively that functional decomposition and localization should themselves be contextualized to behavioral and physiological circumstances (Burnston, 2016a, 2021). On this position, there is no tension between context-sensitivity and mechanistic/reductive explanation (cf. Deleahanty, 2005; for further discussion, see Gillett, 2016). Levy and Bechtel (2016) have suggested that mechanism existence and identity can shift over time – mechanisms may pop into and out of existence, change their organizational properties, etc. The main danger with this dimension is that the dispute risks dissolving into a semantic one, with mechanists and emergentists both recognizing all of the same facts and simply employing different verbiage to describe them (Silberstein, 2022).

My construal of the debate

My construal of the debate begins by focusing on the *definitional* vs. *pragmatic* dimension. In my view, the only productive version of the debate is one that takes pragmatics as its starting point. If the question of emergence is definitional, then there simply *is no debate to be had*. If emergence is co-extensional with complexity, then the presence of complexity entails the presence of emergence. We must either (i) accept that mechanism/reductionism is false full stop, or (ii) redefine mechanism and reductionism to be compatible with emergence. There is no possibility of reconstruing mechanistic/reductionist positions along the lines just discussed, so as to both be compatible with complexity and to be an alternative to emergentist views. Basically, anyone who recognizes complexity in biological systems is an emergentist of some type, anyone who does not recognize it is naïve, and we can all go to the pub.

As fun as the pub sounds, this is not a very productive way to have a philosophical dispute. Hence, the pragmatic phrasing of the debate is the way to go. On this construal, we have all sorts of interesting things to consider, including scientific practice and explanatory frameworks, and these can serve as genuine evidence for theses about emergence and reductionism. Pragmatism also leaves open a lot of room for how one construes the other dimensions. As noted, pragmatic arguments are abductions from scientific practice and explanation, and emergence is affirmed (or denied) as the best explanation for the nature of those practices. This is compatible with having stronger or weaker views of the kind of emergence one must posit to explain those practices, and with whether one thinks that explanation posits ontic or purely epistemic emergence. Importantly, it also gives a way of overcoming the worry that differences between mechanistic and emergentist views are purely semantic. Since the pragmatic approach is based on abduction from certain forms of scientific practice, mechanistic and emergentist views should give genuinely

distinct descriptive and normative readings of scientific investigation.

Given my construal of the debate, I will take as my stalking horse throughout this paper a recent view of emergence developed by Silberstein et al. (Bishop et al., forthcoming). On this view, called “contextual emergence,” widespread context-sensitivity of systems, and the explanations that scientists resort to in order to explain properties of these systems, provide support for an emergentist thesis. Silberstein et al., quite rightly note that context-sensitivity is widespread in biological and physical systems. They describe emergence as necessary to account for the *multi-scale constraints* and the *topological structure* of these systems.

Multi-scale constraints are instances in which organizational properties at higher-levels influence or determine the properties of lower-level entities. Topological structure means that whole systems implement global structures that are characterizable independently of the lower-level components that comprise them – usually, these kinds of explanations make use of the resources of graph theory, and the types of topologies it describes. These can include organizations such as being a *small-world* network, or exhibiting a *rich club* organization (discussed further below), each of which are present in many different kinds of systems with vastly different component parts. In a context-sensitive system, Silberstein et al., argue, topological properties and multi-scale constraints determine how a system can behave in new contexts. As such, “Contextual constraints represent both the screening off and opening up of new areas of modal space, i.e., degrees of freedom, and thereby new patterns” (Silberstein, 2022).

I target this view because it is the first view of emergence that I am aware of that makes context-sensitivity one of its main tenets and sources of argument (although see Huttemann and Love, 2011). Since I agree about the context-sensitivity of biological organization, this is a productive starting point. Moreover, the authors are admirably clear about their position on the dialectical dimensions just discussed. First, like me, they propose to make pragmatics the main argumentative strategy. Their primary argument is that the nature of science shows the context-sensitive, multi-scale, and topological nature of the systems under study. Hence, I agree with them on the way the arguments should proceed.

Silberstein et al., characterize contextual emergence as *moderately strong*, both *ontological and epistemic*, and as *in conflict with mechanistic/reductive analysis*. They are moderately strong in that they think genuine new forms of organization emerge at the global/topological level, and interact with lower-level processes, particularly by constraining them. This is not normal strong emergence in that it posits no breaks in nature, no fundamentally new laws, etc., and it is not inexplicable – there is simply a new type of fact when systems are arranged so as to implement context-sensitivity, multi-level constraints, and topological organization.

But the view is also not among the weakest in that it does not simply posit that any relational or aggregative processes are

emergent. The constraints exerted on gas particles by the wall of a container, for instance, are not emergent on their view. In contrast, Silberstein et al., offer the example of Rayleigh–Bénard convection, in which fluid particles subject to a temperature gradient within a container form subsisting units that move in regular patterns. On this view, it is the context of the container and the temperature gradient which produces a higher-level organization, which then constrains lower-level behavior, canceling out perturbations in individual particles to retain the higher-level structure.

Similarly, while the view is pragmatic, it is not purely epistemic. Silberstein et al. think it is a fact *about nature* that systems are organized in the way they propose, and that this is the best explanation for the multi-scale and topological explanations scientists give. As such, they are against any purely epistemic view that posits emergent properties as the result of explanatory convenience. Topological properties are not, for instance, merely abstractions over lower-level organizations, but are themselves a distinct type of property that systems can instantiate. Lastly, they take contextual emergence to be in conflict with mechanistic explanation, specifically because they think decomposition and localization fail for such systems. They thus suggest a typology of explanations. Multi-scale topological explanations, on this view, are *distinct from* and *explanatorily independent* of mechanistic ones. In particular, if one adopts a topological style of explanation, one eschews decomposition and localization, and vice versa.

Neural systems are among the explanatory targets of contextual emergence. Following on the earlier work of Chemero and Silberstein (2008) and Silberstein and Chemero (2013), Silberstein et al., posit that neural systems meet the classification of contextual emergence, and therefore that mechanistic analysis is either incorrect when applied to these systems or not fruitful. In support of the contextual emergence thesis with regards to neuroscience, Bishop et al. (forthcoming) list a wide range of facts about the multi-scale nature of the brain, including neural modulation at the cell level, neural synchrony at the circuit level, and the dependence of development on social context as evidence in favor of contextual emergence. Silberstein (2021) further discusses the widespread plasticity of neural systems (cf. Zerilli, 2020). In other work, Silberstein and Chemero (2013) and Silberstein (2021) suggest that cognitive phenomena, including those interrupted in psychiatric conditions, are dependent on network organization, and therefore not explicable in terms of localization and decomposition.

Silberstein (2021) has claimed that the attempts of mechanists embrace complexity rob mechanism of any distinctive content. That is, one can only make mechanism compatible with complexity by so weakening decomposition and localization (as well as the conditions on mechanism identity) that they are simply redescribing contextual emergence in mechanist language, hence rendering the debate verbal. So, in order to adjudicate the debate, we need a characterization of mechanistic/reductive explanation that would resist having purely semantic differences with

contextual emergence. And we would need to know what kind of evidence to look for in scientific practice and explanation to determine whether that characterization is met. I propose the following.

I characterize reductionist/mechanistic explanation according to what I call *pragmatic downward pull*. That is, reductive explanation is the normative principle that it is better to understand the lower-level mechanistic organization in one's system of interest, even in the kinds of systems emergentists cite, and that it is not possible to explain phenomena entirely without doing so. We can now use this characterization to re-phrase the debate between the mechanist/reductionist and the contextual emergentist. The question is, are circumstances in which context affects the organization of a system, in which network organization is relevant, etc., inherently circumstances in which mechanistic and reductive frameworks of explanation are either *not possible* or *not desirable*?

It is worth pausing to note the ways in which this formulation of reductionism differs from traditional approaches. This approach is neither atomistic nor an instance of “nothing but-ism.” That is, it does not suggest that explanation can only rely on intrinsic properties of lower level parts; nor does it suggest that we must know all of the relevant lower-level information before we individuate system-level properties; nor does it deny that the resources of, e.g., topological or dynamical models can contribute explanatorily important, distinct information. What it does require, however, is that these system-level properties and models still need to be understood in terms of localization and decomposition. Within a topological organization, for instance, we need to understand how distinct components within that system contribute differentially to the phenomenon of interest; one must link the multiple kinds of explanation, and “connect” functional distinctions within the system to the phenomenon of interest by linking together the causal path that produces the phenomenon (Bickle and Kostko, 2018). So, reductionism construed as pragmatic downward pull offers a substantive view that is genuinely distinct from emergentist ones.

I will only focus on neuroscientific explanation here. In keeping with the pragmatist approach, the success of a position in the debate depends on whether it provides the right descriptive and normative view of how the best neuroscience works. In what follows, I argue that my construal of pragmatic downward pull is the best description of investigation into context-sensitive and network-mediated neural systems.

In particular, I will suggest that scientists seek a particular kind of mechanism when analyzing such systems – that is, they investigate *mechanisms that recognize context and implement new forms of organization*. If these mechanisms can be found, I argue, then we can understand shifts in context in a fully mechanistic way, and indeed we need to investigate these mechanisms in order to understand how the system works. That is, pragmatic downward pull obtains.

In section 4, I will discuss a number of context-recognition and reorganization mechanisms that neuroscientists have

uncovered. Before doing so, however, I want to set the stage a bit by considering some toy examples.

Context, topology, and constraints – Inherently emergent?

This section will be an exercise in deck stacking – or, at least, deck evening. I want to imagine some simple toy systems and ask whether, first, they can exhibit the properties that interest emergentists, and second whether they must be construed as implementing emergence.

One of the longest running daytime TV shows in the US is *The Price is Right*. As part of the show, contestants participate in carnival-style games, one of which is (or at least used to be) *Plinko*. In a *Plinko* system, one drops a ball from the top of a board, and the ball falls through a series of obstacles, ending up in one of several boxes at the bottom, each box representing a prize. The obstacles on the board are set in a lattice organization, so that the movement of the ball is a kind of random walk through the obstacles.

Here, obviously, the lattice affects the movements of the ball. But each of the interactions of the ball with individual obstacles is perfectly well-explained by basic causal interactions between them. The ball exhibits a kind of path dependence. The nature of its interaction with the first obstacle positions it so that it then has a certain interaction with the next obstacle, which positions it for the following one, etc. I submit that if there is emergence in the *Plinko* system, it is only of the weakest kind, where the arrangement of the obstacles shapes the directions in which the ball can go, but every interaction of the ball with the individual obstacles, and each *particular* path of the ball through the system, is fully explainable in terms of local interactions.

Let us imagine some slight variations to the *Plinko* framework. First, there's no reason why gravity can be the only force moving the ball, or downward the only direction. We can imagine a multi-directional *Plinko* board, where fans or vacuums or whatever propel the ball from any side to any other. Second, we do not have to think of the board as constant.

Suppose that, behind the scenes, there is a lever. When someone pulls the lever, a series of gears turn the obstacles so that they are now in a new arrangement. When the lever is thrown, the obstacles move in the following way. First, they closely align into rows, creating corridors through which the ball can quickly move. However, these corridors are frequently punctuated by “clearings,” around which the ball must bounce before finding a new corridor to enter. Further, suppose that some “clearings” are only connected by corridors to a couple of other clearings, but that some are connected to many clearings. In this system, the clearings and corridors roughly mirror the nodes and edges of a network. Clearings that connect to many other clearings will be “hubs” in this network. We can further imagine a distribution of clearings such that most clearings are low in connections or “degree,” and only connected to nodes close to them, but the hub nodes are

heavily interconnected with long range connections. This would be an analogue of a “small world” network. We can even imagine that the hubs are densely connected to *each other*, emulating what is called a “rich club” structure.

Once the lever is thrown, these topological facts will become relevant for the kinds of paths the ball can take. A ball in a rich club system, for instance, will likely move through the board faster, because the motive force will move it down a corridor, soon reaching a hub. Since hubs are richly connected in a rich club system, the ball will more quickly move through the board by hopping from hub to hub. Further, different *specific arrangements* – for instance, distinct spatial groupings of hubs – could each implement a rich club network. Imagine a rich club board, but one where the hub nodes are spatially clustered on one side of the board. Here, not only will ball traversals be slower than on a more spatially diffuse rich club (since balls run the risk of getting “trapped” in the rich club at one end of the board), but which *side* the ball starts on will matter. A ball placed in the rich club side will be more likely to find its way quickly to the other side of the board, due to the long range connections of the hubs, than a ball placed on the other side, that will risk wandering significantly before finding the “highway” corridors connecting hubs.

Here we have a situation where topology and context matter deeply for the “modal possibilities” of ball trajectories. A ball in the rich club board will have a much different distribution of possible trajectories. Still, I submit, there is nothing more than weakly emergent about this system. First, the changes of configuration are fully explained by the lever and gear system. This in turn modulates the way that the ball can move in new contexts (e.g., its direction of travel). But each particular trajectory is just a series of basic of basic mechanical interactions between the ball and the assorted obstacles.

What about the “topological facts” I alluded to earlier, and the fact that they are multiply realizable by distinct spatial layouts of particular boards (not to mention by wooden versus metal obstacles, etc.)? Given the setup of the case, this cannot be sufficient motivation for positing contextual emergence. Of course, different mechanisms can be similar in many respects. Citing a similarity between them is just citing an abstract feature that they share – and, as noted above, contextual emergentists insist that features exhibiting contextual emergence are not best described as useful abstractions of mechanistic properties. Moreover, note that similarities are important until they are not – the fact that the last two boards discussed both implement rich club networks does not mean that they are the same in all relevant respects – in some contexts, the differences between their spatial distributions do matter, for instance depending on the starting point of the ball.

An important aspect of this case is that, when context changes, one can explain that change in context *via a mechanism of contextual reorganization*. In the *Plinko* system, the lever and gear system *explains* the new form of organization, and the paths of the ball through the board are then the result of interactions within

that organization.² These are the two properties which I think are important for assessing the debate in neural systems. If we can explain *both* how contextual changes are implemented mechanistically in a system, *and* can show functional localization and decomposition *within* a context, then mechanistic/reductive explanation is possible despite context-sensitivity. I will argue that both facts obtain in the neural case.

I am not, of course, suggesting that the *Plinko* system is straightforwardly analogous to any biological or neural system. Biological systems have much more complex forms of organization and interaction. For one, they implement connections over a distance (e.g., through signaling) rather than *via* direct physical connection. For another, they often have bi-directional or reciprocal functional connections, wherein two parts influence each other mutually. Further, biological components often respond to *ensemble* properties, such as chemical gradients or, in the neural case, background electrical potentials. But none of these facts *themselves* require that decomposition and localization must fail. It would take an extra argument that localization and decomposition are not possible in these cases.

In the next section, I suggest four different types of mechanisms that neural systems implement to manage contextual change (I’m sure there are more). The emergentist is forced into the awkward position of claiming that *we should not care* about these kinds of mechanisms – i.e., they do not contribute productively to explanation. This, I claim, is wrong.

Context-recognition and implementation mechanisms in the brain

Context re-mapping and invariance mechanisms

The first set of mechanisms that I will consider involve how populations of cells either *re-map* their selectivity in particular contexts, or, just importantly, how they can come to generalize or achieve invariance *within* a type of context. These kinds of mechanisms show that learning and plasticity can *implement* specific forms of functional localization within particular units in the brain, which are themselves sensitive to the context.

² We can of course imagine more complicated, or *themselves varying*, forms of interaction, but the question is, similarly, here, whether these contextual shifts are mechanically mediated. Suppose, in addition to the lever, there is a switch. When the switch is thrown the obstacles exert a slight magnetic attraction on the ball. This will of course change the kinds of paths the ball exhibits. But the switch explains the introduction of the magnetic attraction, and the new paths will be determined by the new type of interaction between the ball and the obstacles. There is nothing here that is not mechanistically explicable.

The first example of re-mapping comes from physiological study of hippocampal neurons in monkeys. It is well-established that hippocampal cells exhibit *mixed-selectivity*, which means that they are selective for multiple parameters of a task context or stimulus (Rigotti et al., 2013). This is true even for place cells, whose responses are dependent primarily on the organism's spatial position. It is also the case that hippocampal cells are *variant* in their responses. This means that their responses can vary depending on the kind of environment that the organism is in, or its position in that environment. Some cells that show place-selectivity for one environment, for instance, will lose it or show a different selectivity in a different environment (Maren et al., 2013). Within a given environment, cells exhibit *phase precession*, which means that they sync to different phases of the theta rhythm in the local field potential depending on the organism's position in the environment.

Baraduc et al. (2019), in a study in *Nature*, explored how hippocampal cells of this type could learn to generalize across superficial changes to behavioral context, where the primarily important structure of that environment stayed the same. To do this, they had monkeys explore a virtual reality maze while recording from hippocampal neurons. They first had the monkeys learn a maze where rewards were "hidden" in different locations, and a primary cue for their locations was the relative spatial position of certain landmarks. So, if (for instance) a tree was to the right of a star, the reward would be in between the two landmarks.

The key manipulation of this study was when the experimenters changed the maze, while keeping the relational positional structure the same. So, for example, rather than a tree being to the left of a star, with the reward to the right of the tree, the tree and the star could be replaced by a triangle and a square, respectively, with the reward to the right of the triangle. Further they "rotated" the maze, such that the starting point varied from the monkey's starting point in the original maze. Intriguingly, once monkeys began to explore these kinds of mazes, versus totally novel mazes, they quickly realized that they had the same structure as the previous maze. This was shown by their rapidly learning the new maze.

Furthermore, *some* cells in the hippocampus exhibited similar selectivity properties in the structurally similar mazes after learning. In particular, these cells were selective to the current position of the monkey in the abstract structure, and its action-possibilities – e.g., re-orienting in a new direction to face the reward. Other cells in the hippocampus did exhibit re-mappings with the novel mazes, even those that shared the same abstract structure. So, the hippocampus exhibits multiple populations with selectivity properties that re-map to new contexts, but also form invariances to higher-order elements of context (e.g., spatial relations) as other aspects change.

A second example of re-mapping of this type involves not the selectivity properties of cells, but instead the structure of the population, i.e., how the population forms functional groups that are appropriate to the context. Cohen and Newsome (2008) performed a study where a sensory stimulus was of the same type

across contexts, but what kind of decision a monkey had to make varied depending on the context. The stimuli were dot-motion stimuli, in which the monkey is shown a pattern of dots moving in different directions. The level of "predominant" motion can be varied depending on the correlation between dots. So, more dots moving together to the left will result in predominant motion to the left, and so on for the other directions. Neurons from area MT, an extrastriate visual area dedicated (partially; see Burnston, 2016a) to motion, were measured while monkeys viewed these stimuli and made decisions about the direction of predominant motion.

The context manipulation involved implementing distinct two-alternative forced choice tasks. One task type involved asking the monkey whether *left* or *right* had more predominant motion. Another involved asking whether *up* or *down* did. This allowed for contrasts in context to be measured within cell populations in MT, based on how they related to the choice situation. Imagine two cells, one with selectivity for "motion upward to the left," and one with selectivity for "motion upward to the right." If the decision that needs to be made is *up* or *down*, then these cells will be cooperating in the decision – each will indicate *up*. However, if their decision is between *left* and *right*, they will be competing – one will indicate *left* and one will indicate *right*.

The idea of the researchers was that the cell population could be differentially recruited to implement these co-operations and competitions in the right setting. In particular, they measured "noise correlation," which is a comparison of the variance between two cells in similar trials. The reasoning here is that if two cells are part of a cooperating circuit, they will tend to vary together even in their noise properties. Intriguingly, they showed just this pattern. Two neurons of the type described above would show *increased* noise correlation in the *up* or *down* decision, and *decreased* noise correlation in the *left* or *right* decision. The authors suggest that (i) this is evidence of the neurons being in cooperative versus competitive circuits in the distinct contexts, and (ii) that the population reorganization may be due to attentional signals from more frontal areas of the cortex, shifting the population between attentional patterns for the different contexts.

In both of these cases, the populations in question exhibit plasticity and context-sensitivity. That is, they show particular selectivity or correlational variance that is sensitive to context. Of course, this is not the whole explanation, since there is still a question of how information *about* the context is relayed to the relevant populations. This brings us to the kind of mechanism discussed in the next subsection.

Context recognition and signaling

In this subsection, I discuss examples in which a system can be decomposed into a part that recognizes the context, compared to parts that provide it input, or which it causally affects downstream.

One example is shown in fMRI studies of humans, specifically with regards to fear conditioning. Context is very important for fear conditioning, since Pavlovian conditioning can be indexed to contexts, for instance when a mouse exhibits freezing in a cage where it has previously experienced a foot shock. The role of the hippocampus in context-based fear conditioning is well established physiologically in animal studies. Maren et al. (2013) cite a range of studies in which aversive fear conditioning is studied in humans, particularly the interaction between the hippocampus and the amygdala. One important finding is that, while the amygdala appears to be sensitive to aversive stimuli generally, the hippocampus is selectively activated for *signaled* as compared to *unsignaled* aversive stimuli. That is, when an organism experiences an aversive stimulus that is paired with a sensory cue, the hippocampus is sensitive to that correlation, whereas the amygdala is active with an aversive stimulus whether it is cued or not.

Further exploration of this circuit has occurred within the phenomenon of *fear extinction*. A previous fear association can be “extinguished” when the cue previously associated with the aversive stimulus is presented without that stimulus. Even further, extinction itself can be context-sensitive; i.e., a stimulus can be unpaired from an aversive response in some contexts but not in all. Fear extinction of this context-sensitive type is interrupted by injury to the hippocampus. Moreover, injuries to the hippocampus after fear extinction inhibit *re-implementation* of the fear in non-extinguished contexts. The interaction between context recognition in the hippocampus and its “gating” of cued associations in the amygdala is posited to be impaired in contextual fear in individuals with PTSD.

Another example comes from lesion studies in mice. Wu et al. (2020) studied a delayed-match-to-sample task in which a mouse must remember a stimulus during a delay period, and then compare it to a second stimulus. One behavior, in this case a lick to a left target, is rewarded if the stimuli match, and another (lick to a right target) is if they do not. This task setting implements a kind of context sensitivity in the association between the second stimulus and the action. Whether the second stimulus needs to be responded to with a left or a right lick depends on the identity of the first stimulus. So, motor areas involved in licking responses must modulate their association between stimulus and response depending on the context.

These kinds of context-dependent behaviors can be used to show where particular aspects of a decision are implemented, and how they specifically are interrupted by injury. For instance, one possibility is that the match or non-match decision is “made” in frontal cortical areas, and then propagated to motor areas such as the ALM, which simply implements the association between second cue and appropriate response. Another possibility is that the ALM itself is involved in computing whether the stimuli match or not, only receiving information about the identity of the first stimulus from other areas.

These alternatives were tested by varying where precise, pharmacologically induced lesions were introduced during

specific trials. For instance, the first possibility mentioned above suggests that lesions to the ALM during stimulus presentation or delay should *not* affect behavior, because the ALM is only relevant after the decision has been made, whereas lesions to frontal areas during the delay would impair performance. But this is the *opposite* of what was found. Lesions to the ALM during stimulus onset and the delay impaired behavior, proportional to the duration of the induced lesion. Conversely, lesions to frontal areas such as the orbitofrontal cortex only affected behavior during onset of the initial stimulus, not during the delay. Further, lesions to the ALM did *not* impair simple associations between stimulus and response, i.e., ones that were not part of a delayed match to sample task.

In each of these two examples we see the difference between a context-recognition element in the system and either an input or an output to that system. In the memory gating system, the hippocampus recognizes the context of a signaled association, or whether an association has now been subject to extinction, and gates the memory in the amygdala accordingly. In the frontal-ALM circuit, the researchers instead discovered that the frontal areas only recognized the inputs and relayed them to ALM, which in turn implemented the context-sensitive decision. In each case, the relevant systems are functionally decomposed.

Of course, each of these systems is only acting within a broader network of brain areas, so we now turn to discuss how such broader networks might be decomposed.

Context-specific network reconfiguration

Senden et al. (2018) performed a network analysis of functional connectivity in cortical areas, specifically with regards to how specific tasks are implemented. Functional connectivity is a measure of the temporal co-activation of brain areas. Each individual area is a node, and when two areas exhibit functional connectivity, this constitutes an edge. This allows for network-theoretic measurements to be applied to neural activity as opposed to bare structural connection, and hence track, as the authors suggest, informational exchange between areas. Importantly, these networks show general topological features of the types we have discussed. For instance, a set of areas, overlapping with but not coextensive with the brain’s “default mode” network, comprises a rich club – recall, this is the kind of network where hub nodes are themselves richly interconnected.

In particular, the researchers studied *changes* in context, including the change between rest and task conditions, and a comparison of the different task conditions. They found an intriguing set of results. Analyzing the temporal sequence of activation – the pattern of how functional connectivity changes over time, can give a sense of the directionality of activity. The predominant directionality of activity in the network changed between rest and task. While the rich club received similar levels of input across conditions, it exerted much more influence on

non-rich club “peripheral” nodes in task conditions. Further, while there was a significant (but not complete) overlap between the areas activated across the different types of task, the interactions *between* those areas varied depending on the task.

Here is an interpretation of these results, in line with that given by the researchers. The rich club serves as a contextual control system. When a particular task, with its particular informational requirements, is being performed, the out-degree of the rich club increases, enforcing a type of functional interaction between the components. These areas then respond in appropriate ways for that type of context. While this explanation is of course sketchy, we can see here both a distinction between control- and task-specific subsystems, with directional interactions between them, organized for the purposes of a specific task.

Slightly more detail can be seen in a study of episodic memory by Watrous et al. (Schedlbauer et al., 2014). They had subjects “navigate” around a virtual environment, dropping off and picking up a virtual friend at a series of stores. Then subjects’ functional connectivity networks were measured while they answered distinct questions about their experience. Some of the questions were *spatial* – e.g., which store was closest to store x? Some were *temporal* – e.g., which store did you visit after store x? While a broad network was activated in each context, some key points distinguished the two. First, while the medial temporal lobe, comprising the hippocampus and associated cortical regions, was an equally significant hub in the functional connectivity networks in both kinds of tasks, different areas – the lateral prefrontal cortex and the posterior parietal cortex, achieved greater network centrality in the temporal and spatial contexts, respectively.³ Again, the interpretation is that the medial temporal lobe serves as a context-reinstating device, organizing the network so as to recall the particular kinds of information needed for the task. Hence, one way that broad networks can be decomposed is in situations of context-sensitivity is to look for the parts of the network that mediate the context, and those that implement task-specific organization.

Dynamic regime shift

The results in the last subsection were discussed at the broad network level, but there is also significant evidence that individual areas vary their behavior to implement the right informational requirements for specific contexts. In addition to the re-mapping results discussed in section 4.1, populations of cells can also change their *dynamical regimes* to represent information in the way required for the context.

To take one example, Warden and Miller (2010) studied working memory in monkeys’ prefrontal cortical cells. They had

two tasks, both of which involved an initial presentation of a sequence of two objects. In the “recognition” task, a delay would be followed by presentation of a second sequence of objects, and the monkey would have to indicate whether the second sequence matched the first. In the “recall” task, after the delay the monkeys were presented with a set of objects and would have to re-create the sequence by making saccades to the two formerly presented objects in the right order. Object-selective cells in the prefrontal cortex behaved differently in these two contexts, specifically in the delay between the presentation of the original sequence and the presentation of the test stimulus.

In the recognition task, activity amongst the cells selective for the second object during the delay was much greater than that of the cells selective for the first object. In the recall task, however, this selectivity was equal. Why would this be? The authors suggest that in the former task, the cells operate with a “passive buffer” type of memory, where the activity of object-selective cells decays over time. This more passive type of memory suffices for the task because there is only one subsequent test stimulus that either matches that selectivity or does not. The recall task, however, requires a more active form of maintenance, since the match must be selected by the monkey out of a number of presented stimuli.

Intriguingly, this change in the dynamics of the representation – from passively decaying to actively maintained – seems to depend on cells dedicated to context-recognition. Particular cells actively represent the task context, and in turn influence the object-selective cells. Here, again, this time within a cell population, we have a part of the system that is recognizing the context, and using it to implement a specific functional change, in this case the passive versus active maintenance of information. Meyers et al. (2012) have in turn shown how task-selective cells develop in the population over the course of task-learning.

This basic idea of changing dynamic regimes has been generalized in theoretical work. Rigotti et al. (2010) modeled a neural population as comprising two sub-networks, a *context network* and an *associative network*. The associative network would learn and implement simple associations between conditioned and unconditioned stimuli. The context network, on the other hand, comprised a fully interconnected group of cells with mixed selectivity for both external events (presentation of stimuli and reward) and the states of the associative network. This allowed the context network to track what combinations of external cues and associative network states led to reward. By providing feedback to the associative network, the context network cells were able to create groupings of associations that were specific to each context. The authors show how these properties capture the kind of physiology observed in prefrontal cortical populations in a reverse-conditioning task, where original learned associations between cues and rewards are flipped in a subsequent learning epoch.

Importantly, this process affects the dynamics in the system, through a process of what the authors refer to as “attractor concretion.” This is the way in which reverberatory activity in the network will be qualitatively distinct in one context rather than

³ This was accompanied by a distinctive change in the background local field potential at which the network synchronized, further dissociating the contexts. See Burnston, 2021, for more details.

another, thus implementing a distinct pattern of activity for each context. In this way, a distributed network with distinct populations can implement context-specific dynamics.

Argument from the cases

I have argued, first, that the question of whether context sensitivity is incompatible with reductive/mechanistic explanation turns on whether we can discover mechanisms that implement new forms of organization for specific contexts. Second, I have argued that there are many types of such mechanisms. The question is how to interpret these cases.

My phrasing of reductionism, espoused in section 2, is that of *pragmatic downward pull*. This is the normative principle that a full understanding of a system requires decomposing the system at lower levels. The cases above suggest that it is possible to do this. In each case, there are specific cells, populations, or sub-networks that implement the contextual changes in the network, and components that shift their function in response to those changes. Of course, these decompositions are not simple, easy, or atomistic. The process of doing functional decomposition in these systems is much more complicated than in the toy *Plinko* case I gave in section 3. But, on the pragmatic view of the debate, complexity does not just equal emergence. It is a claim about what the most successful science does.

Nor have I suggested that any of these explanations are *complete*. There are many more details to fill in, many new contexts to understand, etc. In particular, the way that contextual changes result in particular functional patterns is better understood at the circuit level than at the network level – what, for instance, is it about the new functional organizations of peripheral nodes in the Senden et al. study that enables the specific tasks in which they are implemented? The reductionist picture's *pragmatic downward* aspect suggests that further study of these contexts should proceed until the kinds of explanations that have been given at the cell population level are possible.

The emergentist is forced into an awkward situation with regards to these mechanisms. They must either deny that they really are mechanisms, or they must say that we do not really learn anything from discovering them. I do not think either option is tenable. If we agree that decomposition and localization are definitive of mechanistic/reductionistic explanation, then these analyses which assign particular functions to distinct cells, populations, or subnetworks are mechanistic, despite the fact that these decompositions do not posit immutable, fixed mechanisms, but rather contextually shifting ones. It would be hard to know how to adjudicate the claim that we do not learn anything important from these analyses. Remember, the pragmatic argument is an abduction from successful scientific practices. These practices, if successful in the mechanistic sense, can only be ruled out of bounds by assuming that mechanistic analysis is not productive, which is just what is under consideration.

There are a couple of strategies left open to the emergentist at this point, which I'll call the *shifting domain* strategy, and the *shifting explanandum* strategy. Frequently emergentists make nods to mechanistic analysis – *sometimes*, they admit (e.g., for simple interactions within the system), mechanistic explanation is possible, useful, etc. But they then suggest that for the circumstances that are really interesting for understanding biological function, these approaches must be left aside. So, is there any principled way of defining a series of settings where mechanistic/reductive explanation is not useful, even if it is in explaining the kind of behavioral phenomena I discussed above?

The shifting domain strategy suggests that for certain kinds of phenomena, mechanistic/reductive explanation is bound to fail, even if it is successful in other cases. Psychopathology is one such domain often referred to by Silberstein et al – here the idea is that psychiatry is the kind of domain for which the mechanistic facts about the system drop out of the explanation, and all of the explanatory work is to be done by contextual and topological/dynamical properties. So can such a move help parcel explanations into those that are amenable to mechanism and those that aren't?

There is no doubt that some significant advances in studying psychopathology at the neural level have been achieved by employing network frameworks, including studies of the rich club and the way it is interrupted in such cases as schizophrenia (van den Heuvel et al., 2013). But, given the current state of the dialectic, this bare fact is far from sufficient to establish the emergentist conclusion. Often, the way the argument goes in these cases is that the emergentist contrasts the topological approach with the kind of reductive explanation that seeks, for instance, a single genetic or neural locus for psychiatric disease. If a “biopsychosocial” model is right, they contend, then mechanistic explanation is impossible.

On reflection however, this argument illicitly assumes that that traditional atomistic model of reduction is the only one possible. Nothing about the more sophisticated forms of reductionist/mechanistic explanation denies that topological description can play a role, even a critical role, in explaining the phenomenon. Nor does it ally itself with single-locus analyses of pathology, or deny that social/developmental factors may vitally influence the mechanisms responsible for it. The pragmatic downward pull approach suggests that our explanations will be deeper and better if we seek lower-level explanations in addition to the higher-level ones.

We already saw above how the rich club has been implicated in the way that distinct tasks are mediated by reorganizations of peripheral nodes. A natural hypothesis is that interruption of the rich club network in schizophrenia affects these reorganizations. But in order to understand this, we'd need to *both* understand how informational reorganizations operate in the regular tasks, and how that normal operation is interrupted in schizophrenia. The pragmatic downward pull approach, in my view correctly, normatively suggests searching for those explanations, and that decomposition and localization (of the contextualized sort) are

reasonable explanatory strategies for pursuing them. The emergentist approach rules this out by fiat.

