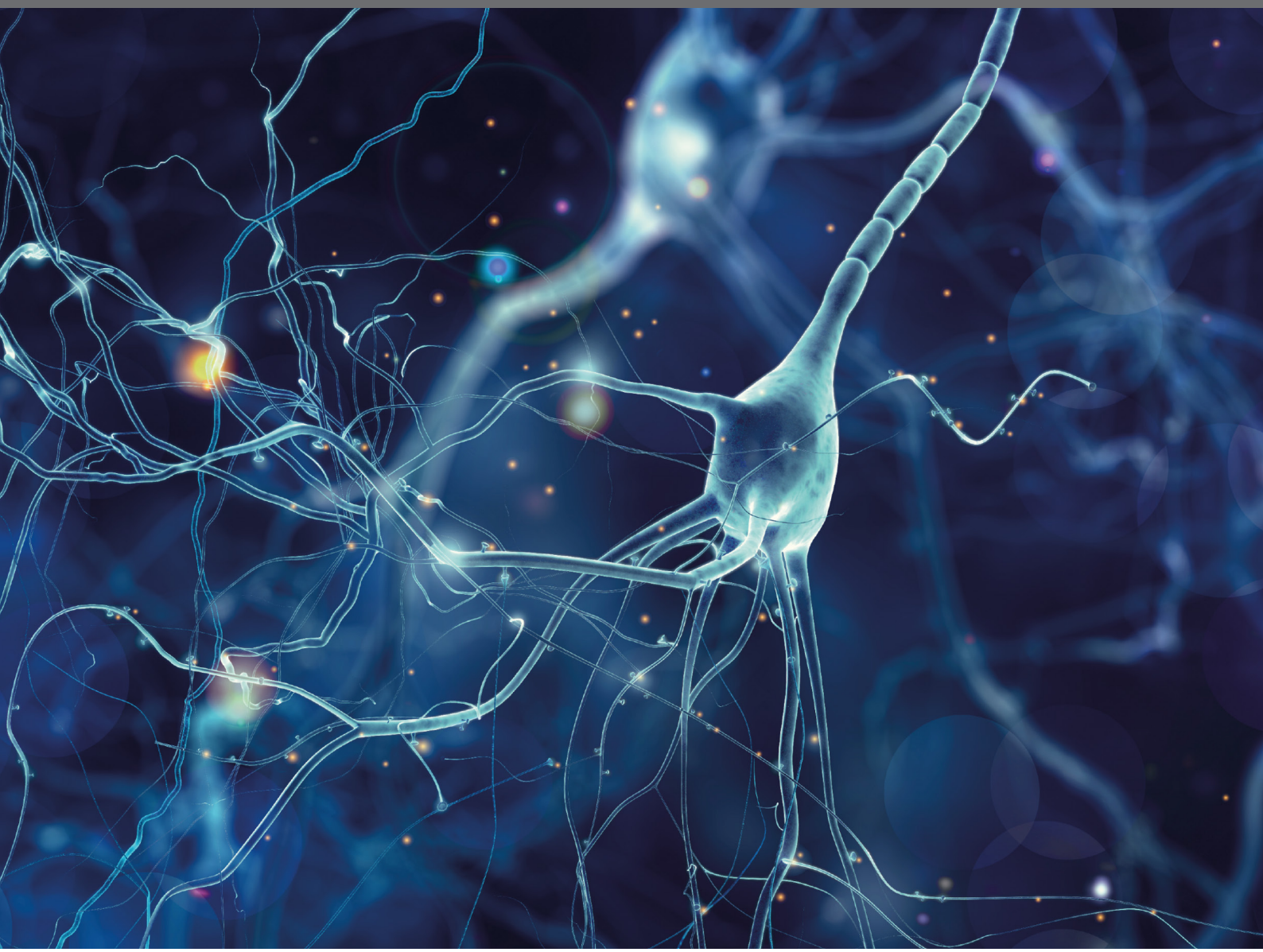
REPRESENTATION IN THE BRAIN

EDITED BY: Asim Roy, Leonid Perlovsky, Tarek Besold, Juyang Weng and Jonathan Edwards
PUBLISHED IN: Frontiers in Psychology



frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88945-596-6
DOI 10.3389/978-2-88945-596-6

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: researchtopics@frontiersin.org

REPRESENTATION IN THE BRAIN

Topic Editors:
Asim Roy, Arizona State University, United States
Leonid Perlovsky, Northeastern University, United States
Tarek Besold, City University of London, United Kingdom
Juyang Weng, Michigan State University, United States
Jonathan Edwards, University College London, United Kingdom

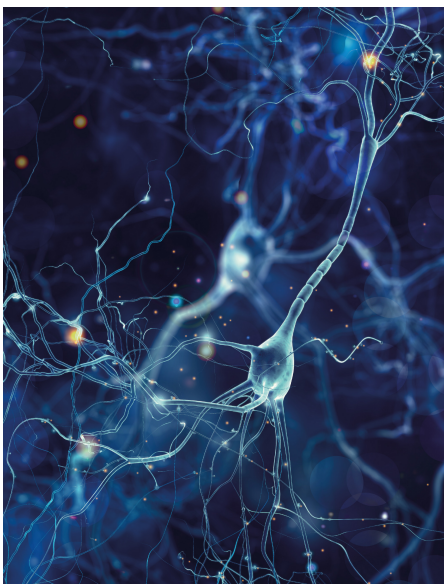


Image: whitehoune/Shutterstock.com

This eBook contains ten articles on the topic of representation of abstract concepts, both simple and complex, at the neural level in the brain.

Seven of the articles directly address the main competing theories of mental representation – localist and distributed. Four of these articles argue – either on a theoretical basis or with neurophysiological evidence – that abstract concepts, simple or complex, exist (have to exist) at either the single cell level or in an exclusive neural cell assembly. There are three other papers that argue for sparse distributed representation (population coding) of abstract concepts.

There are two other papers that discuss neural implementation of symbolic models.

The remaining paper deals with learning of motor skills from imagery versus actual execution.

A summary of these papers is provided in the Editorial.

Citation: Roy, A., Perlovsky, L., Besold, T., Weng, J., Edwards, J., eds. (2018). Representation in the Brain. Lausanne: Frontiers Media.
doi: 10.3389/978-2-88945-596-6

Table of Contents

frontiers in Psychology

EDITORIAL
published: 08 August 2018
doi: 10.3389/fpsyg.2018.01410

Editorial: Representation in the Brain

Asim Roy[1]*, *Leonid Perlovsky*[2], *Tarek R. Besold*[3], *Juyang Weng*[4] and *Jonathan C. W. Edwards*[5]

[1] Department of Information Systems, Arizona State University, Tempe, AZ, United States, [2] Department of Psychology, Northeastern University, Boston, MA, United States, [3] Data Science, City University of London, London, United Kingdom, [4] Department of Computer Science and Engineering, Michigan State University, East Lansing, MI, United States, [5] Division of Medicine, University College London, London, United Kingdom

Keywords: representation in the brain, localist connectionism, distributed representation, abstract concept encoding, symbolic system

Editorial on the Research Topic

Representation in the Brain

Representation of abstract concepts in the brain at the neural level remains a mystery as we argue over the biological and theoretical feasibility of different forms of representations. We have divided the papers in this special topic on "Representation in the brain" broadly into the following sections:

(1) Those arguing, either on a theoretical basis or with neurophysiological evidence, that abstract concepts, simple or complex, exist (have to exist) at the single cell level. Papers by Edwards, Tsotsos, Feldman, and Roy are in this category. However, Feldman and Tsotsos argue that there might be an underlying neural cell assembly (a sub-network) of subconcepts to support a concept at the single cell level. Feldman also stresses action circuits in his paper.
(2) There are three papers that argue for sparse distributed representation (population coding) of abstract concepts. Papers by Balkenius and Gärdenfors, Kajic et al., and Lőrincz and Sárkány are in this category.
(3) There are two papers discussing neural implementation of symbolic models: one by van der Velde et al. and the other by Wolff.
(4) The paper by Frank and Schack, on learning of motor skills from imagery vs. actual execution, is not strictly related to the issue of abstract concept representation, but is about other aspects of learning.

We provide a brief summary of each of the papers next.

ON SINGLE CELL ABSTRACT REPRESENTATION IN THE BRAIN

Edwards argues that both local and distributed representation is present in the brain and explains which occurs when. He explains that distributed representation occurs on the input side of a neuron, but the neuron itself, being the receiver and interpreter of these signals, is localist. This interpretation of brain architecture essentially resolves the fundamental question of who ultimately establishes meaning and interpretation of a collection of signals. In other words, there has to be a "consumer" (a decoder) of such a collection of signals. Without a "consumer," the collection of signals is not "received." In this interpretation, therefore, any signal generated by a neuron has meaning and interpretation. Another neuron, receiving a collection of these signals, then interprets and generates new information. He further argues that this interplay of distributed and localist representation occurs throughout the brain in multiple layers of processing. And he claims that the concept of "representation-as-input" is not in conflict with neuroscience at all.

OPEN ACCESS

Edited and reviewed by:
Bernhard Hommel,
Leiden University, Netherlands

***Correspondence:**
Asim Roy
asim.roy@asu.edu

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *16 June 2018*
Accepted: *19 July 2018*
Published: *08 August 2018*

Citation:
Roy A, Perlovsky L, Besold TR, Weng J and Edwards JCW (2018) Editorial: Representation in the Brain. Front. Psychol. 9:1410. doi: 10.3389/fpsyg.2018.01410

Tsotsos revisits the issue of complexity analysis, mainly of visual tasks, and claims that complexity analysis, accounting for resource constraints, dictates the type of representation required for visual tasks. He argues that complexity analysis could be used as a test to validate theories of the brain. For example, accounting for the resource constraints, certain computational schemes cannot be feasibly implemented in biological systems. For human vision, such resource constraints include numbers of neurons, synapses, neural transmission times, behavioral response times, and so on. He also examines certain abstract representations in the brain and shows how they reduce problem complexity. For example, certain pyramidal processing structures in the brain (which have origins in the work of Hubel and Wiesel) produce abstract representations and thus reduce the problem size and the search space for algorithms. He quotes Zucker (1981) on the need for explicit abstract representation: "*One of the strongest arguments for having explicit abstract representations is the fact that they provide explanatory terms for otherwise difficult (if not impossible) notions*." A key conclusion is that knowledge of the intractability of visual processing in the general case tells us that no single solution can be found that is optimal and realizable for all instances. This forces a reframing of the space of all problem instances into sub-spaces where each may be solvable by a different method. This variety of different solution strategies implies that processing resources and algorithms must be dynamically tunable. An executive controller is important to decide among solutions depending on context and to perform this dynamic tuning, and explicit representations must be available to support these functions.

Feldman focuses on brain activity rather than just structure to explain that action and communication are crucial to neural encoding. The paper starts with a brief review of the localist/distributed issue that was active early in the development of connectionist models. He suggests that there is now a consensus—the main mechanism for neural signaling is frequency encoding in functional circuits of low redundancy, often called sparse coding. The main point of the piece is that the term "representation" presupposes a separation of process and data, which is fine for books and computers, but hopeless for the brain. A related point is that brains are not in the storage or truth business, but compute actions and actionability. Actionability is an agent's internal assessment of the expected utility of its possible actions. In addition, the idea of planning, etc. as programs running against data structures should be replaced by mental "simulations." The final section discusses some mysteries of the mind and suggests that all current theories are incompatible with aspects of our subjective experience. There is evidence for all this, some of which is cited in the short article.

Roy provides extensive evidence for single-cell based simple and complex abstractions from neurophysiological studies of single cells. These single-cell abstractions show up in various forms, but the most significant and complex ones are the category-selective cells, the multisensory neurons and the grandmother-like cells. Category-selective cells encode complex abstract concepts at the highest levels of processing in the brain. There is also extensive evidence for multisensory neurons in the sensory processing areas of the brain. In addition, abstract modality invariant cells (e.g., Jennifer Aniston cells) have been found at higher levels of cortical processing. Overall, according to Roy, these neurophysiology studies reveal the existence of a purely abstract cognitive system in the brain encoded by single cells.

ON SPARSE DISTRIBUTED REPRESENTATION

Topographic representations are used widely in the brain, such as retinotopy in the visual system, tonotopy in the auditory system and somatotopy in the somatosensory system. These topographic representations are projections from a higher dimensional space (of sensory information) to a lower dimensional one. Such abstract, low-dimensional representations also appear in the entorhinal-hippocampal complex (EHC). Lőrincz and Sárkány introduce the concept of Cartesian Factors (they use it to enable localized discrete representation) and use the concept to model and explain the EHC system. They are Cartesian in the sense that they are like coordinates in an abstract space. And these Cartesian Factors can be used like symbolic variables. They conclude that Cartesian Factors provide a framework for symbol formation, symbol manipulation, and symbol grounding processes at the cognitive level.

In Remote Associates Test (RAT), subjects are presented with three cue words (e.g., *fish*, *mine*, and *rush*) and have to find a solution word (e.g., *gold*) related to all cues within a time limit. RAT is commonly used to find an individual's ability to think creatively and finding a novel solution word is usually associated with creativity. Kajic et al. present a spiking neuron model for RAT. Their model shows significant correlation with human performance on such a task. They use distributed representation in their model, but each neuron in such a representation has a preferred stimuli similar to what is found in the visual system and place cells. They used leaky integrate-and-fire spiking neurons in the model. Their RAT model is the first one to link such a cognitive process with neural implementation. However, their current model does not explain how humans learn such word associations. All connection weights and other parameters were determined in an offline mode.

Humans and animals use abstractions (information compression) at different levels of processing in the brain. For example, cones and rods in the retina code for 3-dimensional color perception in humans. Such abstractions to lower dimensional spaces occur explicitly throughout sensory systems. Balkenius and Gärdenfors a, in their paper explain how the brain can abstract from neurocognitive representations to psychological spaces and show how population coding at the neural level can generate these abstractions. They show that radial basis function networks are ideal structures for mapping population codes to such lower dimensional spaces. In their theory, the coding of the low-dimensional spaces need not be explicitly expressed in individual neurons but the spatial structures are emergent properties. They also argue that the

mediation between perception and action occurs through such spatial representations and that this form of mediation results in more efficient learning.

NEURAL IMPLEMENTATIONS OF SYMBOLIC MODELS

van der Velde et al. explore the characteristics of two architectures for representing and processing complex conceptual (sentence-like) structures: (1) the Neural Blackboard Architecture (NBA), which is at the neural level, and (2) the Information Dynamics of Thinking (IDyOT) architecture, which is at the symbolic level. They then explore the combination of these two architectures for the purpose of creating both an artificial cognitive system and to explain representation and processing of such structures in the brain. With IDyOT, one can learn the structural elements from real corpora. NBA provides a way to neurally implement IDyOT, whereas IDyOT itself provides a higher-level formal account and learning abilities. Overall, the combined architecture provides a connection between neural and symbolic levels.

Wolff outlines how his "SP Theory of Intelligence" (where "SP" stands for *Simplicity* and *Power*), can be implemented using connected neurons and signal transmission between them. He calls this neural extension "SP-neural". In the SP theory different kinds of knowledge are represented with *patterns*, where a pattern is an array of atomic symbols in one or two dimensions. In SP-neural, these patterns are realized using an array of neurons, a concept similar to Hebb's cell assembly, but with important differences. The central concept in the SP theory is information compression via "SP-multiple-alignment." A favorable combination of Simplicity and Power is aimed for by trying to maximize compression. In the SP theory, unsupervised learning is the basis for other kinds of learning—supervised, reinforcement, imitation and so on.

LEARNING FROM IMAGERY VS. EXECUTION

Frank and Schack provide an overview of the literature on learning of motor skills by imagery and execution from three different perspectives—performance (actual changes in motor behavior), the brain (changes in the neurophysiological representation of motor action) and the mind (changes in the perceptual-cognitive representation of motor action). Both simulation and execution of motor action leads to functional changes in the motor action system through learning, although perhaps to a different extent. They observe, however, that very little is known about how actual learning takes place under these different forms of motor skill practice, especially in terms of action representation.

AUTHOR CONTRIBUTIONS

AR summarized the topic articles with contributions from LP, TB, JW, and JE.

REFERENCES

Zucker, S. W. (1981). "Computer vision and human perception: an essay on the discovery of constraints," in *Proceedings 7th International Conference on Artificial Intelligence,* eds P. Hayes and R. Schank (Vancouver, BC), 1102–1116.

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

frontiers in Psychology

REVIEW
published: 26 September 2016
doi: 10.3389/fpsyg.2016.01457

Actionability and Simulation: No Representation without Communication

*Jerome A. Feldman**

International Computer Science Institute, University of California, Berkeley, Berkeley, CA, USA

There remains considerable controversy about how the brain operates. This review focuses on brain activity rather than just structure and on concepts of action and actionability rather than truth conditions. Neural Communication is reviewed as a crucial aspect of neural encoding. Consequently, logical inference is superseded by neural simulation. Some remaining mysteries are discussed.

Keywords: actionability, connectionist, fitness, neural code, representation, simulation

OPEN ACCESS

Edited by:
Leonid Perlovsky,
Harvard University and Air Force Research Laboratory, USA

Reviewed by:
Frank Van Der Velde,
University of Twente, Netherlands
Marika Berchicci,
Foro Italico University of Rome, Italy

***Correspondence:**
Jerome A. Feldman
feldman@icsi.berkeley.edu

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *07 June 2016*
Accepted: *12 September 2016*
Published: *26 September 2016*

Citation:
Feldman JA (2016) Actionability and Simulation: No Representation without Communication. Front. Psychol. 7:1457. doi: 10.3389/fpsyg.2016.01457

INTRODUCTION

This Frontiers project on "Representation in the Brain" is extremely timely. Despite significant theoretical and experimental advances, there is still considerable confusion on the topic. Wikipedia says: *Representation:* "A **mental representation** (or **cognitive representation**), in philosophy of mind, cognitive psychology, neuroscience, and cognitive science," is a hypothetical internal cognitive symbol that represents external reality, or else a mental process that makes use of such a symbol: a formal system for making explicit certain entities or types of information, together with a specification of how the system does this. "https://en.wikipedia.org/wiki/Mental_representation, August/8/2016."

The definition above presupposes a separation between data and process that is true of books and computers but is utterly false in neural systems. In this article we use the term "encoding" instead of "representation". The brain is not a set of areas that represent things, but rather a network of circuits that do things. It is the activity of the brain, not just its structure, that matters. This immediately brings focus on actions and thus circuits. This paper will not attempt to describe (the myriad) particular brain circuits but will focus on the mechanisms for coordination among the local information transfer and areas and circuits missing in most discussions of "representation."