The explanandum shifting strategy suggests that there are certain properties of the system that are only explicable by (for example) network frameworks, and not by mechanistic frameworks. A set of properties that has frequently been adduced in this setting includes the system being *robust*, the system having a certain kind of dynamics, or the system exhibiting scale dependence in how it is modeled (Green and Batterman, 2017). The idea here is that when we look at properties of a class of systems themselves, rather than the phenomena they produce, we will have to resort to network and dynamical descriptions at the expense of mechanistic ones. Again, no one should doubt the importance of network description in these contexts. But again, we can question whether this has the overall upshot that the emergentist assumes.

I suggest that this way of arguing implicitly assumes a version of explanatory pluralism that is contestable, and hence the argument does not go through without a prior establishment of a thesis about pluralism, which regularly goes undiscussed in these contexts. On what I'll call "division-of-labor" pluralism (Potochnik, 2017; Rathkopf, 2018; Burnston, 2019), there are distinct explananda we might investigate about a system, and those distinct explananda will require distinct and disjoint types of explanation. When one changes explananda – for instance, in switching from an explanation of how memory occurs, to asking how memories can be robust to patterns of decay and variation – one selects the right kind of explanatory framework for that explanandum. But the division-of-labor view is not the only version of pluralism.

On more "integrative" approaches, distinct kinds of explanations or models need to contribute even to understanding one single explanandum. If one embraces an integrative view, then understanding the relationship between mechanistic description and network description of the system is required. On this kind of view, a system property like robustness will be better explained by taking into account both network features and the particular, functionally distinct roles played by their constituents. So, explaining how a memory can be retained in some contexts while extinguished in others, or how a monkey learns to recognize a type of maze despite superficial variations, can only be achieved by both analyzing the brain networks involved and the causal/functional specifics of the constituents of the network. In any event, the division-of-labor view cannot just be presumed to be more fruitful than more integrative views, and hence the

adjudication of the explanandum-shifting strategy has to be pursued within the larger discussion of scientific pluralism.

Conclusion

Recent projects have argued from the presence of complexity in biological systems to the presence of emergence, and the concomitant failure of mechanistic decomposition, in these systems. I have argued that this argumentative move is by no means obvious, particularly if we focus on the mechanisms involved in contextual reorganization of these systems. If I am right, then there is no easy argument from context-sensitivity to emergence.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

The author confirms being the sole contributor of this work and has approved it for publication.

Conflict of interest

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

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Anderson, M. L. (2014). *After Phrenology: Neural Reuse and the Interactive Brain* Cambridge, MA: MIT Press.
- Baraduc, P., Duhamel, J.-R., and Wirth, S. (2019). Schema cells in the macaque hippocampus. *Science* 363, 635–639. doi: 10.1126/science.aav5404
- Bechtel, W. (2016). Mechanists must be holists too! Perspectives from circadian biology. *J. Hist. Biol.* 49, 705–731. doi: 10.1007/s10739-016-9439-6
- Bechtel, W. (2017). Analysing network models to make discoveries about biological mechanisms. *Brit. J. Phil. Sci.* 70, 459–484. doi: 10.1093/bjps/axx051
- Bechtel, W., and Abrahamsen, A. (2010). Dynamic mechanistic explanation: computational modeling of circadian rhythms as an exemplar for cognitive science. *Stud. Hist. Phil. Sci. Part A* 41, 321–333. doi: 10.1016/j.shpsa.2010.07.003
- Bickle, J., and Kostko, A. (2018). Connection experiments in neurobiology. *Synthese* 195, 5271–5295. doi: 10.1007/s11229-018-1838-0
- Boogerd, F. C., Bruggeman, F. J., Richardson, R. C., Stephan, A., and Westerhoff, H. V. (2005). Emergence and its place in nature: a case study of

- biochemical networks. *Synthese* 145, 131–164. doi: 10.1007/s11229-004-4421-9
- Brigandt, I., Green, S., and O'Malley, M. A. (2018). "Systems biology and mechanistic explanation," in *The Routledge Handbook of Mechanisms and Mechanical Philosophy*. eds. S. Glennan and P. Illari (New York: Routledge), 362–374.
- Burnston, D. C. (2016a). Computational neuroscience and localized neural function. *Synthese* 193, 3741–3762. doi: 10.1007/s11229-016-1099-8
- Burnston, D. C. (2016b). A contextualist approach to functional localization in the brain. *Biol. Phil.* 31, 527–550. doi: 10.1007/s10539-016-9526-2
- Burnston, D. C. (2019). Review of Angela Potochnik's idealization and the aims of science. *Philos. Sci.* 86, 577–583. doi: 10.1086/703574
- Burnston, D. C. (2021). Getting over atomism: functional decomposition in complex neural systems. *Br. J. Philos. Sci.* 72, 743–772. doi: 10.1093/bjps/axz039
- Chemero, A., and Silberstein, M. (2008). After the philosophy of mind: replacing scholasticism with science. *Philos. Sci.* 75, 1–27. doi: 10.1086/587820
- Cohen, M. R., and Newsome, W. T. (2008). Context-dependent changes in functional circuitry in visual area MT. *Neuron* 60, 162–173. doi: 10.1016/j.neuron.2008.08.007
- de Wit, M. M., and Matheson, H. E. (2022). Context-sensitive computational mechanistic explanation in cognitive neuroscience. *Front. Psychol.* 13, 1–13. doi: 10.3389/fpsyg.2022.903960
- Deacon, T. W. (2006). "Emergence: The Hole at the Wheel's Hub," in *The Re-Emergence of Emergence: The Emergentist Hypothesis From Science to Religion*. eds. P. Clayton and P. Davies (Oxford, UK: Oxford University), 111–150.
- Delehanty, M. (2005). Emergent properties and the context objection to reduction. *Biol. Philos.* 20, 715–734. doi: 10.1007/s10539-004-2437-7
- Gillett, C. (2016). *Reduction and Emergence in Science and Philosophy*. Cambridge, UK: Cambridge University Press.
- Green, S., and Batterman, R. (2017). Biology meets physics: reductionism and multi-scale modeling of morphogenesis. *Stud. Hist. Phil. Sci. Part C* 61, 20–34. doi: 10.1016/j.shpsc.2016.12.003
- Green, S., Šerban, M., Scholl, R., Jones, N., Brigandt, I., and Bechtel, W. (2018). Network analyses in systems biology: new strategies for dealing with biological complexity. *Synthese* 195, 1751–1777. doi: 10.1007/s11229-016-1307-6
- Huneman, P. (2018). Diversifying the picture of explanations in biological sciences: ways of combining topology with mechanisms. *Synthese* 195, 115–146. doi: 10.1007/s11229-015-0808-z
- Huttemann, A., and Love, A. C. (2011). Aspects of reductive explanation in biological science: Intrinsicity, fundamentality, and temporality. *Br. J. Philos. Sci.* 62, 519–549. doi: 10.1093/bjps/axr006
- Kaiser, M. I. (2015). *Reductive Explanation in the Biological Sciences*. Dordrecht, NL: Springer.
- Kaplan, D. M., and Craver, C. F. (2011). The explanatory force of dynamical and mathematical models in neuroscience: a mechanistic perspective. *Philos. Sci.* 78, 601–627. doi: 10.1086/661755
- Levy, A., and Bechtel, W. (2016). Towards mechanism 2.0: expanding the scope of mechanistic explanation. *Paper presented at the Biennial Meeting of the Philosophy of Science Association*, Atlanta, November 3–6.
- Maren, S., Phan, K. L., and Liberzon, I. (2013). The contextual brain: implications for fear conditioning, extinction and psychopathology. *Nat. Rev. Neurosci.* 14, 417–428. doi: 10.1038/nrn3492
- Meyers, E. M., Qi, X.-L., and Constantinidis, C. (2012). Incorporation of new information into prefrontal cortical activity after learning working memory tasks. *Proc. Natl. Acad. Sci.* 109, 4651–4656. doi: 10.1073/pnas.1201022109
- Potochnik, A. (2017). *Idealization and the Aims of Science*. Chicago, IL: University of Chicago Press.
- Rathkopf, C. (2018). Network representation and complex systems. *Synthese* 195, 55–78. doi: 10.1007/s11229-015-0726-0
- Rigotti, M., Barak, O., Warden, M. R., Wang, X.-J., Daw, N. D., Miller, E. K., et al. (2013). The importance of mixed selectivity in complex cognitive tasks. *Nature* 497, 585–590. doi: 10.1038/nature12160
- Rigotti, M., Rubin, D. B. D., Morrison, S. E., Salzman, C. D., and Fusi, S. (2010). Attractor concretion as a mechanism for the formation of context representations. *NeuroImage* 52, 833–847. doi: 10.1016/j.neuroimage.2010.01.047
- Sarkar, S. (1992). Models of reduction and categories of reductionism. *Synthese* 91, 167–194.
- Schedlbauer, A. M., Copara, M. S., Watrous, A. J., and Ekstrom, A. D. (2014). Multiple interacting brain areas underlie successful spatiotemporal memory retrieval in humans. *Sci. Rep.* 4, 1–9. doi: 10.1038/srep06431
- Senden, M., Reuter, N., van den Heuvel, M. P., Goebel, R., Deco, G., and Gilson, M. (2018). Task-related effective connectivity reveals that the cortical rich club gates cortex-wide communication. *Hum. Brain Mapp.* 39, 1246–1262. doi: 10.1002/hbm.23913
- Silberstein, M. (2021). "Constraints on localization and decomposition as explanatory strategies in the biological sciences 2.0," in *Neural Mechanisms*. eds. F. Calzavarini and M. Viola (Dordrecht, NL: Springer), 363–393.
- Silberstein, M. (2022). "Context is king: contextual emergence in network neuroscience, cognitive science, and psychology," in *From Electrons to Elephants and Elections: Exploring the Role of Content and Context*. eds. S. Wuppuluri and I. Stewart (Cham: Springer International Publishing), 597–640.
- Silberstein, M., and Chemero, A. (2013). Constraints on localization and decomposition as explanatory strategies in the biological sciences. *Philos. Sci.* 80, 958–970. doi: 10.1086/674533
- van den Heuvel, M. P., Sporns, O., Collin, G., Scheewe, T., Mandl, R. C., Cahn, W., et al. (2013). Abnormal rich club organization and functional brain dynamics in schizophrenia. *JAMA Psychiat.* 70, 783–792. doi: 10.1001/jamapsychiatry.2013.1328
- Warden, M. R., and Miller, E. K. (2010). Task-dependent changes in short-term memory in the prefrontal cortex. *J. Neurosci.* 30, 15801–15810. doi: 10.1523/JNEUROSCI
- Wimsatt, W. C. (2006). Reductionism and its heuristics: making methodological reductionism honest. *Synthese* 151, 445–475. doi: 10.1007/s11229-006-9017-0
- Wu, Z., Litwin-Kumar, A., Shamas, P., Taylor, A., Axel, R., and Shadlen, M. N. (2020). Context-dependent decision making in a premotor circuit. *Neuron* 106, 316–328 e316. doi: 10.1016/j.neuron.2020.01.034
- Zerilli, J. (2020). *The Adaptable Mind: What Neuroplasticity and Neural Reuse Tells Us About Language and Cognition*. Oxford University Press, United States.



OPEN ACCESS

EDITED BY

Francesca Strappini,
Sapienza University of Rome, Italy

REVIEWED BY

Christophe Gauld,
Hospices Civils de Lyon, France

*CORRESPONDENCE

Kerrin A. Jacobs
kjacobs@let.hokudai.ac.jp

SPECIALTY SECTION

This article was submitted to
Personality Disorders,
a section of the journal
Frontiers in Psychiatry

RECEIVED 08 July 2022

ACCEPTED 31 October 2022

PUBLISHED 16 November 2022

CITATION

Jacobs KA (2022) The concept of
Narcissistic Personality
Disorder—Three levels of analysis for
interdisciplinary integration.
Front. Psychiatry 13:989171.
doi: 10.3389/fpsy.2022.989171

COPYRIGHT

© 2022 Jacobs. This is an
open-access article distributed under
the terms of the [Creative Commons
Attribution License \(CC BY\)](#). The use,
distribution or reproduction in other
forums is permitted, provided the
original author(s) and the copyright
owner(s) are credited and that the
original publication in this journal is
cited, in accordance with accepted
academic practice. No use, distribution
or reproduction is permitted which
does not comply with these terms.

The concept of Narcissistic Personality Disorder—Three levels of analysis for interdisciplinary integration

Kerrin A. Jacobs^{1,2*}

¹Department of Philosophy and Ethics, Faculty of Humanities and Human Sciences, University of Hokkaido, Sapporo, Japan, ²Center for Human Nature, Artificial Intelligence, and Neuroscience (CHAIN), University of Hokkaido, Sapporo, Japan

In this paper, I distinguish three different levels for describing, and three corresponding ways for understanding, deficient empathy as the core of NPD (Narcissistic Personality Disorder). On the macro level, deficient empathy can be explained as disturbed interpersonal functioning, and is understood as lack of recognition. On the meso-level, deficient empathy can be described as psychic disintegration, and can be understood specifically in its dissociative aspects. Psychic disintegration in NPD correlates with somatic changes, i.e., dysfunctional affective empathy and mind-reading on the micro level of description, which is the third level. The “core-deficit-model of NPD” that I outline, while not rejecting reductionist approaches outright, argues in favor of integrating (top-down/bottom-up) functionalist descriptions of empathy into a wider conceptual framework of bio-psycho-social functioning. The “core-deficit-model of NPD” is interdisciplinary, can bypass monodisciplinary skepticism, and removes purported barriers between explaining and understanding the “lack” of empathy as the core of pathological narcissism.

KEYWORDS

Narcissistic Personality Disorder, empathy, psychoanalysis, phenomenology, reductionism, integration

Introduction

A scientific explanation of an empathy deficit in terms of biological dysfunctions is usually considered as a bottom up approach, while a hermeneutic understanding of the deficit as a (harmful) impairment of normal interpersonal functioning is usually considered as top down. In this paper, I will first look at well-known symptoms associated with conditions of trait narcissism and pathological narcissism. It turns out that an important symptom is an diminished ability to empathically engage with other persons. This feature stands in need of deeper explanation. If one adopts a gradualist view of psychic capacities, persons with NPD apparently do not lack empathy altogether, but

rather they show a more or less reduced propensity of empathic concern. Instead of empathic concern, persons with severe NPD often develop a rather objectifying view of other persons. This objectifying mode of affective detachment is specific for narcissistic cognition (1).¹

If one wants to bridge the conceptual gap between the world of neuroscientific naturalistic explanations of mental processing in persons with NPD, and hermeneutic explanations of their ways of interacting with others, then bidirectional causal relations between the intrapsychic realm and the interpsychic realm of functioning can serve as the basis for coordinating heterogeneous functional descriptions at different levels of analysis. I will address some expectable methodological doubts about such an integrative attempt later on. The conclusion I want to substantiate is that deficient empathy is at the core of NPD and, moreover, that focusing on deficient empathy is the key for integrating three different, albeit structurally interrelated, levels of functional description, yielding descriptions as (1) a biological dysfunction, (2) an impairment of (intrapersonal) psychic functioning, and (3) as a form of maladjustment of (pro)social relatedness.

The role of deficient empathy in pathological narcissism

The term “narcissist” colloquially (2) refers to the phenotype of self-absorbed, exploitative, egocentric, excessively demanding individuals with a strong tendency toward (predominantly: self-)idealization: they experience themselves as exceptional and grandiose and have little empathy for others. In the literature, articulations of this phenotype vary to some extent (3). However, there is significant overlap in the characteristics of a heightened sense of self and pronounced convictions of entitlement, and corresponding strategies of self-regulation. Assessing pathological narcissism requires not only a focus on intrapersonal functioning but on interpersonal functioning as well (4–6). Relevant for a differentiation of narcissistic modes of relatedness to oneself and to the world is the distinction between *grandiose* narcissism and *vulnerable* narcissism. In this paper, I focus on *pathological narcissism* which conceptually includes aspects of vulnerability and grandiosity as co-occurring symptoms. Both are not distinct traits, but manifestations of the same phenomenon, i.e., it depends on the more basic personality trait of intro- and extraversion whether either vulnerability or grandiosity is displayed (7). A narcissistic person's oscillation between grandiosity and vulnerability indicates a psychic disintegration on the intrapersonal functional level, which

structurally correlates with maladaptive behavior on the level of social functioning. Vulnerability does not count as constitutive for the clinical diagnosis of pathological narcissism, since pathological narcissism is phenomenally not associated with low self-esteem. Nevertheless, vulnerability finds expression in a higher than average tendency of harming oneself (8), in suicide ideation (9), addictions (10), especially substance abuse [e.g., alcohol (11)]. This fact, together with a marked discrepancy between a positive future-orientation and an overall negative outlook (12) amounts to an intrinsic vulnerability factor in NPD. The *vulnerable* phenotype of narcissism appears as introverted, hypersensitive, defensive, with a tendency for withdrawal and lowered self-esteem (13, 14). This is accompanied by increased ratings for anxiety and depression (15). Feelings of shame, inferiority and boredom are relevant symptoms of narcissistic depression as a special type of depression, distinct from ordinary depression (16).

Moreover, empirical findings suggest that threatening situations vary in their relevance for vulnerable narcissism and for grandiose narcissism. While grandiose narcissism is very sensitive to achievement setbacks, vulnerable narcissism is more sensitive experienced shame (17). Narcissism is diagnostically assessed either by means of a structural model (18) or by means of a spectrum model (19) and an assessment of empathy functioning is standardly included in understanding personality disorders in the *DSM-5* (20, 21) and the *ICD-11* (22, 23). While *intrapersonal* functioning is concerned with aspects of identity and the aspect of self-regulation, *interpersonal* functioning addresses affective empathy, mind-reading, and also intimacy as central abilities for functioning well in interpersonal relations (24).

The somatic description of deficient empathy in NPD

There is a staggering wealth of empirical data concerning the neurophysiological causes of mental dysfunction – e.g., causes of altered empathy in narcissism. From a methodological point of view, we have to note that some basic naturalistic underpinnings of empathy capacities have been identified. This fact illustrates how explanatorily strong the naturalistic paradigm has become. One can say without exaggeration that during the last decade, a “somatization” of personality disorders has taken place. This has become the conceptual background for the reassessment particularly of moral-psychological attempts to explain the role of empathy (in mental disorders). Neurophysiological explanations also put some alignment pressure on (social) psychological models that (still) operate with alternative notions of psychic functioning in order to explain psychic impairments in NPD “outside” of the naturalistic discourse of neurophysiological science. My core-deficit hypothesis is

1 This is paradigmatically described in the case review of “Jane” in the study by Bonney Reed-Knight and Sarah Fischer. The patient explains this particular phenomenon by saying: “If they’re not meeting my needs, they’re in my way, if they’re meeting my needs, I need them.” [(1), p.470].

concerned with the *conceptual* analysis of NPD. The question is how to connect functional explanations framed in the naturalistic terms of the languages of neurophysiology to functional explanations framed in non-naturalistic terms of the languages of other disciplines, e.g., clinical psychology, social philosophy, hermeneutic social science. For a comprehensive theory of NPD the power of naturalistic explanations of cognitive and affective empathy are most welcome. However, we should not stop with functional descriptions at the biological and neurophysiological level of analysis. Rather, we should ask how such findings match with functional descriptions at other levels of analysis, e.g., clinical psychological descriptions of *psychic* impairment in persons with NPD. At this level of analysis, we understand empathy deficiencies in terms of maladaptive intersubjective practice. It is an interesting observation that naturalistic analyses often translate their empirical findings on deficient empathy in NPD into the language of *psychic* functioning, usually without caring at all about the proper conceptual clarification of notions like “impairment,” “inability,” “incapacity,” “distortion,” “deficit.”² Typically, naturalistic analyses do not address the question whether we should *identify* neurophysiological events and processes with mental events and processes, or whether we should treat them as merely *correlated*. My impression is that the naturalistic studies that I have referenced resonate with the correlation paradigm (see Section Empathy deficit between *soma* and *psyche*).

From the neuroscientific point of view, empathic processes are grounded in dissociable neural systems (25). Empathy is conceptualized as the ability to affectively experience other persons’ emotional states and as the ability to recognize and understand other persons’ emotional states. A prerequisite for this is the ability to monitor oneself and to maintain and regulate self-other awareness (21) in order to differentiate

between one’s own and others’ experiences (26). *Affective empathy* includes responsiveness to affectivity displayed by others, plus emotion-eliciting stimuli, which is not the same as the ability of mirroring others in one’s responses (27). It is associated with (partially) distinct systems – all require activation of the superior temporal cortex – that show increased activation (amygdala, insula, ventrolateral prefrontal cortex), respectively, when agents respond to emotional expressions of others (28). Affective empathy develops ontogenetically earlier than cognitive empathy. Hence, from a social-philosophical perspective we can say that *recognition* of the other “is ontogenetically (and conceptually) *prior*” to cognition [cf. (29), p. 354]. Empathic concern sets on during the second year of life, and its development depends on whether interactions either hinder or support empathic concern for others and/or support self-other-awareness (30, 31). Genetics (32), temperament and character (33) determine the development of empathic capacity in general. However, how specific genetic and environmental factors contribute to the development of personality disorders, and how genes particularly influence empathy deficiencies in narcissism, these issues continue to be controversially debated and remain a topic for future research (34, 35).

Persons are normally capable of *perspectives-taking*. The development of this ability presupposes an imaginative faculty in order to attribute different emotions, attitudes, and desires to other persons in a given situation (36). Empathy is interpersonally trained and is consequential for psychosocial development (37, 38). Experimental research on empathy in narcissism indicates a stronger deficit in emotional empathy rather than in cognitive empathy, highlighting the factor of psychosocial development. A lack of intersubjective recognition, especially in terms of emotional neglect (39) and abuse, figure prominently in the literature. It is assumed that they constitute a pathogenic potential for the development of NPD (40, 41). *Nota bene*: (primary) narcissism is a normal aspect of children’s development (42) and needs an age-appropriate satisfaction – a reflection of the grandiose self of the child – for a healthy psychic development. If such satisfaction is not forthcoming, i.e., when empathic reactions of the caregivers are missing, or when the child is overwhelmed by a caregiver’s own grandiose self-expectations, according to interactional psychoanalysis (43) this constitutes a causal factor for developing pathological narcissism (44). The specific parental style, the inter-generational consequences of narcissistic relational styles, and the role of distorted self-other-awareness in families establishes a research field of its own (45, 46).

While affective empathy has a subcortical basis, *cognitive empathy* is associated with a network of cortical regions that enable mindreading-related neural processing (47). Empathic ability in the sense of *perspective taking* implies being able to attribute more complex states (thoughts, motives, intentions, etc.) to others, to change perspective and also to take an impartial point of view. This in turn requires overcoming egocentricity

2 The notion of “deficit” suggest a gradual theory of incapacity. While “defects” are structurally manifested, irreversible deficits – biostatistical measurable dysfunctions that can no longer be compensated by other functions (functional networks, modules) – deficits can be captured as reversible and structurally (at least partially) compensable dysfunctions in a network or module of functions. Analogously, these terms are obviously also used on the level of description of abilities. *Incapacity*, as irreversible inability, mark the extreme form of impairment, while deficits are characterized by a changeability in both directions of improvement and deterioration in a corresponding area of ability. Within the field of psychic impairments an application of the gradual view indicates to specify the degree of dysfunctionality and that of inability in equal measure. Moreover, NPD is apparently also characterized by a certain degree of functioning, which can also be taken into account by a gradual view provided with the notion of deficit. Another term is that of *distortion*, which is referred here to explain gradual forms of social *maladjustment*, while *maladjustment* implies an assessment against a normative theory of social behavior and wellbeing.

in perspective-taking in favor of empathizing with another person's situation and mental states. The dissociation between affective and cognitive empathy has been central for criticizing moral psychological approaches that solely rely on perspective-taking for explaining prosocial motivation. Empathy theories that do not take into account that the ability to take perspectives develops later than the ability to respond affectively to the *suffering* of others are open to empirical attack (48). Effective empathy goes, however, "beyond perspective-taking" as Jordan Carpenter and colleagues have suggested with the *mind-reading motivation model (MRM)* (49), according to which the individual differences of agents' willingness to get engaged in understanding the perspectives and mental states of other people is described. Conceptually this closes the gap between a mere registering another person's condition and being *actually* motivated in figuring out what others think, even if it is of no explicit personal relevance. Individuals that score low in MRM show a lower propensity of exposing themselves to others' perspectives. Exactly this is key to describe the *empathy deficit* in narcissism: it is a lack of *real* interest for the thoughts and feelings of others, and often rests on a decision for staying rather detached from others. Low affective empathy scores in pathological narcissism relate strong with grandiosity (50), and a lowered performance in perspective taking has been measured with respect to decision-making tasks (51). Individuals with NPD display also lower levels of perspective taking, when they have to respond in test-scenarios to questions that explicitly ask for *the motivation to become empathically* concerned in with others [cf. (52), p. 7]. These empirical results make a differentiation between a fundamental *lack* (53) and a reduced propensity to *recognize* the feelings and needs of others in NPD [cf. (19), p. 5]. This conceptual change of empathy incapacity has been also adopted to the fourth edition of the *DSM-IV* (54, 55). Individuals with NPD might be capable of processing affective information, but *decide* not to response empathically to others [cf. (52), p. 7], at least, when it is not directly beneficial for them to show concern in a specific situation, which also indicates that displaying empathy may become itself instrumentalized in NPD.

The empathy deficit then seems to be hybrid: albeit individuals with NPD can *register* the affective states of others, the ability to be emotionally *motivated* by them is insufficient and depends on the specific situational conditions. Considering the role of perspective-taking, individuals with NPD are apparently capable of imagining what moves other people, can infer how they might behave and how they themselves *should* behave. Narcissists are also capable of *appearing* compassionate or concerned. If narcissists can imitate ("fake") these reactive attitudes of empathic concern they apparently have a sufficient cognitive understanding of the concept of empathy and/or compassion, and are also able to differentiate when the empathic responses technically should be exhibited. Empathy is (gradually) displayed as long as it severs own interests and can become instrumentalized to the detriment of others (56). The empirical studies are particularly helpful to assess the

impairment of *moral* competence in individuals with NPD. Granted that higher-order cognitive processes (self-reflexivity) are necessary for moral competence (such as the understanding of action guiding maxims or higher-order volitions) it may be especially the insufficient affective motivation that explains the unwillingness of acting in conformity with norms (particularly harms-norms) in NPD. In this respect, the capacity for moral judgment should be further explored according to this bias of cognitive and affective functioning in NPD.

The psychodynamic description of deficient empathy in NPD

Psychodynamic studies on pathological narcissism resonate with the empirical findings, and ever since have stressed the oscillation between the vulnerable and grandiose aspects as main characteristic of pathological narcissism. In contrast, this has been a rather neglected aspect in the diagnostic manuals and the DSM-4 NPD category has been largely criticized for focusing mostly on overt grandiosity and less on the vulnerability dimension (57, 58). It is particularly with respect to psychodynamic approaches that the interrelation of grandiosity and vulnerability becomes explained in terms of psychic functioning and that the narcissistic vulnerability is addressed as potentially increasing the tendencies to devalue and to "act out" onto others in NPD. The crucial symptom of narcissistic personalities is pronounced feelings of insufficiency and these feelings are compensated with fantasies of omnipotence and greatness. Persons with NPD – albeit being relatively unreliable in their empathic responses to others – are themselves highly dependent on how others see them due to their fears of social rejection and worries about threats of their social status (59).

What a psychoanalytic view on NPD exemplarily stresses is the reinterpretations of reality, which particularly manifest in misperceptions of both, the environment and own possibilities [cf. (60), 201ff; transl. KJA]. Erich Fromm – who frees himself from Freud's conceptualization of narcissism within the constricting frame of reference of the libido theory [(61), p. 37–74, 65ff.], accordingly defines narcissism as

"[...] an orientation in which all one's interest and passion are directed to one's own person: one's body, mind, feelings, interests, and so forth. (...) For the narcissistic person, only he and what concerns him are fully real; what is outside, what concerns others, is real only in a superficial sense of perception; that is to say, it is real for one's senses and for one's intellect. But *it is not real in a deeper sense, for our feeling or understanding*. He is, in fact, aware only of what is outside, inasmuch as it affects him. Hence, he has no love, no compassion, no rational, objective judgment. The narcissistic person has built an invisible wall around himself. He is everything, the world is nothing. Or rather: He is the world." [(62), p. 117].

Otto Kernberg who sees pathological narcissism proportionally increasing to the level of aggression mentions that the grandiose self of narcissists is a construct of all the positive and idealized characteristics of themselves and also of others into an unrealistic self-image. Devaluations of this image are split off or projected onto others (63). These psychic defense mechanisms, which serve for self-regulation and maintenance of the grandiose self-image, may be one reason why the grandiose type scores higher in life-satisfaction (64), as criticism of others is not perceived as a signal for self-assessment. On the contrary, a decline in self-esteem due to negative feedback (65), a record of unstable, superficial relationships (66), risk-taking and impulsive behaviors that significantly affect health (67) reveal the not so “happy face” of narcissism (68). With a closer look on empathy distortions as maladjustment, the *dissocial* tendencies associated with NPD have to be mentioned: Others are intentionally harmed in particular, if they are perceived as threatening, because they are scratching the self-ideal that determines the self-image of individuals with NPD. This is a reason for why NPD comes along with a relatively poor compliance to treatment, and what makes NPD one of the most difficult conditions to treat (69, 70) – for instance with respect to processes of countertransference (71) – and which often requires an adjustment of therapy and/or special treatment techniques for NPD (72).

The psychoanalyst Udo Rauchfleisch (60) elaborates the oscillation between the grandiose and vulnerable dimension in close relation to *dissocial behavior* in a variety of psychopathological conditions. His analysis is consistent with the previous mentioned motivational lack of empathic concern in NPD. His analysis allows to focus on the intrapsychic constellation in NPD as characterized by an “oppressive dictate of a hypertrophied ego ideal,” which demands the narcissist is unable to cope with. As a libido-economic consequence resulting feelings of insufficiency are concealed, which is often accompanied by a “diffuse anxiety” and “dysphoria,” and compensated for by impulsive actions [cf. (60), p. 201 ff.]. NPD consequently often includes harming others inasmuch as narcissistic agents engage in interpersonally exploitative behavior (often addressed as narcissistic rage) *because* they have these unrealistic expectations and hypertrophic demands that express in particular claims for loyalty, support, and admiration from others. The narcissistic maladaptive interpersonal functioning structurally corresponds with a “superego pathology (73)” on the intrapersonal functional level of psychic organization. Even if narcissists do not suffer from a maldevelopment of the super-ego as such, they exhibit an *integrational deficit* of the super-ego demands. If we apply this analysis [(60), p. 77ff] of the dissocial personality organization to NPD, the peculiarity of pathological narcissists is that their defense mechanisms are not mainly directed against aggressive and libidinous impulses of the id, but rather become directed against certain parts of the superego instance itself. The

mechanisms of projection and projective identification then play a decisive role: In the projection of the superego demands, for example, onto other individuals or external authorities, externalizing and splitting tendencies – e.g., the split between the exaggerated ego ideal and a negative self-representation “from within,” and that between the “totally good” and “totally bad” objects “outside” – can be maintained. In parallel, the second mechanism of identification (with a superego carrier) prevents the internalization of the conflict-ridden demands or a realistic self-assessment: Albeit the superego demands (e.g., the knowing that one should respond to the needs of others, act in accordance with norms that prohibit harming others, etc.) remain “outside,” these stay nevertheless effective inasmuch as they reappear as threats represented by other persons or institutions (74). As a consequence, narcissists tend to depreciate themselves (vulnerability) or reactivate the greater self (grandiosity). This leads not only to a stance of arrogance, or mere indifference, but to aggressive or otherwise dissocial behavior that is (overtly/covered) displayed toward others, in this case: particularly toward superego carriers.³

The psychodynamic explanation contributes to a deeper understanding of NPD as it resonates with the latest empirical results on the interrelatedness of grandiosity and vulnerability in NPD, and, moreover it conceptually specifies psychic functioning on the meso-level of description of the empathy deficiency: NPD involves a pathological superego constellation – a psychic disintegration – according to which the unemphatic behavior is causally explained within the framework of an alternative model of psychic functioning. This highlights the explanatory power of non-reductionist accounts of mental processing on the one hand, but simultaneously allows for an interdisciplinary integration, as the psychodynamic model stays conceptually open for being empirically further “grounded” in and/or testified against other functional descriptions of empathy distortions provided by neuroscientific explanations of empathy deficiencies, on the other hand. The latter, in contrast, rather tend to be reductionist in their explanations of empathy processing in NPD in terms of biological functioning, but this does not rule out an alternative understanding of the pathogen dynamics of psychic disintegration on a different explanatory level of empathy deficiencies in NPD. Moreover, psychodynamic approaches bring with them the surplus of an explicit *psychosomatic* understanding of pathology, although the liaison between *soma* and *psyche* has been a complicated one for the psychoanalytic discipline, too (75). Nota bene: For the purpose of my analysis it is not needed that the dynamics of the interplay of the psychic instances have to

3 “The adherence to the >>evil<< partial object, which appears again and again as a punishing superego in the outside world, becomes a necessity for the dissocial person, so that he can at least in this way maintain a certain self-identity.” Cf. (74), p. 82.

become one-to-one (re-)translated into the “neuroanatomical” language; but what is, indeed, important for an integrative perspective on the empathy deficit in NPD, is to understand the intra- and interpersonal sphere as *conceptually* inextricably intertwined functional units of the psychic apparatus. From a phenomenological perspective it is clear that empathy distortions affect both, intra- and interpersonal functioning, i.e., the self-*and*-world-relatedness of narcissistic individuals. This is conceptually grasped with the psychodynamic description of impaired psychic functioning. Consequently, *psychic* disintegration on the intrapersonal level conceptually relates to *social* disintegration described on the level of interpersonal functioning.

The philosophical description of deficient empathy in NPD

A third class of functional descriptions of the empathy deficit is provided by social-philosophical view, according to which NPD expresses as a distortion of the intersubjective practice of recognition (76). Empathy enables us to develop a relatively stable stance of an interested involvement with others: It provides an open space of experiential possibilities of relatedness. The philosopher Axel Honneth describes this as primordial to all kinds of objectifying modes of self-*and*-world-disclosure (77). In his discussion of Lukács’ concept of reification he makes a crucial assumption that is not only correct from a (brain-)developmentally perspective, but also true with regard to the sphere of intersubjective action in general: namely, that we – as the relational beings that we are – are always already affectively attuned to this world and are engaged in modes of interested participation in relation with others. This is the opposite to an objectifying mode of self-*and*-world-disclosure.⁴ The empathy deficit of individuals with NPD reveals in the tendency for a (pre-)intentional detachment, an inability to genuinely engage with others (as opposed to merely feigning concern for others in order to appear “social”). This is the socially impairing aspect of the empathy deficit in narcissism. If qualitative relations to others have a priority over objectifying relations to them, then the empathy deficit

hinders an interested involvement with others in terms of (inter-)subjective recognition. The narcissists’ empathy deficit implies the “active” forgetting of this priority (of the other): even if they register the needs of others, their pathological self-centeredness restricts the experiential possibility for empathic concern for others. Quite an opposite view on others comes with this particular mode of active forgetfulness of recognition: The stance of reifying others and to perceive them as mere means to an end. In NPD this can include also modes of “false” recognition (e.g., when others are not recognized for what they truly are, but become “reduced” to a certain function or property that can be valued for a certain purpose). When reification forms the rigid habitual pattern for “relatedness,” this enables the forms of *maladjustment*, that in its typical forms of narcissistic violence shows one significant feature: namely the exploitation and abuse of others (78–80). This is the explicit *dissocial* aspect of severe forms of pathological narcissism. The empathy deficit reveals as a lack of or as false recognition, which practically demolishes what we normally are taking for granted in our relations to others. In one’s daily encounters one is repeatedly confronted with conflicting emotions and commitments, and central to psychic health is the flexibility to cope with these in certain situations. The pathological core of narcissism consist in often not being able to literally feel and to adequately assess these situations as conflicting at all, and/or to readjust in practice, respectively. NPD then comes along with a significant restriction of experiential possibilities for *pro*-social relatedness, for *being-with* others. This can, of course, cause problematic social interactional patterns that are revealing the vulnerability of the interaction partners: one the one hand, people with NPD might feel misunderstood, themselves not appreciated, suspect others to be ill-willed, etc., while their interaction partners are disturbed by their self-centeredness and lack of awareness for the feelings and real needs of others, on the other. If NPD is not seen for what it is – a pathological condition that has socially impairing dimensions and yields a high vulnerability – this might foster stigmatization of patients with NPD.

As an interim conclusion it can be stated: The coherence of an integrative model of the empathy-deficit in NPD requires the structural consistency of the respective different functional explanations provided at the distinct descriptive levels (top down/bottom up). Insofar as the functional descriptions of the affective motivational deficit on the micro level (Section The somatic description of deficient empathy in NPD) is structurally consistent with the analysis of psychic disintegration on the meso-level (Section The psychodynamic description of deficient empathy in NPD), and this is consistent with the description of the empathy deficit as lack of recognition on the macro level (Section The philosophical description of deficient empathy in NPD) my analysis exemplarily shows the compatibility of different functionalistic descriptions within an integrative model of empathy deficit in NPD. In the following section some methodological considerations finally are addressed.

4 Honneth does not rule out that an objectifying stance is begun if it is conducted in a normative permissible manner. What he is targeting with his analysis, is a “forgetfulness of recognition” that happens in reification, and that has the potential to erode the very preconditions for an intersubjective practice based on respect and understanding, thus for an ethical form of social life: Honneth writes “[...]this kind of ‘forgetfulness of recognition’ can now be termed ‘reification.’ I thereby mean to indicate the process by which we lose the consciousness of the degree to which we owe our knowledge and cognition of other persons to an antecedent stance of empathetic engagement and recognition.” cf. (77). p. 52–63, 56.

Discussion: The core-deficit-hypothesis

Empathy deficit between *soma* and *psyche*

The biological approach to narcissism promises a concretization and naturalistic foundation by guaranteeing empirical objectivity, but as a result, might cast doubt on the notion of narcissism as a *mental* disorder. Reductive positions are therefore often associated with a “disqualification” of alternative explanations of mental processes, especially when these are seen as completely reducible to (or even “identifiable” with) physical processes. Instead of abolishing the conceptualization of narcissism as *mental* disorder, it seems much more reasonable to assume a correlation between mental processes and physiological processes *even if* the scientific convincingness of neurophysiological explanatory models might be already considered as providing some grounds for rejecting alternative models as equally reasonable for a conceptualization of mental (dys-)functioning in narcissism. Methodologically it is nevertheless still justified to speak of NPD in terms of *psychic* impairment, even if the interest in an fully objectively accessible “localization” of the narcissistic mind, e.g., in brain-organic explanatory models, would have been already fully satisfied. There have been, indeed, several neurophysiological foundations of pathological narcissism suggested (52, 81–83), as has been proposed also on a larger scale for other types [e.g., for the antisocial personality disorder (APD), or Borderline Personality Disorder (BPD) (84)] of the cluster B-personality disorders (85) with respect to altered empathy processing. Consequently, the clinical studies allow one to trace the characteristics of narcissist’s manifold “relational” problems back to significant changes of predominantly affective empathy [lesser to cognitive empathy (5)].

My analysis stresses the *conceptual* distinction of different descriptive levels of “empathy.” What reductionist approaches conceptually often fail to address is that *psychic functioning* cannot be fully deciphered solely in naturalistic terms – even if one *can* describe mental processing with respect to the *factum brutum* of empirical data provided by brain scans, saliva and blood samples, skin conduction and blood pressure tests, etc.⁵ The notion of ‘psychic functioning’ is, however, neither to be equated with the notion of ‘mental processing’, nor with the notion of ‘physiological functioning’, but rather mediates between both levels of descriptions, and therein serves as an independent category for describing empathy deficiencies in NPD. If this is the case, neither the parlance of the “mental” is fundamentally ruled out with my three-level analysis, nor does the integration of a reductionist view inevitably lead to

a relapse in some sort of “brain-mythology” when we speak of psychic (dys-)functioning. Moreover, a major distinction, namely between *explaining* and *understanding* (86, 87) should generally be kept in mind: The narcissistic brain is something to be explained, but the narcissistic mind is something we have to understand. Naturalistic views on narcissism literally allow to “emphasizes” a core deficit as a *pathology*, but an understanding of it – the *meaning* of empathy impairment – in NPD is provided by an evaluation against the normative backdrop of theories of psychic health, wellbeing, and (pro-)social relatedness. The latter keeps the phenomenal reality of NPD “in mind” from a live-worldly view, without forgetting the former “scientific project” of explaining narcissism in somatic terms within a naturalistic paradigm. As such, reductionist analyses are inevitable useful to objectify certain somatic changes in empathy responses and therein have the explanatory power for additionally empirically “backing” non-reductionist explanations of psychic impairments in NPD. What can be measured is at least neurophysiological reactions [e.g., stress responses (88)] to specific social situations, while it is due to the dynamics of re-enaction that these somatic changes manifest as (rigid) evaluative pattern for self-and-world-disclosure. This already implies an conceptual understanding of the *pathological situatedness* of NPD as irreducible to neurophysiological dynamics, but as ideally well-informed by them (89).

The empathy deficit between conceptual over-complexity and under-complexity

A second methodological doubt might arise with respect to either a *complexity-reduction* or, on the contrary, an *over-complexity* with the focus on empathy as the core deficit of NPD. Generally it is to be assumed that it is a variety of physiological mechanisms that determine the expression of different subtypes and degrees of abilities in and for personality disorders such as narcissism [(52), p. 2]. Moreover, it might be also especially the “plus” of comorbidity that explains *how* empathy-related impairments *exactly* realize in their particular forms in narcissism, and, moreover, in distinct types of Cluster-B-personality disorders. Skeptics could respectively label the core-deficit hypothesis as “naïvely” under-complex or over-complex, thus insufficient to address empathy as the conceptual core of NPD (or to even account for a “core” at all). Starting with the latter, it is obvious that empathy-(cor)related neurophysiological dysfunctions actually provide a very (if not *the* most) promising account for a specification of narcissism as a *disorder* (90–93) – we can objectify the *dys-* of function – and, as such, puts other explanatory models, which rely solely on evaluative criteria in their place. The core-empathy-deficit covers a large number of key symptoms of NPD. Not only the singular functional impairments of affective and cognitive empathy as such, but

⁵ For a systematic overview see: (82), p. 8–11 (Table 2).

also the discrepancy of different levels of this functional units together determine the core of NPD.

Secondly, in lights of the different functional roles empathy has not only for the explanation of NPD, but also for a range of other clinical conditions, this is not an objection against, but rather an argument for empathy as the core-deficit in NPD, because exactly this is the starting point for a more-fine-grained specification (e.g., on the molecular-biological level) for determining the *distinct* impact of empathy for NPD, and explicitly in comparison to other disorders. The core can be conceptually defined according to a differentiation of particular functional *patterns* of empathy-related dysfunctions for and within narcissism as *distinct* disorder type. In narcissism we have the interesting combination of relatively intact cognitive together with impaired affective empathy, and this pattern can be continued to be refined in comparison to other disorder types (such as APD, BPD). This might reduce conceptual over-complexity (here: exemplarily) for the descriptions of the empathy deficit in terms of biological dysfunctions.

Thirdly, under-complexity can be conceptually reduced with respect to the potential of the core-deficit to integrate different levels of description and disciplinary views (neurophysiological, psychological, sociological, philosophical, etc.) that respectively specify the meaning of the empathy deficiencies in NPD due to an interdisciplinary research objective. In order to obtain the conceptual consistency of the three-level analysis, a structural requirement is relying on functionalism (teleological, etiological, system-functional, and propensity functional descriptions of empathy). The explanations provided by different functional explanations then can become approved (testified against each other) with respect to their structural (in-)consistency for each and among different levels of description in my analysis. The higher the consistency, the more coherent is the particular explanatory structure (or pattern) of functional descriptions for empathy distortions in NPD. The basic structural integration must include in my analysis (top down/bottom-up) a reference to (1) biological dysfunctions, (2) intrapersonal impairments of psychic functioning, and (3) a distortion of social relatedness as it has been exemplarily sketched here.

Empathy deficit between mental disorder and social pathology

Finally, one could stipulate that the “personality” of a person is basically nothing that can be addressed in reductionist terms, or should be object of any medical assessment or diagnosis (even if we could objectify the underlying somatic dynamics), because it falls within the protective sphere of privacy, agency and personhood. Such positions could be carried out under the auspices of an (allegedly) pathologizing of narcissistic character traits, and/or even of a “moralization of diagnostics”, especially

when the “harmful” dimension of interpersonal difficulties of narcissism are highlighted with reference to empathy deficiencies. If a normative standard – for instance, norms of prosocial motivation – mutates into a clinically diagnostic yardstick for assessing individuals, this might be untenable from a scientific, (allegedly) value neutral point of view. Reminding on the debate about whether to keep narcissism as a clinical disorder category (94), some skeptics might consider narcissism as mere character “accentuation,” which is – however impairing or otherwise harmfully experienced these traits might be – no reason to suspect a mental disorder; especially not, when related behavioral styles are widely common, or even get promoted for their adaptive potential in certain fields of social practice (95, 96). In the context of methodological considerations of the psychiatric classification systems it can be explicitly pointed out that conflicts between society and the individual alone do not provide a sufficient basis for the attribution of a mental disorder. The classification manuals follow here an important intuition, which also owes itself to a confrontation with psychiatry-skeptical positions, when it is stated in the general definition of mental disorders:

“Whatever its original cause, it must currently be considered a manifestation of a behavioral, psychological, or biological dysfunction in a person. Neither deviant behavior (e.g., political, religious, or sexual) nor conflicts that are primarily between the individual and society are mental disorders unless the deviance or conflict is a symptom of a dysfunction in the individual.” [APA. DSM-IV-TR. (96). p. xxi-xxii.]

Although this theoretical limitation is intended to avoid defining mental disorder solely in terms of social deviance, it does not guarantee that misdiagnoses can always be avoided. These conflicts cannot, in my opinion, sufficiently justify a clinical diagnosis, but are admittedly important indications for a *differential* diagnosis in clinical practice. Moreover, the fact that narcissistic traits are apparently so widespread that they might even have become some standard social norm does not necessarily imply, that a pathology can be fundamentally ruled out (97). Exactly the opposite would have to be assumed, if one takes seriously, for example, studies on *social pathologies* (98, 99). With the psychoanalysts’ and sociologists’ Erich Fromm analysis on the *anatomy of human destructiveness* (100) one could state, that any trying to “normalize” NPD rather would indicate that something can be fundamentally wrong in such a society (or with certain institutions), as it fails to recognize the “pathology of normalcy” (101). At least this is the case, when there is an systemic (even institutionalized) indifference toward, or even a denial of “the intentional use of physical force or power, threatened or actual, against oneself, another person, or against a group or community, that either results in or has a high likelihood of resulting in injury, death, psychological harm, maldevelopment or deprivation” at play [(102), p. 1084,

104]. From a social-cultural diagnostic view provided by Critical Theory such strategies of normalization – in analogy to strategies of pathologizing – could both even be reframed as a signifier of a *second-order social pathology* [(29), p. 347] inasmuch as the second-order sense of an allegedly obviousness of first-order beliefs or normative assessments may contribute to the perpetuation of practice forms that are the relevant causal factors for reproducing these beliefs and assessments (e.g., the power of psychiatric diagnostic politics to declare certain phenomena as (non)-pathological). In severe forms of NPD the condition involves intentional harming of others, and altered empathy certainly contributes to it [excluded here the harm done by empathy induced *altruism* (103, 104)]. A diagnostic view on narcissism that frees itself from the assessment of the complex harm-dimension of NPD denies not only its clinical phenomenality from a live-worldly perspective, but also seems to ignore a scientific understanding of narcissism as *disorder* of interpersonal functioning, for instance, when the association between narcissism and aggression that has been empirically supported in adults and adolescents is denied, or when the particular meaning of harm as referring to individual suffering from vulnerability in NPD is not fully recognized (105). Nota bene: An assessment of actions is logically always different and has to be carefully discriminated from the assessment of personality from an objective clinical diagnosis, but non-trivial self-and other harming actions and behavioral styles must be at least reconsidered as correlated to empathy distortions in pathological narcissism. Considering non-trivial other-harming of additional diagnostic relevance for the diagnostics of NPD appears at least plausible with the focus on an empathy deficit as a causal factor for violence in narcissism, particularly when exactly this simultaneously can be understood in relation to narcissistic vulnerability, i.e., as an expression of social maladjustment due to an altered scope of experiential possibilities to empathically engage with others.

Conclusion

I have examined NPD from an conceptual perspective and focused on its core: the empathy-deficit. This has been reconceptualized with an integrational model that relates different functional descriptions provided by three structurally interrelated descriptive levels: The micro-level of biological dysfunctions, the meso-level of psychic impairment, and the macro-level of distortions of intersubjective practice, that together shape the interdisciplinary view on NPD in this

References

1. Reed-Knight B, Fischer S. Treatment of Narcissistic Personality Disorder Symptoms in a Dialectical Behavior Therapy Framework. In: Campbell W.K, Miller JD, editors. *The Handbook of Narcissism and*

analysis. Although my analysis is restricted in scope, I hope that I have provided some reasons to accept that an integrative approach toward empathy, which stresses on the ‘psyche’ as a mediating category, allows to bridge some trenches between the naturalistic explanation and normative understanding of empathy deficiencies in NPD.

Author contributions

KAJ is the contributing and corresponding author of this article.

Funding

This work was funded by the Center for Human Nature, Artificial Intelligence, and Neuroscience (CHAIN).

Acknowledgments

I would like to express my gratitude to the Frontiers editors and the editors of the special issue for inviting me to make a contribution. Special thanks to Matthias Kettner, to the participants of the panel Psychology and Vulnerability at the International Conference: Vulnerability - Theories and Concepts in Philosophy and the Social Sciences that took place in October 2022 at the University of Graz, to the reviewers for their thoughtful comments on a draft version of this paper, CHAIN, and Hokkaido University for funding this publication.

Conflict of interest

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

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Narcissistic Personality Disorder. Theoretical approaches, empirical findings, and Treatments. New Jersey: John Wiley & Sons. (2011). p. 466–484, 469ff.

2. Koepner T, Jauk E, Kanske P. Lay theories of grandiose and vulnerable narcissism. *Curr Psychol.* (2021). doi: 10.1007/s12144-020-01296-w
3. Pincus AL, Lukowitsky MR. Pathological narcissism and narcissistic personality disorder. *Annu Rev Clin Psychol.* (2010) 6:421–46. doi: 10.1146/annurev.clinpsy.121208.131215
4. Hopwood J, Wright AGC, Ansell EB, Pincus AL. The interpersonal core of personality disorder. *J Pers Disord.* (2013) 27:270–95. doi: 10.1521/pedi.2013.27.3.270
5. Wilson S, Stroud CB, Durbin CE. Interpersonal dysfunction in personality disorders: a meta-analytic review. *Psychol Bull.* (2017) 143:677–734. doi: 10.1037/bul0000101
6. Dickinson KA, Pincus AL. Interpersonal analysis of grandiose and vulnerable narcissism. *J personal disord.* (2003) 17:188–207. doi: 10.1521/pedi.17.3.188.22146
7. Jauk E, Weigle E, Lehmann K, Benedek M, Neubauer AC. The Relationship between Grandiose and Vulnerable (Hypersensitive) Narcissism. *Front. Psychol.* (2017) 8:1600. doi: 10.3389/fpsyg.2017.01600
8. Dawood S, Schroder HS, Donnellan MB, Pincus AL. Pathological Narcissism and Nonsuicidal Self-Injury. *J Pers Disord.* (2018) 32:87–108. doi: 10.1521/pedi.2017.31.291
9. Ponzoni S, Beomonte S, Zobel Rogier G, Velotti P. Emotion dysregulation acts in the relationship between vulnerable narcissism and suicidal ideation. *Scan J of Psych.* (2021) 62:468–75. doi: 10.1111/sjop.12730
10. Jauk E, Dieterich R. Addiction and the dark triad of personality disorder. *Front Psychiatry.* (2019) 17:662. doi: 10.3389/fpsy.2019.00662
11. Naidua ES, Patock-Peckham JA, Ruofa A, Bauman DC, Banovich P, Froeh T, Leamanb RF. Narcissism and devaluing others: An exploration of impaired control over drinking as a mediating mechanism of alcohol-related problems. *Pers Individ Differ.* (2019) 139:39–45. doi: 10.1016/j.paid.2018.10.039
12. Kanske P, Sharifi M, Smallwood J, Dziobek I, Singer T. Where the narcissistic mind wanders: increased self-related thoughts are more positive and future oriented. *J Pers Disord.* (2016) 31:1–24. doi: 10.1521/pedi.2016.30.263
13. Cain NM, Pincus AL, Ansell EB. Narcissism at the crossroads: phenotypic description of pathological narcissism across clinical theory, social/personality psychology, and psychiatric diagnosis. *Clin Psychol Rev.* (2008) 28:638–56. doi: 10.1016/j.cpr.2007.09.006
14. Hyatt CS, Sleep CE, Lamkin J, Maples-Keller JL, Sedikides C, Campbell WK, et al. Narcissism and self-esteem: a nomological network analysis. *PLoS ONE.* (2018) 13:e0201088. doi: 10.1371/journal.pone.0201088
15. Kaufman SB, Weiss B, Miller JD, Campbell WK. Clinical correlates of vulnerable and grandiose narcissism: a personality perspective. *J Pers Disord.* (2020) 34:107–30. p.2. doi: 10.1521/pedi.2018.32.384
16. Zabel L. *Narzisstische Depression: Theorien und Konzepte in Psychiatrie und Psychoanalyse.* Gießen: Psychosozial-Verlag. (2019). p. 93–114.
17. Blasco-Belled A, Rogoza R, Alsinet C. Vulnerable narcissism is related to the fear of being laughed at and to the joy of laughing at others. *Pers and Individ Differ.* (2022) 190:111. doi: 10.1016/j.paid.2022.111536
18. Weiss B, Campbell WK, Lynam DR, Miller JD. A trifurcated model of narcissism. On the pivotal role of trait antagonism. *The handbook of antagonism.* (2019). p. 221–35.
19. Krizan Z, Herlache AD. The narcissism spectrum model: a synthetic view of narcissistic personality disorder. *Person Soc Psychol Rev.* (2017) 22:3–31. doi: 10.1177/108868316685018
20. Alternative DMS-5 Model of Personality Disorder. *Focus.* Reprinted from American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders.* 5th ed Washington, DC: American Psychiatric Publishing. (2013). p. 189–203.
21. Funder DC, Harris MJ. On the several facets of personality assessment: The case of social acuity. *J Person.* (1986) 54:528–50. doi: 10.1111/j.1467-6494.1986.tb00411.x
22. Bender DS, Morey LC, Skodol AE. Toward a model for assessing level of personality functioning in DSM-5, part I: a review of theory and methods. *J Pers Assess.* (2011) 93:332–346. doi: 10.1080/00223891.2011.583808
23. Tyrer P, Mulder, R, Kim, JR, Crawford, MJ. The Development of the ICD-11 Classification of Personality Disorders: An Amalgam of Science, Pragmatism, and Politics. *Annu Rev Clin Psychol.* (2019) 7:481–502. doi: 10.1146/annurev-clinpsy-050718-095736
24. Skodol AE. Personality disorders in DSM-5. *Annu Rev Clin Psychol.* (2012) 8:317–44. doi: 10.1146/annurev-clinpsy-032511-143131
25. Stiez J, Jauk E, Krach S, Kanske, P. Dissociating empathy from perspective taking: Evidence from intra- and inter-individual differences research. *Front Psychiatry.* (2019) 10:126. doi: 10.3389/fpsy.2019.00126
26. Gordon RM. Sympathy, simulation and the impartial spectator. *Ethics.* (1995) 105:727–42. doi: 10.1086/293750
27. Preston SD, de Waal FBM. Empathy: its ultimate and proximate bases. *Behav Brain Sci.* (2002) 25:1–20; discussion: 20–72. doi: 10.1017/S0140525X02000018
28. Gu X, Hof PR, Friston KJ, Fan J. Anterior insular cortex and emotional awareness. *J Comp Neurol.* (2013) 521:3371–88. doi: 10.1002/cne.23368
29. Zurn C. Social Pathology as Second-Order Disorders. In: Petherbridge D, editor. *Axel Honneth: Critical Essays: With a Reply by Axel Honneth.* Leiden, The Netherlands: Brill Academic. (2011). p. 345–70, 54.
30. Decety J, Svetlova M. Putting together phylogenetic and ontogenetic perspectives on empathy. *Dev Cogn Neurosci.* (2012) 2:1–24. doi: 10.1016/j.dcn.2011.05.003
31. Eisenberg N, Fabes RA, Carlo G, Speer AL, Switzer G, Karbon M, et al. The relations of empathy-related emotions and maternal practices to children's comforting behavior. *J Exp Child Psychol.* (1993) 55:131–50. doi: 10.1006/jecp.1993.1007
32. Gutknecht L, Jacob C, Strobel A, Kriegebaum C, Müller J, Zeng Y, et al. Tryptophan hydroxylase-2 gene variation influences personality traits and disorders related to emotional dysregulation. *Intern J Neuropsychopharmacol.* (2007) 10:309–20. doi: 10.1017/S1461145706007437
33. Cloninger CR, Svrakic DM, Przybeck TR. A psychobiological model of temperament and character. *Arch Gen Psychiatry.* (1993) 50:975–90. doi: 10.1001/archpsyc.1993.01820240059008
34. Livesley WJ, Jang LK, Jackson DN, Vernon PA. Genetic and environmental contributions to dimensions of personality disorder. *Am J Psychiatry.* (1993) 150:1826–31. doi: 10.1176/ajp.150.12.1826
35. Vernon PA, Villani VC, Vickers LC, Harris JA. A behavioral genetic investigation of the Dark Triad and the big 5. *Pers Individ Dif.* (2008) 44:445–52. doi: 10.1016/j.paid.2007.09.007
36. Goldman AI. Empathy, mind, and morals. *Proc Addresses Am Philos Assoc.* (1992) 66:17–41. doi: 10.2307/3130659
37. Eisenberg N, Strayer J. *Empathy and its development.* Cambridge: Cambridge UP. (1988).
38. Fonagy, P, Gergely G, Jurist E, Target, M. *Affect Regulation, Mentalization and the Development of the Self.* New York: Other Press. (2002).
39. Rees C. The influence of emotional neglect on development. *Pediatrics Child Health.* (2008) 18:527–34. doi: 10.1016/j.paed.2008.09.003
40. Brummelman E, Thomaes S, Nelemans SA, Orobio de Castro B, Overbeek G, Bushman BJ. Origins of Narcissism in children. *Proc Natl Acad Sci.* (2015) 112:1–4. doi: 10.1073/pnas.1420870112
41. Kernberg P, Weiner A, Bardenstein K. *Personality disorders in children and adolescents.* New York: Basic Books. (2000)
42. Stone MH. Normal narcissism: an etiological and ethological perspective. In: Ronningstam, EF, editor. *Disorders of Narcissism: Diagnostic, Clinical, and Empirical Implications.* Washington, DC: American Psychiatric Press. (1997)
43. Horton RS, Trich T. Clarifying the links between grandiose narcissism and parenting. *J Psychol.* (2014) 148:133–43. doi: 10.1080/00223980.2012.752337
44. Kohut H. *Narzissmus. Eine Theorie der psychoanalytischen Behandlung narzisstischer Persönlichkeitsstörungen.* Frankfurt a. Main: Suhrkamp Taschenbuch Verlag (1973).
45. Manzano J, Palacio Espasa F, Zilkha N. The narcissistic scenarios of parenthood. *Int J Psychoanal.* (1999) 80:465–76. doi: 10.1516/0020757991598855
46. Elkind D. Instrumental narcissism in parents. *Bull Menninger Clin.* (1991) 55:299–307.
47. Eisenberg N, Eggum ND. Empathic responding: sympathy and personal distress. In: Decety J, Ickes W, editors. *The Social Neuroscience of Empathy.* Cambridge, MA: MIT Press. (2009).
48. Nichols S. *Sentimental Rules. On the Natural Foundations of Moral Judgement.* Oxford: Oxford UP. (2004).
49. Carpenter JM, Green MC, Vacharkulksemsuk T. Beyond perspective-taking: Mind-reading motivation. *Motiv Emot.* (2016) 40:358–74. doi: 10.1007/s11031-016-9544-z
50. Pincus AL, Ansell EB, Pimentel CA, Cain NM, Wright AGC, Levy KN. Initial construction and validation of the Pathological Narcissism Inventory. *Psychol Assessm.* (2009) 21:365–79. doi: 10.1037/a0016530

51. Böckler A, Sharifi M, Kanske P, Dziobek I, Singer T. Social decision making in narcissism: Reduced generosity and increased retaliation are driven by alterations in perspective-taking and anger. *Pers Individ Differ.* (2017) 104:1–7. doi: 10.1016/j.paid.2016.07.020
52. Baskin-Sommers A, Krusemark E, Ronningstam E. Empathy in narcissistic personality disorder: from clinical and empirical perspectives. *Pers Disord.* (2014) 5:323–33. doi: 10.1037/per0000061
53. Hepper EG, Hart CM, Sedikides C. Moving Narcissus: can narcissists be empathic? *Pers and Soc Psychol Bull.* (2014) 40:1079–91. doi: 10.1177/0146167214535812
54. Stone M. Normal narcissism: an etiological and ethological perspective, 7–28. In: Ronningstam E, editor. *Disorders of Narcissism: Diagnostic, Clinical and Empirical Implications.* Washington, DC: American Psychiatric Press. (1998).
55. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders. Text Revision (DSM-IV-TRR), 4th Edn.* Washington, DC: American Psychiatric Association (1994).
56. Kauten R, Barry CT. Do you think I'm as kind as I do? The relation of adolescent narcissism with self- and peer-perceptions of prosocial and aggressive behavior. *Pers Individ Differ.* (2014) 61:69–73. doi: 10.1016/j.paid.2014.01.014
57. Karterud, S, Øien M, Pedersen G. Validity aspects of the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, narcissistic personality disorder construct. *Compr Psychiatry.* (2011) 52:517–26. doi: 10.1016/j.comppsy.2010.11.001
58. Ritter K, Dziobek I, Preissler S, Rüter A, Vater A, Fydrich T, et al. Lack of empathy in patients with narcissistic personality disorder. *Psychiatry Res.* (2011) 187:241–7. doi: 10.1016/j.psychres.2010.09.013
59. Ronningstam E, Baskin-Sommers AR. Fear and decision-making in narcissistic personality disorder—a link between psychoanalysis and neuroscience. *Dialogues Clin Neurosci.* (2013) 15:191–201. doi: 10.13187/DCNS.2013.15.2/ronningstam
60. Rauchfleisch U. *Dissozial. Entwicklung, Struktur und Psychodynamik dissozialer Persönlichkeiten.* Göttingen: Vandenhoeck & Ruprecht. (1981).
61. Freud S. Zur Einführung in den Narzissmus. Jahrbuch der Psychoanalyse; Leipzig Bd. In: *Sigmund Freud Studienausgabe: Psychologie des Unbewussten. III.* Frankfurt a. Main: Fischer Taschenbuch Verlag (2000).
62. Fromm E. *The art of being.* 1989 (1974–75). In: Funk R, editor. New York, NY: Continuum (1993).
63. Kernberg OF. *Aggressivity, Narcissism, and Self-Destructiveness in the Psychotherapeutic Relationship: New Developments in the Psychopathology and Psychotherapy of Severe Personality Disorders.* Yale: Yale UP. (2014).
64. Demirci I, Eksi H, Ekşi F. Narcissism, life satisfaction, and harmony: the mediating role of self-esteem and self-compassion. *Eur J Educ Res.* (2019) 84:159–78. doi: 10.14689/ejer.2019.84.8
65. For a longitudinal examining relations between narcissism and several indicators of well-being in a non-clinical population see: Zuckerman M, O'Loughlin, RE. Narcissism and well-being: A longitudinal perspective. *Eur J Soc Psychol.* (2009) 39:957–72. doi: 10.1002/ejsp.594
66. Wurst SN, Gerlach TM, Dufner M, Rauthmann, JF, Grosz MP, et al. Narcissism and romantic relationships: the differential impact of narcissistic admiration and rivalry. *J Pers Soc Psychol.* (2017) 112:280–306. doi: 10.1037/pspp0000113
67. Konrath S, Bonadonna JP. Physiological and health-related correlates of the narcissistic personality. In: Besser A, editor. *Psychology of Narcissism.* Hauppauge, NY: Nova Science Pub Inc (2014).
68. Rose P. The happy and unhappy faces of narcissism. *Pers Individ Differ.* (2002) 33:379–92. doi: 10.1016/S0191-8869(01)00162-3
69. Olsson PA. Psychodrama and the treatment of narcissistic and borderline patients. *Psychodyn Psychiatry.* (2018) 46:252–64. doi: 10.1521/pdps.2018.46.2.252
70. Tanzilli A, Colli L, A, Muzi Lingiardi V. Clinician emotional response toward narcissistic patients: a preliminary report. *Res Psychother Psychopathol Process Outcome.* (2015)181:1–9. doi: 10.4081/ripppo.20.15.174
71. Tanzilli A, Muzi L, Ronningstam E, Lingiardi V. Countertransference when working with narcissistic personality disorder: an empirical investigation. *Psychotherapy.* (2017) 54:184–94. doi: 10.1037/pst0000111
72. Crisp H, Gabbard GO. Principles of Psychodynamic Treatment for Patients With Narcissistic Personality Disorder. *J Pers Disord.* (2020)34 (Suppl):143–58. doi: 10.1521/pedi.2020.34.supp.143
73. Jung CG. Das Gewissen in psychologischer Sicht. (1958). In: Petrilowitsch N, editor. *Gesammelte Werke von C. G. Jung.* Darmstadt: Wissenschaftliche Buchgesellschaft (1981). p. 825–57.
74. Rauchfleisch U. *Außenseiter der Gesellschaft. Psychodynamik und Möglichkeiten zur Psychotherapie Straffälliger.* Göttingen: Vandenhoeck & Ruprecht. (1999).
75. Köpp W, Deter HC. Psychoanalyse und Psychosomatik Anmerkung zur Geschichte einer schwierigen Beziehung (Psychoanalysis and psychosomatics—Notes about the history of a difficult relationship). *Forum Psychoanal.* (2006) 22:297–306. doi: 10.1007/s00451-006-0291-8
76. Honneth A. Der Kampf um Anerkennung: Zur moralischen Grammatik sozialer Konflikte. Frankfurt a.M. Suhrkamp, 1992. *English translated by J. Anderson as The Struggle for Recognition: The Moral Grammar of Social Conflicts.* Cambridge: Polity Press. (2005).
77. Honneth A. Reification and Recognition. In: Martin J, editor. *Reification: A New Look at an Old Idea.* Edited and introduced by Martin Jay. Oxford: Oxford UP. (2018). P. 17–94, 40–52. Originally delivered as the Tanner Lecture in Human Values. Berkeley, CA: Spring. (2005).
78. Barbieri C, Grattagliano I, Catenesi R. Some reflections on the so-called narcissistic crime. *Bassegna Italiana di Criminologica.* (2019) 13:257–67. doi: 10.7347/RIC-042019-p257
79. Hepper EG, Hart CM, Meek R, Cisek SZ, Sedikides C. Narcissism and empathy in young offenders and non-offenders. *Eur J Pers.* (2014) 28:201–10. doi: 10.1002/per.1939
80. Matherne III CF, Credo KR, Gresch EB, Lanier PA. Exploring the relationship between covert narcissism and morality: the mediating influences of self-efficacy and psychological exploring the relationship between covert narcissism and morality. *Am J Manage.* (2019) 19:31–9. doi: 10.33423/ajm.v19i5.2627
81. Decety J, Moriguchi Y. The empathic brain and its dysfunction in psychiatric populations: Implications for intervention across different clinical conditions. *Bio Psycho Social Med.* (2007) 1:1–49. doi: 10.1186/1751-0759-1-22
82. Jauk E, Kanske P. Can neuroscience help to understand narcissism? A systematic review of an emerging field. *Personal Neurosci.* (2021) 4:1–29. doi: 10.1017/pen.2021.1
83. Watson PJ, Grisham SO, Trotter MV, Biderman MD. Narcissism and empathy: validity evidence for the narcissistic personality inventory. *J Pers Assess.* (1984) 48:301–5. doi: 10.1207/s15327752jpa4803_12
84. Blair J, Mitchell D, Blair K. *The Psychopath—Emotion and the Brain.* Malden, MA: Blackwell Publishing. (2005).
85. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders.* 5th edition. Washington, DC: American Psychiatric Association. Arlington, VA: American Psychiatric Publishing. (2013).
86. Dilthey W. *Gesammelte Schriften.* Leipzig: B.G. Teubner. (1923).
87. Boehm R. “Erklären” und “Verstehen” bei Dilthey. *Zeitschrift für philosophische Forschung.* Bd. 5, H. 3. publ. Frankfurt a. Main: Vittorio Klostermann GmbH. (1951). p. 410–17.
88. Coleman SRM, Pincus AL, Smyth JM. Narcissism and stress-reactivity: a biobehavioural health perspective. *Health Psychol Rev.* (2019) 13:35–73. doi: 10.1080/17437199.2018.1547118
89. Jacobs KA. The depressive situation. *Front Psychol.* (2013) 17:429. doi: 10.3389/fpsyg.2013.00429
90. Wakefield JC. The concept of mental disorder: on the boundary between biological facts and social values. *Am Psychol.* (1992) 47:373–88. doi: 10.1037/0003-066X.47.3.373
91. Wakefield JC. Disorder as harmful dysfunction: a conceptual critique of DSM-III-R's definition of mental disorder. *Psychol Rev.* (1992) 99:232–47. doi: 10.1037/0033-295X.99.2.232
92. Wakefield JC. Diagnosing the DSM-IV, Part 1: DSM-IV and the concept of mental disorder. *Behav Res Ther.* (1997) 35:633–50. doi: 10.1016/S0005-7967(97)00018-1
93. Wakefield JC. When is development disordered? Developmental psychopathology and the harmful dysfunction analysis of mental disorder. *Behav Res Ther.* (1997) 35:633–50.
94. Joshua D, Miller JD, Widiger TA, Campbell WK. Narcissistic personality disorder and the DSM-V. *J Abnorm Psychol.* (2010) 119:640–9. doi: 10.1037/a0019529
95. Dutton K. *Psychopathen – Was man von Heiligen, Anwälten und Serienmördern lernen kann.* München: dtv Verlagsgesellschaft mbH & Co. KG (2013).