For concreteness, let's start with a simple, well-known, neural circuit, the knee-jerk reflex shown in **Figure 1**. We are mainly concerned with the simplicity of this circuit; there is a single connection in the spinal cord that converts sensory input to action. The knee-jerk reflex is behaviorally important for correcting a potential stumble while walking upright. The doctor's tap reduces tension in the upper leg muscle and this is detected by stretch receptor in the muscle spindle, sending neural spike signals to the spinal cord. The downward spike signals directly cause the muscle to contract and the leg to "jerk." Not shown here are the many other circuit connections that support coordination of the two legs, voluntary leg jerking, etc.

There are several general lessons to be learned from this simple example. Essentially everyone now agrees that neurons are the foundation of encoding knowledge in the brain. But, as the example above shows, it is the *activity* of neurons, not just their connections, that supports the functionality.

The example involved motor activity, but the basic point is equally valid for perception, thought, and language, they are all based on neural activity. There are three essential considerations in

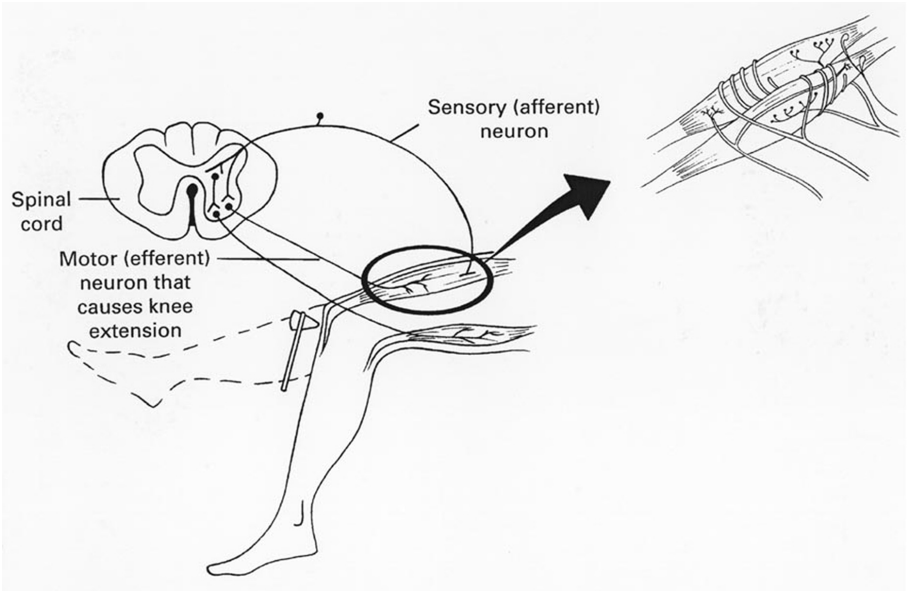


FIGURE 1 | Knee-jerk Reflex Circuit.

discussing neural circuits – the computational properties of individual neurons, the structure of networks and the communication mechanisms involved.

Of these three, it is communication mechanisms that have been studied the least and this fact is the basis for the subtitle "no representation without communication". "Neural Communication and Representation," below is a brief review of what has been called the neural code (Feldman, 2010a). Considerations from neural computation also constrain possible answers to traditional questions like localist vs. distributed representations. Actionability and Simulation goes further and directly addresses the consequences of accepting action and actionability as the core brain function that needs to be explained. The final Conclusions section also considers remaining unsolved mysteries involving the mind-brain problem, some of which are ubiquitous in everyday experience

NEURAL COMMUNICATION AND REPRESENTATION

One key question concerns the basic mechanisms of neural communication. It is now accepted that the dominant method is transmission of voltage spikes along axons and through synapses that are connections to downstream neural processes. Neural spikes are an evolutionary ancient development that remains nature's main technique for fast long distance information passing (Meech and Mackie, 2007). Other neural communication mechanisms are either extremely local (e.g., gap junctions) or much slower (e.g., hormones). Neural spikes serve a wide range of functions.

Much of the chemistry underlying neural spikes goes back even earlier (Katz, 2007; Meech and Mackie, 2007). The earliest use of spiking neurons is to signal coordinated action as in the swimming of jellyfish. This kind of direct action remains one of the main functions of neural spikes as suggested by **Figure 1**. Due to the common chemistry, all neural spikes are of the same duration and size (Katz, 2007; Meech and Mackie, 2007). The basic method of neural information transfer is direct –the information depends on which neurons are linked. Most of the information sent by a sensory neural spike train is based on the sending unit. For output, the result of motor control signaling is largely determined by which fibers are targeted. The other variable is timing; there is a wide range of variation in the firing rate and conduction time of neural spikes.

The other factor on neural computation is resource limitations (Lennie, 2003). The most obvious resource constraint for neural action/decision is time. Many actions need to be fast even at the expense of some accuracy. Some neural systems evolved to meet remarkable timing constraints. Bats and owls make distinctions that correspond to timing differences at the 10 μs level -much faster than neural response times. A second key resource is energy; neural firing is metabolically costly (Lennie, 2003) and brains evolved to conserve energy while meeting performance requirements. The three factors of accuracy, timing, and resources are the elements of a function that conditions neural computation.

We can show why it is not feasible for one neuron to send an abstract symbol (as in ASCII code) to another as a spike pattern (Feldman, 1988). It is known experimentally that the firing of sensory (e.g., visual) neurons is a function of multiple variables, often intensity, position, velocity, orientation, color, etc. It would take an extremely long message to transmit all this as an ASCII like code and neural firing rates are too slow for this, even omitting the stochastic nature of neural spikes. Even if such

a message were somehow encoded and transmitted downstream, it would require a complex computation to decode it and combine the result with the symbolic messages of neighboring cells and then build a new symbolic message for the further levels. Language is a symbolic system that is processed by the brain, but nothing at all like abstract symbols occurs at the individual unit level.

In the past, there have been debates about whether neural representations were basically punctuate with a "grandmother cell" (Bowers, 2009) for each concept of interest. The alternative was basically holographic (with each item encoded by a pattern involving all the units in a large population). It has been understood for decades (Feldman, 1988) that neither extreme could be realized in the neural systems of nature.

Having just a single unit coding the element of interest (concept) is impractical for many reasons. The clearest is that the known death of cells would cause concepts to vanish. Also, the firing of individual units is probabilistic and would not be a stable representation. It is easy to see that there are not nearly enough units in the brain to capture all the possible combinations of sizes, motions, shapes, colors, etc., that we recognize, let alone all the non-visual concepts. The grandmother cell story was always a straw man— using a modest number (∼10) units per concept could overcome all these difficulties.

The holographic alternative was originally more popular because it used the techniques of statistical mechanics. But it is equally implausible. This is easy to see informally and was proved as early as (Willshaw et al., 1969). Suppose a system should represent a collection of concepts (e.g., words) as a pattern of activity over some number M (say 10,000,000) neurons. The key problem is cross-talk: if multiple words are simultaneously active, how can the system avoid interference among their respective patterns. Willshaw et al. (1969) showed that the best solution is to have each concept represented by the activity of only about logM units, which would be about 24 neurons in our example. There are many other computational problems with holographic models (Feldman, 1988). For example if a concept required a pattern over all M units, how would that concept combine with other concepts without cross-talk. Even more basically, there is no way that a holographic representation could be transmitted from one brain circuit to another.

There is a wide range of converging experimental evidence (Quiroga et al., 2008; Bowers, 2009) showing that neural encoding relies on a modest number (10–100) of units. There is also some overlap—the same unit can be involved in the representation of different items. For several reasons, not all of them technical, some papers continue to refer to these structured representations as "sparse population codes." A much more appropriate term would be redundant circuits.

There is now a general consensus on the basis of neural spike signaling and encoding. There are a number of specialized neural structures involving delicate timing. The relative time of spike arrival is also important for plasticity. But the main mechanism for neural signaling is frequency encoding in functional circuits of low redundancy.

ACTIONABILITY AND SIMULATION

Given that knowledge is encoded in the brain as active circuits, the next big question concerns the nature of this embodied knowledge. The key idea is that living things and their brains evolved to **act** in the physical and social world. Action is evolutionarily much older than symbolic thought, belief, etc., and is also developmentally much earlier in people. Sensory actions loops like the knee-jerk reflex (**Figure 1**) significantly pre-date neurons and are crucial even for single celled animals such as amoeba (Katz, 2007). Only living things act (in our sense); natural forces, mechanisms, etc. are said to act by metaphorical extension (Lakoff and Johnson, 1980).

Fitness is the technical term for nature's assessment of agents' actions in context. Natural selection assures that creatures with sufficiently bad choices of actions do not survive and reproduce. The term ***actionability*** has been defined as an organism's internal assessment of its available actions in context (Feldman and Narayanan, 2014). Of course, such an internal calculation will rarely be optimal for fitness, but evolution selects systems where the match is good enough.

Actions, in this formulation, include persistent change of internal state: learning, memory, world models, self-concept, etc. In animals, perception is best-fit, active, and utility/affordance based (Parker and Newsome, 1998). The external world (e.g., other agents) is not static so internal models need **simulation**. Simulation involves imagining actions and estimating their likely consequences before actually entailing the risks of trying them in the real world (Bergen, 2013). Both actionability assessment and simulation rely on good (but not veridical) internal models. This is another fundamental property of neural encoding.

Another important issue concerns the roles of **rules**, including logical rules in the brain. Once a simulation has been done successfully, people can cache (remember) the result as a rule and thus shortcut a costly simulation. Search in a symbolic model can be viewed as a form of simulation. Learning generalizations of symbolic rules is a crucial process and not well understood.

Communication is an important form of action and is needed for cooperation, as discussed in Neural Communication and Representation. Even single-celled animals, like some amoebas, rely on pheromones for survival, particularly for organizing into stable structures in times of environmental stress (Shorey, 2013). Higher plants and animals rely on communication actions for many life functions. And, of course, language is a characterizing trait of people. Much of what we know and what we need to learn about "representation in the brain" is concerned with language.

Actionability, not non-tautological truth, is what an agent/animal can actually compute. We have no privileged access to external truth or to our own internal state. This entails the **operationality** of all living things. In science, operationalism states that theories should be evaluated for their explanatory and predictive power, not as assertions of the reality of their terms, e.g., electrons. Living things incorporate structures that

model the external and internal milieus to enhance fitness. Evolution constrains these structures to be consistent with reality.

The basic actionability story applies to all living things, but there are profound differences between different species. One crucial divide/cline is ***volitional*** action and communication – the boundary is not clear, but birds are above the line; protozoans, plants below. We assume that, in nature, neurons are necessary for volition (Damasio, 1999). Volitional actions have automatic components and influence, e.g., speech. For example, deciding to talk is volitional; the details of articulation are automatic.

Learning is obviously a foundation of intelligent activity and also important in much simpler organisms. The current revolution in big data, deep learning, etc., can help provide insights for this enterprise as well as many others, but is not a model for the mechanisms under study. Structure learning remains to be understood. Observational learning without a model is influenced by the observer's ability to act in the situation (Iani et al., 2013). In Nature, there is no evidence for tabula-rasa learning and massive evidence against it.

Language is a hallmark of human intelligence and its representation in the brain is of major importance. From our actionability perspective, the crucial question is the neural encoding of ***meaning***. A tradition dating literally back to the Greeks identifies meaning with "truth" as defined in formal logics. This historical fact wouldn't matter except that the same definition of meaning dominates much current work in formal linguistics, philosophy, and computer science. But action is evolutionarily much older than symbolic thought, belief, etc., and is also developmentally much earlier in people.

Decades of inter-disciplinary work suggests that the definition of meaning should be expressed in an action-oriented formalism (Narayanan, 1999) that maps directly to embodied mechanisms (Feldman, 2005). For example, the meaning of a word like "push" is captured formally as an action schema that captures the preconditions and resources needed for the action as well as the possible results of the action. Furthermore, all actions inherit from a common control schema (Narayanan, 1999) that models general aspects of action including completion, interrupts, repetition, etc. This action formalism is multi-modal: describing execution, recognition, and planning as well as language.

In addition, the meaning of a word like "push" is assumed to engage neural circuits that produce pushing behavior in people and other animals. There are wide ranging findings that indeed words and images about actions do activate much of the same circuitry as carrying out the action (Garagnani and Pulvermüller, 2016). This is strong evidence about the encoding of actions, action images, and action language in the brain. A further extension of actionability theory accounts for the meaning of metaphorical meanings of words like push in examples like "push for a promotion" (Lakoff and Johnson, 1980). Metaphorical mappings are modeled as mappings from some target domain (here, employment) to an embodied source domain. A remarkable range of phenomena are explained by this theory and, again, there is strong neural support for the connection (Bergen, 2013).

This brings us back to simulation, which was discussed earlier as being necessary for modeling the response of external environment (including other agents) to one's actions. Some automatic simulations (like dreams) are well understood in mammals, but people rely upon volitional (intentional, purposeful) simulation for many functions including planning and language (Feldman, 2005). Some remarkable new experiments (Pfeiffer and Foster, 2013) suggest that rodents might exhibit volitional simulation, but this remains controversial.

More generally, simulation is a cornerstone of an extensive effort on language theory, embodiment, and application. Volitional simulation has been proposed as the mechanism of planning, mind-reading, etc. (Bergen, 2013). With an appropriate formalism, simulation can yield both causal and predictive inferences (Pearl, 2000). Results of simulations can be cached (remembered) and generalized as rules. The NTL theory of language and thought entails additional mechanisms including construction grammar, mental spaces, mappings, etc. (Feldman, 2010b).

CONCLUSIONS AND MYSTERIES

This Frontiers project on "Representation in the Brain" is extremely timely; despite recent theoretical and experimental advances, there is still considerable confusion on the topic. As is often the case, part of the problem arises from the use of anachronistic terms like "representation" to describe neural computation. There are also surviving revivals of old theories (like holographic memory and field theory) that are incompatible with current findings. But for the most part, there is a good scientific consensus on what could be called a standard theory of neural computation (Parker and Newsome, 1998). This is based on the activity of individual neurons that participate in multiple complex circuits and communicate primarily through spikes transmitted through axons to synapses with processes of downstream cells.

In addition to our improved understanding of the computational primitives of the brain, there are promising advances on theories and experiments at the functional level. The ancient idea that meaning should be equated with logical truth is being replaced by theories that emphasize the function of brains in interacting with the physical and social environments (Kahneman, 2011). In a related development, the idea of language and thought as logical deduction is giving way to theory and experiment grounded in bodily experience and simulation (Bergen, 2013).

However, there are fundamental questions on neural computation that remain mysteries in that there is no plausible theory to account for them. The general mind-body problem is known to be intractable and currently mysterious (Chalmers, 1996). This is one of many deep problems, including quantum phenomena, etc., that are universally agreed to be beyond the current purview of science. But all of these famous

unsolved problems are either remote from everyday experience (complementarity, dark matter) or are hard to even define sharply (consciousness, free will, etc.).

There are also problematic ordinary behaviors–recent work (Feldman, 2016) describes some obvious problems in vision that arise every time that we open our eyes and yet are demonstrably **incompatible** with current theories of neural computation, including those presented in this article. The focus was on two related phenomena, known as the *neural binding problem* and the *illusion of a stable visual world*. I, among many others, have struggled with these issues for more than 50 years and I now believe that they are both unsolvable within current neuroscience. By considering some basic facts about how the brain processes image input, (Feldman, 2016) shows that there are not nearly enough brain neurons to compute what we experience as vision. We imagine that we perceive an entire scene at full resolution, but only about 1 degree in the fovea is encoded that precisely. However, the area of visual cortex that encodes the fovea is much too large to be replicated ∼400 times to fully encode a full scene in detail.

I suggest that these facts should induce humility about the prospects for our current neuroscience to yield a complete reductionist account of even concrete aspects of vision and other thought processes. So, "representation in the brain" remains one of the central scientific questions of our time, if not of all time.

AUTHOR CONTRIBUTIONS

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

FUNDING

This work was supported in part by ONR grant N000141110416 and a grant from Google.

ACKNOWLEDGMENTS

This review is obviously based on previous reviews and other work. The original synthetic ideas were developed in collaboration with colleagues and students at ICSI and UC Berkeley. Srini Narayanan has been especially helpful.

REFERENCES

Bergen, B. (2013). *Louder than Words*. New York, NY: Basic Books.

Bowers, J. S. (2009). On the biological plausibility of grandmother cells: implications for neural network theories in psychology and neuroscience. *Psychol. Rev.* 116, 252–282. doi: 10.1037/a0014462

Chalmers, D. J. (1996). *The Conscious Mind: in Search of a Fundamental Theory*. New York, NY: Oxford University Press.

Damasio, A. R. (1999). *The Feeling of What Happens*. New York, NY: Harcourt Brace.

Feldman, J., and Narayanan, S. (2014). "Affordances, actionab ility, simulation," in *Proceedings of the Neural-Symbolic Learning and Reasoning Workshop*, (Wadern: Dagstuhl).

Feldman, J. A. (1988). "Computational constraints on higher neural representations," in *Proceedings on Computational Neuroscience*, ed. E. Schwartz (Cambridge, MA: MIT Press).

Feldman, J. A. (2005). *From Molecule to Metaphor, A Neural Theory of Language*. Cambridge, MA: MIT Press.

Feldman, J. A. (2010a). Ecological expected utility and the mythical neural code. *Cogn. Neurodyn.* 4, 25–35. doi: 10.1007/s11571-009-9090-4

Feldman, J. A. (2010b). Embodied language, best-fit analysis, and formal compositionality. *Phys. Life Rev.* 7, 385–410. doi: 10.1016/j.plrev.2010.06.006

Feldman, J. A. (2016). Mysteries of visual experience. *arXiv*. Available at: http://arxiv.org/abs/1604.08612

Garagnani, M., and Pulvermüller, F. (2016). Conceptual grounding of language in action and perception: a neurocomputational model of the emergence of category specificity and semantic hubs. *Eur. J. Neurosci.* 43, 721–737. doi: 10.1111/ejn.13145

Iani, C., Rubichi, S., Ferraro, L., Nicoletti, R., and Gallese, V. (2013). Observational learning without a model is influenced by the observer's possibility to act: evidence from the simon task. *Cognition* 128, 26–34. doi: 10.1016/j.cognition.2013.03.004

Kahneman, D. (2011). *Thinking Fast and Slow*. New York, NY: Farrar, Straus Giroux.

Katz, P. S. (2007). Evolution and development of neural circuits in invertebrates.*Curr. Opin. Neurobiol.* 17, 59–64. doi: 10.1016/j.conb.2006.12.003

Lakoff, G., and Johnson, M. (1980). *Metaphors We Live By*. Chicago, IL: University of Chicago Press.

Lennie, P. (2003). The cost of cortical computation. *Curr. Biol.* 13, 493–497. doi: 10.1016/S0960-9822(03)00135-0

Meech, R. W., and Mackie, G. O. (2007). "Evolution of excitability in lower metazoans," in *Invertebrate Neurobiology*, eds G. North and R. J. Greenspan (New York, NY: Cold Spring Harbor Laboratory Press).

Narayanan, S. (1999). "Reasoning about actions in narrative understanding," in *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, (San Francisco, CA: Morgan Kaufmann), 350–358.

Parker, A. J., and Newsome, W. T. (1998). Sense and the single neuron: probing the physiology of perception. *Annu. Rev. Neurosci.* 21, 227–277. doi: 10.1146/annurev.neuro.21.1.227

Pearl, J. (2000). *Causality: Models, Reasoning and Inference*. New York, NY: Cambridge University Press.

Pfeiffer, B. E., and Foster, D. J. (2013). Hippocampal place-cell sequences depict future paths to remembered goals. *Nature* 497, 74–79. doi: 10.1038/nature 12112

Quiroga, R. Q., Mukamel, R., Isham, E. A., Malach, R., and Fried, I. (2008). Human single-neuron responses at the threshold of conscious recognition. *Proc. Natl. Acad. Sci. U.S.A.* 105, 3599–3604. doi: 10.1073/pnas.0707043105

Shorey, H. H. (2013). *Animal Communication by Pheromones*. New York, NY: Academic Press.

Willshaw, D. J., Buneman, O. P., and Longuet-Higgins, H. C. (1969). Non-holographic associative memory. *Nature* 222, 960–962. doi: 10.1038/222960a0

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

frontiers in Psychology

HYPOTHESIS AND THEORY
published: 16 February 2017
doi: 10.3389/fpsyg.2017.00186

The Theory of Localist Representation and of a Purely Abstract Cognitive System: The Evidence from Cortical Columns, Category Cells, and Multisensory Neurons

*Asim Roy**

Department of Information Systems, Arizona State University, Tempe, AZ, USA

The debate about representation in the brain and the nature of the cognitive system has been going on for decades now. This paper examines the neurophysiological evidence, primarily from single cell recordings, to get a better perspective on both the issues. After an initial review of some basic concepts, the paper reviews the data from single cell recordings – in cortical columns and of category-selective and multisensory neurons. In neuroscience, columns in the neocortex (cortical columns) are understood to be a basic functional/computational unit. The paper reviews the fundamental discoveries about the columnar organization and finds that it reveals a massively parallel search mechanism. This columnar organization could be the most extensive neurophysiological evidence for the widespread use of localist representation in the brain. The paper also reviews studies of category-selective cells. The evidence for category-selective cells reveals that localist representation is also used to encode complex abstract concepts at the highest levels of processing in the brain. A third major issue is the nature of the cognitive system in the brain and whether there is a form that is purely abstract and encoded by single cells. To provide evidence for a single-cell based purely abstract cognitive system, the paper reviews some of the findings related to multisensory cells. It appears that there is widespread usage of multisensory cells in the brain in the same areas where sensory processing takes place. Plus there is evidence for abstract modality invariant cells at higher levels of cortical processing. Overall, that reveals the existence of a purely abstract cognitive system in the brain. The paper also argues that since there is no evidence for dense distributed representation and since sparse representation is actually used to encode memories, there is actually no evidence for distributed representation in the brain. Overall, it appears that, at an abstract level, the brain is a massively parallel, distributed computing system that is symbolic. The paper also explains how grounded cognition and other theories of the brain are fully compatible with localist representation and a purely abstract cognitive system.

Keywords: localist representation, distributed representation, amodal representation, abstract cognitive system, theory of the brain, cortical columns, category cells, multisensory neurons

OPEN ACCESS

Edited by:
George Kachergis, Radboud University Nijmegen, Netherlands

Reviewed by:
Dipanjan Roy, Allahabad University, India
Bruno Lara, Universidad Autonoma del Estado de México, Mexico

***Correspondence:**
Asim Roy
asim.roy@asu.edu

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *12 August 2016*
Accepted: *30 January 2017*
Published: *16 February 2017*

Citation:
Roy A (2017) The Theory of Localist Representation and of a Purely Abstract Cognitive System: The Evidence from Cortical Columns, Category Cells, and Multisensory Neurons. Front. Psychol. 8:186. doi: 10.3389/fpsyg.2017.00186

INTRODUCTION

We have argued for decades about how features of the outside world (both abstract and concrete) are encoded and represented in the brain (Newell and Simon, 1976; Newell, 1980; Smith, 1982; Hinton et al., 1986; Earle, 1987; Smolensky, 1987, 1988; Fodor and Pylyshyn, 1988; Rumelhart and Todd, 1993). In the 70s and 80s, however, when the various theories were proposed and most of the fundamental arguments took place, study of the biological brain was still in its infancy. We, therefore, didn't have much neuroscience data to properly evaluate the competing theories. Thus, the arguments were mainly theoretical. Fortunately, that situation has changed in recent years with a significant amount of research in neurophysiology. We are, therefore, in a better position now to evaluate the competing theories based on real data about the brain.

Freeman and Skarda (1990) have argued that the brain does not need to encode or represent features of the outside world in any explicit way. Representation, however, is a useful abstraction for computer and cognitive sciences and for many other fields and neurophysiology continues to search for correlations between neural activity and features of the external world (Logothetis et al., 1995; Chao and Martin, 2000; Pouget et al., 2000; Freedman et al., 2001; Wang et al., 2004; Quiroga et al., 2005; Samejima et al., 2005; Averbeck et al., 2006; Martin, 2007; Patterson et al., 2007; Kriegeskorte et al., 2008). In fact, the two Nobel prizes in physiology for ground-breaking discoveries about the brain have been about encoding and representation: (1) Hubel and Wiesel's discovery of a variety of fundamental visual processing cells in the primary visual cortex, such as line, edge, color and motion detector cells (Hubel and Wiesel, 1959, 1962, 1968, 1977), and (2) the discovery of place cells by O'keefe and grid cells by Mosers (O'Keefe and Dostrovsky, 1971; O'keefe and Nadel, 1978; Moser et al., 2008). Thus, in this paper, I focus primarily on the two main competing theories of representation – localist vs. distributed.

The cortical column – a cluster of neurons that have similar response properties and which are located physically together in a columnar form across layers of the cortex – is now widely accepted in neuroscience as the fundamental processing unit of the neocortex (Mountcastle, 1997; Horton and Adams, 2005; DeFelipe, 2012). There are some very interesting findings from studies of the cortical columns and it makes sense to understand the nature and operation of cortical columns from a representational and computational point of view. So that is a major focus of this paper.

Encoding of complex abstract concepts is the second major focus of this paper. Distributed representation theorists have always questioned whether the brain is capable of abstracting complex concepts and encoding them in single cells (neurons) or in a collection of cells dedicated to that concept. There was an article in *MedicalExpress* (Zyga, 2012) on localist representation following the publication of Roy (2012). That article includes an extensive critique of localist representation theory by James McClelland. I quote here a few of his responses regarding encoding of complex concepts:

(1) "*what basis do I have for thinking that the representation I have for any concept – even a very familiar one – as associated with a single neuron, or even a set of neurons dedicated only to that concept*?"

(2) "*A further problem arises when we note that I may have useful knowledge of many different instances of every concept I know – for example, the particular type of chicken I purchased yesterday evening at the supermarket, and the particular type of avocados I found to put in my salad. Each of these is a class of objects, a class for which we may need a representation if we were to encounter a member of the class again. Is each such class represented by a localist representation in the brain*?"

As one can sense from these arguments, the nature and means of encoding of complex abstract concepts is a major issue in cognitive science. A particular type of complex abstract concept is the concept of a category. There are several neurophysiological studies on category representation in the brain and they provide some new insights on the nature of encoding of abstract concepts. I review some of those studies that show that single cells can indeed encode abstract category concepts.

I also address the issue of modality-invariant (or amodal) representation, which is also a form of abstraction, and provide evidence for the extensive use of an amodal cognitive system in the brain where such abstractions are encoded by single cells. Finding these different kinds of abstractions in the brain (from categorization to modality-invariance) resolves a long standing dispute within cognitive science – between grounded cognition, which is modality-based, and the traditional cognitive system defined on the basis of abstractions (Borghi and Pecher, 2011). Given the evidence for grounded cognition (Barsalou, 2008) and the various forms of abstractions encoded by single cells, it is fair to claim that both a purely abstract form of cognition and modality-dependent cognition co-exist in the brain providing different kinds of information and each is supported by localist representation.

Finally, I address the issue of distributed representation or population coding (Panzeri et al., 2015) and its conflict with the evidence for localist representation. I essentially argue that there is no evidence for distributed representation because there is no evidence for dense distributed coding. And dense distributed coding is the essential characteristic of distributed representation as claimed by some of the original proponents (McClelland et al., 1995).

The paper has the following structure. In Section "Localist vs. Distributed Representation," I provide the standard definitions for localist and distributed representations and explain the difference between distributed processing and distributed representation. In Section "Columnar Organization in the Neocortex," I explore the neuroscience of columnar organization in the neocortex and what it implies for representational theories. In Section "Category Cells," I review neurophysiological studies that relate to encoding of category concepts in the brain. Section "Multisensory Integration in the Brain" has the evidence for multi-sensory integration and modality-invariant single cells

in the brain. In Section "The Existence of a Single Cell-Based Purely Abstract and Layered Cognitive System and Ties to Grounded Cognition," I argue that there's plenty of evidence for a purely abstract, single-cell based cognitive system in the brain. In addition, I argue that a sensory-based (grounded) non-abstract and a purely abstract cognitive system co-exist and support each other to provide cognition in its various forms. In Section "On the "Meaning and Interpretation" of Single Neuron Response," I explain what "meaning and interpretation" implies for a single cell response. Section "Localist Representation and Symbols" explains why localist neurons are symbols in a computational and cognitive sense. Section "No Evidence for Distributed Representation" argues that there is no neurophysiological evidence for distributed representation because distributed representation is about dense representation. Section "Conclusion" has the conclusions.

LOCALIST VS. DISTRIBUTED REPRESENTATION

Definitions and What They Mean

Distributed representation is generally defined to have the following properties (Hinton et al., 1986; Plate, 2002):

- A concept is represented by a pattern of activity over a collection of neurons (i.e., more than one neuron is required to represent a concept).
- Each neuron participates in the representation of more than one concept.

By contrast, in localist representation, a single neuron represents a single concept on a stand-alone basis. But that doesn't preclude a collection of neurons representing a single concept. The critical distinction between localist units and distributed ones is that localist units have "meaning and interpretation" whereas the distributed ones don't. Many authors have pointed out this distinction.

- Elman (1995, p. 210): "*These representations are distributed, which typically has the consequence that interpretable information cannot be obtained by examining activity of single hidden units.*"
- Thorpe (1995, p. 550): "*With a local representation, activity in individual units can be interpreted directly... with distributed coding individual units cannot be interpreted without knowing the state of other units in the network.*"
- Plate (2002):"*Another equivalent property is that in a distributed representation one cannot interpret the meaning of activity on a single neuron in isolation: the meaning of activity on any particular neuron is dependent on the activity in other neurons (Thorpe, 1995).*"

Thus, the fundamental difference between localist and distributed representation is only in the interpretation and meaning of the units, nothing else. Therefore, any and all kinds of models can be built with either type of representation; there are no limitations as explained by Roy (2012).

Reviewing single cell studies, Roy (2012) found evidence that single cell activations can have "meaning and interpretation," starting from the lowest levels of processing such as the retina. Thus, localist representation is definitely used in the brain. Roy (2013) found that multimodal invariant cells exist in the brain that can easily identify objects and concepts and such evidence supports the grandmother cell theory (Barlow, 1995, 2009; Gross, 2002). This paper builds on those previous ones and provides further evidence for widespread use of localist representation by examining columnar organization of the neocortex and the evidence for category cells.

Other Characteristics of Distributed Representation

(a) Representational efficiency – Distributed representation is computationally attractive because it can store multiple concepts using a small set of neurons. With n binary output neurons, it can represent 2^n concepts because that many different patterns are possible with that collection of binary neurons. With localist representation, n neurons can only represent n concepts. In Section "Columnar Organization in the Neocortex," I explain that this property of distributed representation could be its greatest weakness because such a representation cannot be a feasible structure for processing in the brain, given the evidence for columnar organization of the neocortex.

(b) Mapping efficiency – Distributed representation allows for a more compact overall structure (mapping function) from input nodes to the output ones and that means less number of connections and weights to train. Such a mapping function requires less training data and will generalize better.

(c) Resiliency – A distributed representation based mapping function is resilient in the sense that degradation of a few elements in the network structure may not disrupt or effect the overall performance of the structure.

(d) Sparse distributed representation – A distributed representation is sparse if only a small fraction of the n neurons is used to represent a subset of the concepts. Some argue that representation in the brain is sparse (Földiak, 1990; Olshausen and Field, 1997; Hromádka et al., 2008; Yu et al., 2013).

McClelland et al. (1995), however, have argued that sparse distributed representation doesn't generalize very well and that the brain uses it mainly for episodic memories in the hippocampus. They also argue that dense distributed representation is the only structure that can generalize well and that the brain uses this dense form of representation in the neocortex to learn abstract concepts. Bowers (2009) summarizes this particular theory of McClelland et al. (1995) in the following way: "*On the basis of this analysis, it is argued that sparse coding is employed in the hippocampus in order to store new episodic memories following single learning trials, whereas dense distributed representations are learned slowly and reside in cortex in order to support word, object, and face identification (among other functions), all of which require generalization (e.g.,*

to identify an object from a novel orientation)." The essence of this theory is that only dense representations can generalize and learn abstract concepts. And thus the only form of distributed representation to consider is the dense one.

Distributed Processing vs. Distributed Representation

The interactive activation (IA) model of McClelland and Rumelhart (1981), shown in **Figure 1**, is a classic localist model. The IA model is a localist model simply because the letter-feature, letter and word units have labels on them, which implies that they have "meaning and interpretation." Although the model is localist, it uses distributed and parallel processing. For example, all of the letter units are computed in parallel with inputs from the letter-feature layer. Similarly, all of the word units are computed in parallel with inputs from the letter units layer. Thus, both localist and distributed representation can exploit parallel, distributed processing. The representation type, therefore, does not necessarily place a restriction on the type of processing. And localist representation can indeed parallelize computations.

COLUMNAR ORGANIZATION IN THE NEOCORTEX

Although the neocortex of mammals is mainly characterized by its horizontal layers with different cell types in each layer, researchers have found that there is also a strong vertical organization in some regions such as the somatosensory, auditory, and visual cortices. In those regions, the neuronal responses are fairly similar in a direction perpendicular to the cortical surface, while they vary in a direction parallel to the surface (Goodhill and Carreira-Perpiñán, 2002). The set of neurons in the perpendicular direction have connections between them and form a small, interconnected column of neurons. Lorente de Nó (1934) was the first to propose that the cerebral cortex is formed of small cylinders containing vertical chains of neurons and that these were the fundamental units of cortical operation. Mountcastle (1957) was the first to discover this columnar organization (that is, the clustering of neurons into columns with similar functional properties) in the somatosensory cortex of cats. Hubel and Wiesel (1959, 1962, 1968, 1977) also found this columnar organization in the striate cortex (primary visual cortex) of cats and monkeys.

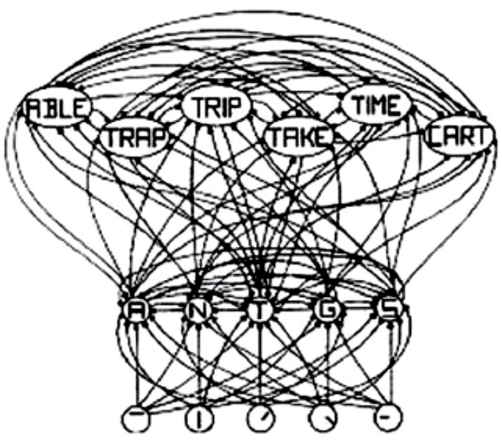


FIGURE 1 | Schematic diagram of a small subcomponent of the interactive activation model. Bottom layer codes are for letter features, second layer codes are for letters, and top layer codes are for complete words, all in a localist manner. Arrows depict excitatory connections between units; circles depict inhibitory connections. Adapted from Figure 3 of McClelland and Rumelhart (1981), by permission of American Psychological Association.

A *minicolumn*, a narrow vertical chain of interconnected neurons across the cortical layers, is considered the basic unit of the neocortex. The number of neurons in these minicolumns generally is between 80 and 100, but can be more in certain regions like the striate cortex. A *cortical column* (or module) consists of a number of minicolumns with horizontal connections. A cortical column is a complex processing unit that receives input and produces outputs. In some cases, the boundaries of these columns are quite obvious (e.g., barrels in the somatosensory cortex and ocular dominance columns in the visual cortex), but not always (e.g., orientation columns in the striate cortex).

Figure 2 shows the "ice cube" models that explain the spatial structure of orientation columns, ocular dominance columns and hypercolumns across layers of the striate cortex. An orientation column has cells that have the same orientation (i.e., they respond to an edge or bar of light with the same orientation) and this columnar structure is repeated in the striate cortex for different orientations and different spatial positions [receptive fields (RFs)] on the retina. Tanaka (2003) notes that: "*Cells within an orientation column share the preferred orientation, while they differ in the preferred width and length of stimuli, binocular disparity, and the sign of contrast.*" *Hypercolumn* (*macrocolumn*) cells, on the other hand, respond to the same spatial position (RF) in the retina, but have different orientation preferences. Orientation preferences generally changes linearly from one column to the next, but can have jumps of 90 or 180°. A hypercolumn (macrocolumn) contains about 50–100 minicolumns. According to Krueger et al. (2008), the neocortex has about 100 million minicolumns with up to 110 neurons in each.