96. Dutton K, McNab A. *Der gute Psychopath in dir – Entdecke deine verborgenen Stärken!* Frankfurt am Main: Fischer Verlag. (2015).
97. Maaz HJ. *Die narzisstische Gesellschaft*. München: CH Beck. (2013).
98. Honneth A. Pathologies of the social: the past and present of social philosophy. In: Rasmussen DM, editor: *Handbook of Critical Theory*. Cambridge/MA: Blackwell. (1996).
99. Jacobs KA, Kettner M. Zur Theorie “sozialer Pathologien” bei Freud, Fromm, Habermas und Honneth. In: Clemenz M, Zitko H, Büchsel M, Pflichthofer D, editors. *IMAGO. Interdisziplinäres Jahrbuch für Psychoanalyse und Ästhetik, Band 4*. Gießen: Psychosozial-Verlag (2017). p. 119–46.
100. Fromm E. Aggression und Narzissmus, 225–230. In: Fromm E. *Anatomie der menschlichen Destruktivität*. See also: Fromm E. *Anatomie der menschlichen Destruktivität*. Hamburg: Open Publishing Rights GmbH. (2015).
101. Fromm E. *The Pathology of Normalcy*. Its Genius for Good and Evil. Riverdale, New York: AMHF. (2010[1991]).
102. Krug E, Mercy JA, Dahlberg LL, Zwi AB. World report on violence and health. *Lancet*. (2002) 369:1083–8. doi: 10.1016/S0140-6736(02)11133-0
103. Batson CD, Klein TR, Highberger L, Shaw LL. Immorality from empathy-induced altruism: when compassion and justice conflict. *J Pers Soc Psychol*. (1995) 68:1042–54. doi: 10.1037/0022-3514.68.6.1042
104. Batson CD, Duncan B, Ackerman P, Birch K. Is empathic emotion a source of altruistic motivation? *J Pers Soc Psychol*. (1981) 40:290–302. doi: 10.1037/0022-3514.40.2.290
105. Sophie L, Kjærvik SL, Bushman BJ. The link between narcissism and aggression: a meta-analytic review. *Psychol Bull*. (2021) 147:477–503. doi: 10.1037/bul0000323



OPEN ACCESS

EDITED BY

Francesca Strappini,
Sapienza University of Rome,
Italy

REVIEWED BY

William Bechtel,
University of California,
San Diego,
United States
Mark Couch,
Seton Hall University,
United States

*CORRESPONDENCE

David Parker
✉ Djp27@cam.ac.uk

SPECIALTY SECTION

This article was submitted to
Consciousness Research,
a section of the journal
Frontiers in Psychology

RECEIVED 05 July 2022

ACCEPTED 28 November 2022

PUBLISHED 22 December 2022

CITATION

Parker D (2022) Neurobiological reduction:
From cellular explanations of behavior to
interventions.
Front. Psychol. 13:987101.
doi: 10.3389/fpsyg.2022.987101

COPYRIGHT

© 2022 Parker. This is an open-access
article distributed under the terms of the
[Creative Commons Attribution License](#)
(CC BY). The use, distribution or
reproduction in other forums is permitted,
provided the original author(s) and the
copyright owner(s) are credited and that
the original publication in this journal is
cited, in accordance with accepted
academic practice. No use, distribution or
reproduction is permitted which does not
comply with these terms.

Neurobiological reduction: From cellular explanations of behavior to interventions

David Parker*

Department of Physiology, Development and Neuroscience, University of Cambridge, Cambridge, United Kingdom

Scientific reductionism, the view that higher level functions can be explained by properties at some lower-level or levels, has been an assumption of nervous system analyses since the acceptance of the neuron doctrine in the late 19th century, and became a dominant experimental approach with the development of intracellular recording techniques in the mid-20th century. Subsequent refinements of electrophysiological approaches and the continual development of molecular and genetic techniques have promoted a focus on molecular and cellular mechanisms in experimental analyses and explanations of sensory, motor, and cognitive functions. Reductionist assumptions have also influenced our views of the etiology and treatment of psychopathologies, and have more recently led to claims that we can, or even should, pharmacologically enhance the normal brain. Reductionism remains an area of active debate in the philosophy of science. In neuroscience and psychology, the debate typically focuses on the mind-brain question and the mechanisms of cognition, and how or if they can be explained in neurobiological terms. However, these debates are affected by the complexity of the phenomena being considered and the difficulty of obtaining the necessary neurobiological detail. We can instead ask whether features identified in neurobiological analyses of simpler aspects in simpler nervous systems support current molecular and cellular approaches to explaining systems or behaviors. While my view is that they do not, this does not invite the opposing view prevalent in dichotomous thinking that molecular and cellular detail is irrelevant and we should focus on computations or representations. We instead need to consider how to address the long-standing dilemma of how a nervous system that ostensibly functions through discrete cell to cell communication can generate population effects across multiple spatial and temporal scales to generate behavior.

KEYWORDS

reductionism, psychiatry, neuroeducation, cognitive enhancement, volume transmission, ephapse, neuron doctrine

Introduction

There is extensive debate on reductionism in the philosophy of science (Van Riel and Van Gulick, 2019), and in psychology and neuroscience (Selverston, 1980; Barlow, 1990; Gold and Stoljar, 1999; Endicott, 2001; Bickle, 2003; Bechtel, 2007; Craver, 2007; Parker, 2010; Krakauer et al., 2017). These debates consider whether one field can be eliminated by reducing it to another, and if and how component properties relate to mechanistic explanations (even the definition of mechanism is debated; see Silberstein and Chemero, 2013). These debates have continued for decades and show no sign of ending (Ingo and Love, 2022), which questions whether definitive answers are likely. Resistance from fields to being eliminated by those below is to be expected: psychology resists the claim it is a placeholder science that will be eliminated once the physiology of the brain is understood, and physiology the claim that it can be reduced to molecular biology (Noble and Boyd, 1993). There may be some professional defense in this resistance, but it is right to question what a reductionist approach can offer.

Reductionism is not a unitary phenomenon (Ingo and Love, 2022). The biologist Ernst Mayr defined three types (Mayr, 1988): constitutive reduction (functions reflect their underlying parts and their properties); explanatory reduction (mechanisms can be explained from their constitutive details); and intertheoretical reduction (a theory can be reduced to another, more inclusive theory, e.g., psychology to neurophysiology, and neurophysiology to molecular biology; Noble and Boyd, 1993; Bickle, 2003). Mayr considered constitutive reduction the simplest and least controversial, while explanatory and intertheoretical reduction were more contentious.

While debate in the philosophy of science has traditionally focused on intertheoretical reduction, constitutive, and explanatory reductionism have come to the fore (Ingo and Love, 2022). These forms of reduction have been related to a mechanical or “machine model” where outputs are generated by parts that perform specific functions. Ingo and Love (2022) wrote, “mechanisms are understood as akin (though not equivalent) to machines with interconnected, organized parts operating to produce regular or expected outcomes,” following Alberts (1998) who compared a cell to a factory where specific functions are performed sequentially along chains of protein machines (see also Reynolds, 2007). Hanahan and Weinberg (2000, p. 67) updated the machine analogy by claiming, “Two decades from now...it will be possible to lay out the complete integrated circuit of the cell... we will then be able to apply the tools of mathematical analysis to explain.” The integrated circuit analogy is notable as Jonas and Kording (2017) have shown that current reductive approaches applied to actual integrated circuits fail to explain their function. Hanahan and Weinberg’s two decades have now passed, but instead of a complete integrated circuit features have been identified that negate the integrated circuit analogy. These include a fluid cytoskeleton, “intrinsically disordered proteins,” enzymes with numerous substrates or that perform non-enzymatic

functions, pleomorphic molecular assemblies with “probability clouds” of interactions, and probabilistic gene expression (see Nicholson, 2019). These aspects do not negate reductionist approaches in principle but show that previous assumptions and metaphors were simplistic.

Broadly speaking, given the controversy over definitions (Ingo and Love, 2022), psychoneural, or neurobiological reduction sees psychology and behavior explained mechanistically in terms of the constituent molecules and cells, and can include some combination of constitutive, explanatory, and intertheoretical reduction. Constitutive and explanatory reduction has been dominant aspects in neurobiology for several decades (e.g., Selverston, 1980; Gettings, 1989; Ito, 2006; Yuste, 2008). These analyses have led to significant insight into cellular and synaptic properties. Some have claimed causal explanations of behavior from these analyses (see Parker, 2006, 2019 for examples and critique), and where gaps in mechanistic schemes are acknowledged it is assumed that reductive approaches will ultimately be successful. For example, in reviewing the link between the long-term potentiation (LTP) of hippocampal synapses and memory, Bliss et al. (2018, p. A105) admitted “definitive proof that the mechanisms of LTP subserve learning and memory in the behaving animal is still lacking,”...but they went on to say that “few neuroscientists doubt that such proof will eventually be forthcoming.”

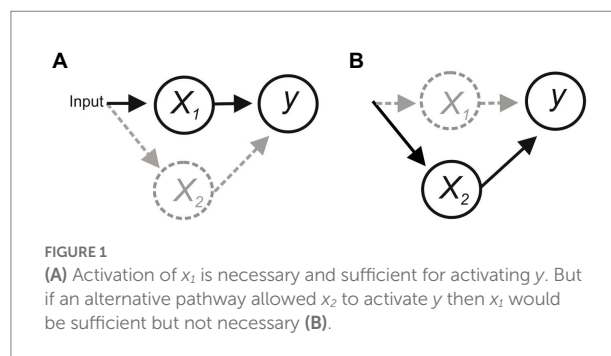
To consider constitutive and explanatory reduction in neurobiology, I will start with the basic issue of experimentally identifying component neurons, and then consider how the organization of neurons in neural circuits affects our ability to offer reductive or mechanistic explanations. I will finish by considering claims that our mechanistic knowledge of the nervous system obtained in reductionist analyses is sufficient to be translated into practical uses. These claims extend beyond interventions in traditional areas like neurology and psychopathology to include aspects of normal cognition and behavior, with some claiming that not only can we safely and effectively intervene in the normal brain, but also that we should.

The identification of components and their roles

The reductionist belief that molecular and cellular properties underlie cognition and behavior has been called the neuron doctrine (e.g., Barlow, 1972; Gold and Stoljar, 1999). This doctrine takes different forms with different implications: the trivial neuron doctrine sees psychological explanations remaining autonomous despite being implemented by neuronal properties, while the radical doctrine sees psychological aspects explained by neuronal properties (see Gold and Stoljar, 1999). The term neuron doctrine originated at the end of the 19th century with acceptance that the brain is made of discrete cells rather than being a continuous reticulum (Shepherd, 1991). This became an experimental focus in the 1950s with the development of techniques for intracellular

recordings from single cells (Bickle and Parker, 2022) and is referenced in the terms like command neuron, place cell, grandmother cell, gnostic unit, and feature detector (e.g., Barlow, 1972; Hubel, 1974; Edelman, 1989; Zeki, 1993; Crick, 1994; Changeux, 1997; Kandel, 1998). Gold and Stoljar (1999, p. 2–3) quote several philosophers and neuroscientists who claim that nervous system functions can be explained from cellular components. For example, Churchland and Sejnowski claimed “it is highly improbable that emergent properties cannot be explained by low-level properties”; Semir Zeki wrote that “It is only through a knowledge of neurobiology that philosophers of the future can hope to make any substantial contribution to understanding the mind”; Gerald Edelman said a theory of the brain needs “a description based on the neuronal and phenotypic organization... formulated solely in terms of physical and chemical mechanisms giving rise to that organization”; and Francis Crick that “A person’s mental activities are entirely due to the behavior of nerve cells, glial cells, and the atoms, ions, and molecules that make them up.” Crick made the definitive reductionist statement, “All approaches at a higher level are suspect until confirmed at the molecular level” (Crick, 1988, p. 61). These views suggest that once all the relevant component molecules, cells, and interactions have been characterized we will understand function, a neuroscience version of Laplace’s demon (Laplace, 1902).

The experimental criteria claimed for a reductive explanation of behavior in neurobiology have been outlined several times (they have been repeated, albeit using different terms, by philosophers of neuroscience in their discussions of various reductive approaches; Ingo and Love, 2022). Neurobiological criteria reflect the need to identify the component neurons involved in a behavior, their direct synaptic connections, and the functional properties of specific classes of neurons and synapses; this information is ultimately integrated to try to provide a unified explanation of the behavior (e.g., Bullock, 1976; Selverston, 1980; Gettling, 1989; Yuste, 2008; Braganza and Beck, 2018). These criteria have been applied to experimental analyses in invertebrate and vertebrate nervous systems, traditionally using electrophysiological and anatomical techniques and now also using various imaging, molecular genetics, and optogenetic approaches. They remain a major focus of neuroscience research and tool development. Thus, the US BRAIN initiative explicitly aims to develop new tools for reductionist analyses, stating “By accelerating the development and application of innovative technologies, researchers will be able to produce a revolutionary new dynamic picture of the brain that, for the first time, shows how individual cells and complex neural circuits interact in both time and space.”¹ These analyses and their assumptions have also influenced our views of psychopathology, with aberrant functions being considered to reflect genes, neurotransmitters, and other signaling molecules that can be targeted in interventions: thus, the



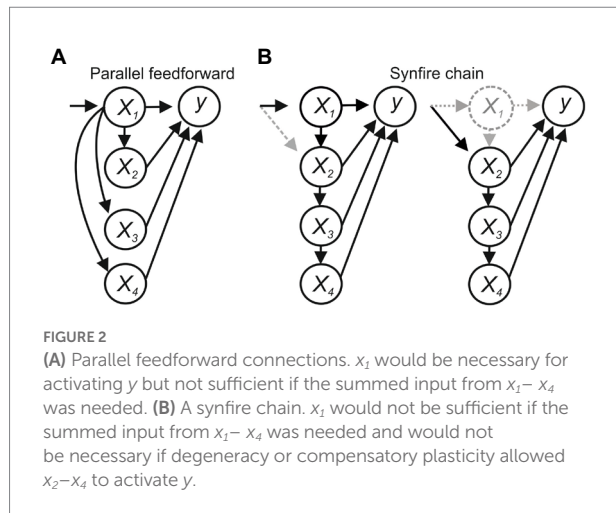
BRAIN initiative promises “new ways to treat, cure, and even prevent brain disorders.”¹

Much of the debate around reductionist approaches assume that we can obtain or have obtained the necessary component detail, debate focusing on what this data can explain. But even the correct identification of component cells, a crucial step in a mechanistic explanation, is far from trivial (Selverston, 1980; Parker, 2006). Components, either molecules, cells, or brain regions, underlying different functions have been identified using the criteria of necessity and sufficiency. This traditionally used lesions, electrical stimulation, or pharmacological activation or inhibition, and now includes molecular genetic and optogenetic loss and gain of function approaches, with outputs assessed from behavior or by imaging or recording from neurons or brain regions. A necessary condition must be present for an effect to occur, shown by the correlated activity of a component with an effect and the absence of the effect when the component is silenced; sufficiency is shown when the activation of a component can evoke the effect. While these criteria have been used experimentally for many years, there are long-standing issues with them. These have been discussed in the context of simpler nervous systems where direct links, or the lack of them, are easier to examine (e.g., see debate over the command neuron concept in Kupfermann and Weiss, 1978).

Consider the scheme in Figure 1. In (a), x_1 is the only functional connection onto y , and activity of y is evoked/abolished when x_1 is activated/inactivated, suggesting that x_1 is necessary and sufficient for y . However, degeneracy (i.e., different components can perform the same function; Tononi et al., 1999) or compensatory plasticity that can rapidly adapt to a perturbation to maintain function (Davis and Bezprozvanny, 2001; Frank et al., 2006) could allow x_2 to substitute for x_1 , making x_1 sufficient but not necessary for evoking y (b).

Feedforward connections between x_{1-4} introduce additional issues. In Figure 2A, x_1 sends parallel feedforward projections to x_2 – x_4 , which sum to activate y . Activating x_1 will evoke y , and blocking x_1 will block y , suggesting x_1 is necessary and sufficient for y , but x_1 would not be sufficient if the summed input from x_1 – x_4 was needed to activate y . In a synfire-like chain (Figure 2B), activating x_1 will evoke y and inhibiting x_1 will block y , again suggesting x_1 is necessary and sufficient. But x_1 may again not be sufficient if the summed input from x_1 to x_4 was needed, and in

¹ <https://braininitiative.nih.gov/>



this case would also not be necessary if degeneracy or compensatory effects allowed x_2 to recruit x_3-x_4 to evoke y .

Feedback connections add further issues. Figure 3 shows effects in a simple computer simulation (Jia and Parker, 2016). Here, x_1 sends parallel excitatory inputs to output neurons y_1 and y_2 , and to interneuron x_2 . Assume y_1 generates the output underlying a behavior we are investigating, and we positively and negatively manipulate x_2 to test the hypothesis that it inhibits y_1 . Without feedback connections (Figure 3A), activating or inactivating x_2 reduces or increases y_1 activity, respectively, consistent with the hypothesized inhibitory role of x_2 (albeit subject to the provisos of degeneracy and compensation outlined above). However, with feedback excitation from y_1 to x_1 (Figure 3B), removing x_2 will increase y_1 activity, as hypothesized, but increasing x_2 activity will cause oscillation rather than inhibition because (1) increased inhibition from x_2 reduces y_1 activity; which (2) reduces feedback excitation of x_1 ; which (3) reduces x_2 activation and disinhibits y_1 ; (4) this increases y_1 activity and thus feedback excitation of x_1 and x_2 activity to reduce y_1 activity; and (5): the cycle repeating to cause oscillation. With feedback inhibition from y_1 to x_1 (Figure 3C), removing x_2 will increase y_1 activity, but as this inhibits x_1 the excitatory drive to y_1 and y_1 feedback inhibition of x_1 will be reduced, again causing oscillation in y_1 as x_1 activity increases and decreases. Finally, as x_1 connects to y_2 , any changes in x_1 will alter y_2 , even though neither x_1 nor y_2 is directly affected by x_2 and y_2 has no role in the function. This is an example of “diaschisis” (Carrera and Tononi, 2014; Otchy et al., 2015), a neurological term seemingly less appreciated experimentally that means “shocked throughout” to represent the widespread system changes evoked by even very precise manipulations of system components.

Changes in y_1 thus occur that are not predicted from manipulation of x_2 . An added issue is that if x_2 directly affects y_2 then diaschisis could also result in y_1 activity being unaffected despite widespread changes in functionally relevant system components. In Figure 3D, x_2 inhibits both y_1 and y_2 , and y_2 provides feedforward inhibition of y_1 . Removing x_2 inhibition will

increase y_1 activity, but it will also increase y_2 activity through disinhibition. This could evoke inhibition of y_1 that leaves y_1 activity unchanged, a negative result that could erroneously suggest no influence of x_2 in the circuit.

Degeneracy, compensatory plasticity, diaschisis, and feedforward and feedback connections, all established aspects of nervous systems, can thus complicate interpretations of even totally precise and controlled manipulations of component parts (note that even the most advanced molecular techniques are promiscuous and can affect more than the intended target (Newton et al., 2019), negating the “surgical” analogy that they allow molecular dissection of circuits; Kiehn and Kullander, 2004). Misinterpretations can lead to the erroneous inclusion or omission of components in mechanistic schemes, with obvious consequences to claimed explanations, understanding, and interventions. We could claim that with sufficient (Laplacian?) knowledge these issues would be recognized and correct explanations would be provided, but while easily seen in these cartoon examples could we readily identify these features in more complex circuits? The practical and conceptual challenges of reductive approaches that link component parts to functions have been highlighted several times in invertebrate and lower vertebrate nervous systems containing relatively few, often large and uniquely identifiable cells (Kupfermann and Weiss, 1978; Selverston, 1980, 2010; Getting, 1989; Parker, 2010). These features should make the linking of components to functions easier than analyses of cognition in mammals, but even in these systems (tellingly referred to as “simpler” rather than “simple”) errors have been made and gaps remain in explanatory schemes after decades of analysis (e.g., Selverston, 1980, 2010; Parker, 2006).

Relational aspects in reductionist schemes

Identifying components is only the first step in a mechanistic explanation. Neurobiological criteria for reductive explanations highlight the requirement of knowing how component cells are synaptically connected in a system organization or architecture, and the functional properties of the cells and synapses that allow them to perform their functions. Crick (1994, p. 3) claims that “your joys and your sorrows, your memories and ambitions, your sense of personal identity, and freewill, are in fact *no more than* [my italic] the behavior of a vast assembly of nerve cells,” downplays the importance of the assembly. For example, sodium channel function does not simply reflect a vast assembly of molecules but requires cooperativity between appropriately arranged parts (e.g., voltage-sensitive S4 regions; Marban et al., 1998). This reflects the folding of the channel polypeptide chain, which depends on the amino acid sequence, interactions between amino acids, extrinsic factors (“chaperone” proteins), and the physico-chemical properties of the environment (hydrophobic amino acids orientate internally), with channel function ultimately reflecting the properties of the whole cell (e.g., voltage and

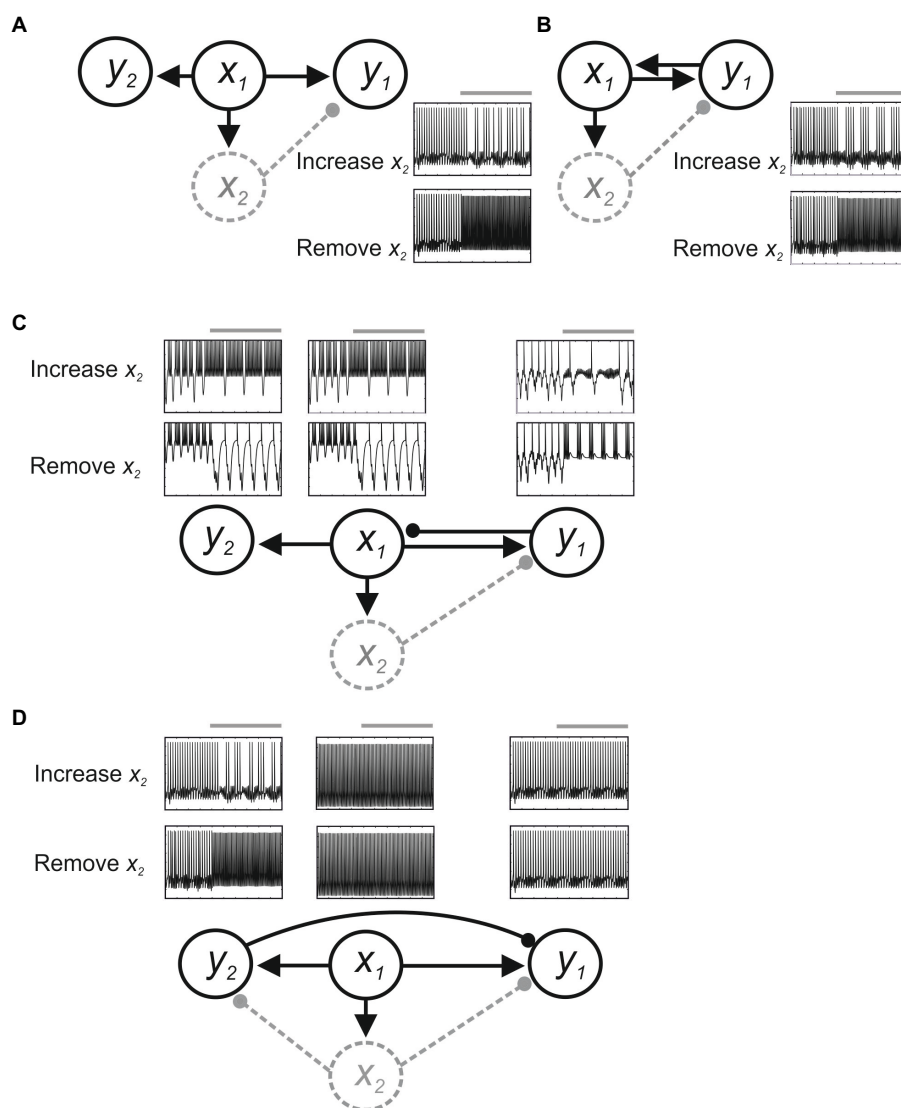


FIGURE 3

Feedback effects in a simple model. Neurons are modeled using Hodgkin-Huxley kinetics, and inhibitory (filled circle) and excitatory synapses (arrow) are modeled using alpha functions. The circuit is driven by a constant excitatory input to x_1 . (A) With only feedforward connections, positive and negative manipulations of x_2 decrease or increase y_1 activity. (B) Feedback excitation from y_2 to x_1 causes oscillation of y_2 activity when x_2 inhibition is increased. (C) Feedback inhibition from y_1 to x_1 can also evoke oscillation in y_1 , x_1 , and in y_2 as a result of diaschisis. (D) With inhibition from y_2 to y_1 positive and negative manipulations of x_2 may cause no change in y_1 .

electrochemical gradients). The functions that Crick poetically refers to reflect specific populations of neurons that make specific synaptic connections and have neuron and synapse-specific functional properties. To claim a reductionist decomposition of a system of n cells requires characterizing some $>n$ synaptic connections and some $\gg n$ cellular and synaptic properties. This was highlighted by Selverston (1980), Getting (1989), and later by Koch (2012) who (albeit in a straw man assumption of all-to-all connectivity) used Bell's number to calculate that it would take 2,000 years to completely characterize the direct connectivity of a system of 1,000 fully interacting components.

This demonstrates why reductionist analyses even in nervous systems that contain only 100s or even 10s of neurons do not

examine every component and interaction. Analyses instead define neurons as belonging to populations, either by the region they are in or some cellular marker (e.g., GAD2 as a claimed marker of inhibitory; i.e., GABAergic, neurons; Quina et al., 2020), or focus on more tractable larger cells like motor neurons, hippocampal and cortical pyramidal neurons, or cerebellar Purkinje cells instead of the smaller and often more numerous interneurons. Connectivity is often examined using indirect methods, extracellular stimulation of presynaptic neurons, and statistical models of connectivity (Horwitz, 2003), using criteria that can fail to correctly identify direct connections (Berry and Pentreath, 1976; Parker, 2010). Grouping neurons and synapses into populations characterized by mean values is a necessary and

acceptable approach providing that we appreciate that this may remove functionally-relevant variability (Parra et al., 1998; Aradi and Soltesz, 2002; Golowasch et al., 2002; Soltesz, 2006; Parker and Srivastava, 2013).

While these approaches are necessary, they leave mechanistic descriptions lacking detail on identified component neurons and their specific synaptic interactions, features highlighted as necessary criteria for reductive analyses in neurobiology (Bullock, 1976; Selverston, 1980; Getting, 1989). We can of course debate whether this level of detail is needed (see the commentaries in Selverston, 1980 for an early debate; Parker and Srivastava, 2013), but we cannot simply appeal to experimental convenience or tractability in the components we include (i.e., we cannot just ignore aspects that we cannot currently examine, but need to highlight the absence of potentially important details). Approaches may be field-dependent: neurophysiological analyses will typically attempt to identify specific neurons and interactions but may pay little attention to molecular aspects or behavior (Krakauer et al., 2017), while very detailed molecular analyses and manipulations may be examined at the neurophysiological only on unidentified or crudely characterized neurons or those that are experimentally tractable. A causal mechanism would seem to require that we know how the simultaneous integrated activity of specific types of cell, their properties, and their specific interactions in circuits generate a behavior, otherwise we can only correlate some molecular or cellular property to a behavior (a correlation is not necessarily uninformative).

The lack of relevant detail does affect explanations. In my field, the claimed experimentally characterized the lamprey spinal cord locomotor network in reality uses several assumptions and extrapolations to cover missing details and uncertainties over components, their connectivity, and functional properties (Parker, 2006, 2010). Likewise, the analysis of the ~200 interneurons involved in the *Aplysia* gill-withdrawal reflex was described as “forbidding in its complexity” by Hawkins et al. (1981), most subsequent work focusing on the experimentally tractable sensory neurons, an analytically convenient approach that fails to provide the claimed, and widely accepted, causal account of the behavior because it ignores known changes in motor neurons and interneurons (see Glanzman, 2010 and Trudeau and Castellucci, 1993; see Parker, 2019 for review).

New techniques promise to overcome analytical limitations. For example, the BRAIN initiative claims understanding will follow from recording “from ever more cells over larger brain regions” (Mott et al., 2018, p. 3); connectomic analyses claim that functional explanations will follow from more detailed brain mircoanatomy (Morgan and Lichtman, 2013; Schroter et al., 2017); and the originator of the Human Brain Project, Henry Markram, claimed that a more detailed cortical column model will cause a Copernican revolution in neuroscience (see Lehrer, 2008). These claims reflect an illusion of depth (Ylikoski, 2009). An explanation, let alone a Copernican revolution, is not a reflection of how many components we monitor or manipulate but of knowing what this data means in an explanatory scheme.

An example is provided by Greenberg and Manor (2005) who modeled the pyloric network of the crustacean stomatogastric ganglion, a system containing relatively very few neurons that is arguably the best understood neural circuit. Greenberg and Manor went beyond modeling the usual three neuronal groups to include five types of circuit neurons and their connections. They showed that an interaction resulting from the combination of an A-type potassium current and short-term synaptic depression was needed to generate the normal pyloric rhythm, but the complexity of the model, which consisted of almost 50 coupled differential equations, prevented them from explaining the underlying mechanism. As a result, they reverted to the use of a simpler model, stating “The reduced model emphasizes a result that is difficult to discern in the detailed model because of its complexity” (Greenberg and Manor, 2005, p. 676).