Direction of motion selectivity columns have been found in the middle temporal (MT) visual area of macaque monkeys (Albright et al., 1984; DeAngelis and Newsome, 1999). **Figure 3** shows the distribution of preferred directions of 95 direction-selective lateral intraparietal area (LIP) neurons of two male rhesus monkeys from the study by Fanini and Assad (2009). Out of the 614 MT direction selective neurons monitored by Albright et al. (1984), 55% responded to moving stimuli independent of color, shape, length, or orientation. The response magnitude and tuning bandwidth of the remaining cells depended on stimulus length, but not the preferred direction. They also found that "*cells with a similar direction of motion preference are also organized in vertical columns and cells with opposite direction preferences are located in adjacent columns within a single axis of motion column.*" Diogo et al. (2002) found direction selective clusters of

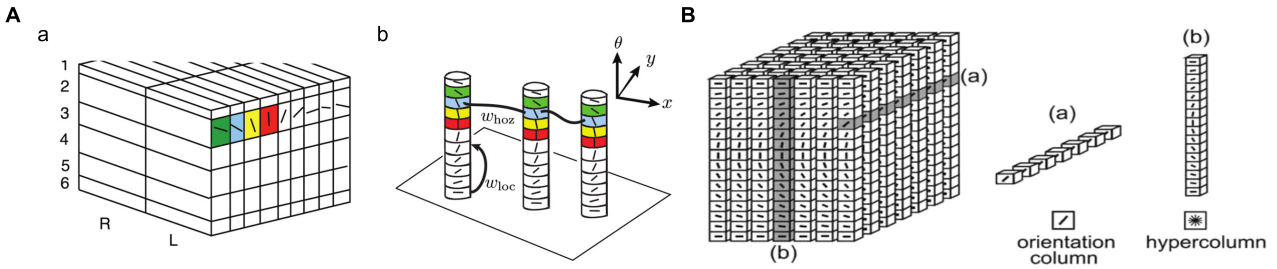


FIGURE 2 | Orientation columns, ocular dominance columns, hypercolumns, and layers of the striate cortex. (A) Adapted from Figure 1 of Bressloff and Carroll (2015). **(B)** Reprinted from Ursino and La Cara (2004), with permission from Elsevier.

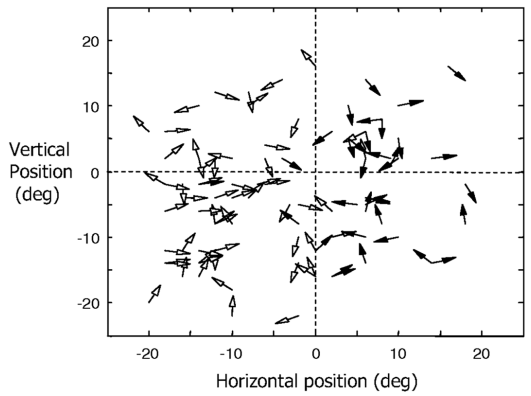


FIGURE 3 | Distribution of preferred directions for 95 direction-selective LIP neurons of two male rhesus monkeys (filled arrowheads for *monkey H* and open arrowheads for *monkey R*). Adapted from Figure 6 of Fanini and Assad (2009), by permission of The American Physiological Society.

cells in the visual area MT of the *Cebus apella* monkey that change gradually across the surface of MT but also had some abrupt 180° discontinuities.

Tanaka (2003) found cells in the inferotemporal cortex (area TE) that selectively respond to complex visual object features and those that respond to similar features cluster in a columnar form. For example, he found cells in a TE column that responded to star-like shapes, or shapes with multiple protrusions in general. Tanaka (2003) notes: "*They are similar in that they respond to star-like shapes, but they may differ in the preferred number of protrusions or the amplitude of the protrusions.*" **Figure 4** shows types of complex objects (complex features) found (or hypothesized) by Tanaka in TE columnar modules. He also notes: "*Since most inferotemporal cells represent features of object images but not the whole object images, the representation of the image of an object requires a combination of multiple cells representing different features contained in the image of the object.*"

In general, neuroscientists have discovered the columnar organization in many regions of the mammalian neocortex. According to Mountcastle (1997), columnar organization is just one form of modular organization in the brain. Mountcastle (1997) notes that the modular structure varies *"in cell type and number, in internal and external connectivity, and in mode of neuronal processing between different large entities."* DeFelipe (2012) states that "*The columnar organization hypothesis is currently the most widely adopted to explain the cortical processing of information. . .*" although there are area and species specific variations and some species, such as rodents, may not have cortical columns (Horton and Adams, 2005). However, Wang et al. (2010) found similar columnar functional modules in laminated auditory telencephalon of an avian species (*Gallus gallus*). They conclude that laminar and columnar properties of the neocortex are not unique to mammals. Rockland (2010) states that columns (as modules) are widely used in the brain, even in non-cortical areas.

Columnar Organization – Its Functional Role and as Evidence for Localist Representation

Neuroscience is still struggling to understand the functional role of columnar organization in cortical processing (Horton and Adams, 2005; DeFelipe, 2012). Here I offer a macro level functional explanation for columnar organization and the way it facilitates fast and efficient processing of information. I also explain why distributed representation (population coding) is inconsistent with and infeasible for the type of superfast processing required in certain parts of the neocortex (and perhaps for other parts of the brain also), where such superfast processing is facilitated by the columnar organization. And columnar organization could be the most extensive neuroscience evidence we have so far for the widespread use of localist representation in the brain.

What the columnar organization reveals is a massively parallel search mechanism – a mechanism that, given an input, searches in parallel for a match within a discrete set of explicitly coded features (concepts). In other words, it tries to match the input, in parallel, to one of the component features in the discrete set, where each such component feature is encoded separately by one or more minicolumns. And the search is parallelized for all similar inputs that arrive simultaneously at a processing stage. That is, each input that arrives at the same time at a processing stage, is processed immediately and separately in a parallel mode. To make this type of parallelized search feasible for multiple inputs, it provides a dedicated macrocolumn (such

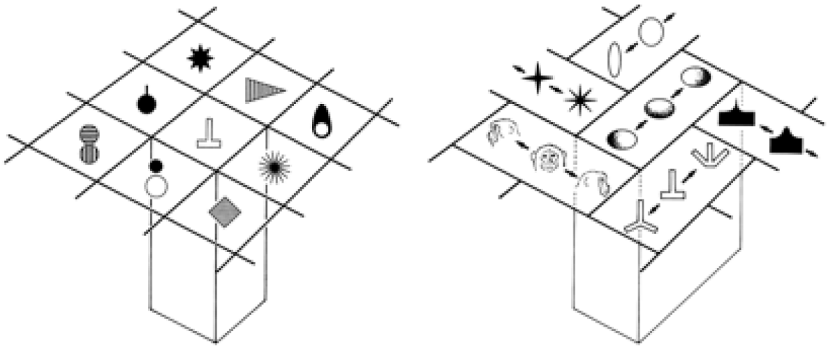


FIGURE 4 | Columnar modules of region TE. Adapted from Figures 3 and 7 of Tanaka (2003), by permission of Oxford University Press.

as a hypercolumn), that encodes the same set of discrete features in its minicolumns, to each and every input (e.g., a RF) so that it can be processed separately in parallel. Horton and Adams (2005) describe a hypercolumn as a structure that contains "*a full set of values for any given set of receptive field parameters*." The discrete set of explicit features (concepts) – which range from simple features (e.g., line orientation) to complex and invariant ones (e.g., a star-like shape) and where the set of features depends on the processing level – is, of course, learned over time.

Thus, the defining principle of columnar organization is this parallel search for a matching explicit feature within a discrete set, given an input, and performing such searches for multiple inputs at the same time (in parallel), where such parallel searches for multiple inputs are facilitated by deploying separate dedicated macrocolumns for each input. This same parallel search mechanism is used at all levels of processing as necessary. This mode of processing is, without question, very resource intensive. However, this mode of processing is an absolute necessity for the neocortex (and elsewhere in the brain) wherever there is a need for incredibly fast processing.

What's really unique about columnar organization is the fact that it creates a discrete set of features (concepts) that are explicit. The features are explicit in the sense that they are interpretable and can be assigned meaning. And that organizing principle provides direct evidence for widespread use of localist representation in the cortex and perhaps other areas of the brain (Page, 2000; Roy, 2012, 2013). Here's an explanation from a computational point of view why columnar organization works that way and why distributed representation, especially dense distributed representation which is hypothesized to be used in the neocortex (McClelland et al., 1995; Poggio and Bizzi, 2004; Bowers, 2009), is not compatible with the processing needs. In dense distributed representation, concepts are coded by means of different patterns of activation across several output units (neurons) of a network. If such a pattern vector, which can code for any number of concepts, is transmitted to another system, that system would have to know how to decode that pattern vector and determine what the concept is. That means that the receiving system would require a decoding processor (a decoder) to understand an incoming pattern vector encoded by signals from a population of neurons. If the columnar organization were to use dense distributed representation to code for features and concepts, it would have to deploy millions of such decoders. That obviously would add layers of processing and slow down the processing of any stimulus. Explicit features, encoded by one or more neurons in cortical columns, make the interpretation (decoding) task simple for subsequent processes. Thus, learning of explicit features by the columnar organization could be mainly about simplification of computations and to avoid a complex decoding problem at every stage of processing.

CATEGORY CELLS

There is significant evidence at this point that animal brains, from insects to humans, have the ability to generalize and create abstract categories and concepts and encode and represent them in single cells or multiple cells, where each group of such cells is dedicated to a single category or concept. This reveals a lot about mental representation in the brain. This aspect of abstraction and representation of such abstractions has been ignored and denied in the distributed representation theory.

The Evidence for Abstract Category Cells

Regarding the ability to create abstract categories, Freedman and Miller (2008) notes (p. 312): "*Categorization is not an ability that is unique to humans. Instead, perceptual categorization and category-based behaviors are evident across a broad range of animal species, from relatively simple creatures like insects to primates*." Researchers have found such abstraction capability in a variety of studies of animals and insects. Wyttenbach et al. (1996), for example, found that crickets categorize the sound frequency spectrum into two distinct groups – one for mating calls and the other for signals of predatory bats. Schrier and Brady (1987), D'amato and Van Sant (1988) and others have found that monkeys can learn to categorize a large range of natural stimuli. Roberts and Mazmanian (1988) found that pigeons and monkeys can learn to distinguish between animal and non-animal pictures. Wallis et al. (2001) recorded from single neurons in the prefrontal cortex (PFC) of monkeys that learned to distinguish whether

two successively presented pictures were same or different. Fabre-Thorpe et al. (1998) found that monkeys can accurately categorize images (food vs. non-food, animal vs. non-animal) with remarkable speed in briefly flashed stimuli. They conclude: "*Overall, these findings imply that rapid categorization of natural images in monkeys must rely, as in humans, on the existence of abstract categorical concepts*."

Merten and Nieder (2012) found single neurons in the PFC of two rhesus monkeys that encoded abstract "yes" and "no" decisions from judgment about the presence or absence of a stimulus. They note the following (p. 6291): "*we report a predominantly categorical, binary activation pattern of "yes" or "no" decision coding*." Rolls et al. (1997) found viewpoint-independent spatial view cells in the vicinity of the hippocampus in monkeys. These cells responded when the monkey looked toward a particular view, independent of the place where the monkey is or its head direction. Vogels (1999) found single cells in the anterior temporal cortex of two rhesus monkeys that were involved in distinguishing trees from non-trees in color images. About a quarter of those neurons responded in a category-specific manner (that is, either trees or non-trees). And the responses were mostly invariant to stimulus transformation, e.g., to changes in position and size.

Lin et al. (2007) report finding "nest cells" in the mouse hippocampus that fire selectively when the mouse observes a nest or a bed, regardless of the location or the environment. For example, they found single cells that drastically increased the firing rate whenever the mouse encountered a nest. If the mouse looked away from the nest, that single cell became inactive. In testing for invariance, they note (p. 6069): "*Together, the above experiments suggest that the responses of the nest cell remained invariant over the physical appearances, geometric shapes, design styles, colors, odors, and construction materials, thereby encoding highly abstract information about nests. The invariant responses over the shapes, styles, and materials were also observed in other nest cells*."

Other single cell studies of the monkey visual temporal cortex have discovered neurons that respond selectively to abstract patterns or common, everyday objects (Fujita et al., 1992; Logothetis and Sheinberg, 1996; Tanaka, 1996; Freedman and Miller, 2008). Freedman and Miller (2008) summarize these findings from single cell recordings quite well (p. 321): "*These studies have revealed that the activity of single neurons, particularly those in the prefrontal and posterior parietal cortices (PPCs), can encode the category membership, or meaning, of visual stimuli that the monkeys had learned to group into arbitrary categories*."

Different types of faces, or faces in general, represent a type of abstract categorization. Face-selective cells have been a dominant area of investigation in the last few decades. Bruce et al. (1981) were the first ones to find face selective cells in the monkey temporal cortex. Rolls (1984) found face cells in the amygdala and Kendrick and Baldwin (1987) found face cells in the cortex of the sheep. Gothard et al. (2007) studied neural activity in the amygdala of monkeys as they viewed images of monkey faces, human faces and objects on a computer monitor. They found single neurons that respond selectively to images from each category. They also found one neuron that responded to threatening monkey faces in particular. Their general observation is (p. 1674): "*These examples illustrate the remarkable selectivity of some neurons in the amygdala for broad categories of stimuli*." Tanaka (2003) also observed single cell representation of faces and observes: "*Thus, there is more convergence of information to single cells for representations of faces than for those of non-face objects*."

On the human side, in experiments with epileptic patients, Fried et al. (1997) found some single medial temporal lobe (MTL) neurons that discriminate between faces and inanimate objects and others that respond to specific emotional expressions or facial expression and gender. Kreiman et al. (2000), in similar experiments with epileptic patients, found MTL neurons that respond selectively to categories of pictures including faces, houses, objects, famous people and animals and they show a strong degree of invariance to changes in the input stimuli. Kreiman et al. (2000) report as follows: "*Recording from 427 single neurons in the human hippocampus, entorhinal cortex and amygdala, we found a remarkable degree of category-specific firing of individual neurons on a trial-by-trial basis. . .. Our data provide direct support for the role of human medial temporal regions in the representation of different categories of visual stimuli*." Recently, Mormann et al. (2011) analyzed responses from 489 single neurons in the amygdalae of 41 epilepsy patients and found that individual neurons in the right amygdala are particularly selective of pictures of animals and that it is independent of emotional dimensions such as valence and arousal.

In reviewing these findings, Gross (2000) observes: "*Electrophysiology has identified individual neurons that respond selectively to highly complex and abstract visual stimuli*." According to Pan and Sakagami (2012), "*experimental evidence shows that the PFC plays a critical role in category formation and generalization*." They claim that the prefrontal neurons abstract the commonality across various stimuli. They then categorize them on the basis of their common meaning by ignoring their physical properties. These PFC neurons also learn to create boundaries between significant categories.

Can We Believe these Studies? Are They Truly Category-Selective Cells?

These studies, that claim category-selective response of single cells, are often dismissed because, in these experiments, the cells are not exhaustively evaluated against a wide variety of stimuli. Desimone (1991) responds to that criticism with respect to face cell studies: "*Although they do not provide absolute proof, several studies have tried and failed to identify alternative features that could explain the properties of face cells*." For example, many studies tested the face cells with a variety of other stimulus, including textures, brushes, gratings, bars and edges of various colors, and models of complex objects, such as snakes, spiders, and food, but there was virtually no response to any such stimulus (Bruce et al., 1981; Perrett et al., 1982; Desimone et al., 1984; Baylis et al., 1985; Rolls and Baylis, 1986; Saito et al., 1986). In fact, each such face cell responded to a variety of faces, including

real ones, plastic models, and photographs of different faces (e.g., monkey, human). Rolls and Baylis (1986) found that many face cells actually respond to faces over more than a 12-fold range in the size. Others report that many face cells respond over a wide range of orientations in the horizontal plane (Perrett et al., 1982, 1988; Desimone et al., 1984; Hasselmo et al., 1989). Desimone (1991) concludes: "*Taken together, no hypothesis, other than face selectivity, has yet been advanced that could explain such complex neuronal properties.*"

Are Category-Selective Cells Part of a Dense Distributed Representation? If So, Do We Need Exhaustive Testing to Find that Out?

A dense distributed representation uses a small set of neurons to code for many different concepts. The basic idea is compressed encoding of concepts using a small physical structure. This also means that different levels of activations of these neurons will code for different concepts. In other words, for any given concept, most of the neurons in such a representation should be active at a certain level. If that is the case and if a so-called "category-selective" cell is actually a part of a dense representation, then stimuli that belong to different abstract concepts should activate the so-called "category-selective" cell quite often. There is no need for exhaustive testing with different stimuli to find that the "category-selective" cell is part of a dense representation. Testing with just a few different types of stimuli should be sufficient to verify that a cell is either part of a dense representation that codes for complex concepts or codes for a lower level feature. And that's what is usually done in these neurophysiological studies and that should be sufficient. That doesn't mean that rigorous testing is not required. It only means that we don't need exhaustive testing to establish that a cell is selective of certain types of stimuli.

MULTISENSORY INTEGRATION IN THE BRAIN

Research over the last decade or so has produced a large body of evidence for multisensory integration in the brain and even in areas that were previously thought to be strictly unisensory or unimodal. Ghazanfar and Schroeder (2006) claim that multisensory integration extend into early sensory processing areas of the brain and that neocortex is essentially multisensory. Stein and Stanford (2008) observes that many areas that were previously classified as unisensory contain multisensory neurons. This has been revealed by anatomical studies that show connections between unisensory cortices and by imaging and ERP studies that reveal multisensory activity in these regions. Klemen and Chambers (2012), in a recent article, notes that there is now "*broad consensus that most, if not all, higher, as well as lower level neural processes are in some form multisensory.*" The next two sections examine some specific evidence for multisensory integration.