Simon (1962) suggested that mechanistic explanations are in principle possible irrespective of the number of components and interactions if systems are decomposable or nearly decomposable, namely have a fixed hierarchy of components where intracomponent interactions are strong but intercomponent interactions are relatively weak (but non-negligible), and each component processes the input it receives from the component above it in the hierarchy: this makes the behavior of each component approximately independent of the behavior of the others. The decomposability of nervous systems was examined by Bassett et al. (2010) using the connectome of the *C. elegans* nervous system and human brain fMRI data. They claimed that both showed “some” degree of hierarchical organization, and cited Simon in claiming that they are thus nearly decomposable. But Simon did not say that even fully hierarchical systems are nearly decomposable, just that “some kinds of hierarchical systems can be approximated successfully as nearly decomposable systems” (Simon, 1962, p. 474).

Being decomposable or nearly decomposable is a core assumption of reductionist approaches. For example, when we manipulate a system component we assume that the resulting effect reflects the function of that component. We should consider the validity of our assumptions, and an obvious consideration is whether nervous systems are decomposable and whether we can directly link a manipulation to an observed effect. A system is minimally or non-decomposable if interactions between components are many or strong and the function of a component reflects not only its intrinsic properties but also its relationships with other components. This seems to better describe nervous systems, which consist of multiple parallel feedforward, lateral, and feedback pathways (Sporns, 2011; Pessoa, 2014). For example, cortical areas, including primary sensory regions, receive parallel convergent inputs from various sources that make the regions multifunctional, while feedback connections from these regions can influence the nature of the incoming inputs that they process (Anderson, 2010; Schroeder and Foxe, 2005). Conversely, specific functions can be performed by multiple regions. A classic example of this is Lashley’s equipotentiality hypothesis that suggested that memory is stored diffusely in multiple cortical areas (Lashley,

1929). Re-analysis of Lashley's data resulted in some quantitative modifications but his general conclusions have held (see Thomas, 1971), and despite the localization of the molecular and cellular mechanisms of memory being a major focus of neuroscience research over the last 4 decades (Bliss et al., 2018), Josselyn et al. (2015, p. 521) wrote that the failure to localize the engram reflects the “widely distributed and dynamic nature of memory representations in the brain.”

These distributed and multifunctional effects necessarily complicate the mapping of specific functions to specific regions or components. This was highlighted by the neurophysiologist Charles Sherrington who called reflexes a “convenient fiction...a simple reflex is probably a purely abstract conception, because all parts of the nervous system are connected together and no part of it is probably ever capable of reaction without affecting and being affected by various other parts” (Sherrington, 1906, p. 8). McCulloch (1945) highlighted that parallel feedback and feedforward connections in what he called heterodromic systems were necessary to co-ordinate behavioral responses, and claimed that even simple heterarchical systems are unpredictable because their connectivity allows component relationships, their independence, and their importance and ordering can change. This view has been supported experimentally by the heterarchical organization of spinal cord sensorimotor systems (see Cohen, 1992) and by the switching of molecular and cellular components between functions (e.g., Meyrand et al., 1991; Daaka et al., 1997; Fahoum and Blitz, 2021). These aspects do not deny hierarchical processing occurs, but there is not a fixed hierarchy as a component's role and position can change depending on context. A heterarchical organization does not prevent explanations of non-decomposable systems but does require that explanations consider variable relational aspects rather than seeking 1-to-1 links. This was highlighted by Rashevsky (1954) who referred to “metric biology” for analyses of system components, and “relational biology” for aspects dependent on system organization. Specific analyses are needed because new properties can appear at different levels (Anderson, 1972).

Relational aspects oppose substantialist views that see functions represented in components, expressed in references to memory molecules, inhibitory or excitatory neurotransmitters or neurons, and mood or reward neurotransmitters. Neurons and neurotransmitters are not intrinsically inhibitory or excitatory, as can often be claimed (e.g., Eckstein et al., 2020). Inhibition and excitation as well as functions like mood (serotonin), reward (dopamine), and pain (substance P) are not intrinsic to a neuron or neurotransmitter but depend on the transmitter receptor activated, the cells the receptors are in, and the circuits/regions where the cells are located (dopamine is also involved in retinal processing (Korshunov et al., 2020), 5-HT in motor control (Jacobs and Fornal, 1993), and substance P in breathing (Pilowsky, 2014)). Consider the neurotransmitter GABA. Identification of GAD2, the enzyme that synthesizes GABA is often used to identify “inhibitory” neurons (see Quina et al., 2020). But GABA itself is not inhibitory: ionotropic GABA_A receptors are permeable to Cl⁻,

but whether this evokes inhibition or excitation depends on whether chloride enters or leaves to hyperpolarize or depolarize the cell (shunting inhibition can occur if there is no net movement of Cl⁻). This depends on the equilibrium potential for chloride, which in turn depends on the activity of Cl⁻-pumps that determine the intracellular Cl⁻-levels and the membrane potential of the cell, all of which will simultaneously change as the neuron and receptor are activated. Even if a GABAergic neuron was known to hyperpolarize and inhibit a postsynaptic cell this may still not describe its functional effect as inhibition of other inhibitory neurons (disinhibition) will evoke excitation.

Relational aspects are illustrated by homeostatic plasticity where parameter values vary as a function of the variability of other components to maintain an output. In single cells this can reflect variations in different classes of ion channels or synaptic inputs to maintain a certain level of cellular excitability (Turrigiano et al., 1998; Swensen and Bean, 2005), while in neural circuits variability in neuronal and synaptic properties can maintain a particular circuit output providing that the ratios between the different functional components are in appropriate balance (see Prinz, 2010). Examples of the latter include the 4,000,000 combinations of eight types of ion channels and seven types of synapse that could generate the modeled output of a three neuron stomatogastric ganglion circuit (Prinz et al., 2004); compensations in basal ganglia circuitry that delay Parkinson's disease symptoms until 80% of the dopaminergic neurons in the substantia nigra have degenerated (Bezard et al., 2004), and in functional recovery from spinal cord lesions where locomotor behavior matching that in unlesioned animals according to various behavioral measures can be generated using spinal cord systems with markedly different anatomical and functional properties (Davis et al., 1993; Edgerton et al., 2001; Parker, 2017).

These variable relational effects show that multiple neurophysiological states ($N_1 \vee N_2 \dots N_n$) can realize a single behavior or cognitive process ($P_1 \leftrightarrow N_1 \vee N_2 \dots N_n$). This could be considered a neurobiological example of multiple realizability, although this is a contentious issue (see Aizawa and Gillett, 2009). But the evidence above suggests that nervous system outputs are linked to multiple, not single neurophysiological states, even when the cellular properties and the output are both measured in comparable detail (Aizawa and Gillett, 2009). Multiple realization is claimed to prevent reductive explanations (see Aizawa and Gillett, 2009 for discussion) but this seems not to be the case in the examples above. But it does require that variable relational effects between components are known, and that we know when and why one or other particular neurophysiological state is used.

Relational aspects are not confined to interactions within the nervous system but also reflect interactions of the nervous system with the body (e.g., proprioceptive, neural-immune, and gut-brain interactions) acting in the environment (see Dreyfus, 2012). This has been called embodied cognition (Shapiro and Spaulding, 2021) and has been inspired by ecological psychology and neuroethological analyses (Chiel and Beer, 1997). These effects move away from the view that the nervous system sequentially

processes inputs to plan and generate outputs to one where adaptive behavior results from the continuous two-way relationships between the nervous system, the body, and the environment. These are referred to as levels, but while this may provide a simplifying concept for organizing data and analyses it also gives the erroneous impression of effects working up or down through separate stages when all effects are occurring simultaneously (Noble, 2012). There are many examples of behavioral and environmental influences on brain function: adult neurogenesis is enhanced in enriched environments (Altman, 2011); the availability of receptive females influences male primate and human testosterone levels and sexual behavior (Anonymous, 1970); amphetamine effects on primate behavior reflect position in the social hierarchy (see Cacioppo and Berntson, 1992); and hyperactivity and low blood levels of serotonin, a correlation that could have led to a claimed causal link, were both normalized in children when they were hospitalized (Coleman (1971).

It could be argued that as embodied effects are represented in the molecular transduction of sensory neurons and through various sub-cortical and cortical sensory processing stages, they can be incorporated into the neurobiological explanation. But this would require consideration of heterarchical processing that is continuously altered by internal and external relationships between ongoing functions and behaviors. As McCulloch (1945) suggested, relational contexts mean that even though an input is represented by a pattern of sensory activity, this activity won't necessarily predict the resulting effect. The placebo effect in pain perception would be an example, where an external context, expectation of analgesia, leads to an alteration in nervous system processing through activation of endogenous opioid systems that alters the perception of the sensory input and the resulting behavior (Benedetti, 2007).

Noble (2012) uses the heart to illustrate these relational effects. Even though genes for various cardiac ion channels specify heart cells over other cell types, the heart rhythm is not determined by these genes but by the component ion channels, cellular properties (electrochemical potentials) that affect ion channel activity, gross heart structure, the ongoing heart rhythm, and internal and the external environmental factors that influence it. Relational aspects were demonstrated when computing advances in the 1990s allowed the detailed information obtained on cardiac ion channels to be incorporated into multicellular models of the sino-atrial node pacemaker. In the model, cells at the edge of the node depolarized first and activity spread inwards, but in the heart the activity originates near the center of the node and spreads outwards. When the sino-atrial node was dissected from the atrium it behaved like the computer model, normal activity thus reflecting relational influences arising from the organization of the heart (see Noble et al., 2019).

Simon (1969, p. 52) summarized these wider relational effects, "A man [*sic*], viewed as a behaving system, is quite simple. The apparent complexity of his behavior over time is largely a reflection of the complexity of the environment in which he finds himself." But this can be extended because environments are also continuously modified by ongoing

behavior (see Chiel and Beer, 1997), a circular interaction. There is nothing mystical or metaphysical about these higher-level context-dependent effects. They can be expressed mechanistically and mathematically with the same precision as lower-level mechanisms, the latter using differential equations, and higher-level context-dependent influences by the initial and boundary conditions of these equations (Noble et al., 2019).

As mentioned above, relational effects complicate experimental approaches that attempt a 1-to-1 mapping of components to functions (e.g., reverse inferences in brain imaging studies), and the interpretation of system manipulations. Newer techniques like gene knock-outs and optogenetics claim greater precision than traditional approaches (e.g., physical lesions or pharmacological approaches), but no matter how surgical a manipulation is, relational effects mean that there will necessarily be changes in the properties of other components (i.e., diaschisis). This is of course the aim of a manipulation, to identify a component and its role from how the system changes after its manipulation. In decomposable or nearly-decomposable systems where parts are relatively independent we could relate the resulting system changes to the manipulated component, but relational effects in non-decomposable systems mean that system effect will reflect changes in more than the manipulated component (or no system effect despite key components being altered; Figure 3D) thus requiring us to consider multiple causes for a system effect. Also, while we assume that we can at least be confident that we can precisely control the component we have manipulated, feedback pathways in relational systems can affect the manipulated component and thus the system is not manipulated in the way we intended. Both of these aspects necessarily complicate attempts to localize system functions to specific component parts.

An additional aspect of relational effects is that when a system is inactive or its normal organization is disturbed (e.g., in cell cultures or tissue slices, routine experimental approaches used because they provide us with greater access and control over a system) the properties that we characterize can differ to those in the intact, active system. Claude Bernard wrote, "the phenomena of a living body are in such reciprocal harmony one with another that it seems impossible to separate any part without at once disturbing the whole organism," quoting Georges Cuvier, "All parts of a living body are interrelated; they can act only in so far as they act all together; trying to separate one from the whole means transferring it to the realm of dead substances; it means entirely changing its essence" (see Normandin, 2007). An example of a change in essence in dissected or quiescent systems is the absence of functional properties normally established by relationships in the intact, active system. These are not components in the traditional sense that can be isolated; they do not exist in specific locations with specific values or even exist at all under some conditions. These effects include volume transmission and ephapses (Faber and Pereda, 2018; Svensson et al., 2019), both of which negate the claim that "wired" axonal and synaptic connections determine functional interactions in nervous systems (e.g., Price and Friston, 2005).

Ephaptic signals reflect changing electrical fields in the extracellular space generated by summed neuronal activity. The membrane potential (V_m) is the difference between the intracellular (V_i) and extracellular potential (V_e), $V_m = V_i - V_e$. Current flow to or from the local extracellular environment caused by cellular or synaptic activity generates local field potentials that change V_e and thus V_m . These effects are anisotropic, the magnitude and direction of the change in V_e reflecting a complex interaction of several variables including the number of active cells, the pattern of their activity, the packing and orientation of neurons and processes, the geometry of the extracellular space, and the properties of the “postephaptic” cell (e.g., location of ion channels in an ephaptic field). Ephapses provide a concrete neurobiological example of an emergent effect. While these effects can be modeled, the equations currently rely on assumptions of several unknowns. We know field effects occur, they are measured in EEGs, but are they functionally-relevant signals (e.g., Bullock, 1959) or an epiphenomenon of neural activity? The latter view has seemingly dominated with the experimental focus on single cells and synaptic connections.

To consider the implications of ephapses to reductionist explanations, assume that field effects are important. They are generated by neuronal activity and neuronal activity is altered by field effects, a circular interaction. But neuronal activity also alters the geometry of the extracellular space (Østby et al., 2009), meaning that even if all the variables listed above were characterized, they will all continually change during system activity: a change in the extracellular space will alter the magnitude and spread of field effects, which will alter neuronal and system activity, and thus alter the extracellular space... a circular interaction influencing another circular interaction.

We cannot explain these effects by describing individual cellular or system properties but must consider the relationships between local and global effects simultaneously. This was expressed in Lashley's dilemma, “Nerve impulses are transmitted from cell to cell through definite intercellular communication. Yet all behavior seems to be determined by masses of excitation” (Lashley, 1942, p. 306). We can appreciate this from the intuitive sense of our own behaviors, which does not support a mechanistic sequence of effects passed from one element to another along axons and across synaptic connections. Take movement: robotic systems split movements into sequences of distinct parts, but in a natural movement like reaching for a cup you do not first move your shoulder, then elbow, then wrist, then fingers: movement at the beginning (shoulder) and end (hand/finger) may change, but as the shoulder moves the wrist or fingers are shaped to be in position when the hand reaches the cup. Bullock (1959, p. 999) offered a potential solution by extending the neuron doctrine in saying “perhaps much of the normal functioning is carried out without nerve impulses...by means of graded and decrementally spreading activity,” and proposed, like Sherrington had for reflexes, that circuits of wired interacting components are an “oversimplified abstraction involving a limited subset of communicated signals...in fact, there are many parallel types of signals” (Bullock, 1981, p. 281). Despite his optimism that “in the

near future we will gain significant new insight” (Bullock, 1959), these ephaptic signals have received very little attention compared to single molecules, cells, and wired synaptic connections. This is starting to change as functional ephaptic effects have been shown and studied in several systems (Faber and Pereda, 2018).

Volume transmission is the diffusion of neurotransmitters through the extracellular space to affect targets distant from their release sites (μm for amines and mm for neuropeptides; Svensson et al., 2019): anatomical localization thus does not determine a transmitter's effects (cf Price and Friston, 2005). Volume signaling is not simply a synaptic signal spread over a wider area but like ephaptic effects is anisotropic, the direction and extent depending on the size and charge of the transmitter, the activity-dependent geometry of the extracellular space, the presence of or efficacy of uptake or breakdown mechanisms, charges on extracellular proteins or ephaptic field potentials that attract or repel molecules, and even “tidal” effects caused by blood pulsing in arteries. Like ephapses, volume effects thus reflect changing spatial and temporal relationships between components.

Volume transmission also allows two or more transmitters released from spatially distant regions to interact (interactions can also occur locally through co-release from single synaptic terminals or vesicles; Svensson et al., 2019). Transmitter interactions are a highly conserved basic feature from invertebrate to mammalian nervous systems that can generate additive, subtractive, non-linear, or emergent effects (i.e., effects not associated with any individual transmitter; Brezina, 2010; Svensson et al., 2019). Amines and neuropeptides act on G protein-coupled receptors and intracellular pathways to modulate the functional properties of cells and synapses from seconds to hours (Svensson et al., 2019). They can thus evoke a background “modulatory tone” that allows interactions between transmitters whose release is not only spatially but also temporally divorced.

Consider the well-described ascending modulatory systems to the cortex (McCormick, 1992; Hasselmo, 1995; McCormick et al., 2020; Figure 4). These are typically presented as separate pathways with specific roles (e.g., arousal and learning). The traditional view that these systems diffusely modulate the cortex has been challenged by the presence of specific neuronal populations in each system that project to distinct cortical regions: for example, Breton-Provencher et al. (2022, p. 732) say “locus coeruleus-noradrenergic (LC-NA) activity was causal for both task execution and optimization [during learning].” But these ascending systems are connected to each other by direct lateral connections and by feedback connections from the cortex which makes it difficult to decompose and causally link them to specific functions like learning. Even if they could be activated independently by inhibiting lateral and feedback connections, volume transmission, and the modulatory tone resulting from G protein-coupled receptor activation can still generate context-dependent interactions driven by internal or external events (e.g., sensory inputs, learning, and arousal) between transmitters released at different times from different ascending systems that prevent a functional effect being causally linked to a single transmitter.

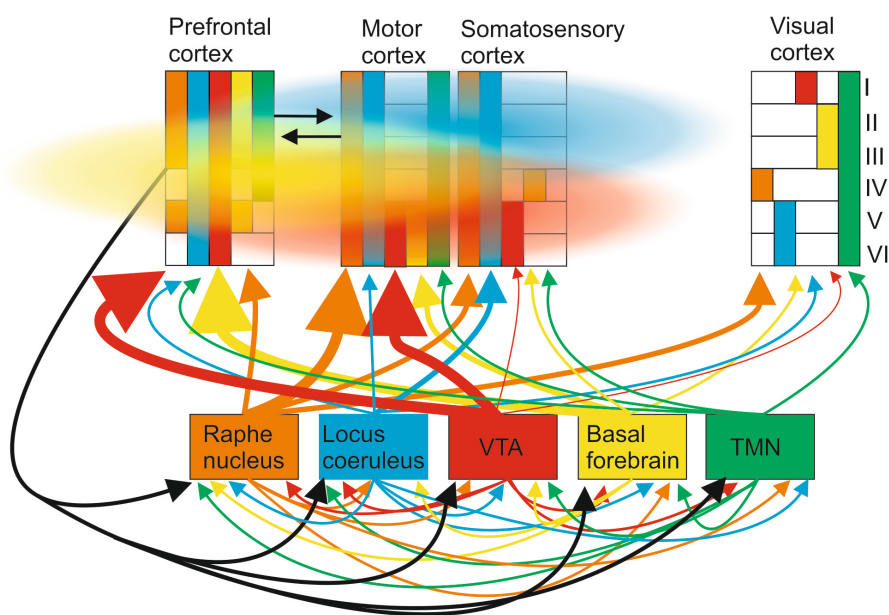


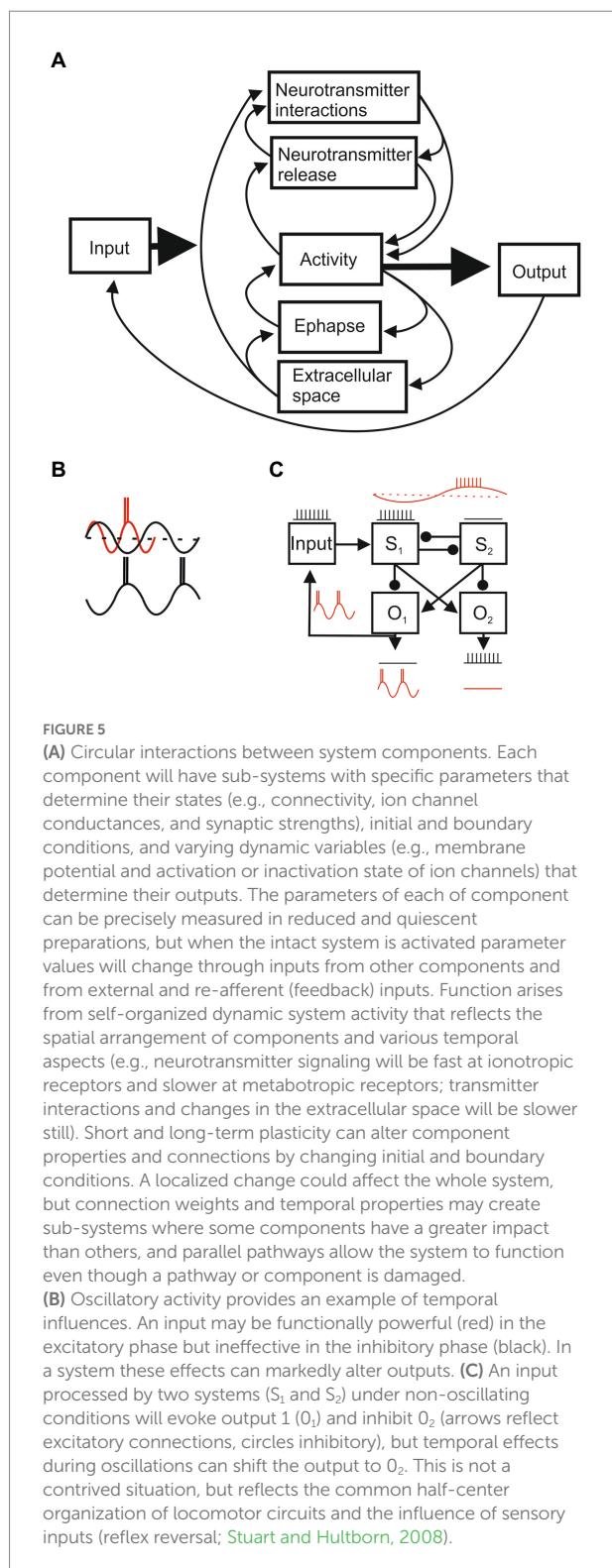
FIGURE 4

Ascending modulatory pathways make wired connections to multiple cortical areas (regional projections to different cortical areas and layers are indicated by the size of the ascending arrow and the colored blocks in the different cortical regions; layers are indicated by the roman numeral on the right). These include cholinergic inputs from the basal forebrain, noradrenergic inputs from the locus coeruleus, serotonergic inputs from the raphe nuclei, dopaminergic inputs from the substantia nigra and ventral tegmental area, and histaminergic inputs from the hypothalamus. These systems also connect directly to each other and receive cortical feedback. Cortical signaling occurs through wired axonal connections (black arrows) and volume transmission (colored clouds), the direction and extent of volume signals reflecting ease of diffusion in different directions.

The idea that we can relate cognitive effects or behaviors to the actions of single, specific transmitters seems naïve given the evidence that multiple transmitters can interact in even simpler nervous systems. Amino acids, amines, and neuropeptides are released by successively higher rates of presynaptic activity (Verhage et al., 1991; Svensson et al., 2019). This frequency-dependence multiplexes synapses (a terminal co-localizing five transmitters could generate over 100 different combinations/signals). Nervous system activity thus alters the complement of neurotransmitters released into the extracellular space, the geometry of the extracellular space influencing the diffusion and potential for interactions along volume transmission pathways, while transmitter release and interactions will alter nervous system activity, transmitter release, the geometry of the extracellular space, and the potential for interactions, adding further circular interactions to those outlined for field effects (field effects and volume transmission are also not dissociable: transmitter-mediated or ephaptic changes in activity will alter field effects, neuronal activity, the geometry of the extracellular space, and transmitter release, diffusion, and interactions). Even without considering embodied and environmental influences, nervous system activity will reflect an equilibrium between multiple wired and non-wired circular interactions (Figure 5A) that is affected by various spatial and temporal factors (Figures 5B,C). As highlighted by McCulloch (1945), these nested circular interactions allow an equilibrium to be shifted to a new one by very small changes in activity, matching James Clerk Maxwell's claim that life differs to physics because a "strictly infinitesimal force may determine the

course of the system to any one of a finite number of equally possible paths" (see Van Strien, 2015).

While the astronomical number of cellular and synaptic components in heterarchical organizations and their degeneracy and variability offer significant practical challenges to reductionist approaches, they are in principle, if not currently in practice, achievable using reductive current approaches, albeit with the requirement that these approaches consider more than the decomposition of systems into parts. But ephaptic effects and volume transmission differ in that they not only present practical but also conceptual challenges as they reflect transient "non-wired" signals that are not reflected in anatomically defined neurons, axons, or synaptic terminals, and they require the simultaneous analysis of multiple components during ongoing activity in intact functioning systems rather than a focus on single components in the reduced quiescent or non-behaving systems often used experimentally. It could be argued that highlighting these aspects adds complexity for complexities sake and invites a pessimistic or nihilistic view of our chances of understanding. But while ephaptic signaling and volume transmission are intangible, they are not hypothetical or mysterious but are established features of nervous systems identified in reductionist analyses over several decades (Faber and Pereda, 2018; Svensson et al., 2019). It is claimed we can appeal to the decomposability of systems into component parts as a knowingly "fallible" heuristic (Bechtel, 2002; Wimsatt, 2006) or "fat-handed" approach (Romero, 2015) to gain entry to a system. There is obvious merit in simplicity, but not in being too simplistic: Occam's razor is not that the simplest explanation is



best, but that entities should not be added beyond those necessary. We may need to ask if the non-local and relational effects of ephapses and volume transmission are necessary considerations given their demonstration in multiple nervous systems and our limited ability to explain cognition and behavior based on reductionist assumptions.

Reductionist assumptions and the basis for interventions

The philosophy of neuroscience considers the “bounds of sense... what can coherently be thought and said” (Bennett and Hacker, 2003). Considering these bounds, which include the practical and conceptual aspects of reductive approaches outlined above, is important in mechanistic schemes to avoid erroneous claims to explanation that can prevent or delay genuine advances, not least by adding the requirement to an already difficult task of identifying and undoing errors before a correct explanation can be reached. While important to explanations, considering the bounds of sense is essential for any intervention as transgressing these bounds can, and has, resulted in significant negative consequences. This section will consider approaches to intervening based on reductionist analyses and assumptions of the insight it has given us into nervous system function, using psychopathology and education as examples.

Psychoanalysis dominated early 20th century psychiatry until the mid-1950s when advances in psychopharmacology were considered to provide a more scientific basis by relating psychopathologies to abnormalities in neurotransmitter systems (e.g., the monoamine hypothesis of depression; Healy, 2015; Kendler, 2015), with interventions targeted on correcting these abnormalities. Examining biological mechanisms is as potentially useful in psychopathology as in any clinical condition providing that an explanation considers causal factors at multiple levels (see Proctor, 2012), including the aspects outlined above, and that effectiveness is established before claims for interventions are made. But this is not always the case. A spokesperson for a Huntington's disease advocacy group expressed the dismay the group felt when a recent gene targeting trial was canceled because it worsened outcomes: “There has been so much positive noise around it; both from researchers and clinicians and from the drug company themselves. I think the community was really swept up by that hope” (see Kwon, 2021, p. 180). Similarly, claimed treatments for spinal cord injury have routinely failed (Steward et al., 2012), and the optimism that completion of the Human Genome Project will “very quickly” bring new treatments so “whole families are relieved, forever, of the curse of genetic disease” (Blakemore, 2000, p. 3) has not been realized.

An issue for mechanistic explanations and interventions in psychopathology is that despite extensive effort in trying to find them, the anatomical, cellular, or molecular features that act as diagnostic markers for neurological disorders are absent in psychiatric conditions, and diagnoses are instead made from behavior and cognition (Wang and Krystal, 2014, p. 3 wrote “there is not a single symptom of a single psychiatric disorder for which we fully understand its physiologic basis”). Absence of a biological mechanism means there is no specific biological aspect to target (e.g., normalization of an aberrant neurotransmitter level) and no end point to reach (chronic anti-depressant use and chronic depression both reduce life expectancy; Warren, 2020). Biological interventions that generate beneficial effects are still possible without a mechanistic explanation. In Parkinson's disease, stimulation of the subthalamic nucleus can improve motor function despite current

models suggesting it should worsen it (McIntyre et al., 2004). This contradiction would not matter if the stimulation reliably worked, but it is only effective in a proportion of patients, and effectiveness could presumably be improved by better understanding of the underlying neurobiological mechanisms (Piasecki and Jefferson, 2004). Penicillin provides an example: it was successfully used for some years before its mechanism was understood, but greater understanding of antibiotic mechanisms has allowed treatments to be optimized (Lobanovska and Pilla, 2017).

While interventions can be made blind to mechanisms, the history of these interventions in psychiatry is poor. In the 20th century these included ice baths, malaria-induced fevers, insulin-induced comas, electrical or drug-induced seizures, removal of teeth or parts of the digestive tract (Khazan, 2014; Davidson, 2016), and lobotomy (Breggin, 1993). These shock approaches had little or no evidence to support their use, but they were still confidently used in mainstream psychiatry: the psychiatrist R.D. Laing wrote, “I am still more frightened by the fearless power in the eyes of my fellow psychiatrists, than by the powerless fear in the eyes of their patients” (Laing, 1985, p. 18). Developments in psychopharmacology rightly removed these approaches. But although psychopharmacology assumes to target causally-relevant mechanisms, its 1950s developments reflected chance observations: the antidepressant iproniazid was developed as a treatment for tuberculosis but was coincidentally found to improve mood, its assumed inhibition of monoamine oxidases leading to the monoaminergic-hypothesis of depression; while the first antipsychotic, chlorpromazine, originated from the search for new anti-histamines, its calming effect and action on dopamine receptors promoting the dopamine-hypothesis of schizophrenia (Lehmann, 1993). Early meta-analyses suggested that psychopharmacology was the most effective approach (see Shorter, 2021), the merits of psychotherapy being questioned by claims that it did not matter what type of therapy was given, for how long, or the credentials of the therapist (see Smith and Glass, 1977). However, recent meta-analyses suggest that psychotherapy and psychopharmacology are equally effective for major depression, panic disorder, and seasonal affective disorder (Cuijpers et al., 2013; Warren, 2020; see Cipriani et al. (2018) and Munkholm et al. (2019) for discussion).

Even though we can use beneficial interventions without knowing the mechanisms underlying their effects, it is still important to consider whether relationships are causal or correlational. For example, assume that a negative life event is processed in the brain through a known cellular mechanism that lowers serotonin levels and that this in turn acts on a known mechanism that affects mood circuitry. This mechanism and the associated reduction in serotonin levels could be claimed as the cause of the depression, but at best (i.e., serotonin levels do causally influence mood) this only says how the depression occurred, not why. Knowing why is necessary to determine the optimal intervention; do we act on serotonin levels or address the negative life event? An analogy would be that hemorrhage may cause death through loss of blood volume and blood pressure leading to insufficient oxygen delivery to the brain and heart, but treatment for this would not be continual blood transfusions to maintain blood volume but treating the hemorrhage.

Even if neurobiological causation was determined, this still may not necessarily make a neurobiological intervention better than non-biological approaches (e.g., coping strategies for those with memory deficits following head injury; Tsoulosides and Gordon, 2009). Phenylketonuria provides a textbook example of a causal genetic factor associated with profound psychological and neurological impairments that is successfully managed through diet, a reflection of behavior influencing lower-level effects (Rampon et al., 2000). But current views of psychopathology, as with attempts to explain normal functions, can have a neurobiological focus. For example, the perception in the autism community of a neurobiological focus in the Wellcome Trust-funded Spectrum 10K autism genetics study, led to concerns that saw the study being paused (see Sanderson, 2021). Another example comes from a Royal Society report that claimed “neuroscience provides concrete evidence of biological differences between children with ADHD and others,” despite then seemingly contradicting this by saying “There is no biological test at present” (p. 11) and that assessment is based on behavior.² Pharmacological use in ADHD has increased markedly without concomitant understanding of drug mechanisms (Bachmann et al., 2017), but as with phenylketonuria there are non-biological interventions that reflect behavior in particular environments, including cognitive approaches that train children in self-evaluation (identifying issues, setting goals) and give behavioral management strategies to parents and teachers (time outs and chart/point systems; Miranda et al., 2002; Howard-Jones, 2008). These approaches require investment rather than generating profit, but the latter is not a factor that should be considered in the bounds of sense.

In addition to promoting biological explanations and approaches, constitutive and explanatory reductive views have also altered assumptions of psychopharmacology mechanisms from a drug-centered approach where drugs have some net beneficial effect on brain states underlying cognition and behavior, to a disease-centered view that sees drugs normalizing function by targeting specific biological mechanisms (e.g., excess dopamine in schizophrenia; Middleton and Moncrieff, 2019). This generates a potentially fallacious circular argument: because drugs target biological mechanisms, the mechanism is biological. Given the identification of volume transmission and neurotransmitter interactions in reductive analyses in a range of nervous systems (Svensson et al., 2019), neurobiological considerations seem to make a drug rather than disease-centered mechanism far more likely. Consider depression again: assuming that serotonin was the causal factor for depression (see Kendler, 2015) and that serotonergic drugs only affect serotonergic systems, the effect of these drugs would not necessarily reflect a serotonin-specific effect in the brain as changes in serotonin levels along volume transmission pathways would affect numerous circular and other interactions to generate new equilibrium states (the time to establish this with global rather local physiological changes in serotonin

² https://royalsociety.org/-/media/Royal_Society_Content/policy/publications/2011/4294975733.pdf

levels may influence the delay in psychological effects despite changes in serotonin levels; Healy, 2015). Unless reasons were found to negate the need to consider volume transmission and transmitter interactions, psychopharmacological approaches won't need drugs that more specifically target transmitter systems but knowledge of what constitutes a normal or pathological brain state, what intrinsic (e.g., personality) and extrinsic factors (social conditions) influence these states, and how (or if) we should intervene using a drug-centered approach to shift the state to one we identify as desirable.

While interventions have traditionally been poor, just as new techniques promise insight into nervous system functions new "neurotechnologies" promise better reductive interventions by using genetic engineering, stem cells, brain implants ("nanobiochips"), smart drugs ("emoticeuticals"), or downloading, "straightening out," and re-uploading information from the brain (Geake and Cooper, 2003; Lynch, 2004; Tancredi, 2005). These claims were called the "lobotomy attitude" to reflect their limited scientific basis (Dudai, 2004), and the claims in these older references have not been realized. Proponents have made the fallacious a fortiori appeal to success in other areas, vaccination, cardiac pacemakers, control of diabetes or blood pressure, and cochlear implants (Tancredi, 2005), which offers no logical basis from which to claim success for neurotechnological interventions. The uncertainty surrounding the serotonin-hypothesis of depression and other mental disorders (Kendler, 2015) highlight that psychopathology differs to physiological conditions like diabetes where the disease-centered approach applies. Although causality is difficult to establish, understanding factors like volume transmission and transmitter and other circular interactions in heterarchic systems should provide a better basis for interventions.

Education is a recent focus for translational neuroscience. Neuroeducation claims that neuroscience can inform educational practices. This could reflect multiple approaches (Goswami, 2009), but there is again a focus by some on neurobiological mechanisms. A Royal Society report³ claimed that "Biological factors play an important role in accounting for differences in learning ability between individuals," despite admitting that this conclusion is made even though "high quality information is scarce" (summary p. 5). Neuroimaging of brain areas activated in tasks like reading, speaking, writing, and counting (see Ansari and Coch, 2006) are claimed to offer insight into optimal teaching methods by facilitating specific neural mechanisms, but not what these are or how they could be targeted. The Royal Society report also says, "the brain changes constantly as a result of learning and remains 'plastic' throughout life" (summary p. 5). Plasticity invokes neurobiological mechanisms driven by specific inputs that alter the nervous system while emphasizing the potential influence of external or higher-level factors that drive these changes. This is highlighted in the Royal Society report which states "education is the most powerful and successful cognitive enhancer of all" (p. 1). Plasticity

has been promoted as a concept that teachers can use, but plasticity just means that children learn rather than giving novel insight that would shift the emphasis from the child or school to the brain, even if we did causally understand how plasticity mechanism affect cognition (Bliss et al., 2018; Parker, 2019). Educational achievement, not a change in the brain, is the aim.

Behavioral genetics illustrates a dominant reductive neurobiological focus on cognitive abilities. This is a long-standing and contentious issue that uses heritability estimates derived from family studies of identical and non-identical twins raised together (same or different genetics in the same environment) or after adoption (different environments; Rose et al., 1985; Plomin et al., 1996) to assess the relative contribution of genetic mechanisms and the environment. These contributions are not separable and they do not have fixed values. For example, height reflects genetic and environmental influences (nutrition), but plentiful food will reduce the environmental variability and increase heritability. Various aspects complicate measures of the heritability of cognitive abilities: children alter the behavior of those around them meaning that first-born children have different environments to their siblings; adoption studies can include twins not separated at birth (allowing early environmental influences) and separation can mean one twin living with the mother and one with a relative (Rose et al., 1985). Even with complete separation at birth adoption studies usually have a restricted environmental range as adoptive parents tend to come from higher socioeconomic groups, and heritability estimates decrease when a broader socioeconomic range and thus greater environmental variability is considered (Turkheimer et al., 2003). This is mirrored in animal studies where genetically-influenced behavioral differences can disappear in enriched environments (Crabbe et al., 1999; Rampon et al., 2000).

Genetic influences on cognitive abilities are unlikely to be simple: half the genome is expressed in the brain during development and genetic effects are subject to environmental influences. External influences on cognition and behavior were thought to be limited to genetically-determined "critical periods" associated with neurogenesis and synaptogenesis, leading to claims in policy papers, the media, and brain-based education literature that neurobiological evidence suggests children should be taught before school age (Huttenlocher, 2002; McCoy et al., 2019). Nobody would deny a positive early environment is advantageous, and pre-school educational interventions are beneficial, although it is unclear what aspects are improved and for how long effects last (McCoy et al., 2019). But the claimed neurobiological mechanism needs updating: neurogenesis and synaptogenesis persist into adulthood (Lledo et al., 2006; Gould, 2007; Thompson and Wolpaw, 2014) supporting "sensitive" rather than critical periods, and pre-school interventions also reflect higher-level influences of classroom environment and teacher-child interactions (McCoy et al., 2019). If we base interventions on erroneous or simplistic mechanistic claims then beneficial effects may not occur, but a worst-case scenario is that these interventions may be deleterious. An example comes from animal studies where normally beneficial rehabilitative training given

3 https://royalsociety.org/-/media/Royal_Society_Content/policy/publications/2011/4294975733.pdf

prematurely after experimentally-induced stroke can increase lesion areas and worsen functional recovery (Schallert et al., 2000).

A neurobiologically-inspired approach that has attracted significant recent interest is pharmacological cognitive enhancement in the absence of pathology. Lifestyle drugs like these can be sought even when their efficacy or safety is questioned: the withdrawn appetite suppressor fenfluramine was sought by dieters even though it caused fatal heart disease (Flower, 2004), and the ADHD drug methylphenidate is widely used as a cognitive enhancer by non-ADHD students (Koren and Korn, 2021) despite evidence that it may worsen performance (Farah et al., 2004).

Bostrom and Sandberg (2009, p. 316) appeal to reductionist neurobiological mechanisms by claiming that cognitive enhancers work by “increasing neuronal activation or by releasing neuromodulators,” a very vague mechanistic statement, but they then say that they work by “facilitating the synaptic changes that underlie learning,” and that “intervening in the permanent encoding at synapses, a process which has been greatly elucidated in recent years, [they presumably mean LTP the significance of which remains uncertain; Queenan et al., 2017; Bliss et al., 2018; Parker, 2019] is a promising target for drug development... that not only allow the brain to learn quickly, but which also facilitate selective retention of the information that has been learned” (Bostrom and Sandberg, 2009, p. 317). Very vague mechanistic claims, necessarily so given that we lack the necessary neurobiological detail, are thus turned into concrete physiological mechanisms that promise cognitive improvements by acting on memory encoding and retention.

Drugs can improve memory. Effects seem greater in poorer performers exposed to more difficult tasks, but they are modest and currently difficult to attribute to any specific biological mechanism (chewing gum can also evoke memory improvements; Wilkinson et al., 2002). A common cognitive enhancer, modafinil, directly or indirectly affects multiple transmitter systems, has varied effects on memory and other cognitive systems, and varied side-effects (Ackerman and Kanfer, 2009). Even if modafinil significantly improved real-world memory (i.e., beyond statistical effects under laboratory conditions) the bounds of sense requires asking if pharmacological interventions targeting unknown mechanisms should take priority? In addition to chewing gum, taking breaks significantly improves cognitive performance in nurses, doctors, and air traffic controllers (Smith-Coggins et al., 2006; Signal et al., 2009), a safer and more cost-effective approach. Claiming that a pharmacological cognitive enhancer is no different to using contact lenses to improve performance is a trivially false analogy⁴ (only one of these is non-invasive, readily reversible, with a known mechanism, safety, and effectiveness providing an appropriate prescription that matches the intervention to the features of the individual).

Ethical issues have been discussed extensively in cognitive enhancement, principally the unfair advantage given to those who can access and afford the drugs. But given the lack of mechanistic understanding and limited effects these discussions beg the question by assuming that significant benefits exist. But strong claims are made: Lynch (2004, p. 229) claimed that neurotechnology targeting neurobiological mechanisms will generate a “post-industrial post-informational neurosociety,” where learning and memory will be enhanced to improve competitive advantage in the workplace, sensory abilities will be improved to extend artistic expression, and emotional stability will be increased to improve personal relationships, political opinions, and cultural beliefs (what political or cultural norms are we aiming for?). Bostrom and Sandberg (2009) go further and claim cognitive enhancers could solve societal problems by making people “smarter, wiser, or more creative,” and given “the potentially enormous gains from even moderately effective general cognitive enhancements, this area deserves large-scale funding” (p. 332). In arguably, the most remarkable of the reductive “lobotomy attitude” statements they conclude by saying, “The societal benefits of effective cognitive enhancement may even turn out to be so large and unequivocal that it would be Pareto optimal to subsidize enhancement for the *poor* [my italic] just as the state now subsidizes education” (Bostrom and Sandberg, 2009, p. 334). Ignoring the ample evidence that wealth does not equal intelligence, this is some claim for drugs that lack mechanistic understanding and whose effects are mimicked by chewing gum or taking a nap.

These claims may be loosely based on the scientific approach of neurobiological reductionism, but not on science, and the bounds of sense should negate the science fiction statements and false analogies. Should we, apart from profit and convenience, appeal to pharmacological interventions with limited efficacy and unknown mechanisms (and risks?) over education that we know enhances cognition and has benefits beyond job status and salary in improving overall health and quality of life (Johnston, 2004)? Claiming that pharmacological enhancement and education are equivalent as both cause physiological changes in the brain is another false analogy (Bostrom and Sandberg, 2009): education changes the brain through the gradual integration of experiences in specific neural systems, whereas drugs instantly impose largely unknown global effects on nervous systems.

This hyperbole is balanced by Goswami (2009, p. 182), who in a paper cited only one-tenth as often as Bostrom and Sandberg (2009), considers the scientific basis and the bounds of sense of applying neuroscience to education by saying we “must proceed with caution. We cannot afford to ignore the nature of what is (and is not) possible to measure using current neuroscience techniques when framing our research questions about the brain,” and goes on to say that we should “start small, using the outcome measures that are actually possible given the current state of the art, and then to adapt educational questions to variables that we can meaningfully measure” (i.e., not try to engineer society by cognitively enhancing the poor). Bruer (2002) claims that we do

⁴ <https://www.theguardian.com/society/2007/nov/08/health.lifeandhealth>

not know enough about the relationship between brain physiology and learning to form meaningful links to education, yet these links are promoted. Premature neuroscience translations to education will make the classroom a laboratory. Penicillin again shows that we do not need a complete mechanism for effective interventions (Lobanovska and Pilla, 2017), but penicillin use was based on knowledge of bacterial infections and demonstrated effectiveness, a basis that pharmacological cognitive enhancements lack.

Conclusion

Reductionist analyses that examine component parts to provide mechanistic schemes have been successful in many areas of science, including neuroscience where over several decades experimental tools have allowed increasingly precise molecular and cellular analyses and manipulations that have given insight into various aspects of nervous system function and dysfunction (e.g., the identification of biomarkers in neurology that have supplemented traditional behavioral descriptions; Anthony et al., 2014). Despite this, our success in terms of explanations or understanding of cognition and behavior and the ability to intervene has arguably been limited.

Knowledge of parts, their organization, and the functions they perform can in principle explain any system, including relational and emergent effects, providing that the necessary parts, interactions, and functions are considered. What constitutes necessary and sufficient detail remains debated (see Selverston, 1980 and the debates in the commentaries). Even if this was debate was settled in favor of a reductive approach, reductive explanations are affected by the practical difficulties of the large number of components and interactions to examine in even relatively small systems, their amenability to analysis, and limitations introduced by experimental approaches [e.g., the use of quiescent (non-behaving) and dissected or dissociated preparations]. These issues can lead to components that are less experimentally tractable being ignored for experimental convenience and functionally-relevant aspects like feedback pathways, ephapses, and volume transmission being lost. This can leave explanations based on the information available rather than the information that may be needed.

Explanations can also selectively use available information. In discussing the neuron doctrine, Gold and Stoljar (1999, p. 821) used Kandel's sensory neuron mechanism for associative learning of the gill-withdrawal reflex in *Aplysia* as an exemplar of a psychoneural reduction, saying "we take it to be a sociological fact that Kandel's theory is widely regarded in the neuroscientific community as the best that neuroscience can now offer in the way of explanation of behavior or the mind in fundamental neuroscientific terms." They evaluated Kandel's explanation at some length and concluded that it was not a successful psychoneural reduction because it still relies on psychological concepts. But the claim of a successful neurobiological reduction can be negated on far simpler grounds as it begs the question in

only considering the sensory neurons and ignores known and relatively well-characterized changes in motor neurons and interneurons (see Parker, 2019). A successful neurobiological reduction would require either that the non-sensory changes were shown to be irrelevant to the explanation, or that the relative contributions of all of the effects were determined. This would require significant time and effort given the claimed forbidding complexity of the interneuronal connections (Hawkins et al., 1981), but aspects should not be ignored for convenience.

Highlighting the limitations and challenges of reductive analyses should not be taken as support for the opposing view that lower-level detail is irrelevant and we should instead focus on higher-level computations and representations (Silberstein and Chemero, 2013; Barack and Krakauer, 2021). The latter offer descriptions of population effects that reductive do not usually provide, but they also offer limited explanations (only approximately 30% of the variance in visual cortex responses to natural stimuli can be accounted for by current computational coding models; Bertalmio et al., 2020). One obvious benefit of reductive analyses is to provide detail that can inform and constrain higher-level abstract or phenomenological models. Hodgkin and Huxley (1952, p. 541) cautioned their action potential model, "must not be taken as evidence that our equations are anything more than an empirical description...An equally satisfactory description of the voltage clamp data could no doubt have been achieved with equations of very different form": their model was ultimately supported by molecular analyses of channel properties over three decades later (Catacuzzeno and Franciolini, 2022).

Dichotomies like that between reductionist and representational approaches have stymied various fields (e.g., sensory vs. centrally driven locomotion, and presynaptic vs. postsynaptic expression of LTP; Stuart and Hultborn, 2008; Lomo, 2018). The need to consider effects at multiple levels has been raised repeatedly. Bernard (1927) wrote, "Admitting that vital phenomena rest upon physico-chemical activities, which is the truth, the essence of the problem is not thereby cleared up...when we wish to ascribe to a physiological quality its value and true significance, we must always refer to this whole." Sherrington made a similar claim: although he recognized the importance of relational interactions in nervous systems in calling reflexes a "convenient fiction," he highlighted the benefits of a reductive approach in saying "it is helpful in analyzing complex reflexes to separate from them components which we may consider apart and therefore treat as though they were simple reflexes" (Sherrington, 1906, p8). From this reductive approach, he provided functional evidence for synapses and rules of synaptic integration still relevant today. Bullock also followed a reductionist approach in his neuroethological analyses (see Zupanc and Zupanc, 2008): he was the first to examine synaptic transmission using paired recordings in the squid and identified electrical synapses in the crustacean cardiac ganglion. But he also examined sensory and motor principles at behavioral levels, using a neuroethological focus on the species-dependent differences that reflected

adaptations to ecological and behavioral requirements. Bullock criticized the “mutual disparagement” between single neuron and population approaches, saying “Each of these approaches is a window and a quite inadequate one. We need both and the combination of the two and still others to untangle this most complex of known systems” (Bullock, 1995, p. 231).

In addition to being counterproductive, debate, or for Bullock the mutual disparagement, over the relative merits of representational/computational and reductionist approaches seems premature given the lack of necessary detail and clarity of definitions. For example, definitions of representations vary (Barack and Krakauer, 2021) and numerous abstract computational terms and analytical approaches are used that have only tangential links to each other and to neurobiology (Silberstein and Chemero, 2013). Neurobiological details need to be considered to prevent computational aspects becoming “descriptive conveniences” (Warren, 2012). Bennett and Hacker (2003, p. 147) wrote, “To say that the mind has ‘access’ to the ‘internal representation’ produced by the brain is no less mysterious than the Cartesian claim that the mind has access to an image on the pineal gland.” Does it matter that a synapse is a complicated molecular system of multiple protein–protein interactions (Wilhelm et al., 2014) rather than a number in a matrix: it probably does. Conversely, claims of mechanistic explanations of cognitive functions and behaviors from neurobiological analyses seem premature as they are predicated on data that fails to satisfy the minimal neurobiological criteria for understanding (e.g., Selverston, 1980), criteria that need to be updated and expanded to include variable relational effects in heterarchical systems, ephaptic fields, volume transmission, and transmitter interactions.

Claiming that cognitive explanations need to account for state spaces across many spatio-temporal scales (e.g., Churchland and Churchland, 1990; Barack and Krakauer, 2021) repeats Lashley’s dilemma (see above; Lashley, 1942). Whether the non-local relational aspects discussed here could help link representations and state spaces across different spatio-temporal scales, as Bullock (1959) suggested, remains an open question given the limited consideration of these phenomenon. Ephapses will provide spatially and temporally varying activity in neuronal populations, while volume transmission and transmitter interactions will allow spatially and temporally varying context-dependent effects driven by changes in internal or external conditions (e.g., sensory or cortical activity evoking modulator release from brainstem modulatory systems). These effects should also be considered by those who claim functions “bottom-out” in genes, molecules, neurotransmitters and neurons. Kaplan and Craver (2011, p. 603) write, “we oppose strong dynamicist and functionalist views according to which mathematical and computational models can explain a phenomenon without embracing commitments about the causal mechanisms,” but the same applies to mechanistic views that fail to embrace known mechanisms that alter simple mechanistic views and complicate causal claims.

Placing representational or computational aspects in neurobiological terms is not impossible: a visual receptive field is a representation of external space that can be reduced, although not yet completely, to the connectivity of retinal neurons; analyses of synaptic information transfer consider representational aspects in neurobiological terms (Laughlin et al., 1998), and graph theoretical approaches group neurons into functional assemblies or motifs (Hadjiabadi and Soltesz, 2022). While the latter are presented as novel insights, these motifs have been considered in neurobiology for many years albeit under the original term of building-blocks (Getting, 1989). Despite claims that the identification of an anatomical motif can predict function (Morgan and Lichtman, 2013), we know from reductive analyses that this is not possible from identification of a motif alone: Elson et al. (2002) showed that a single two-neuron motif can generate alternating or synchronous activity depending on the functional properties of their connections. But by combining computational approaches with connectomic data and imaging cell populations at single cell resolution (e.g., zebrafish or hippocampal slices) links are now being made between single cell and population effects (see Hadjiabadi and Soltesz, 2022).

Linking lower and higher-level effects nevertheless remains the major open question in neuroscience. Claims to Kuhnian paradigm shifts and scientific revolutions (Kuhn, 1962), which are generally rare events, are frequently made in neuroscience (Parker, 2018). These claims could, in principle reflect genuine revolutionary advances; a reflection of the pre-paradigm state as neuroscience tries to find its optimal approach from among the various reductionist or representational approaches suggested; or evidence of a field in a scientific crisis as claimed or promised explanations and interventions have failed to materialize (Parker, 2019). A scientific revolution does not occur when current views face anomalies (cf Barack and Krakauer, 2021), anomalies can instead entrench views, but when an alternative approach is offered that overcomes the addresses the issues that have held a field back. Attention focused on the relational aspects originally highlighted by Lashley (1942), McCulloch (1945), and Bullock (1959) may provide insight that suggests alternatives to current paradigms and dichotomies that move the field forward.

Author contributions

The author confirms being the sole contributor of this work and has approved it for publication.

Conflict of interest

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

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated

References

- Ackerman, P. L., and Kanfer, R. (2009). Test length and cognitive fatigue: an empirical examination of effects on performance and test-taker reactions. *J. Exp. Psychol. Appl.* 15, 163–181. doi: 10.1037/a0015719
- Aizawa, K., and Gillett, C. (2009). "Levels, individual variation, and massive multiple realization in neurobiology," in *The Oxford Handbook of Philosophy and Neuroscience*. ed. J. Bickle (Oxford: Oxford University Press).
- Alberts, B. (1998). The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cells* 92, 291–294.
- Altman, J. (2011). "The discovery of adult mammalian neurogenesis," in *Neurogenesis in the Adult Brain I: Neurobiology*. eds. T. Seki, K. Sawamoto, J. Parent and A. Alvarez-Buylla (Japan: Springer), 3–46.
- Anderson, P. (1972). More is different. *Science* 177, 393–396.
- Anderson, M. L. (2010). Neural reuse: A fundamental organizational principle of the brain. *Behavioral and Brain Sciences* 33, 245–266.
- Anonymous (1970). Effects of sexual activity on beard growth in man. *Nature* 226, 869–870.
- Ansari, D., and Coch, D. (2006). Bridges over troubled waters: education and cognitive neuroscience. *Trends Cogn. Sci.* 10, 146–151. doi: 10.1016/j.tics.2006.02.007
- Anthony, K., Arechavala-Gomez, V., Taylor, L. E., Vulin, A., Kaminoh, Y., Torelli, S., et al. (2014). Dystrophin quantification. *Neurology* 83, 2062–2069. doi: 10.1212/WNL.0000000000001025
- Aradi, I., and Soltesz, I. (2002). Modulation of network behavior by changes in variance in interneuronal properties. *J. Physiol. Lond.* 538, 227–251. doi: 10.1113/jphysiol.2001.013054
- Bachmann, C. J., Wijlaars, L. P., Kalverdijk, L. J., Burcu, M., Glaeske, G., Schuiling-Veninga, C. C. M., et al. (2017). Trends in ADHD medication use in children and adolescents in five western countries, 2005–2012. *Eur. Neuropsychopharmacol.* 27, 484–493. doi: 10.1016/j.euroneuro.2017.03.002
- Barack, D. L., and Krakauer, J. W. (2021). Two views on the cognitive brain. *Nat. Rev. Neurosci.* 22, 359–371. doi: 10.1038/s41583-021-00448-6
- Barlow, H. B. (1972). Single units and sensation: a neuron doctrine for perceptual psychology? *Perception* 1, 371–394.
- Barlow, H. (1990). The mechanical mind. *Annu. Rev. Neurosci.* 13, 15–24. doi: 10.1146/annurev.ne.13.030190.000311
- Bassett, D. S., Greenfield, D. L., Meyer-Lindenberg, A., Weinberger, D. R., Moore, S. W., and Bullmore, E. T. (2010). Efficient physical embedding of topologically complex information processing networks in brains and computer circuits. *PLoS Comput. Biol.* 6:e1000748. doi: 10.1371/journal.pcbi.1000748
- Bechtel, W. (2002). Decomposing the brain: a long-term pursuit. *Brain Mind* 3, 229–242. doi: 10.1023/A:1019980423053
- Bechtel, W. (2007). "Reducing psychology while maintaining its autonomy via mechanistic explanation," in *The Matter of the Mind: Philosophical Essays on Psychology, Neuroscience and Reduction*. eds. M. Schouten and H. Looren de Jong (Oxford: Basil Blackwell).
- Benedetti, F. (2007). Placebo and endogenous mechanisms of analgesia. *Handb. Exp. Pharmacol.* 177, 393–413. doi: 10.1007/978-3-540-33823-9_14
- Bennett, M., and Hacker, P. (2003). *Philosophical Foundations of Neuroscience*. London, UK: Wiley-Blackwell.
- Bernard, C. (1927). *An Introduction to the Study of Experimental Medicine Translated by HC Greene*. Pennsylvania, USA: Macmillan & Co.
- Berry, M., and Pentreath, V. (1976). Criteria for distinguishing between monosynaptic and polysynaptic transmission. *Brain Res.* 105, 1–20.
- Bertalmio, M., Gomez-Villa, A., Martin, A. N., Vazquez-Corral, J., Kane, D., and Malo, J. (2020). Evidence for the intrinsically nonlinear nature of receptive fields in vision. *Sci. Rep.* 10:16277. doi: 10.1038/s41598-020-73113-0
- Bezard, E., Gross, C., and Brochie, J. (2004). Presymptomatic compensation in Parkinson's disease is not dopamine-mediated. *Trends Neurosci.* 26, 215–221. doi: 10.1016/S0166-2236(03)00038-9
- Bickle, J. (2003). *Philosophy and Neuroscience: A Ruthlessly Reductive Account*. Netherlands: Springer.
- Bickle, J., and Parker, D. (2022). "Revolutions in 'wet' neurobiology," in *The SAGE Handbook of Cognitive and Systems Neuroscience*. ed. G. Boyle (Abingdon, UK: Routledge).
- Blakemore, C. (2000). Achievements and challenges of the decade of the brain. *Euro. Brain* 2, 1–4.
- Bliss, T. V. P., Collingridge, G. L., Morris, R. G. M., and Reymann, K. G. (2018). Long-term potentiation in the hippocampus: discovery, mechanisms and function. *e-Neuroforum* 24, A103–A120. doi: 10.1515/nf-2017-A059
- Bostrom, N., and Sandberg, A. (2009). Cognitive enhancement: methods, ethics, regulatory challenges. *Sci. Eng. Ethics* 15, 311–341. doi: 10.1007/s11948-009-9142-5
- Braganza, O., and Beck, H. (2018). The circuit motif as a conceptual tool for multilevel neuroscience. *Trends Neurosci.* 41, 128–136. doi: 10.1016/j.tins.2018.01.002
- Breggin, P. (1993). *Toxic Psychiatry*. London; Fontana, CA: Flamingo.
- Breton-Provencher, V., Drummond, G. T., Feng, J., Li, Y., and Sur, M. (2022). Spatiotemporal dynamics of noradrenaline during learned behavior. *Nature* 606, 732–738. doi: 10.1038/s41586-022-04782-2
- Brezina, V. (2010). Beyond the wiring diagram: signaling through complex neuromodulator networks. *Philos. Trans. Roy. Soc. London B. Biol. Sci.* 365, 2363–2374. doi: 10.1098/rstb.2010.0105
- Bruer, J. T. (2002). Avoiding the pediatrician's error: how neuroscientists can help educators (and themselves). *Nat. Neurosci.* 5, 1031–1033. doi: 10.1038/nn934
- Bullock, T. H. (1959). Neuron doctrine and electrophysiology. *Science* 129, 997–1002. doi: 10.1126/science.129.3355.997
- Bullock, T. H. (1976). "In search of principles in neural integration," in *Simple Networks and Behavior*. ed. J. D. Fentress (Sunderland: Sinauer Assoc), 52–60.
- Bullock, T. H. (1981). Spikeless neurons: where do we go from here? in *Neurons Without Impulses*. (eds.) A. Roberts and Bush, B. M. H. (Cambridge, UK: Cambridge Univ. Press), 269–284.
- Bullock, T. H. (1995). Neural integration at the mesoscopic level: the advent of some ideas in the last half century. *J. Hist. Neurosci.* 4, 216–235.
- Cacioppo, J., and Berntson, G. (1992). Social psychological contributions to the decade of the brain. *Am. Psychol.* 47, 1019–1028.
- Carrera, E., and Tononi, G. (2014). Diaschisis: past, present, future. *Brain* 137, 2408–2422. doi: 10.1093/brain/awu101
- Catacuzzeno, L., and Franciolini, F. (2022). The 70-year search for the voltage sensing mechanism of ion channels. *J. Physiol.* 600, 3227–3247. doi: 10.1113/jp282780
- Changeux, J.-P. (1997). *Neuronal Man: The Biology of Mind*. Princeton: Princeton University Press.
- Chiel, H., and Beer, R. (1997). The brain has a body: adaptive behavior emerges from interactions of nervous system, body and environment. *Trends Neurosci.* 20, 553–557. doi: 10.1016/S0166-2236(97)01149-1
- Churchland, P. M., and Churchland, P. S. (1990). "Intertheoretic reduction: A neuroscientist's field guide. The neurosciences 2:249–56. [TS] (1994) Intertheoretic reduction: A neuroscientist's field guide," in *The Mindbody Problem*. eds. R. Warner and T. Szubka (Oxford, UK: Blackwell)
- Cipriani, A., Furukawa, T. A., Salanti, G., Chaimani, A., Atkinson, L. Z., Ogawa, Y., et al. (2018). Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: a systematic review and network meta-analysis. *Lancet* 391, 1357–1366. doi: 10.1016/S0140-6736(17)32802-7
- Cohen, A. (1992). The role of heterarchical control in the evolution of central pattern generators. *Brain Behav. Evol.* 40, 112–124. doi: 10.1159/000113907
- Coleman, M. (1971). Serotonin concentrations in whole blood of hyperactive children. *J. Pediatr.* 78, 985–990.
- Crabbe, J. C., Wahlsten, D., and Dudek, B. C. (1999). Genetics of mouse behavior: interactions with laboratory environment. *Science* 284, 1670–1672. doi: 10.1126/science.284.5420.1670