The Evidence for Multisensory Integration in Various Parts of the Brain

Neurons in the lateral intraparietal (LIP) area of the PPC are now known to be multisensory, receiving a convergence of eye position, visual and auditory signals (Andersen et al., 1997). Ventral intraparietal area (VIP) neurons have been found to respond to visual, auditory, somatosensory and vestibular stimuli, and for bi- or tri-modal VIP neurons, RFs driven through different modalities usually overlap in space (Duhamel et al., 1998). Graziano et al. (1999) found neurons in the premotor cortex that responded to visual, auditory and somatosensory inputs. Maier et al. (2004) found that the function of these neurons appear to be 'defense' related in the sense that monkeys (and humans) are sensitive to visual, auditory and multisensory looming signals that indicate approaching danger. Morrell (1972) reported that up to 41% of visual neurons could be driven by auditory stimuli. Single unit recordings in the IT cortex of monkeys performing a crossmodal delayed-match-to-sample task shows that the ventral temporal lobe may represent objects and events in a modality invariant way (Gibson and Maunsell, 1997). Saleem et al. (2013) recorded from mice that traversed a virtual environment and found that nearly half of the primary visual cortex (V1) neurons were part of a multimodal processing system that integrated visual motion and locomotion during navigation. In an anatomical study, Budinger and Scheich (2009) show that the primary auditory field AI in a small rodent, the Mongolian gerbil, has multiple connections with auditory, non-auditory sensory (visual, somatosensory, olfactory), multisensory, motor, "higher order" associative and neuromodulatory brain structures. They observe that these connections possibly mediate multimodal integration processes at the level of AI. Some studies have shown that auditory (Romanski and Goldman-Rakic, 2002), visual (Wilson et al., 1993; O'Scalaidhe et al., 1999; Hoshi et al., 2000), and somatosensory (Romo et al., 1999) responsive neurons are located within the ventrolateral prefrontal cortex (VLPFC), suggesting that VLPFC is multisensory.

The Evidence for Modality-Invariant Single Cell Representation in the Brain

Here, I review some of the evidence for modality-invariant single cells in the brain of humans and non-human.

Fuster et al. (2000) were the first to find that some PFC cells in monkeys integrate visual and auditory stimuli across time by having them associate a tone of a certain pitch for 10 s with a color. PFC cells responded selectively to tone and most of them also responded to colors as per the task rules. They conclude that PFC neurons are part of an integrative network that represent cross modal associations. Romanski (2007) recorded from the VLPFC of rhesus macaques as they were presented with audiovisual stimuli and found that some cells in VLPFC are multisensory and respond to both facial gestures and corresponding vocalizations. Moll and Nieder (2015) trained carrion crows to perform a bimodal delayed paired associate task in which the crows had to match auditory stimuli to delayed visual items. Single-unit recordings from the area

nidopallium caudolaterale (NCL) found memory signals that selectively correlated with the learned audio-visual associations across time and modality. Barraclough et al. (2005) recorded from 545 single cells in the temporal lobe (upper and lower banks of the superior temporal sulcus (STS) and IT) from two monkeys to measure the integrative properties of single neurons using dynamic stimuli, including vocalizations, ripping paper, and human walking. They found that 23% of STS neurons that are visually responsive to actions are modulated significantly by the corresponding auditory stimulus. Schroeder and Foxe (2002), using intracranial recordings, have confirmed multisensory convergence in the auditory cortex in macaque monkeys. Using single microelectrode recordings in anesthetized monkeys, Fu et al. (2003) confirmed that such convergence in the auditory cortex occurs at the single neuron level.

In some experiments, reported in Quian Quiroga et al. (2009) and others, they found that single MTL neurons can encode an object-related concept irrespective of how it is presented – visual, textual, or sound. They checked the modality invariance properties of a neuron by showing the subjects three different pictures of the particular individual or object that a unit responds to and their spoken and written names. In these experiments, they found a neuron in the left anterior hippocampus that fired selectively to three pictures of the television star Oprah Winfrey and to her written and spoken name (Quian Quiroga et al., 2009, p. 1308). The neuron also fired to a lesser degree to a picture of actress Whoopi Goldberg. And none of the other responses of the neuron were significant, including to other text and sound presentations. They also found a neuron in the entorhinal cortex of a subject that responded (Quian Quiroga et al., 2009, p. 1308) "*selectively to pictures of Saddam Hussein as well as to the text 'Saddam Hussein' and his name pronounced by the computer. . ... There were no responses to other pictures, texts, or sounds.*"

Quian Quiroga (2012, p. 588) found a hippocampal neuron which responded selectively to pictures of Halle Berry, even when she was masked as Catwoman (a character she played in a movie). And it also responded to the letter string "HALLE BERRY," but not to other names. They also found that a large proportion of MTL neurons respond to both pictures and written names of particular individuals (or objects) and could also be triggered by the name of a person pronounced by synthesized voice. Hence, they conclude: "*These and many other examples suggest that MTL neurons encode an abstract representation of the concept triggered by the stimulus.*" Quian Quiroga et al. (2008) estimate that 40% of MTL cells are tuned to such explicit representation.

Suthana and Fried (2012, p. 428) found an MTL neuron that responded to a picture of the Sydney Opera House but not to 50 other landmarks. It also responded to "*many permutations and physically different representations of the Sydney Opera House, seen in color, in black and white, or from different angles.*" The same neuron also responded to the written words "Sydney Opera." Nieder (2013) found single neurons in a parieto-frontal cortical network of non-human primates that are selectively tuned to number of items. He notes that: "*Such 'number neurons' can track items across space, time, and modality to encode numerosity in a most abstract, supramodal way.*"

THE EXISTENCE OF A SINGLE CELL-BASED PURELY ABSTRACT AND LAYERED COGNITIVE SYSTEM AND TIES TO GROUNDED COGNITION

Sections "Category Cells and Multisensory Integration in the Brain" on category cells and multisensory, modality-invariant cells provide significant biological evidence for the existence of a single cell-based purely abstract cognitive system in the brain. The multisensory cells are abstract in the sense that they integrate information from more than one sensory process. And since the multisensory neurons are also present in what are generally considered to be unisensory areas, such an abstract cognitive system is well-spread out in various parts of the brain and not confined to a few areas. This does not mean that cognition in appropriate cases is not grounded in sensory-motor processes (Barsalou, 2008, 2010; Pezzulo et al., 2013). In this section, I extend a well-known abstract model of cognition and show how abstract cognition could be connected to modality-based representations, memory and sensory processes and invoke them as necessary. And it is fair to claim, based on the biological evidence, that both the abstract and non-abstract systems co-exist in the brain and are tightly integrated.

Let's now examine an often referenced abstract model of cognition from Collins and Quillian (1969) shown in **Figure 5**. Rogers and McClelland (2004, 2008) uses the same model to illustrate how distributed representation might be able to create the same semantic structure. **Figure 5** shows a possible way of storing semantic knowledge where semantics are based on a hierarchy of abstract concepts and their properties. Given the evidence for category and multisensory abstract cells, this model now looks fairly realistic. In this tree structure, nodes represent abstract categories or concepts and arrows reflect properties of that category or concept. For example, the node *bird* has arrows for the properties *feathers*, *fly,* and *wings*. The arrows point to other nodes that represent these properties, which are also abstract concepts. The semantic tree shows the hierarchical relationship of these abstract concepts and categories. For example, *plant* and *animal* are subcategories of *living thing*. Here, nodes pass down their properties to the descendant nodes. For example, *salmon* inherits all the properties of *fish* (*scales*, *swim,* and *gills*) and also the properties of *animal* (*move*, *skin*) and *living thing* (*grow*, *living*). The properties of higher level concepts reflect the common properties of lower level concepts. The tree produces propositions such as: *living things grow*; *a plant is a living thing*; *a tree is a plant*; and *an oak is a tree*. It therefore follows that *an oak can grow*.

This model can be easily extended to include modality-based representations, memory and sensory processes including simulations. For example, the *robin* node could be a multimodal invariant abstraction that is activated by the physical appearance of a robin (or its picture), by its singing and by the written or spoken name "robin." However, multisensory integration exists at many levels of processing. For example, there could be a multisensory neuron that integrates information from just the visual and auditory systems. That is, it fires with the physical

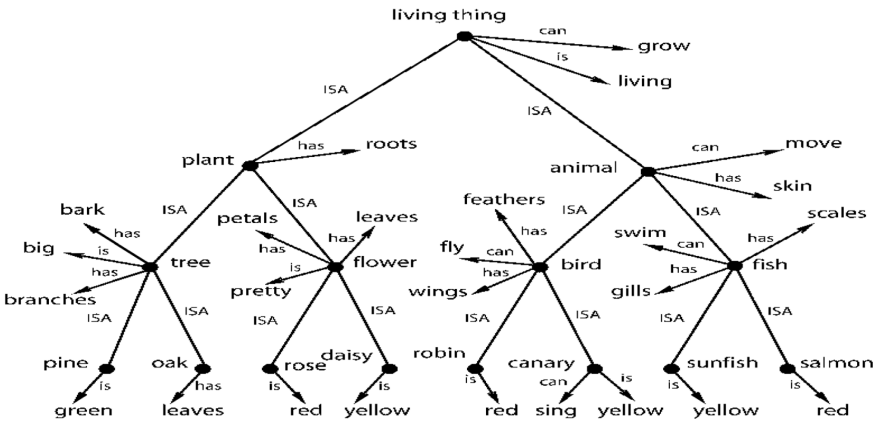


FIGURE 5 | A taxonomic hierarchy of the type used by Collins and Quillian (1969). Adapted from Figure 2, Rogers and McClelland (2008) reproduced with permission.

appearance of a robin (or its picture) and/or when it sings. Many other combinations of sensory information are possible – two at a time, three at a time and so on.

Thus, there could be a layered structure of abstractions in the brain, starting with bi-modals, then tri-modals and so on. And Section "Multisensory Integration in the Brain" cites evidence for such different levels of abstractions. One can think of this layered structure of abstractions in terms of an inverted tree (similar to **Figure 5**) culminating in a single, high-level multimodal abstraction such as the *robin* node of **Figure 5**. Inversely, one can think of the *robin* node having deep extensions into lower levels of modality invariant neurons through an extended tree structure. The lowest level bi-modal invariant nodes, in turn, could be coupled with modal-based representations, memories and sensory processes. A modal representation of a robin in the visual system could have links to a memory system that has one or more generic pictures of robins in different colors and thereby provide access to the imagery part of cognition (Kosslyn et al., 2006). A visual system can also trigger a simulation of the bird flying (Goldman, 2006).

In summary, a purely abstract cognitive system could be tightly integrated with the sensory system and the integration could be through the layered level of abstractions that various multisensory neurons provide. In other words, the conjecture is that a purely abstract cognitive system co-exists with a sensory-based cognition system and perhaps is mutually dependent. For example, the fastest way to trigger the visualization of robins on hearing some robins singing in the background could be through the multisensory (bi-modal) neurons embedded in the sensory systems. The abstract cognitive system could, in fact, provide the connectivity between the sensory systems and be the backbone of cognition in its various forms. So the second part of this Barsalou (2008, p. 618) statement is very consistent with the claims in this section: "*From the perspective of grounded cognition, it is unlikely that the brain contains amodal symbols; if it does, they work together with modal representations to create cognition*." And Sections "Multisensory Integration in the Brain and The Existence of a Single Cell-Based Purely Abstract and Layered Cognitive System and Ties to Grounded Cognition" answers another Barsalou question (p. 631): "*Can empirical evidence be found for the amodal symbols still believed by many to lie at the heart of cognition*?"

ON THE "MEANING AND INTERPRETATION" OF SINGLE NEURON RESPONSE

I come back to the issue of "meaning and interpretation" of the response of a single neuron, an issue that is crucial to the claims of both localist representation and a purely abstract cognitive system. Instead of getting into a philosophical discussion on meaning of the term "meaning," it would be better if we grounded the discussion in neurophysiology. In neurophysiology, the purpose of testing single neurons with different stimuli is to find the correlation between the response and the collection of stimuli that causes it. This is the "meaning and interpretation" of the response to an external observer such as a scientist. From an internal point of view of the brain, the firing of a neuron can have a cascading effect and trigger other neurons to fire and this generates extra information or knowledge. This is best explained with reference to **Figure 5** and the discussions in Sections "Multisensory Integration in the Brain and The Existence of a Single Cell-Based Purely Abstract and Layered Cognitive System and Ties to Grounded Cognition." For example, when we see a robin, it would fire a bi-modal neuron that associates the physical appearance of a robin with its singing. This and other multisensory neurons would, in turn, cause the multimodal invariant *robin* node of **Figure 5** to fire. That firing, in turn, would cause the other associated nodes of **Figure 5** to fire, such as the nodes *bird*, *animal*, *living thing* and their associated properties. What this means is that the brain activates and collects a body of knowledge after seeing the robin. And that body of knowledge, from multiple cell activations, is the composition of internal meaning of robin in the brain. And that whole body of knowledge

can be activated by any and all of the sensory modalities. And that body of knowledge is the sense of "meaning" internal to the brain. And we observe this body of knowledge when we find the multisensory and abstract neurons in the brain. Of course, a simple line orientation cell or a color detection cell may not activate such a large body of abstract knowledge internally in the brain. But these cells still have both internal and external meaning in a similar sense.

LOCALIST REPRESENTATION AND SYMBOLS

An obvious question is, in what way is localist representation symbolic? I explain it here in a computational sense without getting into a philosophical discussion of symbols. One can think of the neurons, in parts of the brain that use localist representation, as being a unit of memory in a computing system that is assigned to a certain variable. The variables in this case range from a purely abstract concept (e.g., a bird) to something as concrete as a short line segment with a certain orientation. And when any of these neurons fire, it transmits a signal to another processor. These processors could, in turn, be neurons in the next layer of a sensory cortex, in the working memory of the PFC or any other neurons it is connected to. Thus, a localist neuron not only represents a variable in the computing sense, but also does processing at the same time. And, in this computational framework, the so-called variables represented by the localist neurons have meaning inside the brain and are also correlated with stimuli from the external world, as explained in Section "Localist Representation and Symbols." Hence, these localist neurons are symbols both in the computing sense and because they are correlated with certain kinds of external stimuli.

NO EVIDENCE FOR DISTRIBUTED REPRESENTATION

As mentioned in Section "Other Characteristics of Distributed Representation," McClelland et al. (1995) have argued that sparse distributed representation does not generalize very well and that the brain uses it mainly for episodic memories in the hippocampus. They also argue that dense distributed representation is the only structure that can generalize well and that the brain uses this dense form of representation in the cortex to learn abstract concepts. And thus the only form of distributed representation to consider is the dense one. But no one has found a dense form of coding anywhere in the brain. In a recent review article, Panzeri et al. (2015) summarize the findings of population coding studies as follows (p. 163): "*. . . a small but highly informative subset of neurons is sufficient to carry essentially all the information present in the entire observed population.*" They further observe that (pp. 163–164): "*This picture is consistent with the observed sparseness of cortical activity* (Barth and Poulet, 2012) *(at any moment only a small fraction of neurons are active) and is compatible with studies showing that perception and actions can be driven by small groups of neurons* (Houweling and Brecht, 2008)." These observations are also supported by other studies (Olshausen and Field, 1997; Hromádka et al., 2008; Ince et al., 2013; Yu et al., 2013). And these findings are quite consistent with findings on multisensory neurons that indicate that a lot of information can be coded in a compact form by a small set of neurons.

CONCLUSION

Neurophysiology has provided a significant amount of information about how the brain works. Based on these numerous studies, one can generalize and claim that the brain uses single cells (or a collection of dedicated cells) to encode particular features and abstract concepts at various levels of processing. One can also claim, based on the evidence for multisensory neurons and category cells, that the brain has a purely abstract and layered cognitive system that is also based on single cell encoding. And that abstract cognitive system, in turn, is connected to the sensory processes and memory. The combined abstract and non-abstract cognitive systems provide the backbone for cognition in its various forms. Parts of the abstract system are also embedded in the sensory systems and provide fast connectivity between the non-abstract systems. This kind of architecture has real value in terms of simplification, concreteness, automation, and computational efficiency. It essentially automates the recognition of familiar patterns at every processing layer and module and delivers such information to other layers and modules in a simplified form.

Cells that encode features and abstract concepts have meaning and interpretation at the cognitive level. Thus, these cells provide easy and efficient access to cognitive level information. Thus far, we have had no clue where cognitive level information was in the brain. These neurophysiological studies are slowly revealing that secret. It could be claimed that these feature and abstract concept cells provide the fundamental infrastructure for cognition and thought.

From these neurophysiological studies, it appears that, at an abstract level, the brain is a massively parallel, distributed computing system that is symbolic. It employs symbols from the earliest levels of processing, such as with discrete sets of feature symbols for line orientation, direction of motion and color, to the highest levels of processing, in the form of abstract category cells and other modality-invariant concept cells.

AUTHOR CONTRIBUTIONS

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

REFERENCES

Albright, T. D., Desimone, R., and Gross, C. G. (1984). Columnar organization of directionally selective cells in visual area MT of the macaque. *J. Neurophysiol.* 51, 16–31.

Andersen, R. A., Snyder, L. H., Bradley, D. C., and Xing, J. (1997). Multimodal representation of space in the posterior parietal cortex and its use in planning movements. *Annu. Rev. Neurosci.* 20, 303–330. doi: 10.1146/annurev.neuro.20.1.303

Averbeck, B. B., Latham, P. E., and Pouget, A. (2006). Neural correlations, population coding and computation. *Nat. Rev. Neurosci.* 7, 358–366. doi: 10.1038/nrn1888

Barlow, H. (2009). "Grandmother cells, symmetry, and invariance: how the term arose and what the facts suggest," in *The Cognitive Neurosciences*, 4th Edn, ed. M. Gazzaniga (Cambridge, MA: MIT Press), 309–320.

Barlow, H. B. (1995). "The neuron doctrine in perception," in *The Cognitive Neurosciences*, ed. M. S. Gazzaniga (Cambridge, MA: MIT Press), 415–434.

Barraclough, N. E., Xiao, D., Baker, C. I., Oram, M. W., and Perrett, D. I. (2005). Integration of visual and auditory information by superior temporal sulcus neurons responsive to the sight of actions. *J. Cogn. Neurosci.* 17, 377–391. doi: 10.1162/0898929053279586

Barsalou, L. W. (2008). Grounded cognition. *Annu. Rev. Psychol.* 59, 617–645. doi: 10.1146/annurev.psych.59.103006.093639

Barsalou, L. W. (2010). Grounded cognition: Past, present, and future. *Top. Cogn. Sci.* 2, 716–724. doi: 10.1111/j.1756-8765.2010.01115.x

Barth, A. L., and Poulet, J. F. (2012). Experimental evidence for sparse firing in the neocortex. *Trends Neurosci.* 35, 345–355. doi: 10.1016/j.tins.2012.03.008

Baylis, G. C., Rolls, E. T., and Leonard, C. M. (1985). Selectivity between faces in the responses of a population of neurons in the cortex in the superior temporal sulcus of the monkey. *Brain Res.* 342, 91–102. doi: 10.1016/0006-8993(85)91356-3

Borghi, A. M., and Pecher, D. (2011). Introduction to the special topic embodied and grounded cognition. *Front. Psychol.* 2:187. doi: 10.3389/fpsyg.2011.00187

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

Bressloff, P. C., and Carroll, S. R. (2015). Laminar neural field model of laterally propagating waves of orientation selectivity. *PLoS Comput. Biol.* 11:e1004545. doi: 10.1371/journal.pcbi.1004545

Bruce, C., Desimone, R., and Gross, C. G. (1981). Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. *J. Neurophysiol.* 46, 369–384.

Budinger, E., and Scheich, H. (2009). Anatomical connections suitable for the direct processing of neuronal information of different modalities via the rodent primary auditory cortex. *Hear. Res.* 258, 16–27. doi: 10.1016/j.heares.2009.04.021

Chao, L. L., and Martin, A. (2000). Representation of manipulable man-made objects in the dorsal stream. *Neuroimage* 12, 478–484. doi: 10.1006/nimg.2000.0635

Collins, A., and Quillian, M. (1969). Retrieval time from semantic memory. *J. Verbal Learn. Verbal Behav.* 8, 240–247. doi: 10.1016/S0022-5371(69)80069-1

D'amato, M. R., and Van Sant, P. (1988). The person concept in monkeys (*Cebus apella*). *J. Exp. Psychol.* 14, 43–55.

DeAngelis, G. C., and Newsome, W. T. (1999). Organization of disparity-selective neurons in macaque area MT. *J. Neurosci.* 19, 1398–1415.

DeFelipe, J. (2012). The neocortical column. *Front. Neuroanat.* 6:22. doi: 10.3389/fnana.2012.00022

Desimone, R. (1991). Face-selective cells in the temporal cortex of monkeys. *J. Cogn. Neurosci.* 3, 1–8. doi: 10.1162/jocn.1991.3.1.1

Desimone, R., Albright, T. D., Gross, C. G., and Bruce, C. (1984). Stimulus-selective properties of inferior temporal neurons in the macaque. *J. Neurosci.* 4, 2051–2062.

Diogo, A. C. M., Soares, J. G., Albright, T. D., and Gattass, R. (2002). Two-dimensional map of direction selectivity in cortical visual area MT of *Cebus* monkey. *An. Acad. Bras. Ciênc.* 74, 463–476. doi: 10.1590/S0001-37652002000300009

Duhamel, J. R., Colby, C. L., and Goldberg, M. E. (1998). Ventral intraparietal area of the macaque: congruent visual and somatic response properties. *J. Neurophysiol.* 79, 126–136.

Earle, D. C. (1987). On the differences between cognitive and noncognitive systems. *Brain Behav. Sci.* 10, 177–178. doi: 10.1017/S0140525X00047397

Elman, J. (1995). "Language as a dynamical system," in *Mind as Motion: Explorations in the Dynamics of Cognition*, eds R. Port and T. van Gelder (Cambridge, MA: MIT Press), 195–223.

Fabre-Thorpe, M., Richard, G., and Thorpe, S. J. (1998). Rapid categorization of natural images by rhesus monkeys. *Neuroreport* 9, 303–308. doi: 10.1097/00001756-199801260-00023

Fanini, A., and Assad, J. A. (2009). Direction selectivity of neurons in the macaque lateral intraparietal area. *J. Neurophysiol.* 101, 289–305. doi: 10.1152/jn.00400.2007

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

Földiak, P. (1990). Forming sparse representations by local anti-Hebbian learning. *Biol. Cybern.* 64, 165–170. doi: 10.1007/BF02331346

Freedman, D., and Miller, E. (2008). Neural mechanisms of visual categorization: insights from neurophysiology. *Neurosci. Biobehav. Rev.* 32, 311–329. doi: 10.1016/j.neubiorev.2007.07.011

Freedman, D. J., Riesenhuber, M., Poggio, T., and Miller, E. K. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. *Science* 291, 312–316. doi: 10.1126/science.291.5502.312

Freeman, W. J., and Skarda, C. A. (1990). "Representations: who needs them," in *Brain Organization and Memory: Cells, Systems, and Circuits*, eds J. L. McGaugh, N. M. Weinberger, and G. Lynch (Oxford: Oxford University Press), 375–380.

Fried, I., McDonald, K., and Wilson, C. (1997). Single neuron activity in human hippocampus and amygdala during recognition of faces and objects. *Neuron* 18, 753–765. doi: 10.1016/S0896-6273(00)80315-3

Fu, K. M. G., Johnston, T. A., Shah, A. S., Arnold, L., Smiley, J., Hackett, T. A., et al. (2003). Auditory cortical neurons respond to somatosensory stimulation. *J. Neurosci.* 23, 7510–7515.

Fujita, I., Tanaka, K., Ito, M., and Cheng, K. (1992). Columns for visual features of objects in monkey inferotemporal cortex. *Nature* 360, 343–346. doi: 10.1038/360343a0

Fuster, J. M., Bodner, M., and Kroger, J. K. (2000). Cross-modal and cross-temporal association in neurons of frontal cortex. *Nature* 405, 347–351. doi: 10.1038/35012613

Ghazanfar, A. A., and Schroeder, C. E. (2006). Is neocortex essentially multisensory? *Trends Cogn. Sci.* 10, 278–285. doi: 10.1016/j.tics.2006.04.008

Gibson, J. R., and Maunsell, J. H. (1997). Sensory modality specificity of neural activity related to memory in visual cortex. *J. Neurophysiol.* 78, 1263–1275.

Goldman, A. I. (2006). *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. Oxford: Oxford University Press.

Goodhill, G. J., and Carreira-Perpiñán, M. Á. (2002). "Cortical columns," in *Encyclopedia of Cognitive Science*, ed. L. Nadel (London: Macmillan).

Gothard, K. M., Battaglia, F. P., Erickson, C. A., Spitler, K. M., and Amaral, D. G. (2007). Neural responses to facial expression and face identity in the monkey amygdala. *J. Neurophysiol.* 97, 1671–1683. doi: 10.1152/jn.00714.2006

Graziano, M. S., Reiss, L. A., and Gross, C. G. (1999). A neuronal representation of the location of nearby sounds. *Nature* 397, 428–430. doi: 10.1038/17115

Gross, C. G. (2000). Coding for visual categories in the human brain. *Nat. Neurosci.* 3, 855–855. doi: 10.1038/78745

Gross, C. G. (2002). The genealogy of the grandmother cell. *Neuroscientist* 8, 512–518. doi: 10.1177/107385802237175

Hasselmo, M. E., Rolls, E. T., Baylis, G. C., and Nalwa, V. (1989). Object-centered encoding by face- selective neurons in the cortex in the superior temporal sulcus of the monkey. *Exp. Brain Res.* 75, 417–429. doi: 10.1007/BF00247948

Hinton, G. E., McClelland, J., and Rumelhart, D. E. (1986). "Distributed representations," in *Parallel Distributed Processing*, eds J. McClelland, D. Rumelhart, and the PDP Research Group (Cambridge, MA: MIT Press).

Horton, J. C., and Adams, D. L. (2005). The cortical column: a structure without a function. *Philos. Trans. R. Soc. Lond.* 360, 837–862. doi: 10.1098/rstb.2005.1623

Hoshi, E., Shima, K., and Tanji, J. (2000). Neuronal activity in the primate prefrontal cortex in the process of motor selection based on two behavioral rules. *J. Neurophysiol.* 83, 2355–2373.

Houweling, A. R., and Brecht, M. (2008). Behavioural report of single neuron stimulation in somatosensory cortex. *Nature* 451, 65–68. doi: 10.1038/nature06447

Hromádka, T., DeWeese, M. R., and Zador, A. M. (2008). Sparse representation of sounds in the unanesthetized auditory cortex. *PLoS Biol.* 6:e16. doi: 10.1371/journal.pbio.0060016

Hubel, D., and Wiesel, T. (1968). Receptive fields and functional architecture of monkey striate cortex. *J. Physiol.* 195, 215–243. doi: 10.1113/jphysiol.1968.sp008455

Hubel, D. H., and Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *J. Physiol.* 148, 574–591. doi: 10.1113/jphysiol.1959.sp006308

Hubel, D. H., and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* 160, 106–154. doi: 10.1113/jphysiol.1962.sp006837

Hubel, D. H., and Wiesel, T. N. (1977). Ferrier lecture: Functional architecture of macaque monkey visual cortex. *Proc. R. Soc. Lond. B* 198, 1–59. doi: 10.1098/rspb.1977.0085

Ince, R. A., Panzeri, S., and Kayser, C. (2013). Neural codes formed by small and temporally precise populations in auditory cortex. *J. Neurosci.* 33, 18277–18287. doi: 10.1523/JNEUROSCI.2631-13.2013

Kendrick, K. M., and Baldwin, B. A. (1987). Cells in temporal cortex of conscious sheep can respond preferentially to the sight of faces. *Science* 236, 448–450. doi: 10.1126/science.3563521

Klemen, J., and Chambers, C. D. (2012). Current perspectives and methods in studying neural mechanisms of multisensory interactions. *Neurosci. Biobehav. Rev.* 36, 111–133. doi: 10.1016/j.neubiorev.2011.04.015

Kosslyn, S. M., Thompson, W. L., and Ganis, G. (2006). *The Case for Mental Imagery*. Oxford: Oxford University Press.

Kreiman, G., Koch, C., and Fried, I. (2000). Category-specific visual responses of single neurons in the human medial temporal lobe. *Nat. Neurosci.* 3, 946–953. doi: 10.1038/78868

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., et al. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron* 60, 1126–1141. doi: 10.1016/j.neuron.2008.10.043

Krueger, J. M., Rector, D. M., Roy, S., Van Dongen, H. P., Belenky, G., and Panksepp, J. (2008). Sleep as a fundamental property of neuronal assemblies. *Nat. Rev. Neurosci.* 9, 910–919. doi: 10.1038/nrn2521

Lin, L. N., Chen, G. F., Kuang, H., Wang, D., and Tsien, J. Z. (2007). Neural encoding of the concept of nest in the mouse brain. *Proc. Natl. Acad. Sci. U.S.A.* 104, 6066–6071. doi: 10.1073/pnas.0701106104

Logothetis, N., and Sheinberg, D. (1996). Visual object recognition. *Annu. Rev. Neurosci.* 19, 577–621. doi: 10.1146/annurev.ne.19.030196.003045

Logothetis, N. K., Pauls, J., and Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Curr. Biol.* 5, 552–563. doi: 10.1016/S0960-9822(95)00108-4

Lorente de Nó, R. (1934). Studies on the structure of the cerebral cortex. II. Continuation of the study of the ammonic system. *J. Psychol. Neurol.* 46, 113–177.

Maier, J. X., Neuhoff, J. G., Logothetis, N. K., and Ghazanfar, A. A. (2004). Multisensory integration of looming signals by rhesus monkeys. *Neuron* 43, 177–181. doi: 10.1016/j.neuron.2004.06.027

Martin, A. (2007). The representation of object concepts in the brain. *Annu. Rev. Psychol.* 58, 25–45. doi: 10.1146/annurev.psych.57.102904.190143

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

McClelland, J. L., McNaughton, B. L., and O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychol. Rev.* 102, 419–457. doi: 10.1037/0033-295X.102.3.419

Merten, K., and Nieder, A. (2012). Active encoding of decisions about stimulus absence in primate prefrontal cortex neurons. *Proc. Natl. Acad. Sci. U.S.A.* 109, 6289–6294. doi: 10.1073/pnas.1121084109

Moll, F. W., and Nieder, A. (2015). Cross-modal associative mnemonic signals in crow endbrain neurons. *Curr. Biol.* 25, 2196–2201. doi: 10.1016/j.cub.2015.07.013

Mormann, F., Dubois, J., Kornblith, S., Milosavljevic, M., Cerf, M., Ison, M., et al. (2011). A category-specific response to animals in the right human amygdala. *Nat. Neurosci.* 14, 1247–1249. doi: 10.1038/nn.2899

Morrell, F. (1972). Visual system's view of acoustic space. *Nature* 238, 44–46. doi: 10.1038/238044a0

Moser, E. I., Kropff, E., and Moser, M. B. (2008). Place cells, grid cells, and the brain's spatial representation system. *Annu. Rev. Neurosci.* 31, 69–89. doi: 10.1146/annurev.neuro.31.061307.090723

Mountcastle, V. B. (1957). Modality and topographic properties of single neurons of cat's somatic sensory cortex. *J. Neurophysiol.* 20, 408–434.

Mountcastle, V. B. (1997). The columnar organization of the neocortex. *Brain* 120, 701–722. doi: 10.1093/brain/120.4.701

Newell, A. (1980). Physical symbol systems. *Cogn. Sci.* 4, 135–183. doi: 10.1207/s15516709cog0402_2

Newell, A., and Simon, H. (1976). Computer science as empirical inquiry: symbols and search. *Commun. ACM* 1, 113–126. doi: 10.1145/360018.360022

Nieder, A. (2013). Coding of abstract quantity by 'number neurons' of the primate brain. *J. Comp. Physiol. A* 199, 1–16. doi: 10.1007/s00359-012-0763-9

O'Keefe, J., and Dostrovsky, J. (1971). The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Res.* 34, 171–175. doi: 10.1016/0006-8993(71)90358-1

O'keefe, J., and Nadel, L. (1978). *The Hippocampus as a Cognitive Map*. Oxford: Oxford University Press.

Olshausen, B. A., and Field, D. J. (1997). Sparse coding with an overcomplete basis set: a strategy employed by V1? *Vis. Res.* 37, 3311–3325. doi: 10.1016/S0042-6989(97)00169-7

O'Scalaidhe, S. P., Wilson, F. A. W., and Goldman-Rakic, P. S. (1999). Face-selective neurons during passive viewing and working memory performance of rhesus monkeys: evidence for intrinsic specialization of neuronal coding. *Cereb. Cortex* 9, 459–475. doi: 10.1093/cercor/9.5.459

Page, M. (2000). Connectionist modelling in psychology: a localist manifesto. *Behav. Brain Sci.* 23, 443–467. doi: 10.1017/S0140525X00003356

Pan, X., and Sakagami, M. (2012). Category representation and generalization in the prefrontal cortex. *Eur. J. Neurosci.* 35, 1083–1091. doi: 10.1111/j.1460-9568.2011.07981.x

Panzeri, S., Macke, J. H., Gross, J., and Kayser, C. (2015). Neural population coding: combining insights from microscopic and mass signals. *Trends Cogn. Sci.* 19, 162–172. doi: 10.1016/j.tics.2015.01.002

Patterson, K., Nestor, P. J., and Rogers, T. T. (2007). Where do you know what you know? The representation of semantic knowledge in the human brain. *Nat. Rev. Neurosci.* 8, 976–987. doi: 10.1038/nrn2277

Perrett, D. I., Mistlin, A. J., Chitty, A. J., Smith, P. A. J., Potter, D. D., Broennimann, R., et al. (1988). Specialized face processing and hemispheric asymmetry in man and monkey: evidence from single unit and reaction time studies. *Behav. Brain Res.* 29, 245–258. doi: 10.1016/0166-4328(88)90029-0

Perrett, D. I., Rolls, E. T., and Caan, W. (1982). Visual neurones responsive to faces in the monkey temporal cortex. *Exp. Brain Res.* 47, 329–342. doi: 10.1007/BF00239352

Pezzulo, G., Barsalou, L. W., Cangelosi, A., Fischer, M. H., McRae, K., and Spivey, M. (2013). Computational grounded cognition: a new alliance between grounded cognition and computational modeling. *Front. Psychol.* 3:612. doi: 10.3389/fpsyg.2012.00612

Plate, T. (2002). "Distributed representations," in *Encyclopedia of Cognitive Science*, ed. L. Nadel (London: Macmillan).

Poggio, T., and Bizzi, E. (2004). Generalization in vision and motor control. *Nature* 431, 768–774. doi: 10.1038/nature03014

Pouget, A., Dayan, P., and Zemel, R. (2000). Information processing with population codes. *Nat. Rev. Neurosci.* 1, 125–132. doi: 10.1038/35039062

Quian Quiroga, R. (2012). Concept cells: the building blocks of declarative memory functions. *Nat. Rev. Neurosci.* 13, 587–597. doi: 10.1038/nrn3251

Quian Quiroga, R., Kraskov, A., Koch, C., and Fried, I. (2009). Explicit encoding of multimodal percepts by single neurons in the human brain. *Curr. Biol.* 19, 1308–1313. doi: 10.1016/j.cub.2009.06.060

Quian Quiroga, R., Kreiman, G., Koch, C., and Fried, I. (2008). Sparse but not 'Grandmother-cell' coding in the medial temporal lobe. *Trends Cogn. Sci.* 12, 87–94. doi: 10.1016/j.tics.2007.12.003

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., and Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature* 435, 1102–1107. doi: 10.1038/nature03687

Roberts, W. A., and Mazmanian, D. S. (1988). Concept learning at different levels of abstraction by pigeons, monkeys, and people. *J. Exp. Psychol.* 14, 247–260.

Rockland, K. S. (2010). Five points on columns. *Front. Neuroanat.* 4:22. doi: 10.3389/fnana.2010.00022

Rogers, T., and McClelland, J. (2004). *Semantic Cognition: A Parallel Distributed Processing Approach*. Cambridge, MA: MIT Press.

Rogers, T., and McClelland, J. (2008). Precis of semantic cognition: a parallel distributed processing approach. *Behav. Brain Sci.* 31, 689–749. doi: 10.1017/S0140525X0800589X

Rolls, E. T. (1984). Neurons in the cortex of the temporal lobe and in the amygdala of the monkey with responses selective for faces. *Hum. Neurobiol.* 3, 209–222.

Rolls, E. T., and Baylis, G. C. (1986). Size and contrast have only small effects on the responses to faces of neurons in the cortex of the superior temporal sulcus of the monkey. *Exp. Brain Res.* 65, 38–48. doi: 10.1007/BF00243828

Rolls, E. T., Robertson, R. G., and Georges-Françoiset, P. (1997). Spatial view cells in the primate hippocampus. *Eur. J. Neurosci.* 9, 1789–1794. doi: 10.1111/j.1460-9568.1997.tb01538.x

Romanski, L. M. (2007). Representation and integration of auditory and visual stimuli in the primate ventral lateral prefrontal cortex. *Cereb. Cortex* 17(Suppl. 1), i61–i69. doi: 10.1093/cercor/bhm099

Romanski, L. M., and Goldman-Rakic, P. S. (2002). An auditory domain in primate prefrontal cortex. *Nat. Neurosci.* 5, 15–16. doi: 10.1038/nn781

Romo, R., Brody, C. D., Hernández, A., and Lemus, L. (1999). Neuronal correlates of parametric working memory in the prefrontal cortex. *Nature* 399, 470–473. doi: 10.1038/20939

Roy, A. (2012). A theory of the brain: localist representation is used widely in the brain. *Front. Psychol.* 3:551. doi: 10.3389/fpsyg.2012.00551

Roy, A. (2013). An extension of the localist representation theory: grandmother cells are also widely used in the brain. *Front. Psychol.* 4:300. doi: 10.3389/fpsyg.2013.00300

Rumelhart, D., and Todd, P. (1993). "Learning and connectionist representations," in *Attention and Performance XIV: Synergies in Experimental Psychology, Artificial Intelligence, and Cognitive Neuroscience*, eds D. Meyer and S. Kornblum (Cambridge, MA: MIT Press), 3–30.

Saito, H. A., Yukie, M., Tanaka, K., Hikosaka, K., Fukada, Y., and Iwai, E. (1986). Integration of direction signals of image motion in the superior temporal sulcus of the macaque monkey. *J. Neurosci.* 6, 145–157.

Saleem, A. B., Ayaz, A., Jeffery, K. J., Harris, K. D., and Carandini, M. (2013). Integration of visual motion and locomotion in mouse visual cortex. *Nat. Neurosci.* 16, 1864–1869. doi: 10.1038/nn.3567

Samejima, K., Ueda, Y., Doya, K., and Kimura, M. (2005). Representation of action-specific reward values in the striatum. *Science* 310, 1337–1340. doi: 10.1126/science.1115270

Schrier, A. M., and Brady, P. M. (1987). Categorization of natural stimuli by monkeys (*Macaca mulatta*): effects of stimulus set size and modification of exemplars. *J. Exp. Psychol.* 13, 136–143.

Schroeder, C. E., and Foxe, J. J. (2002). The timing and laminar profile of converging inputs to multisensory areas of the macaque neocortex. *Cogn. Brain Res.* 14, 187–198. doi: 10.1016/S0926-6410(02)00073-3

Smith, B. C. (1982). "Prologue to reflection and semantics in a procedural language," in *Readings in Knowledge Representation*, eds R. J. Brachman and H. J. Levesque (Los Altos, CA: Morgan Kaufmann).

Smolensky, P. (1987). The constituent structure of mental states: a reply to fodor and pylyshyn. *South. J. Philos.* 26, 137–160. doi: 10.1111/j.2041-6962.1988.tb00470.x

Smolensky, P. (1988). On the proper treatment of connectionism. *Behav. Brain Sci.* 11, 1–74. doi: 10.1016/S0034-7450(14)60076-7

Stein, B. E., and Stanford, T. R. (2008). Multisensory integration: current issues from the perspective of the single neuron. *Nat. Rev. Neurosci.* 9, 255–266. doi: 10.1038/nrn2331

Suthana, N., and Fried, I. (2012). Percepts to recollections: insights from single neuron recordings in the human brain. *Trends Cogn. Sci.* 16, 427–436. doi: 10.1016/j.tics.2012.06.006

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annu. Rev. Neurosci.* 19, 109–139. doi: 10.1146/annurev.neuro.19.1.109

Tanaka, K. (2003). Columns for complex visual object features in the inferotemporal cortex: clustering of cells with similar but slightly different stimulus selectivities. *Cereb. Cortex* 13, 90–99. doi: 10.1093/cercor/13.1.90

Thorpe, S. (1995). "Localized versus distributed representations," in *The Handbook of Brain Theory and Neural Networks*, ed. M. Arbib (Cambridge, MA: MIT Press).

Ursino, M., and La Cara, G. E. (2004). A model of contextual interactions and contour detection in primary visual cortex. *Neural Netw.* 17, 719–735. doi: 10.1016/j.neunet.2004.03.007

Vogels, R. (1999). Categorization of complex visual images by rhesus monkeys. Part 2: single-cell study. *Eur. J. Neurosci.* 11, 1239–1255. doi: 10.1046/j.1460-9568.1999.00531.x

Wallis, J. D., Anderson, K. C., and Miller, E. K. (2001). Single neurons in prefrontal cortex encode abstract rules. *Nature* 411, 953–956. doi: 10.1038/35082081

Wang, Y., Brzozowska-Prechtl, A., and Karten, H. J. (2010). Laminar and columnar auditory cortex in avian brain. *Proc. Natl. Acad. Sci. U.S.A.* 107, 12676–12681. doi: 10.1073/pnas.1006645107

Wang, Z., Singhvi, A., Kong, P., and Scott, K. (2004). Taste representations in the *Drosophila* brain. *Cell* 117, 981–991. doi: 10.1016/j.cell.2004.06.011

Wilson, F. A. W., O'Scalaidhe, S. P., and Goldman-Rakic, P. S. (1993). Dissociation of object and spatial processing domains in primate prefrontal cortex. *Science* 260, 1955–1958. doi: 10.1126/science.8316836

Wyttenbach, R. A., May, M. L., and Hoy, R. R. (1996). Categorical perception of sound frequency by crickets. *Science* 273, 1542–1544. doi: 10.1126/science.273.5281.1542

Yu, Y., McTavish, T. S., Hines, M. L., Shepherd, G. M., Valenti, C., and Migliore, M. (2013). Sparse distributed representation of odors in a large-scale olfactory bulb circuit. *PLoS Comput. Biol.* 9:e1003014. doi: 10.1371/journal.pcbi.1003014

Zyga, L. (2012). *Do Brain Cells Need to be Connected to Have Meaning?* Available at: http://medicalxpress.com/news/2012-12-brain-cells.html

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

frontiers in Psychology

HYPOTHESIS AND THEORY
published: 30 September 2016
doi: 10.3389/fpsyg.2016.01537

Distinguishing Representations as Origin and Representations as Input: Roles for Individual Neurons

*Jonathan C. W. Edwards**

University College London, London, UK

It is widely perceived that there is a problem in giving a naturalistic account of mental representation that deals adequately with the issue of meaning, interpretation, or significance (semantic content). It is suggested here that this problem may arise partly from the conflation of two vernacular senses of representation: representation-as-origin and representation-as-input. The flash of a neon sign may in one sense represent a popular drink, but to function as a representation it must provide an input to a 'consumer' in the street. The arguments presented draw on two principles – the neuron doctrine and the need for a venue for 'presentation' or 'reception' of a representation at a specified site, consistent with the locality principle. It is also argued that domains of representation cannot be defined by signal traffic, since they can be expected to include 'null' elements based on non-firing cells. In this analysis, mental representations-as-origin are distributed patterns of cell firing. Each firing cell is given semantic value in its own right – some form of atomic propositional significance – since different axonal branches may contribute to integration with different populations of signals at different downstream sites. Representations-as-input are patterns of local co-arrival of signals in the form of synaptic potentials in dendrites. Meaning then draws on the relationships between active and null inputs, forming 'scenarios' comprising a molecular combination of 'premises' from which a new output with atomic propositional significance is generated. In both types of representation, meaning, interpretation or significance pivots on events in an individual cell. (This analysis only applies to 'occurrent' representations based on current neural activity.) The concept of representations-as-input emphasizes the need for an internal 'consumer' of a representation and the dependence of meaning on the co-relationships involved in an input interaction between signals and consumer. The acceptance of this necessity provides a basis for resolving the problem that representations appear both as distributed (representation-as-origin) and as local (representation-as-input). The key implications are that representations in the brain are massively multiple both in series and in parallel, and that individual cells play specific semantic roles. These roles are discussed in relation to traditional concepts of 'gnostic' cell types.

Keywords: mental representation, percept, grandmother cell, pontifical cell, gnostic cell

OPEN ACCESS

Edited by:
Bernhard Hommel,
Leiden University, Netherlands

Reviewed by:
Roland Thomaschke,
University of Regensburg, Germany
Raphael Fargier,
University of Geneva, Switzerland

***Correspondence:**
Jonathan C. W. Edwards
jo.edwards@ucl.ac.uk

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *10 June 2016*
Accepted: *21 September 2016*
Published: *30 September 2016*

Citation:
Edwards JCW (2016) Distinguishing Representations as Origin and Representations as Input: Roles for Individual Neurons. Front. Psychol. 7:1537. doi: 10.3389/fpsyg.2016.01537

INTRODUCTION

Concepts of mental representation are widely invoked in neurobiology, linguistics, artificial intelligence, and philosophy. Yet, as Seager and Bourget (2007) note: "there is no acknowledged theory of mental representation." This appears to be partly because people differ in terms of the explanatory work they want such a theory to do (Stich, 1992). It also reflects an impasse in reaching a consensus on how mental representations could fit into a naturalistic account of the brain; what sort of substrate, or causal nexus could support a mental representation, and how? I shall argue that these are interdependent questions and that a careful assessment of the logical constraints on substrate, in terms of physical dynamics and their location, may clarify the ways in which mental representation may be a useful concept, as well as vice versa.

From the outset I wish to emphasize that the problem I address relates only to what may be called 'occurrent' or 'active' representations in which signals are sent and received on specific occasions. There is another use of the term that might be called a 'dispositional representation' – an acquired pattern of cellular connectivity underlying memory, knowledge, or concept acquisition, that disposes the brain to generate occurrent representations in response to stimuli (Simmons and Barsalou, 2003). I will be using 'representation' to mean 'occurrent representation.'

The naturalization problem is not so much about whether a representation is to the right, left, front or back of the brain, or what connection tracts are involved. The more basic problem is defining the *type,* or level, of biophysical location that could support a fitting causal role, and with appropriate information capacity ('bandwidth'). There are those who would argue that we have a rough answer: that representations can be equated with patterns of neural activity, or firing. However, as discussed below, this fails to address key problems, justifiably of concern to philosophers of mind. Meaning is *not* to be solved so easily.

It might be argued that searching for a detailed substrate type for mental representation is overly reductionist or, in theoretical modeling terms, simply premature. It might even be considered immaterial to understanding of how a representation can have a meaning, either in terms of external referents or internal 'meaning to the subject.' However, I think the search is justified on the following grounds. Firstly, spatial pattern is about the only way meaning can be encoded in a brain at any point in time, as far as we know, so at least *type* of spatial pattern and location is likely to be central to a theory of meaning. Secondly, recognizing that reductive analysis of mechanism is only part of the story does not mean that fruitful progress in neural mechanisms should be abandoned half-finished and replaced by hand-waving. Rather than, as Marr (1982) advocated, treating the biophysical and 'functional' levels of analysis as incommensurable, to be able to test viability of theories I believe, with Trehub (1991), that we need some idea of how and where they could correspond.

Moreover, the ability to suggest at least one plausible physical example for any theoretical model is a requirement that is arguably never premature. A search for such examples can render explicit contradictions in popular concepts. The key proposal here is that neuropsychology may benefit from a greater focus on the input aspect of mental representation. The author's background is in immunology. It was not until we insisted on a grounding in a dynamics of integration of signals into individual cells that we began to understand leucocyte behavior in immune recognition and memory (Male et al., 2012). Hypotheses that could not be so grounded were discarded. The gap between work on post-synaptic integration (e.g., Branco and Häusser, 2011; Smith et al., 2013; Ishikawa et al., 2015) and psychology may still be harder to bridge but the possibility of grounding in plausible input mechanisms should be an acid test of all models of mental representation.

THE NATURE OF MENTAL REPRESENTATIONS

Representation is a term used in a variety of ways that are not always transparent. It is not simply 're-presentation,' and not just because 'presentation' might be a better label. It can also imply 'proxy' or 'symbol.' In the mental case, where representations do not resemble their referents in any simple way, the meaning of the term will be preconditioned not only by presumptions about how brains work but also metaphysical standpoint. A materialist may think in terms of brain states representing external 'things' whereas someone taking a dynamist or structural realist approach (as I do) may think in terms of internal dynamic relations representing external dynamic relations (Ladyman and Ross, 2007). There will also be different views on how these concepts relate to subjectivity or phenomenality. To clarify the way 'representation' relates to meaning it may help to consider two main purposes to which the term 'mental representation' is put.

Mental representation may be invoked simply as part of an account of the human brain as a machine that generates outputs from inputs. A mental representation can be seen as the equivalent of local currents or magnetizations in a computer. As long as we accept that brain cells send messages around in a way vaguely similar to computer components, we can consider the nature of mental representations in this context as just a technical issue, like the difference between Microsoft Windows and Mac OS-X, without raising too many philosophical questions. 'Representation' is being used here purely to imply some internal dynamics that co-vary usefully with external world dynamics.

There is, nevertheless, even here, a need to define a representation more precisely than just that total pattern of brain activity that arises in a specific context, whether the presence of a red square or blue circle, or when thinking 'I suspect the recession will double-dip.' A representation is not just a pattern of events; it is a pattern with a causal role. A red square will trigger patterns in the retinae, geniculate bodies, primary and secondary visual cortices, temporal, parietal and frontal lobes, all with different causal roles. To function, the content of any individual representation must be available to some functional component at a causal nexus: what Millikan calls a 'consumer' (Ryder et al., 2012). Thus we may need to talk of many mental representations at many levels rather than a single representation.

That then begs the question of which mental representations are those envisaged by philosophers and linguists such as Fodor (1985) or Dretske (1986) and what their consumers are.

The second motivation for talking about mental representations is in the context of questions about first person experience, as conceived from positions on the nature of 'mentality' ranging from Cartesian to eliminitivist (Stich, 1992). Thus 'mental representation' is often used to imply an associated experience, in which operational meaning is somehow 'interpreted.' This may be as a 'percept,' as when something is viewed or heard, or a 'mental image,' as when retrieving memories, thinking of a scene or sound, or in dreams (Fodor, 1975; Kosslyn, 1994).

There is a general assumption that there is only *one instance* of this 'percept' type of representation in a brain at a time, and there has been extended debate over whether this is local or distributed (e.g., Barlow, 1972; Fodor and Pylyshyn, 1988; Marcus, 2001), which remains unresolved. It is suggested here that this may reflect confusion about what we should expect the biophysical processes underlying a representation, of the 'percept' type, to consist of and where they might be – and that the assumption that there is only one such representation needs challenging.

There are those who, probably rightly, point out that a first person account of mental representation will ultimately be redundant to a description of its physical dynamics (e.g., Churchland, 1992). The mistake, I believe, is to take this as a reason for discounting the first person account. Even granted that representations of the percept type may form a tiny minority of the total, and quite apart from the desire to know how there comes to be a first person account, it is likely that without heuristic clues from experience and the language we use to describe it the causal dynamics of all our representations will remain intractable. However tidy it may feel to regard talk of 'phenomenality' as outside physical science, I follow those who argue that there is a strong case for accepting that 'phenomenal experience' plays a crucial role in all science, as the medium of observation, and that we should be happy to make all use of it we can. Thus, mental representations associated with experience or 'feel,' whether percepts or 'current belief states' (Crane, 2014) are not only those of greatest philosophical interest but may also be particularly worth exploring for their potential to shed light on mental processes in general. I shall therefore focus on such representations from now on, taking sensory percepts as the paradigm.

GENERAL CAUSAL PRINCIPLES

Unless there are good reasons otherwise, an account of a representation-as-percept in a brain should follow causal principles used elsewhere in physical science, where possible confirmed by experimental neurophysiology. Two such principles are particularly relevant. The first is the neuron doctrine. The second is that the content of a percept will be encoded in signals that form inputs to some physical domain.

The neuron doctrine, in essence, is the principle that brain function (*qua* 'thinking') can be explained by the interactions of separate neuronal units (Gold and Stoljar, 1999). Each neuron is a discrete computational (in the broad sense of having rule-based input–output relations) unit, conforming to biophysical laws. The timing of firing of a neuron is determined by chemical and electrical interactions between the cell and its immediate environment. All cause and effect relations occur locally. The neuron doctrine does not preclude other levels of explanation in terms of groups of cells or macroscopic brain domains, but holds that these can be broken down, without residue, to an account of individual cell interactions.

Some have suggested that the neuron doctrine should be replaced by a description of brain function at a 'global' level (Gold and Stoljar, 1999). However, since the causal biophysical pathways of the neuron doctrine are not seriously in doubt it is unclear that a global description can be an alternative, rather than just a higher-level analysis grounded in the same local dynamics. There may be a temptation to suggest that some of the perplexing aspects of mental representation can only be accounted for using approaches such as systems theory or non-linear dynamics that might be seen to give an 'emergent' dynamic 'greater than the sum of the parts.' However, without clear evidence it seems safer to assume that, as Barlow (1994) says, all causal relations pass through the bottlenecks of individual neurons.

The second premise is as fundamental but less often articulated. It underlies Rosenberg's (2004) concept of receptivity and Millikan's idea of 'consumer' and is laid out in explicit neurological terms by Orpwood (2007). The representations we call percepts must be based on the co-availability of certain signals to some neuron-based domain, i.e., they must be inputs to such a domain, which will also generate outputs in response that allow the percept to be 'reported.' (Reporting may be a complex indirect process but the basic point is unaffected.) Something has to receive the signals that encode a percept, whether these are derived originally from sense organs or other sources as in dreams. An un-received signal does not even qualify as a signal, since reception is entailed in the concept.

This might seem self-evident. However, this second premise is worth emphasizing because literature on consciousness often appears to take a different view. Representations may be seen as associated with computational or 'information processing' operations, which involve not inputs but input–output relations, or 'roles in the world' – the essence of 'functionalism' (Fodor, 1975; Block, 1996). The 'content' of the representation is then seen as being dependent not only on the effect of the world on the computational unit but also on the effect of the unit on the world. This appears to imply that if percepts belong to physical domains then those domains are in some way *acquainted with, or informed by, their outputs* (effects on the world) as well as their inputs. This is self-contradictory for any computational system that obeys standard concepts of causality – what something has access to *is* its input – and neuroscience consistently indicates that these concepts of causality hold good.

I must emphasize that this is a low-level analysis dealing with individual neuro-computational steps. Events within feedback systems taken as whole, as in anticipatory models of perception (Hommel, 2009) can, in a broader sense, be considered as representing a certain action/perception scenario but even here

it is not the input/output relation that gives the content, but the particular pattern of signals ('the data'), considered either as cellular outputs or inputs.

Both in neuroscience and philosophy, representations are often considered in terms of patterns of cell activity, with no specific reference to input or output. The problem here is that to consider a pattern as an operant representation implies that the total activity pattern is accessible to something. A pattern of activity of 73,456 out of a bank of 1,000,000 right occipital cells might seem to represent a scene. However, each of these cells may have 10,000 branches to its axonal output, some feeding forward, some back. Only 6,228 cells may send branches to each of a bank of temporal cells, and 18,992 to a bank of prefrontal cells (through any one direct or indirect route) and, moreover, there will be variation (and plasticity) in this between individual sending and receiving cells in each bank. Although the activity of the 73,456 cells is a representation in a certain legitimate vernacular sense, there seems to be another important sense in which it underpins, together with whatever other 'null cells' whose non-firing may contribute critically to the content being conveyed, not one, but many, representations-as-inputs, diverse in content and function.

In other words, an act of representation must ultimately imply an input *to* something specified. It is sometimes implied that there are no 'inner receiving entities' for representations in a brain, but, again, this is inconsistent with our understanding of causality. To be part of a causal chain, and thus reportable, the information encoded in a representation must be made available to something that generates a response. A word of text in a forgotten language embedded in an opaque medium that cannot be removed without destroying the text cannot function as a representation. Similarly, a pattern of lines of cellular activity in my visual cortex that bears a homotopic relation to a pattern of tree trunks I am viewing is not acting as a spatial representation *by dint of homotopy*, since no part of me, including the cells themselves, is informed of the spatial relations of active and inactive cells. The cells provide a representation in the form of presenting sensory data to other parts of my brain through patterns of downstream synaptic transmission, but the homotopic spatial relation of their cell bodies is itself of no consequence. Representation must be linked to a causal path.

Inner receiving entities are often rejected as 'homuncular' and criticized on grounds that shifting the problem of the input/percept relationship for a brain to a subdomain of brain leaves the problem unchanged and therefore invokes infinite regress. The implication of regress is, however, *non sequitur*. If the problem is the same as for the whole brain then that must surely also suffer from the regress. The reverse conclusion applies: if the problem has *any* solution for the brain it may *also* have a solution for a homuncular subdomain and it may only have a solution there (see also Fodor, 1975, p. 189) Thus, even Dennett's (1988) homunculi that 'repeat entirely the talents they are rung in to explain' are only straw bogeymen. Homunculi are in fact usefully rung in to deal with practical computational issues.

There is no doubt that treating representations as inputs to specific neural structures raises difficulties. However, nothing in neuroscience so far conflicts with the idea that a representation-as-percept is an input to something. It might be argued that standard causal principles only apply at the periphery of the system and not centrally, but it is unclear why or how. We have no reason to postulate an invisible envelope that divides an external or peripheral world from an inner 'animate' world (perhaps Fodor's organism) with novel (i.e., supernatural) non-local properties, at any structural level. Neurobiology has shown that we can push the concept of 'input' as far in as interpretable empirical observation will allow, and well within the confines of the human body or brain. Pressure from an intervertebral disk on a lumbar nerve root gives pain in the foot. Cochlear implants give deaf people an experience of sound. Stimulation of cerebral cortex in the awake individual can evoke sensations and memories. The evidence indicates that sensory pathways, at all points up to that where a percept is experienced, are simply providing an input to the next stage, which often can be mimicked artefactually.

The work of Hubel and Wiesel (2005) and others has shown that detailed mechanisms of acquisition and collation of sensory data can be tracked far into the brain. Cells that respond to lines at particular angles, lines of limited length, or color contrasts can be demonstrated. It might be argued that the absence of precise analysis beyond this level could indicate that signals enter a 'black box' in which percepts are no longer associated with inputs, but rather with input-output relations. However, the simpler explanation is that beyond this level computation is so sophisticated that analysis requires very sophisticated experimental approaches. The more recent work of Quian Quiroga et al. (2005) showing that individual cortical cells respond to specific faces suggests that this is so.

In summary, despite speculations in other directions in some fields of study, the two assumptions of the neuron doctrine and the doctrine of percepts as based on inputs to perceiving entities appear to be worth retaining.

POSSIBLE DOMAINS FOR REPRESENTATIONS AS PERCEPTS

Armed with this basic causal standpoint, it is possible to ask general questions about the location of the representations as percepts and the nature of the entities to which these are available. The starting premise is that at least one domain exists in a waking brain that supports an experience correlated with input from sense organs, contextualized by anticipations derived from kinesthetic monitoring, etc. We want to describe such a domain in dynamic physical terms. The prima facie case is that it will be a dynamic domain comprising part or all of one or more neurons, receiving inputs derived from all sensory modalities, and other internally generated signals, like names and concepts retrieved from memory (i.e., anything and everything we can experience), and capable of sending a sequence of outputs that can connect to all, or most, motor pathways. Conventional neuroscience indicates that the input will be of signals leading to patterns of depolarization of cell membrane. Since we are considering input this ought to be a pattern within dendrites (i.e., input projections).

It might be questioned that any single domain has inputs of all perceptual modalities and also concepts. However, our

ability to mix raw sensory data and concepts in use of language indicates that somewhere in the brain signals with these disparate types of meaning are integrated – i.e., are co-inputs to some computational unit. Moreover, introspection indicates that human perceiving subjects experience them concurrently in a meaningful relationship and do so alongside the use of relevant language. Synchronization of signals may be important for optimal computation but as von der Malsburg (1981) pointed out when first suggesting that synchrony of signals was important, it can only be important *because it determines synchronized arrival* at some site of input.

I agree with Orpwood's (2007) reasoning that percepts must be based on inputs that somehow are 'interpreted' on arrival at the perceiving domain and thereby have meaning to the perceiving subject. As this meaning belongs to the input itself, rather than any computational input–output relation, it seems that it too should be located at the site of input in dendrites. 'Interpretation' is not meant here in the sense that sensory signals encoding four legs, a bushy tail, pointed ears, and a toothy snout are converted to a signal meaning fox. That would imply at least one computation involving an input–output relation. The identification label 'fox' would be the input to the next domain along. Interpretation is used here to mean simply the correspondence of an input, (of electrical or chemical signals based on collation amongst sensory data and with data from memory) to a 'percept' that 'is like something' for, or has a meaning to, the receiving entity (in the above case legs, tail, ears, and snout). 'Manifestation' might be an alternative term, since it implies no additional physical interaction, but simply a correspondence between physical input and its meaning to the receiving entity.

Absence of a mechanism for this sense of interpretation may seem puzzling. However, immediate local correspondence between physical dynamics and meaningful experience seems to be something that, like Descartes, we have to take as brute fact. Ascribing it to processes prior to the point of input to the perceiving entity makes things no easier. There is no means by which to carry interpretation forward from previous events, since we have no evidence for anything other than the physical input itself being available to the receiving unit. Moreover, the idea of 'carrying meaning forward' generates an absurdity. Since the history of past events contributing to any causal interaction is immeasurably complex an immeasurably large number of 'interpretations' from earlier events should be carried forward in causal chains and that is not what we experience. Both the existence and the richness of the meanings inputs have to human perceiving subjects may be things for us to wonder at, but trying to delegate richness elsewhere is no solution.

It seems that representations as meaningful percepts ought to occur in neural dendrites.

REPRESENTATIONAL DOMAINS CANNOT BE BASED ON TRAFFIC

A further consideration is helpful in narrowing down options for the domain of a percept. The content of a percept almost certainly has to be an interpretation of both signals associated with membrane excitation and 'null signals' corresponding to where membrane might have been excited but was not. Unless signals are interpreted in the context of all possible signals in a domain we lose what appears to be essential for a complex percept: encoding of information in patterns of inter-relation. A summation of all and only the black spots of a set of printed words can have only one meaning: black. (Or if black is coded null the sum of white areas just means white.) Moreover, it is indeterminate whether the spots included are on one page, or in a whole library. Only if both active and null signals and their relations are included do we have diverse meaning and bounded domains of meaning. In visual cortex, a 'line' of uniform color within a block of the same color is not interpreted as a line. The interpretation of 'a line' implies the absence of signals encoding similar color on either side of the line.

This means that the domain that supports a representation with meaning cannot be defined by a pattern of active signal traffic; it cannot be defined in terms of where signals are occurring. It must include null signals, so there must be some intrinsically defined structural domain within which signals and null signals are co-interpreted. The domain receiving signals interpreted as a percept cannot be an 'active circuit' in the sense of a set of pathways *currently* carrying signal traffic.

There is a distinction here between the processing units in a brain and in a computer. In a computer there are 'gates' in which electrical signals 'open' or 'close' connections between units, forming and breaking electrical circuits. The brain does not have gates in this sense. Connections remain unchanged, at least over periods of hours, regardless of traffic. The processing units are integrators, but not gates. Something akin to gating will occur during refractory periods and if input signals show differential synchronization in relation to refractory periods there may be triage, so that some active signals are 'let through' and others not. However, these signals will still operate in the context of null signals within the non-refractory time window.

LOCALIZED VERSUS DISTRIBUTED REPRESENTATIONS

Representations-as-percepts, if only in a degraded form, survive damage to large areas of cerebral cortex. Damage to certain areas produces predictable defects, but does not appear to remove the capacity for some sort of perceptual experience, even if there is agnosia in the sense of not being aware that the percept is defective. The inference is that if the type of domain receiving representations as percepts is indeed cortical then there is no *single and local* domain. That leaves options of one very extended domain or multiple local domains.

The idea that a percept is an interpretation of the inputs to cells over a wide area of brain generates a range of problems, quite apart from the basic problem noted by James (1890/1983) that each cell's input is separate. Many neurons are involved in 'housekeeping,' such as suppression of vision during saccades, or motor co-ordination. The inputs to such cells do not appear to figure in percepts, which reflect the input to a select cell population involved in a field of attention. It is unclear, in a

distributed model, why the inputs to certain cells and not others should figure in a reportable percept. Nor is it clear why we should perceive a single 'copy' of sensory data if cells over a wide area contribute, since most if not all signals arising from cellular activity in sensory pathways are sent as inputs to many cells through widely ramifying axonal branches. When we see a red tomato early signals referring to a red tomato are sent to 1000s of cells further forward in the brain. Why should we consider these thousands of 'copies' a single representation? If a company sends out 1000 Christmas cards, each with a photo of head office in the snow, do we consider this 'one representation' of head office?

These and related concerns may have motivated the proposal by Pribram (1991) that the cortex carries information somewhat in the manner of a hologram, in which every part of a spatial array carries a copy of the entire pattern of information being handled. Although often thought of as a model of distributed representation, the holographic model provides a means for having very many 'copies' of a pattern at many sites rather than a single copy available to one extended site. A simpler and neurologically reasonable version of the idea is just that sensory data are sent to many locations in the cortex and each of these has the potential to interpret its input as percept. This would seem to be in keeping with the experiments of Quian Quiroga et al. (2005) in which visual sense data often gave rise to excitation in many sampled cortical cells. In some cases cells were highly restricted in their responses to images, but others are more promiscuous. At least there is little doubt that sensory stimuli lead to signals being sent widely to many cells.

In summary, although the discussion so far might suggest that the question of what domain supports a perceptual representation is just what it must always have been – which cell or cells – it may need a subtler formulation. How many of which sort of neuron have a perceptual representation encoded in their input(s) and do they constitute one domain of one representation of this type at any one time or are there multiple domains, with multiple representations based on the same sensory data? It is important to note that the latter should not be expected to evoke a *sense* of multiplicity (of the perception of being one of many subjects) since multiplicity would not itself be encoded, represented or, therefore, perceived by anything, being a fact about parallel reception events, not a property of the receiving unit, or the content of its input.

At this point the reader may sense that the concept of representation is too confusing to be useful, and there is a case for that position! I would argue, however, that if some historic confusions in the literature are unpacked it is possible to restore the usefulness of the idea, with some riders that add significant explanatory power.

PONTIFICAL, GRANDMOTHER AND CARDINAL CELLS

The simplest hypothesis for the domain of a percept, now universally taken as a null hypothesis, is that of a single pontifical cell, as discussed by James (1890/1983) and dating back at least to ideas raised by Leibniz (Woolhouse and Franks, 1998), writing shortly after cells were first observed. This form of pontifical cell is a single cell that supports all 'my' representational percepts of all sensory inputs. It is the 'me' cell. Other cells act as conduits to and from this central cell, collating inputs and delegating outputs. James considers that they might also support 'percepts,' but of a meaner sort than those I report as 'mine.' (He includes the point that none of these percepts need involve any sense of multiplicity or presence of others.) The attraction of this idea is that the cell is the brain's integrating unit, with an intrinsically delimited input domain, and the contents of human experience appear to be integrated and delimited. However, the idea that just one neuron should have this specialized function is implausible on a range of grounds and, as indicated above, the argument that experience *seems* 'single' is immaterial, since there would be no reason for there to be representation (and thus perception) of multiplicity, or a sense of 'other copies,' within each of multiple representations.

It is useful to raise here a potential confusion in terminology between sites of representation and sites of recognition. Sherrington (1940) invoked a concept of a *quite different* sort of 'pontifical' cell to explain recognition. Sensory data relating to an object such as a dog enters through many 1000s of receptors. Recognition would appear to require sequential stages of discrimination, each leading to a reduced number of possible interpretations. This might be expected to form of a 'pyramid' with fewer cells at each stage until the input finally converged on one cell responsible of recognizing dogs. There would be a pontifical cell for a dog, another for a cat and another for grandmother.

A key point here is that we have no reason to think that only the cell with the job of recognizing dogs will receive input signals encoding doggy features. If 100 cells each recognized a different mammal we would not expect the presence of a dog to lead to input to only one of these. We would expect all the cells to receive signals encoding doggy features but only one (or some) to fire. It could be argued that synapses receiving signals encoding long snouts will atrophy on koala-recognizing cells but at least to be able to learn to recognize new animals we have to assume that cells with catholic inputs exist.

Thus if a representation is based on an input pattern we do not expect sites of representation and recognition to be commensurate. This emphasizes the need to consider a causal chain as potentially involving many levels of representation with multiplicity at each level (**Figure 1**). It highlights the fact that a representation is always a step in a causal chain and is thus always a representation *to* a domain at a particular point in that chain. Thus a pattern of data, perhaps encoding legs, fur and muzzle, would represent a dog *to* a 'dog-pontifical cell' as well as to a lot of other cells, untuned, or tuned to other creatures. In turn the firing of the dog-pontifical cell *and not* its neighbors would denote 'dog' to the rest of the brain. The two types of representation would be quite different. Moreover, intuition tells us that whatever domain has a percept of a dog of the sort normally discussed it must have an input encoding both the key features of a dog – legs, fur, etc. – *and* the sense of these being part of a dog, apparently putting the relevant domain downstream of the site of dog-recognition

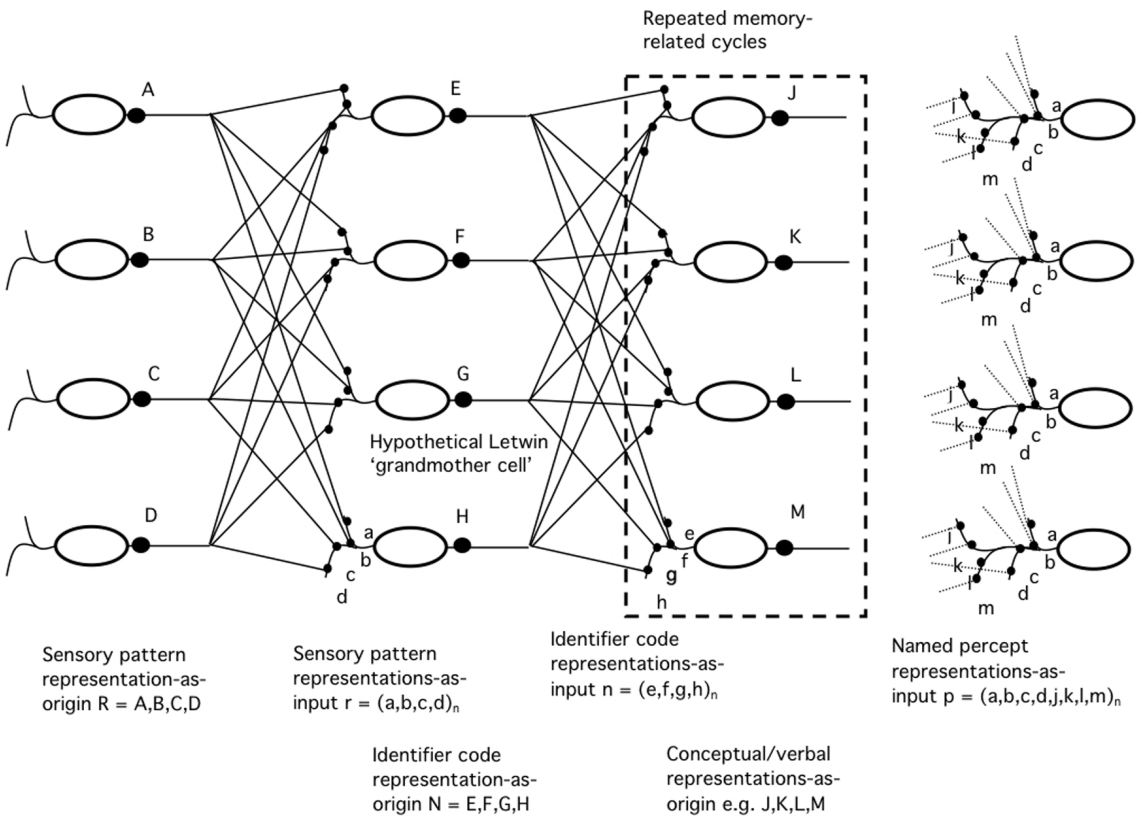


FIGURE 1 | Simplified schema of successive representations-as-origin and representations-as-input, starting with sensory patterns, followed by recognition with encoding as identifiers that allow recall from memory of concepts and finally the generation of percepts including both sensory patterns and conceptual/naming features.

with additional parallel input encoding the original upstream context-dependent sensory data.

Empirical studies indicate that recognition does not use a pyramidal system with fewer and fewer cells at each stage (Barlow, 1972). Sequential stages involve as many, if not more, cells as at the beginning – as implied by the above discussion. Recognition is signaled by the firing of one or a few cells in the context of non-firing of many more cells. At all stages representations are thus widespread, but it needs to be established