- Craver, C. (2007). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford: Clarendon Press.
- Crick, F. (1988). *What Mad Pursuit*. London, UK: Penguin.
- Crick, F. (1994). *The Astonishing Hypothesis*. New York, NY: Scribners.
- Cuijpers, P., Sijbrandij, M., Koole, S. L., Andersson, G., Beekman, A. T., and Reynolds, C. F. 3rd. (2013). The efficacy of psychotherapy and pharmacotherapy in treating depressive and anxiety disorders: a meta-analysis of direct comparisons. *World Psychiatry* 12, 137–148. doi: 10.1002/wps.20038
- Daaka, Y., Luttrell, L., and Lefkowitz, R. (1997). Switching of the coupling of the B2-adrenergic receptor to different G proteins by protein kinase A. *Nature* 390, 88–91. doi: 10.1038/36362
- Davidson, J. (2016). Bayard Holmes (1852–1924) and Henry cotton (1869–1933): surgeon-psychiatrists and their tragic quest to cure schizophrenia. *J. Med. Biogr.* 24, 550–559. doi: 10.1177/0967772014552746
- Davis, G., and Bezprozvany, I. (2001). Maintaining the stability of neural function: a homeostatic hypothesis. *Annu. Rev. Neurosci.* 63, 847–869. doi: 10.1146/annurev.physiol.63.1.847
- Davis, G., Troxel, M., Kohler, V., Grossman, E., and McClellan, A. (1993). Time course of locomotor recovery and functional regeneration in spinal-transected lamprey: kinematics and electromyography. *Exp. Brain Res.* 97, 83–95.
- Dreyfus, H. L. (2012). A history of first step fallacies. *Mind. Mach.* 22, 87–99. doi: 10.1007/s1023-012-9276-0
- Dudai, Y. (2004). “The neurosciences: the danger that we will think that we have understood it all,” in *The New Brain Sciences: Perils and Prospects*. eds. S. Rose and D. Rees (Cambridge: Cambridge University Press).
- Eckstein, N., Bates, A. S., Du, M., Hartenstein, V., Jefferis, G. S. X. E., and Funke, J. (2020). Neurotransmitter classification from electron microscopy images at synaptic sites in drosophila. bioRxiv [Preprint]. doi: 10.1101/2020.06.12.148775v1
- Edelman, G. (1989). *The Remembered Present: A Biological Theory of Consciousness*. New York: Basic Books.
- Edgerton, V., Leon, R., Harkema, S., Hodgson, J., London, N., Reinkensmeyer, D., et al. (2001). Retraining the injured spinal cord. *J. Physiol.* 533, 15–22. doi: 10.1111/j.1469-7793.2001.0015b.x
- Elson, R. C., Selverston, A. I., Abarbanel, H. D. I., and Rabinovich, M. I. (2002). Inhibitory synchronization of bursting in biological neurons: dependence on synaptic time constant. *J. Neurophysiol.* 88, 1166–1176. doi: 10.1152/jn.2002.88.3.1166
- Endicott, R. (2001). Post-structuralist angst – critical notice: John Bickle, *Psycheal reduction: the new wave*. *Philos. Sci.* 68, 377–393. doi: 10.1086/392890
- Faber, D., and Pereda, A. (2018). Two forms of electrical transmission between neurons. *Front. Mol. Neurosci.* 11:427. doi: 10.3389/fnmol.2018.00427
- Fahoum, S.-R. H., and Blitz, D. M. (2021). Neuronal switching between single-and dual-network activity via modulation of intrinsic membrane properties. *J. Neurosci.* 41, 7848–7866. doi: 10.1523/JNEUROSCI.0286-21.2021
- Farah, M. J., Illes, J., Cook-Deegan, R., Gardner, H., Kandel, E., King, P., et al. (2004). Neurocognitive enhancement: what can we do and what should we do? *Nat. Rev. Neurosci.* 5, 421–425. doi: 10.1038/nrn1390
- Flower, R. (2004). Lifestyle drugs: pharmacology and the social agenda. *Trends Pharmacol. Sci.* 25, 182–185. doi: 10.1016/j.tips.2004.02.006
- Frank, C., Kennedy, M., Goold, C., Marek, K., and Davis, G. (2006). Mechanisms underlying the rapid induction and sustained expression of synaptic homeostasis. *Neuron* 52, 663–677. doi: 10.1016/j.neuron.2006.09.029
- Geake, J., and Cooper, P. (2003). Cognitive neuroscience: implications for education? *Westminst. Stud. Educ.* 26, 7–20. doi: 10.1080/0140672030260102
- Getting, P. (1989). Emerging principles governing the operation of neural networks. *Annu. Rev. Neurosci.* 12, 185–204.
- Glanzman, D. L. (2010). Common mechanisms of synaptic plasticity in vertebrates and invertebrates. *Curr. Biol.* 20, R31–R36. doi: 10.1016/j.cub.2009.10.023
- Gold, I., and Stoljar, D. (1999). A neuron doctrine in the philosophy of neuroscience. *Behav. Brain Sci.* 22, 809–869.
- Golowasch, J., Goldman, M., Abbott, L., and Marder, E. (2002). Failure of averaging in the construction of a conductance-based neuron model. *J. Neurophysiol.* 87, 1129–1131. doi: 10.1152/jn.00412.2001
- Goswami, U. (2009). Mind, brain, and literacy biomarkers as usable knowledge for education. *Mind Brain Educ.* 3, 176–184. doi: 10.1111/j.1751-228X.2009.01068.x
- Gould, E. (2007). How widespread is adult neurogenesis in mammals? *Nat. Rev. Neurosci.* 8, 481–488. doi: 10.1038/nrn2147
- Greenberg, I., and Manor, Y. (2005). Synaptic depression in conjunction with A-current channels promote phase Constancy in a rhythmic network. *J. Neurophysiol.* 93, 656–677. doi: 10.1152/jn.00640.2004
- Hadjiabadi, D., and Soltesz, I. (2022). From single-neuron dynamics to higher-order circuit motifs in control and pathological brain networks. *J. Physiol.* doi: 10.1113/JP282749, (Epub ahead of print)
- Hanahan, D., and Weinberg, R. A. (2000). The hallmarks of cancer. *Cells* 100, 57–70.
- Hasselmo, M. E. (1995). Neuromodulation and cortical function: modeling the physiological basis of behavior. *Behav. Brain Res.* 67, 1–27. doi: 10.1016/0166-4328(94)00113-T
- Hawkins, R. D., Castellucci, V. F., and Kandel, E. R. (1981). Interneurons involved in mediation and modulation of gill-withdrawal reflex in Aplysia. II. Identified neurons produce heterosynaptic facilitation contributing to behavioral sensitization. *J. Neurophysiol.* 45, 315–328. doi: 10.1152/jn.1981.45.2.315
- Healy, D. (2015). Serotonin and depression. *Br. Med. J.* 350:h1771. doi: 10.1136/bmj.h1771
- Hodgkin, A., and Huxley, A. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol.* 117, 500–544. doi: 10.1113/jphysiol.1952.sp004764
- Horwitz, B. (2003). The elusive concept of brain connectivity. *NeuroImage* 19, 466–470. doi: 10.1016/s1053-8119(03)00112-5
- Howard-Jones, P. (2008). Philosophical challenges for researchers at the Interface between neuroscience and education. *J. Philos. Educ.* 42, 361–380. doi: 10.1111/j.1467-9752.2008.00649.x
- Hubel, D. (1974). “Neurobiology: a science in need of a Copernicus,” in *The Heritage of Copernicus: Theories “Pleasing to the Mind”*. ed. J. Neyman (Cambridge: MIT Press).
- Huttenlocher, P. (2002). *Neural Plasticity: The Effects of Environment on the Development of the Cerebral Cortex*. Cambridge, MA: Harvard University Press
- Ingo, B., and Love, A. (2022). Reductionism in biology. The Stanford Encyclopedia of philosophy. Available at: <https://plato.stanford.edu/archives/sum2022/entries/reduction-biology/> Summer 2022 Edition.
- Ito, M. (2006). Cerebellar circuitry as a neuronal machine. *Prog. Neurobiol.* 78, 272–303. doi: 10.1016/j.pneurobio.2006.02.006
- Jacobs, B., and Fornal, C. (1993). 5-HT and motor control: a hypothesis. *Trends Neurosci.* 16, 346–351. doi: 10.1016/0166-2236(93)90090-9
- Jia, Y., and Parker, D. (2016). Short-term synaptic plasticity at Interneuronal synapses could sculpt rhythmic motor patterns. *Front. Neur. Circuit.* 10:4. doi: 10.3389/fncir.2016.00004
- Johnston, M. (2004). Clinical disorders of brain plasticity. *Brain and Development* 26, 73–80. doi: 10.1016/S0387-7604(03)00102-5
- Jonas, E., and Kording, K. P. (2017). Could a neuroscientist understand a microprocessor? *PLoS Comput. Biol.* 13:e1005268. doi: 10.1371/journal.pcbi.1005268
- Josselyn, S. A., Kohler, S., and Frankland, P. W. (2015). Finding the engram. *Nat. Rev. Neurosci.* 16, 521–534. doi: 10.1038/nrn4000
- Kandel, E. (1998). A new intellectual framework for psychiatry. *Am. J. Psychiatry* 155, 457–469.
- Kaplan, D., and Craver, C. (2011). The explanatory force of dynamical and mathematical models in neuroscience: A mechanistic perspective. *Philos. Sci.* 78, 601–627. doi: 10.1086/661755
- Kendler, K. (2015). “The dopamine hypothesis of schizophrenia: an updated perspective,” in *Philosophical Issues in Psychiatry III: The Nature and Sources of Historical Change*. eds. K. KS and J. Parnas (New York: Oxford University Press), 283–294.
- Khazan, O. (2014). Pulling teeth to treat mental illness. *The Atlantic*, October 22.
- Kiehn, O., and Kullander, K. (2004). Central pattern generators deciphered by molecular genetics. *Neuron* 41, 317–321. doi: 10.1016/s0896-6273(04)00042-x
- Koch, C. (2012). Modular biological complexity. *Science* 337, 531–532. doi: 10.1126/science.1218616
- Koren, G., and Korn, L. (2021). The use of methylphenidate for cognitive enhancement in young healthy adults: the clinical and ethical debates. *J. Clin. Psychopharmacol.* 41, 100–102. doi: 10.1097/JCP.0000000000001336
- Korshunov, K. S., Blakemore, L. J., and Trombley, P. Q. (2020). Illuminating and sniffing out the Neuromodulatory roles of dopamine in the retina and olfactory bulb. *Front. Cell. Neurosci.* 14:275. doi: 10.3389/fncel.2020.00275
- Krakauer, J. W., Ghazanfar, A. A., Gomez-Marín, A., MacIver, M. A., and Poeppel, D. (2017). Neuroscience needs behavior: correcting a reductionist bias. *Neuron* 93, 480–490. doi: 10.1016/j.neuron.2016.12.041

- Kuhn, T. (1962). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.
- Kupfermann, I., and Weiss, K. (1978). The command neuron concept. *Behav. Brain Sci.* 1, 3–39.
- Kwon, D. (2021). Failure of genetic therapies for Huntington's devastates community. *Nature* 593:180. doi: 10.1038/d41586-021-01177-7
- Laing, R. (1985). *Wisdom, Madness and Folly: The Making of a Psychiatrist*. London: MacMillan.
- Laplace, P. (1902). *A Philosophical Essay on Probabilities (translated by FW Truscott and FL Emory)*. New York: Wiley.
- Lashley, K. S. (1929). *Brain Mechanisms and Intelligence*. Chicago: Appleton-Century-Crofts.
- Lashley, K. S. (1942). "The problem of cerebral organization in vision." in *Biological Symposia, VII, Visual Mechanisms*. Lancaster, UK, Jaques Cattell Press, 301–322.
- Laughlin, S., de Ruyter, R. R., van Steveninck, R. R., and Anderson, J. (1998). The metabolic cost of neural information. *Nat. Neurosci.* 1, 36–41.
- Lehmann, H. (1993). Before they called it psychopharmacology. *Neuropharmacology* 8, 291–303.
- Lehrer, J. (2008). Can a thinking, remembering, decision-making, biologically accurate brain be built from a supercomputer? *SeedMagazine* March 3.
- Lledo, P.-M., Alonso, M., and Grubb, M. S. (2006). Adult neurogenesis and functional plasticity in neuronal circuits. *Nat. Rev. Neurosci.* 7, 179–193. doi: 10.1038/nrn1867
- Lobanovska, M., and Pilla, G. (2017). Penicillin's discovery and antibiotic resistance: lessons for the future? *Yale J. Biol. Med.* 90, 135–145. doi: 10.3390/ph15080942
- Lomo, T. (2018). Discovering long-term potentiation (LTP) – recollections and reflections on what came after. *Acta Physiol.* 222:e12921. doi: 10.1111/apha.12921
- Lynch, Z. (2004). Neurotechnology and society. *Ann. N. Y. Acad. Sci.* 1013, 229–233. doi: 10.1196/annals.1305.016
- Marban, E., Yamagishi, T., and Tomaselli, G. F. (1998). Structure and function of voltage-gated sodium channels. *J. Physiol.* 508, 647–657. doi: 10.1111/j.1469-7793.1998.647bp.x
- Mayr, E. (1988). *Toward a New Philosophy of Biology*. Cambridge: Harvard University Press.
- McCormick, D. (1992). Neurotransmitter actions in the thalamus and cerebral cortex and their role in neuromodulation of thalamocortical activity. *Prog. Neurobiol.* 39, 337–388. doi: 10.1016/0301-0082(92)90012-4
- McCormick, D. A., Nestvogel, D. B., and He, B. J. (2020). Neuromodulation of brain state and behavior. *Annu. Rev. Neurosci.* 43, 391–415. doi: 10.1146/annurev-neuro-100219-105424
- McCoy, D. C., Gonzalez, K., and Jones, S. (2019). Preschool self-regulation and Preacademic skills as mediators of the long-term impacts of an early intervention. *Child Dev.* 90, 1544–1558. doi: 10.1111/cdev.13289
- McCulloch, W. S. (1945). A heterarchy of values determined by the topology of nervous nets. *Bull. Math. Biophys.* 7, 89–93. doi: 10.1007/BF02478457
- McIntyre, C., Savata, M., Goff, L. K.-L., and Vitek, J. (2004). Uncovering the mechanism(s) of action of deep brain stimulation: activation, inhibition, or both. *Clin. Neurophysiol.* 115, 1239–1248. doi: 10.1016/j.clinph.2003.12.024
- Meyrand, P., Simmers, J., and Moulins, M. (1991). Construction of a pattern generating circuit with neurons of different networks. *Nature* 351, 60–63. doi: 10.1038/351060a0
- Middleton, H., and Moncrieff, J. (2019). Critical psychiatry: a brief overview. *BJ Psychiatry Adv.* 25, 47–54. doi: 10.1192/bja.2018.38
- Miranda, A., Presentacion, M. J., and Soriano, M. (2002). Effectiveness of a school-based multicomponent program for the treatment of children with ADHD. *J. Learn. Disabil.* 35, 547–563. doi: 10.1177/00222194020350060601
- Morgan, J. L., and Lichtman, J. W. (2013). Why not connectomics? *Nat. Methods* 10, 494–500. doi: 10.1038/nmeth.2480
- Mott, M. C., Gordon, J. A., and Koroshetz, W. J. (2018). The NIH BRAIN initiative: advancing neurotechnologies, integrating disciplines. *PLoS Biol.* 16:e3000066. doi: 10.1371/journal.pbio.3000066
- Munkholm, K., Paludan-Muller, A. S., and Boesen, K. (2019). Considering the methodological limitations in the evidence base of antidepressants for depression: a reanalysis of a network meta-analysis. *BMJ Open* 9:e024886. doi: 10.1136/bmjopen-2018-024886
- Newton, M. D., Taylor, B. J., Driessen, R. P. C., Roos, L., Cveticic, N., Allyjaun, S., et al. (2019). DNA stretching induces Cas9 off-target activity. *Nat. Struct. Mol. Biol.* 26, 185–192. doi: 10.1038/s41594-019-0188-z
- Nicholson, D. J. (2019). Is the cell really a machine? *J. Theor. Biol.* 477, 108–126. doi: 10.1016/j.jtbi.2019.06.002
- Noble, D. (2012). A theory of biological relativity: no privileged level of causation. *Interface Focus* 2, 55–64. doi: 10.1098/rsfs.2011.0067
- Noble, D., and Boyd, C. (1993). "The challenge of integrative physiology," in *The Logic of Life*. ed. C. B. D. Noble (Oxford: Oxford University Press), 1–14.
- Noble, R., Tasaki, K., Noble, P. J., and Noble, D. (2019). Biological relativity requires circular causality but not symmetry of causation: so, where, what and when are the boundaries? *Front. Physiol.* 10:827. doi: 10.3389/fphys.2019.00827
- Normandin, S. (2007). Claude Bernard and an introduction to the study of experimental medicine: physical Vitalism, dialectic, and epistemology. *J. Hist. Med. Allied Sci.* 62, 495–528. doi: 10.1093/jhmas/jrm015
- Østby, I., Øyehaug, L., Einevoll, G., Nagelhus, E., Plahte, E., Zeuthen, T., et al. (2009). Astrocytic mechanisms explaining neural-activity-induced shrinkage of extraneuronal space. *PLoS Comput. Biol.* 5:e1000272. doi: 10.1371/journal.pcbi.1000272
- Otchy, T., Wolff, S., Rhee, J., Pehlevan, C., Kawai, R., Kempf, A., et al. (2015). Acute off-target effects of neural circuit manipulations. *Nature* 528, 358–363. doi: 10.1038/nature16442
- Parker, D. (2006). Complexities and uncertainties of neuronal network function. *Philos. Trans. Roy. Soc. B. Biol. Sci.* 361, 81–99. doi: 10.1098/rstb.2005.1779
- Parker, D. (2010). Neuronal network analyses: premises, promises and uncertainties. *Philos. Trans. R. Soc. Lond. B* 365, 2315–2328. doi: 10.1098/rstb.2010.0043
- Parker, D. (2017). The lesioned spinal cord is a "new" spinal cord: evidence from functional changes after spinal injury in lamprey. *Front. Neur. Circuit.* 11:84. doi: 10.3389/fncir.2017.00084
- Parker, D. (2018). Kuhnian revolutions in neuroscience: the role of tool development. *Biol. Philos.* 33:17. doi: 10.1007/s10539-018-9628-0
- Parker, D. (2019). Psychoneural reduction: a perspective from neural circuits. *Biol. Philos.* 34:44. doi: 10.1007/s10539-019-9697-8
- Parker, D., and Srivastava, V. (2013). Dynamic systems approaches and levels of analysis in the nervous system. *Front. Physiol.* 4:15. doi: 10.3389/fphys.2013.00015
- Parra, P., Gulyas, A., and Miles, R. (1998). How many subtypes of inhibitory cells in the hippocampus? *Neuron* 20, 983–993.
- Pessoa, L. (2014). Understanding brain networks and brain organization. *Phys Life Rev* 11, 400–435. doi: 10.1016/j.plrev.2014.03.005
- Piasecki, S. D., and Jefferson, J. W. (2004). Psychiatric complications of deep brain stimulation for Parkinson's disease. *J. Clin. Psychiatry* 65, 845–849. doi: 10.4088/JCP.v65n0617
- Pilowsky, P. (2014). Peptides, serotonin, and breathing: the role of the raphe in the control of respiration. *Prog. Brain Res.* 209, 169–189. doi: 10.1016/B978-0-444-63274-6.00009-6
- Plomin, R., Petrill, S., and Cutting, A. (1996). What genetic research on intelligence tells us about the environment. *J. Biosoc. Sci.* 28, 587–606. doi: 10.1017/S0021932000022604
- Price, C. J., and Friston, K. J. (2005). Functional ontologies for cognition: the systematic definition of structure and function. *Cogn. Neuropsychol.* 22, 262–275. doi: 10.1080/02643290442000095
- Prinz, A. (2010). Computational approaches to neuronal network analysis. *Philos. Trans. R. Soc. Lond. B* 365, 2397–2405. doi: 10.1098/rstb.2010.0029
- Prinz, A., Bucher, D., and Marder, E. (2004). Similar network activity from disparate circuit parameters. *Nat. Neurosci.* 7, 1345–1352. doi: 10.1038/nn1352
- Proctor, R. N. (2012). The history of the discovery of the cigarette-lung cancer link: evidentiary traditions, corporate denial, global toll. *Tob. Control.* 21, 87–91. doi: 10.1136/tobaccocontrol-2011-050338
- Queenan, B. N., Ryan, T., Gazzaniga, M., and Gallistel, C. R. (2017). On the research of time past: the hunt for the substrate of memory. *Ann. N. Y. Acad. Sci.* 1396, 108–125. doi: 10.1111/nyas.13348
- Quina, L. A., Walker, A., Morton, G., Han, V., and Turner, E. E. (2020). GAD2 expression defines a class of excitatory lateral Habenula neurons in mice that project to the raphe and Pontine Tegmentum. *eNeuro* 20:7. doi: 10.1523/ENEURO.0527-19.2020
- Rampon, C., Jiang, C., Dong, H., Tang, Y.-P., Lockhart, D., Schultz, P., et al. (2000). Effects of environmental enrichment on gene expression in the brain. *PNAS* 97, 12880–12884. doi: 10.1073/pnas.97.23.12880
- Rashevsky, N. (1954). Topology and life: in search of general mathematical principles in biology and sociology. *Bull. Math. Biophys.* 16, 317–348. doi: 10.1007/BF02484495
- Reynolds, A. (2007). The cell's journey: from metaphorical to literal factory. *Endeavour* 31, 65–70. doi: 10.1016/j.endeavour.2007.05.005

- Romero, F. (2015). Why there isn't inter-level causation in mechanisms. *Synthese* 192, 3731–3755. doi: 10.1007/s11229-015-0718-0
- Rose, S., Kamin, L., and Lewontin, R. (1985). *Not in Our Genes: Biology, Ideology and Human Nature*. New York, NY: Pantheon Books.
- Sanderson, K. (2021). High-profile autism genetics project paused amid backlash. *Nature* 598, 17–18. doi: 10.1038/d41586-021-02602-7
- Schallert, T., Bland, S., Leasure, J., Tillerson, J., Gonzales, R., Williams, L., et al. (2000). "Motor rehabilitation, use-related neural events, and reorganization of the brain after injury," in *Cerebral Reorganization of Function After Brain Damage*. eds. J. Grafman and H. Levin (New York: Oxford University Press), 145–167.
- Schroeder, C., and Foxe, J. (2005). Multisensory contributions to low-level, 'unisensory' processing. *Curr. Opin. Neurobiol.* 4, 454–458. doi: 10.1016/j.conb.2005.06.008
- Schroter, M., Paulsen, O., and Bullmore, E. T. (2017). Micro-connectomics: probing the organization of neuronal networks at the cellular scale. *Nat. Rev. Neurosci.* 18:131. doi: 10.1038/nrn.2016.182
- Silverston, A. (1980). Are central pattern generators understandable. *Behav. Brain Sci.* 3, 535–571. doi: 10.1017/S0140525X00006580
- Silverston, A. (2010). Invertebrate central pattern generator circuits. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 365, 2329–2345. doi: 10.1098/rstb.2009.0270
- Shapiro, L., and Spaulding, S. (2021). Embodied cognition, the Stanford Encyclopedia of philosophy. Available at: <https://plato.stanford.edu/archives/win2021/entries/embodied-cognition/>
- Shepherd, G. (1991). *Foundations of the Neuron Doctrine*. Oxford: Oxford University Press.
- Sherrington, C. (1906). *The Integrative Action of the Nervous System*. New Haven: Yale University Press.
- Shorter, E. (2021). *The Rise and Fall of the Age of Psychopharmacology*. Oxford: Oxford University Press.
- Signal, T. L., Gander, P. H., Anderson, H., and Brash, S. (2009). Scheduled napping as a countermeasure to sleepiness in air traffic controllers. *J. Sleep Res.* 18, 11–19. doi: 10.1111/j.1365-2869.2008.00702.x
- Silberstein, M., and Chemero, A. (2013). Constraints on localization and decomposition as explanatory strategies in the biological sciences. *Philos. Sci.* 80, 958–970. doi: 10.1086/674533
- Simon, H. (1962). The architecture of complexity. *Proc. Am. Phil. Soc.* 106, 467–482.
- Simon, H. (1969). *The Sciences of the Artificial*. 3rd Edn. Cambridge, MA: The MIT Press.
- Smith, M., and Glass, G. (1977). Meta-analysis of psychotherapy outcome studies. *Am. Psychol.* 32, 752–760.
- Smith-Coggins, R., Howard, S. K., Mac, D. T., Wang, C., Kwan, S., Rosekind, M. R., et al. (2006). Improving alertness and performance in emergency department physicians and nurses: the use of planned naps. *Ann. Emerg. Med.* 48, 596–604.e593. doi: 10.1016/j.annemergmed.2006.02.005
- Soltesz, I. (2006). *Diversity in the Neuronal Machine*. Oxford: Oxford University Press.
- Sporns, O. (2011). *Networks of the Brain*. Cambridge, MA: MIT Press.
- Steward, O., Popovich, P. G., Dietrich, W. D., and Kleitman, N. (2012). Replication and reproducibility in spinal cord injury research. *Exp. Neurol. Special Issue* 233, 597–605. doi: 10.1016/j.expneurol.2011.06.017
- Stuart, D., and Hultborn, H. (2008). Thomas Graham Brown (1882–1965), Anders Lundberg (1920–), and the neural control of stepping. *Brain Res. Rev.* 59, 74–95. doi: 10.1016/j.brainresrev.2008.06.001
- Svensson, E., Aspergis-Schoute, J., Burnstock, G., Nusbaum, M., Parker, D., and Schioth, H. (2019). General principles of neuronal co-transmission: insights from multiple model systems. *Front. Neur. Circuit.* 12:117. doi: 10.3389/fncir.2018.00117
- Swensen, A. M., and Bean, B. P. (2005). Robustness of burst firing in dissociated Purkinje neurons with acute or long-term reductions in sodium conductance. *J. Neurosci.* 25, 3509–3520. doi: 10.1523/JNEUROSCI.3929-04.2005
- Tancredi, L. (2005). *Hardwired Behavior: What Neuroscience Reveals About Morality*. New York: Cambridge University Press
- Thomas, R. (1971). Mass function and Equipotentiality: A reanalysis of Lashley's retention data. *Psychol. Rep.* 27, 899–902.
- Thompson, A., and Wolpaw, J. (2014). Operant conditioning of spinal reflexes: from basic science to clinical therapy. *Front. Integr. Neurosci.* 8:25. doi: 10.3389/fnint.2014.00025
- Tononi, G., Sporns, O., and Edelman, G. (1999). Measures of degeneracy and redundancy in biological networks. *Proc. Natl. Acad. Sci.* 96, 3257–3262. doi: 10.1073/pnas.96.6.3257
- Trudeau, L.-E., and Castellucci, V. (1993). Sensitisation of the gill and siphon withdrawal reflex of Aplysia: multiple sites of change in the neuronal network. *J. Neurophysiol.* 70, 1210–1220. doi: 10.1152/jn.1993.70.3.1210
- Tsaousides, T., and Gordon, W. (2009). Cognitive rehabilitation following traumatic brain injury: assessment to treatment. *Mt Sinai J. Med.* 76, 173–181. doi: 10.1002/msj.20099
- Turkheimer, E., Haley, A., Waldron, M., D'Onofrio, B., and Gottesman, I. I. (2003). Socioeconomic status modifies heritability of IQ in young children. *Psychol. Sci.* 14, 623–628. doi: 10.1046/j.0956-7976.2003.psci_1475.x
- Turrigiano, G., Leslie, K., Desai, N., Rutherford, L., and Nelson, S. (1998). Activity-dependent scaling of quantal amplitude in neocortical neurons. *Nature* 391, 892–896. doi: 10.1038/36103
- Van Riel, R., and Van Gulick, R. (2019). "Scientific reduction," in *The Stanford Encyclopedia of Philosophy*. ed. E. N. Zalta. Available at: <https://plato.stanford.edu/cgi-bin/encyclopedia/archinfo.cgi?entry=scientific-reduction&archive=spr2019>
- Van Strien, M. (2015). Vital instability: life and free will in physics and physiology, 1860–1880. *Ann. Sci.* 72, 381–400. doi: 10.1080/00033790.2014.935954
- Verhage, M., McMahon, H., Ghijzen, W., Boomsa, F., Scholten, G., Wiegant, V., et al. (1991). Differential release of amino acids, neuropeptides and catecholamines from isolated nerve terminals. *Neuron* 6, 517–524. doi: 10.1016/0896-6273(91)90054-4
- Wang, X.-J., and Krystal, J. H. (2014). Computational psychiatry. *Neuron* 84, 638–654. doi: 10.1016/j.neuron.2014.10.018
- Warren, W. H. (2012). Does this computational theory solve the right problem? Marr, Gibson, and the goal of vision. *Perception* 41, 1053–1060. doi: 10.1068/p7327
- Warren, J. B. (2020). The trouble with antidepressants: why the evidence overplays benefits and underplays risks: an essay by John B Warren. *Br. Med. J.* 370:m3200. doi: 10.1136/bmj.m3200
- Wilhelm, B. G., Mandad, S., Trukenbrodt, S., Kröhnert, K., Schäfer, C., Rammner, B., et al. (2014). Composition of isolated synaptic boutons reveals the amounts of vesicle trafficking proteins. *Science* 344, 1023–1028. doi: 10.1126/science.1252884
- Wilkinson, L., Scholey, A., and Wesnes, K. (2002). Chewing gum selectively improves aspects of memory in healthy volunteers. *Appetite* 38, 235–236. doi: 10.1006/appe.2002.0473
- Wimsatt, W. C. (2006). Reductionism and its heuristics: making methodological reductionism honest. *Synthese* 151, 445–475. doi: 10.1007/s11229-006-9017-0
- Ylikoski, P. K. (2009). "The illusion of depth of understanding in science," in *Scientific Understanding: Philosophical Perspectives*. eds. H. W. S. Leonelli and K. Eigner (Pittsburgh, PA: University of Pittsburgh press), 100–119.
- Yuste, R. (2008). Circuit neuroscience: the road ahead. *Front. Neurosci.* 2, 6–9. doi: 10.3389/neuro.01.017.2008
- Zeki, S. (1993). *A Vision of the Brain*. Oxford: Blackwell Scientific Publications.
- Zupanc, G. K. H., and Zupanc, M. M. (2008). Theodore H. Bullock: pioneer of integrative and comparative neurobiology. *J. Comp. Physiol. A.* 194:119. doi: 10.1007/s00359-007-0286-y



OPEN ACCESS

EDITED BY

Antonino Raffone,
Sapienza University of Rome, Italy

REVIEWED BY

John Bickle,
Mississippi State University, United States
Gualtiero Piccinini,
University of Missouri–St. Louis, United States
Daniel C. Burnston,
Tulane University, United States

*CORRESPONDENCE

Mark Couch
✉ mark.couch@shu.edu

RECEIVED 02 July 2022

ACCEPTED 05 April 2023

PUBLISHED 25 April 2023

CITATION

Couch M (2023) Clarifying the relation between
mechanistic explanations and reductionism.
Front. Psychol. 14:984949.
doi: 10.3389/fpsyg.2023.984949

COPYRIGHT

© 2023 Couch. This is an open-access article
distributed under the terms of the [Creative
Commons Attribution License \(CC BY\)](#). The
use, distribution or reproduction in other
forums is permitted, provided the original
author(s) and the copyright owner(s) are
credited and that the original publication in this
journal is cited, in accordance with accepted
academic practice. No use, distribution or
reproduction is permitted which does not
comply with these terms.

Clarifying the relation between mechanistic explanations and reductionism

Mark Couch*

Department of Philosophy, Seton Hall University, South Orange, NJ, United States

The topic of mechanistic explanation in neuroscience has been a subject of recent discussion. There is a lot of interest in understanding what these explanations involve. Furthermore, there is disagreement about whether neurological mechanisms themselves should be viewed as reductionist in nature. In this paper I will explain how these two issues are related. I will, first, describe how mechanisms support a form of antireductionism. This is because the mechanisms that exist should be seen as involving part-whole relations, where the behavior of a whole is more than the sum of its parts. After this, I will consider mechanistic explanations and how they can be understood. While some people think the explanations concern existing entities in the world, I will argue that we can understand the explanations by viewing them in terms of arguments. Despite the fact that it is possible to understand mechanistic explanations in this manner, the antireductionist point remains.

KEYWORDS

mechanistic explanation, reduction, wholes, neuroscience, action potential

1. Introduction

The topic of mechanistic explanations has been an issue of recent interest among philosophers and scientists. It is evident to many researchers that an appeal to mechanisms plays an important role in the sciences. For instance, neuroscientists have explained the signaling by the action potential in the neuron in terms of the physical mechanism that underlies this phenomenon. The action potential is taken to be a result of the components and their behaviors in the neuron that give rise to this distinctive capacity. Furthermore, we can understand the action potential not in terms of any individual component, but as in some way a product of a set of components working together in an organized manner. In this way the phenomenon can be viewed as a higher-level behavior of a neurological mechanism that is not reducible to its lower-level components. While the behavior of the mechanism cannot be reduced to the individual components, it is still dependent upon them.

Just how to think of mechanisms like this and how they should be understood is the topic of this paper. My aim will be to describe how we should think about mechanistic explanations and how this relates to reductionism. After introducing the subject in this first part, I will go on in section 2 to describe how I think we should understand mechanisms. I will offer an account of mechanisms that explains the features they have, including the idea that mechanisms should be viewed as wholes that are made from a collection of parts. In section 3, I will explain why this is a nonreductive way of thinking about mechanisms when considered in terms of how mechanisms exist in the world. In section 4, I will develop this by discussing a variety of reasons for why mechanisms are nonreductive. After this in section 5, I will turn from mechanisms as they exist to the notion of mechanistic explanation and describe how this too should

be understood. In my view the notion of explanation should be understood as including both representational features and ontological features that are needed for characterizing mechanisms. I will explain why this view is in contrast to other views that are more ontologically focused. Furthermore, I will offer a view of the explanations which takes them to be expressible in terms of arguments that consist of statements. While this view is not as common as it once was I think it can still be useful. In section 6, I will describe the implications of this way of thinking about mechanistic explanation for the notion of reductionism, and suggest that the antireductionist view of mechanisms described before is consistent with this perspective. In the last section 7 I will draw some conclusions for how these two issues are related.

2. Understanding mechanisms

It will help to begin with an account of mechanisms and how they should be understood. Here I am talking about mechanisms themselves and, as we might say, how they exist in nature.

While the notion of a mechanism is commonly appealed to in the sciences it is not entirely clear how this notion should be analyzed. There are different ways that people have offered for thinking about this. These different accounts sometimes emphasize different features, or include subtle differences to note about what makes something a mechanism. Since this is not the place to review these discussions in detail what I will do is begin with an account that I think captures the main features that need to be included. This is a way of thinking about mechanisms that has been presented by [Bechtel and Abrahamsen \(2005\)](#) and is often appealed to by others. As they write, “A mechanism is a structure performing a function in virtue of its component parts, component operations, and their organization. The orchestrated functioning of the mechanism is responsible for one or more phenomena” (423). To understand this characterization of a mechanism, we will need to explain each of the notions it mentions.

We can begin with the notion of a “function,” which can be understood in different ways that need to be distinguished from each other. I have said that a mechanism is a structure that consists of components working together to produce a behavior of the mechanism. This behavior is what I mean by the functioning of the mechanism. The notion of functioning at work here refers to what is sometimes called the “causal role” exhibited by the mechanism ([Cummins, 1975](#)), or what I will refer to as its “behavior” ([Glennan, 2017](#), p. 24). For instance, because the behavior of the nerve cell is to transmit electrical signals through the neural system we say that “it functions to transmit electrical signals.” Notice that with this approach to the notion we are not including any biological purposes or goals that are due to the evolutionary history of a mechanism. The evolutionary notion of function is important in those disciplines concerned with why a trait evolved due to some kind of selective pressures, but this sort of notion is distinct from the one we are concerned with. We can describe the behavior of the nerve cell in the neural system independently from saying anything about its evolutionary history.

One thing to note is that there’s an ambiguity that occurs in how we talk about behaviors, since we sometimes talk about the performance of a behavior, or the capacity to perform a behavior. The

performance of a behavior involves its actual occurrence, while a capacity concerns the presence of an ability that can be manifested. We say, for instance, that the nerve cell has the capacity to produce a signal in the nervous system even in its resting state. I will follow Cummins in allowing that both kinds of notions should be included in our account.¹

After this the next notion to consider concerns the “components” of a mechanism and their behaviors; these can be understood to be the parts of a mechanism that contribute to its operation. A mechanism will typically contain a number of parts but not all of these will serve as working components. An example of this would be chemicals introduced into a neuron that do not affect its operation, which are in a sense “a part” contained within the brain. But these are not working parts whose behaviors help them to contribute to a behavior of the mechanism. The notion of a component refers to those parts within the mechanism that contribute to its behavior, not those parts which are merely present in the system in some way. The nerve cells in the brain that contribute to the signaling system will count as components in the system by this criterion.

The last notion to consider in characterizing a mechanism is the “organization” of the components. A mechanism is not merely a set of components taken by themselves, but concerns a set of components that have been organized somehow to produce a behavior. The components of a nerve cell do not produce the cell’s overall behavior individually, but work together in an organized manner to produce the signals. This organization of the components should be understood to include things like the causal, spatial, and temporal organization they exhibit. For instance, the physical events that constitute the action potential in the nerve cell concern the signal, which involves a sequence of steps, beginning with the opening of channels, an influx of ions across the cell membrane, a change in the resting potential, and then a response signal. Each of these events occurs in order and explaining the behavior of the cell requires describing how the components present behave in an organized way. So this is an important feature of the components to include in describing the features of the mechanism.

Understood this way, a mechanism should be viewed as a set of components whose coordinated behavior results in the behavior of the system as a whole. In this sense, a mechanism consists of a collection of interacting parts that underlies a particular behavior. This notion of a mechanism appears to play a central role in fields like neuroscience, which is concerned with investigating the systems of the brain and how they underlie our mental capacities. An understanding of our mental capacities leads researchers to be interested in the details of the mechanisms and how they should be understood. Given their role, it is important to be clear about the features of mechanisms and what their study can teach us about this area of the sciences.

1 A related clarification to make is that in describing mechanisms I will talk of entities and properties, rather than (as some prefer) entities and activities. I think that activity language can be cashed out in terms of the manifestation of capacities when understood properly (*cf.* [Psillos, 2004](#), p. 311), though I do not think this issue will be important in what follows. For discussion of this topic one can see [Kaiser \(2018\)](#).

3. Antireductionism in mechanisms

Up to this point, I have described the behavior of a mechanism as a whole in relation to the behavior of its components. I need to say something more about this issue, since our concern is with trying to understand what the study of mechanisms tells us about reductionism in the area of the sciences we are concerned with. Let us begin with the idea that mechanisms as a whole are constituted by the components that make them up. The components are the individual parts that contribute to a behavior of a mechanism. The mechanism as a whole can be understood as the group of parts-plus-their-organization that serves as the behaving unit. They are related to one another in the sense that there is a form of part-whole relation between the components and the mechanism as a whole.² This point needs to be described carefully to make sense of the features of mechanisms with which we are concerned.

This way of describing mechanisms has been characterized by Craver (2007, p. 188) in terms of the notion of “levels of mechanisms.” The idea is that the components of a mechanism should be seen as at a lower level than the mechanism as a whole, and that organizing the components together results in a higher level of mechanism.³ For instance, the intracellular components in the nerve cell that were described are the lower-level components that serve to make up the higher-level mechanism of the cell as a whole. In this approach to mechanisms, the intracellular components should be understood as individual entities with their behaviors. These entities and their behaviors constitute the mechanism as a whole, which consists in another individual with its behaviors. So in this approach the mechanism should be understood to involve a relation between different individuals that exist (an individual is an entity that is capable of independent existence). It should also be mentioned that there is another notion that is sometimes appealed to in this area by philosophers that concerns a relation between the properties of an object. This notion is called “levels of realization” (Craver, 2007, p. 165) and refers to a different notion. This notion is different from the one we are concerned with about individuals since it concerns relations between properties. As Craver suggests, the right way to think about the mechanisms we are considering is to view them as complex systems constituted by individual components that work together to produce a behavior of the whole.

On this way of thinking it follows that mechanisms as a whole have behaviors that their individual components lack; we can see how this works in terms of the example being used. The nerve cell has the behavior of sending an electrochemical signal to other neural cells in the brain, which is a behavior of the whole cell. But this behavior of the cell is not a behavior of any of its intracellular components individually. The ions that flow into the channels of the cell do not themselves have the behavior of sending electrochemical signals through the axon; they are merely one component that (partially) contributes to this behavior. It is also important to see that the behavior of the cell is not a result of adding the contributions of the ions and other parts together in a simple way. In some structures, when we add the components together the result is a property that differs from the components. An illustration would be the weight of a pile of sand that simply results from adding together the weights of the individual grains. But the behavior one finds in the nerve cell is not like this since it depends on the different interrelations among the components that include the channels, ions, and changing resting potential. The channels have to open and allow the ions to enter, which produces a change in the resting potential, and as this changes new channels open and close to facilitate the signal through the axon. It is not a simple relationship that is involved like with the grains of sand but a situation where the behavior that results from the organization of the complex is more than the sum of its parts (Craver, 2007, p. 216; cf. Biem Graben, 2016).⁴ In this respect, the cell as a whole should be seen to have behaviors that are distinct and novel from the behaviors of its components. It is this aspect of the nerve cell that distinguishes it from other kinds of cases that is characteristic of the mechanisms we are examining.

The behaviors of a whole are important to recognize for understanding mechanisms. This is because they enable the mechanisms to make new kinds of causal contributions. As Craver puts it, “wholes have causal powers that their parts individually do not have” (Craver, 2007, p. 214). There are causal powers at the level of the whole mechanism that are distinct from the causal powers at the level of the components. In terms of our example, due to its organization the causal powers of the nerve cell as a whole are distinct from the causal powers of the components that make it up. The cell as a whole causally contributes to the transmission of information through the signaling system. But the ions in the cell do not directly do this. In this respect, the causal powers of the entities are different because the causal relationships in which they participate are different. Because of this the mechanism is capable of entering into different interactions and so causally contributes something new to the world aside from the components.

One thing to add is that, in saying the behavior of the whole mechanism is more than the sum of its parts, I am not intending to deny that mechanisms are constituted by the physical entities and behaviors that make them up. There is a notion of antireductionism according to which the higher-level behaviors of a system go beyond

² I say “form of” because there are various part-whole relations that exist and the only one I am concerned with is a part-whole relation involving mechanisms of the type I’ve described.

³ I will not stop to consider the notion of levels being used (which is a local notion only in contrast to more global ones) because there is recently a large debate about this notion, and considering this would take me too far afield. What I mean to invoke is the notion of levels used by Craver according to which “X’s ϕ -ing is at a lower mechanistic level than S’s ψ -ing if and only if X’s ϕ -ing is a component in the mechanism for S’s ψ -ing” (Craver, 2007, p. 189). It’s possible the notion of levels could be reframed and the arguments of this paper would still go through, as long as there is some appropriate notion of mereological relationships that applies to mechanisms and components existing in a hierarchy. For discussion of different notions of levels one can see Craver (2007) and Potochnik (2017).

⁴ In describing his view Craver says, “lower-level components are made up into higher-level [mechanisms] by organizing them spatially, temporally, and actively into something greater than a mere sum of the parts” (Craver, 2007, p. 189).

the organized interactions of the parts [cf. strong emergentism (Craver, 2007, p. 216)], but that is not what I am claiming. The idea is that there are complex interactions among the components of the system that produce a new behavior of the whole, but where this whole is constituted out of parts and their behaviors that make it up. So the form of antireductionism being described is consistent with the idea that the resulting behavior is dependent on the components. What matters to the account offered is the idea that various lower-level entities in the world can become organized together in certain ways, giving rise to new properties and behaviors at higher levels. These higher-level mechanisms are made out of lower-level constituents, but they cannot be reduced to the constituents.

4. Versions of reductionism

To clarify this point it will help to consider some notions of reductionism and explain whether any of these notions apply to the notion of mechanisms I have described. Here I will consider three common ways of thinking about ontological reductionism one hears.

On one of the common ways of thinking about reductionism over the years, this concerns a relation between different types of properties. The idea is that we have a reduction when a higher-level property of an entity is shown to be the same as some lower-level property of the entity (Sklar, 1967; Kim, 1998). On this view a property can be reduced to another property just in case the former is type identical with the latter. For instance, if we can identify the higher-level property of “being water” with the lower-level property of “being H₂O,” then we have shown that being water can be reduced to being H₂O and there is really no difference between these properties.

The view I have presented of mechanisms is inconsistent with this point. The account offers a way of thinking about the relation between higher and lower-level mechanisms in which they involve distinct individuals and properties. It was noted before, for instance, that the behavior of wholes involves distinct individuals from the behaviors of the components. What is going on in the neural cell as a whole when it signals is distinct from what is going on with the individual ions. The lower-level behaviors of the components contribute to but are distinct from the higher-level behavior of the cell. This point has been explained by Gillett in terms of the “qualitative distinctness” of the properties of different individuals. In his account of mechanisms there are distinct levels of individuals and these come with corresponding distinct levels of behaviors (Gillett, 2010, 2022). Because there are qualitatively distinct behaviors like this that are not shared we should not think of the behaviors of the whole as just a subset (or part) of the behaviors of the components (for an alternative account see Piccinini (2022a,b)). As a result of all this, there is no identification to make among the higher and lower-level properties and so no reduction which exists. This point can also be combined with the point that there are sometimes different lower-level mechanism types which can underlie the same type of higher-level mechanism in the sense of multiple realization (cf. Piccinini, 2020). For instance, it is possible that the type “neural signal” can be produced in different ways in different neurons, say with different numbers of ions and channels that have different spatial organization throughout the cell. In this respect there is no identification to make between the types present.

A second notion of reduction involves the idea that lower levels of mechanisms explain higher-level behaviors without intermediate

explanatory levels, in the sense that the lower levels directly account for the higher levels. These lower-level components are what matter fundamentally and scientists should focus their attention on these in their research. An approach like this is represented by ruthlessly reductionist views of neuroscience (Bickle, 2003, 2020) that hold that lower levels of mechanisms are what matter for how mechanisms work.

The view I presented of mechanisms does not fit with this either. The account offers a way of thinking about the notion of mechanisms in which they consist of wholes that are different from the components. Due to their organization higher-level mechanisms can do things that cannot be accounted for in terms of the behavior of lower-level components alone and need to be studied in their own terms. The neural signal, for instance, results from the components operating together at the level of the whole and this needs to be cited for a full explanation. This point has been made by Bechtel who notes that “typically the behavior of the whole system must be studied at its own level with appropriate tools for that level. Research at the level of whole systems ... studies, using its own modes of investigation, phenomena different from those studied at the level of the component parts” (Bechtel, 2008, p. 129). Think, for instance, of how researchers might electrically stimulate a whole neural cell to see how it behaves in response. Accounting for the behavior of this sort of case will be done in terms of interventions upon the mechanism as a whole and is not a strictly lower-level affair about individual components. The lower-level components have a contribution to make, but this does not replace the contribution of the whole mechanism.

The third notion I will consider is that mechanisms are reducible in the sense that there is a decomposition of a mechanism’s behavior into the components and their behaviors, so that there is a one-to-one mapping that is preserved. This notion of strong decomposition may also include the idea that in a mechanism individual components and their behaviors can be studied separately from other components in the mechanism (Kaiser and Krickel, 2017).

The problem with this way of thinking about reductionism is that there are often facts about the interrelations of components in a mechanism that affect the behaviors of the components that occur. What happens in a neural cell is not a simple sequence of steps within the cell but a complex set of interacting components behaving together. For instance, the channels in the cell membrane behave by both opening and closing, and this occurs at different rates, and which behavior is performed depends on what the different concentrations of ions are elsewhere in the cell. As a result these other components affect the behavior of the channels and their properties. To know why a channel behaves the way it does one thus has to know about what else is going on in the cell. Because of this the strong notion of decomposition does not seem to apply in this sort of case (Andersen, 2014; Burnston, 2021; Silberstein, 2021). Accepting this is not to deny that a mechanism’s behavior can be explained more weakly in some sense in terms of the behavior of the components and their affects on each other. But this notion of decomposition does not require the stronger notion which is sometimes associated with reductive ways of thinking about mechanistic explanations.

There is more to say about the notion of reductionism than I have said so far and I am not suggesting that what I’ve said on this is complete. What I have been trying to do is to describe how to think about neural mechanisms and their behaviors in a way that I think can be supported by the examples. It seems to me that when we consider the mechanisms that exist, they are best described as involving new

behaviors from their components and require study in their own terms, and in this sense we cannot reduce the mechanisms to their components' properties and behaviors. It should be allowed that there may be other notions of reduction that have different implications in this area since there are different notions that have been offered by people.⁵ Some of the concerns with other ways of thinking about reductionism will be considered at a later point.

5. The explanation of mechanisms

So far I have been describing how I think we should view mechanisms as they exist in the world. The account has been concerned with the features of mechanisms, and the entities and behaviors that make them up. I think it is helpful to be concerned with this aspect of mechanisms because we want an adequate account of mechanisms as they exist. What I want to do at this point, though, is turn from questions about how to understand mechanisms to questions about how to explain them. To do this, I will need to say something about the notion of explanation and how it should be understood in this context.

At a general level, when we are concerned with the explanation of a mechanism, we are concerned with providing the reasons why something has occurred in the mechanism. The explanation involves accounting for why that something has occurred. When we apply this sort of idea to explaining the behavior of a mechanism, this means the behavior will be explained in terms of the features that bring it about. We have seen that this consists in referring to the components and their behaviors and how they are organized to produce the behavior. In this sense, it is the reference to the details of the components and their organization that provide the explanation.

This way of describing the explanation comes from a way of thinking about how they should be characterized that's become widely accepted more recently (Craver, 2007). In the account Craver presents, he is interested in describing the notion of explanation and how it applies to mechanisms in connection to earlier work from Salmon (1984). In the approach Craver takes, an explanation occurs when we have exhibited the entities in the world that serve to bring the phenomenon about. The world consists of entities that stand in causal and other relations to one another, in a temporal and spatial framework. To explain a phenomenon in this framework is to situate it in this causal structure. For instance, think of how we might explain the presence of water on the street after it rains. The explanation would consist of referring to the factors in the environment that served to bring the rain about, which include things like the condensation in the atmosphere and the effects of gravity. We have explained why the street is wet when we have exhibited the factors in the world whose presence led to this phenomenon occurring.

This way of talking about explanation sometimes leads Craver to say that an explanation concerns objective features of the world. To explain why something occurs we have to describe how it fits within the objective structure that exists. But Craver does not limit himself

to these objective aspects in talking about the notion of explanation, since he sometimes seems to allow that there is also a role for representations to play. This is because in giving an explanation humans make use of representations of different kinds. This can be understood to mean that explanations involve the use of representations (or conceptual vehicles) that are part of the explanation being offered by someone. In the example of explaining why the street is wet, for example, we have to characterize the phenomenon in terms of the representations "gravity" and "atmospheric condensation," and describe how these are related to each other to produce the "rain." This seems to be a common feature of giving explanations since we exchange information with others by means of representations. To include this other aspect in the account we should accept that the activity of giving explanations involves reference to features of the world and includes a means for representing them in language or other forms of representation. In this way of viewing the notion of explanation I described it has both objective and representational aspects [for discussion of this approach see Illari (2013)]. Though it has not always been clear in his account, I think this sort of approach is consistent with what Craver says since he makes reference in his work to "explanatory texts" in places (Craver, 2007, p. 27) that he takes to be representational. While he tends to emphasize the world having objective structure, there is more to explanation than this. I will follow him in including these representational aspects since it is helpful to view explanation as involving both of these together.

Having said this about the explanation of mechanisms, there is a further issue to be addressed. Something needs to be said about the kinds of representations that one might use. There are different types of representations which one may want to make use of in an explanation, which include linguistic, visual, and other forms of representation. The approach I will take on this departs from Craver and comes from an earlier way of thinking about explanation associated with Hempel which characterizes them in terms of a type of argument (Hempel and Oppenheim, 1948). The idea is that we can characterize the explanatory factors of a mechanism in terms of the premises of an argument, from which a conclusion describing the phenomenon to be explained can be derived. The premises will consist of sentences describing the features of the mechanism, and the conclusion will consist of a sentence describing the phenomenon at issue. The explanation will then consist in showing how the conclusion concerning the phenomenon follows from the information contained in the premises. This way of viewing an explanation descends from earlier work which has been influential. But we need to be careful here since not everyone agrees with the idea that explanations should be understood as arguments made of sentences. My approach to this issue will be to follow Hausman (1998) in thinking that explanations can at least be represented in this way, and that there is something helpful in doing this.⁶ This is because it will show how this common form of representation can be used. Furthermore, it is not always clear

⁵ Another notion holds that a reduction occurs if a mechanism's behavior is explained merely in relation to its components and their behaviors (Bechtel, 2008, p. 151). This is a rather weak notion which I do not oppose at some level.

⁶ Note that others have suggested that we can have an argument-based approach to explanation in a way that differs from Hempel's account. For example, beyond Hausman, both Kitcher (1981) and Strevens (2008) describe the notion of explanation in terms of arguments, but in ways that depart from Hempel.

to everyone what such an approach would look like and it may help to see this laid out carefully. In saying this I am not taking myself to have settled whether this is the only way of thinking about the notion of explanation one might accept. Discussing this would require more time than I can devote to this issue in this setting and a full account will have to be left for another occasion. What I will do is merely show that there is a way of describing mechanistic explanations in this manner that is plausible and illustrate the form such an approach might take.

It will help to provide a more specific example of what a mechanistic explanation will look like along these lines. The basic idea will involve explaining why a mechanism *O* has a behavior. The explanation will involve analyzing *O* in terms of the behavior of its components and their organization in the mechanism that enables it to perform the behavior, and representing this in terms of an argument (*cf.* Levine, 2001, p. 74). Here is what this might look like with the example that we have been using. Suppose we are interested in explaining why an action potential is propagated down the axon in a particular cell. We can say that the behavior to be explained is the behavior for having an action potential. The first step in the explanation is to characterize the properties that define the behavior, which consist of the precipitating and manifestation conditions for the behavior. In the example that we are discussing, being an action potential is a behavior of a structure that results from inputs to some components and their behaviors that leads to signals being propagated down a cell. Once we have specified the behavior in this way, the next step is to describe the particular components and their behaviors in a mechanism that lead to this behavior. This is done by identifying the components and behaviors in the mechanism and how they are organized to result in the behavior. Once this is done we have explained why the behavior occurs.

We can lay out the steps of such an explanation in the following way:

1. Having an action potential =df having some components and behaviors caused by inputs to a cell, and that leads to a signal down the axon.
2. The presence of input states causes components and behaviors in organization *S*, and this leads to a signal down the axon.
3. Mechanism *O* has components and behaviors in organization *S*.
4. Thus, mechanism *O* has the behavior of an action potential.

In the nervous system, the components and behaviors in organization *S* will consist of the opening of channels, an influx of ions across the membrane, an increase in resting potential, and the initial signal. When these are present they lead to the propagation of the electrical signal down the cell.⁷ The explanation that is offered consists of an argument whose conclusion is that the mechanism has

the behavior for an action potential. The explanation is such that the information described in the premises leads to the conclusion regarding the presence of the behavior that is at issue. The explanation works by describing the sequence of events in the mechanism and their order that result in the behavior.

Note that this way of characterizing an explanation is different from Hempel's earlier account of explanation in certain ways. In particular, notice that there is no requirement that the premises of the explanation include a law of nature, as Hempel required. The first line of the explanation in the account is not a law of nature in the traditional sense, but merely serves to specify a behavior that a mechanism can have. So the account is different from Hempel's Deductive-Nomological approach that was concerned with explanation in terms of laws. One of the reasons for this is that Hempel was interested in causal explanations between events, which are different from the examples I am considering. The examples I am considering are concerned with explaining how a mechanism underlies a behavior or capacity. With this form of explanation it is not important to describe laws of nature which may (or may not) apply to the mechanism and how they are involved. The explanation is merely concerned with referring to the features within the mechanism whose occurrence underlie the behavior at issue. Representing this information in the explanation enables us to see why the behavior follows from the features described. In doing this, the explanation makes use of arguments to present this information, but in other respects it is different from Hempel's account.

A further feature of the account to note is that it is consistent with the earlier point that there are both objective and representational aspects to the explanation. On the one hand, there is the mechanism with its features in the world which exists independently from us. The behavior of the mechanism occurs in the world and depends on the other features that make up the mechanism. On the other hand, the explanation is presented in the form of an argument that conveys the information about how the different features of the mechanism are related. By describing the components and how they are organized to bring about the mechanism's behavior in the premises, we can make sense of why the behavior occurs. This way of thinking about the explanation is useful because the form of argument makes clear the sequence of changes the mechanism undergoes that enables us to understand why the behavior occurs. Furthermore, it should be apparent that the explanation is distinct from the mechanism and merely provides a means for representing information about the mechanism. In this respect, the account is different from Hempel's approach since he appeared to think that the causal relations some thought existed in the world could be captured entirely in terms of the explanatory information presented in an explanation. This is not a feature of the account I have offered. The account holds that there is a difference between the mechanism in the world and the information in the explanation which serves to represent it.

I think this approach can provide us with a way of understanding the explanation of mechanisms that is useful for thinking about how the explanations work. It allows us to describe the explanation in terms of a common form of representation, and makes clear the different features that are involved in the explanation. There are other aspects of the notion of explanation that one may want to consider in thinking about this notion and I have not tried to address all the

⁷ It should be observed that the expression "components and behaviors in organization *S*" in lines 2 and 3 is intended as a summary of whatever components and behaviors and their order exist in the (actual) mechanism in question. These could be listed out with more detail if preferred, although it would make the explanation more complex in certain ways that I would like to avoid here.

concerns that may exist.⁸ Rather than take up all of these issues which need separate treatment, what I want to do is consider how the approach relates to the previous account of mechanisms offered. If the account of mechanistic explanations that was presented can be made to work, what implications does this have with respect to the issue of reductionism?

6. Some implications

Let me return to the issue of reductionism in relation to these concerns. There are several implications that would appear to follow from the approach that was offered for this issue.

The first point to observe has to do with the character of the explanations given. We have seen that it was a feature of the approach that an explanation consists in the information in the premises leading to the information in the conclusion listed. The idea is that we have explained a mechanism's behavior when we have shown how a description of it follows from the information about the mechanism's components, behaviors, and organization. In this respect, the information about the mechanism's behavior can be derived from information about the different features of the mechanism that are referenced. Given this, one might think that the account is in tension with the earlier point that the mechanism as a whole is distinct from the components and their behaviors and cannot be reduced to them. The fact that one can *derive* information about one from information about the other may suggest to someone that they are not really distinct. But this way of thinking does not follow from the approach I have offered. Recall the approach I have presented holds that the behavior of the mechanism as a whole is not explained in terms of the behaviors of its components individually. The way to see this is to observe that the features appealed to in the explanation in line 2 concern the components' behaviors *and their organization*. It is not merely the components that do the work in the explanation. The explanation appeals to how the components have been organized together in such a way that they result in the behavior. This organizational property of the mechanism is not an aspect of the lower-level itself but exists with the higher-level parts-plus-their-organization.⁹ So the explanatory scheme provided is consistent with the earlier point that the mechanism's behaviors should be understood nonreductively. We do not derive this organizational property from the lower-level itself and so this is not a feature to which the mechanism's behavior can be reduced.

One may think that an approach to explanation that views them as I have described will have to view mechanisms reductively. But this

would be a concern only if one overlooked the role of organization among the components. While it may be true that the components are involved in the behaving mechanism, this is not enough to show that the behavior of the whole is due merely to the behavior of the components. As Bechtel notes in one place, "an understanding of the parts alone is not sufficient to understand why the mechanism behaves as it does scientists need to consider how the parts and operations are organized" (Bechtel, 2008, p. 151). What the above explanatory scheme helps to illustrate is that this aspect of the explanation is in addition to the reference to the lower-level components. So it is not an explanation of the mechanism's behavior just in terms of lower-level features.

A second observation is related to this point and concerns the history of debates over reductionism in this area. It should be evident that the account offered is different from an earlier, influential approach to reductionism presented by Nagel (1961) that has been widely discussed (Silberstein, 2002; Ney, 2022). In his account, the notion of reductionism is characterized in terms of the relations between aspects of different theories. Nagel conceived of theories as collections of statements which include laws, and thought that the right way to understand issues about reductionism was to consider theories from different sciences (psychology vs. neuroscience, say) and how these were related. A reduction occurs only if we can state bridge principles connecting the kind terms in the laws of the two theories to one another. In addition, one has to show how the laws of the reduced theory, T1, can be logically derived from the laws of the reducing theory, T2, together with any appropriate boundary conditions that may be involved. So this form of intertheoretic reduction is characterized in terms of the derivation of one set of laws from those of another. This way of thinking about reductionism is different than the account I have presented. While it is true that the derivation of information is important to the explanation of a mechanism's behavior on my account, this is not a matter of an derivation between theories or laws. Moreover, Nagel's concern with theories as being the relevant phenomenon is not how I have characterized the notion of reductionism. The account I have offered is not concerned with theories but views things differently.

In the history of debates over reductionism many people have, in fact, moved past approaches focused on relations between theories because there are problems with this sort of approach. For instance, one of the concerns with this approach that was noted is that, if we construe reductionism as requiring correlations between kind terms from different theories (which is a common way the account has been interpreted that I will follow¹⁰), this would appear too weak to underwrite a genuine form of reduction (Sklar, 1967; Kim, 1998). Knowing that term K1 from T1 can be correlated with term K2 from T2 merely establishes a biconditional relationship between the terms, which is consistent with views that are nonreductionist. This is because even a dualist who accepts the presence of correlations between mental and neurological state kinds can satisfy such an

⁸ I'm thinking of such concerns as the problem of symmetry, the problem of relevance, etc. Another concern with the account is about other forms of representation than sentences. For instance, sometimes researchers describe mechanisms in terms of diagrams. One response is to say that, if the diagrams reveal the relevant components and their behaviors of a mechanism (Bechtel and Abrahamsen, 2005, p. 425), then this information should be translatable into the form of explanation that was described (for an alternative view see Burnston (2016)). But I will not try to develop this point further.

⁹ Cf. Craver's claim that "lower-level components are organized together to form higher-level components" (Craver, 2007, p. 189).

¹⁰ There has been some disagreement about how to describe the bridge principles at issue. Richardson (1979) argues that Nagel only required one-way conditionals in his account. But many people have thought that the correlations would have to be at least as strong as bi-conditionals to work. I will follow this approach, though I do not think it affects anything that follows.

account, though such a person would not be considered a reductionist. To make a claim of reductionism work it seems what is needed is something stronger than a requirement of mere correlations. It was this sort of concern that helped people to see that reductionism should not be conceived as a relationship between theories, but is better characterized in terms of relations between entities that exist in the world. The account I have offered is consistent with this more recent way of thinking about reductionism.

The account offered is concerned with how entities are related to each other. But this should not be taken to mean that there is not a role for theories to play in understanding how entities are structured in the world. The account allows that we can still accept that we need theories at different levels, corresponding to the different levels of entities. To understand this point remember that the account of mechanisms offered holds that there are higher-level behaviors of mechanisms as a whole, and the lower-level components and their behaviors that make them up. The fact that there are different levels of entities helps to explain why there are theories that have been developed in different areas of the sciences. When we are trying to understand why a mechanism behaves as it does, we will sometimes be concerned with the lower-level constituents that contribute to its operation. Understanding the components and their behaviors helps us to understand why the overall behavior of the mechanism occurs; or it may sometimes be that we are just interested in understanding how the components operate in themselves or in relation to others. But knowing about the components does not prevent us from having to study the mechanism at higher levels of organization. The behavior of the mechanism as a whole needs to be studied in its own terms¹¹ and in relation to other mechanisms in the environment it interacts with, and with respect to whatever principles are at work at higher levels. These features of the mechanism are not something that can be understood merely by looking to the lower-level components and their behaviors. As a result there is a need for different theories to be offered at different levels because this will help us to make sense of the different aspects that exist.

Finally, it should be noted that this picture of how theories are understood might be further developed to explain what's useful about having such theories. I have suggested that part of the explanation for this has to do with theories that might be developed at higher levels, which we need to know for an understanding of the various aspects of the mechanisms. These theories might concern how the mechanisms are causally related to other mechanisms in the environment, or they might concern ways of picking out the mechanisms that are of interest, or something else. I think there is more that one would need to say to explain just what these theories are concerned with and how they are able to be useful in the sciences. I would suggest that we can recognize this point without worrying that we need to have all of this worked out at this point to make sense of the account. Given that there are mechanisms in the world with entities and behaviors that exist at different levels, there will be a need for researchers to develop different theories to describe them adequately. The account I have presented can be developed to fit with this point about the features of mechanisms and there is no reason to think the details will change this fact. Regardless of such issues, there

will be a need for theories at different levels because of the structure the world exhibits.

7. Conclusion

Issues about how mechanistic explanations and reductionism are related have raised a lot of concerns. In this paper, I have tried to offer an account of mechanisms as systems constituted by parts that make them up and say something about how mechanisms so understood can be explained. Once these views have been presented, it helps us to clarify some of the relationships at work in talking about reductionism and mechanisms. The account I have presented suggests that the proper way to understand the mechanisms I have been concerned with is nonreductively. A mechanism should be understood to have behaviors that exist which cannot be reduced to the behaviors of the parts. A behavior of the mechanism is based on the behaviors of the parts present but goes beyond them. The explanation of a mechanism's behavior has also to include reference to the organizational properties of the mechanism. We can accept that mechanistic explanations refer to components without thinking that is all there is to the explanation.

It is hoped that this way of thinking about these issues provides us with some clarification of mechanisms. Needless to say, I have not attempted to say everything that has to be said about how to understand mechanisms or how they should be explained. Both of these are topics about which more could certainly be said. For example, one issue I noted I have not examined concerns the way one should understand the notion of "levels" used and how this notion can be made more precise. There are different ways of thinking about this notion and it may be useful to consider this more carefully at some point. There are also questions I have not considered about the notion of explanation and how it connects to other issues like the "pragmatics" of explanation (is explanation a contrastive notion, say?), among others. Rather than consider these issues, what I have tried to do is to present a way of thinking about mechanistic explanations and reductionism that offers a way of helping us understand their relationship. It is thought that improving our understanding of their relationship will be useful for addressing these other sorts of issues in the area.

Author contributions

The author confirms being the sole contributor of this work and has approved it for publication.

Acknowledgments

The author thanks Ken Aizawa, Daniel Burnston, and the reviewers for useful comments.

Conflict of interest

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

¹¹ Cf. section 4.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated

organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Andersen, H. (2014). A field guide to mechanisms: part II. *Philos Compass* 9, 284–293. doi: 10.1111/phc3.12118
- Bechtel, W. (2008). *Mental Mechanisms*. New York, NY: Taylor and Francis.
- Bechtel, W., and Abrahamsen, A. (2005). Explanation: a mechanistic alternative. *Stud. Hist. Philos. Biol. Biomed. Sci.* 36, 421–441. doi: 10.1016/j.shpsc.2005.03.010
- Bickle, J. (2003). *Philosophy and Neuroscience*. Dordrecht: Kluwer.
- Bickle, J. (2020). Laser lights and designer drugs: new techniques for descending levels of mechanisms “in a single bound”? *Top. Cogn. Sci.* 12, 1241–1256. doi: 10.1111/tops.12452
- Biem Graben, P. (2016). Contextual emergence in neuroscience, ed. Hady A. El. *Closed Loop Neuroscience*. Amsterdam: Elsevier, 171–184.
- Burnston, D. (2016). Data graphs and mechanistic explanation. *Stud. Hist. Phil. Biol. Biomed. Sci.* 57, 1–12. doi: 10.1016/j.shpsc.2016.01.002
- Burnston, D. (2021). Getting over atomism: functional decomposition in complex neural systems. *Br. J. Philos. Sci.* 72, 743–772. doi: 10.1093/bjps/axz039
- Craver, C. (2007). *Explaining the Brain*. Oxford: Oxford University Press.
- Cummins, R. (1975). Functional analysis. *J. Philos.* 72, 741–764. doi: 10.2307/2024640
- Gillett, C. (2010). Moving beyond the subset model of realization. *Synthese* 177, 165–192. doi: 10.1007/s11229-010-9840-1
- Gillett, C. (2022). Engaging the plural parts of science. *J. Conscious. Stud.* 29, 195–217. doi: 10.53765/20512201.29.7.195
- Glennan, S. (2017). *The New Mechanical Philosophy*. Oxford: Oxford University Press.
- Hausman, D. M. (1998). *Causal Asymmetries*. Cambridge: Cambridge University Press.
- Hempel, C. G., and Oppenheim, P. (1948). Studies in the logic of explanation. *Philos. Sci.* 15, 135–175. doi: 10.1086/286983
- Illari, P. (2013). Mechanistic explanation: integrating the ontic and epistemic. *Erkenntnis* 78, 237–255. doi: 10.1007/s10670-013-9511-y
- Kaiser, M. (2018). “The components and boundaries of mechanisms” in *The Routledge Handbook of Mechanisms and Mechanical Philosophy*. eds. S. Glennan and P. Illari (New York, NY: Routledge), 116–130.
- Kaiser, M., and Krickel, B. (2017). The metaphysics of constitutive mechanistic phenomena. *Br. J. Philos. Sci.* 68, 745–779. doi: 10.1093/bjps/axv058
- Kim, J. (1998). *Mind in a Physical World*. Cambridge, MA: Harvard University Press.
- Kitcher, P. (1981). Explanatory unification. *Philos. Sci.* 48, 507–531. doi: 10.1086/289019
- Levine, J. (2001). *Purple Haze*. New York: Oxford University Press.
- Nagel, E. (1961). *The Structure of Science*. New York: Harcourt, Brace & World, Inc, 29, 716.
- Ney, A. (2022). *Reductionism*. The Internet Encyclopedia of Philosophy. ISSN 2161-0002. Available at: <https://iep.utm.edu/> (Accessed May 28, 2022).
- Piccinini, G. (2020). *Neurocognitive Mechanisms: Explaining Biological Cognition*. Oxford: Oxford University Press.
- Piccinini, G. (2022a). Neurocognitive mechanisms. *J. Conscious. Stud.* 29, 167–174. doi: 10.53765/20512201.29.7.167
- Piccinini, G. (2022b). Neurocognitive mechanisms: some clarifications. *J. Conscious. Stud.* 29, 226–250. doi: 10.53765/20512201.29.7.226
- Potochnik, A. (2017). *Idealization and the Aims of Science*. Chicago: University of Chicago Press.
- Psillos, S. (2004). A glimpse of the secret connexion. *Perspect. Sci.* 12, 288–319. doi: 10.1162/1063614042795426
- Richardson, R. C. (1979). Functionalism and reductionism. *Philos. Sci.* 46, 533–558. doi: 10.1086/288895
- Salmon, W. C. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton, NJ: Princeton University Press.
- Silberstein, M. (2002). “Reduction, emergence and explanation” in *The Blackwell Guide to the Philosophy of Science*. eds. P. Machamer and M. Silberstein (Cambridge, MA: Blackwell Press), 80–107.
- Silberstein, M. (2021). “Constraints on localization and decomposition as explanatory strategies in the biological sciences 2.0” in *Neural Mechanisms*. eds. F. Calzavarini and M. Viola (Switzerland: Springer), 363–393.
- Sklar, L. (1967). Types of inter-theoretic reduction. *Br. J. Philos. Sci.* 18, 109–124. doi: 10.1093/bjps/18.2.109
- Strevens, M. (2008). *Depth: An Account of Scientific Explanation*. Cambridge, MA: Harvard University Press.

Frontiers in Psychology

Paving the way for a greater understanding of human behavior

The most cited journal in its field, exploring psychological sciences - from clinical research to cognitive science, from imaging studies to human factors, and from animal cognition to social psychology.

Discover the latest Research Topics

[See more →](#)

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact

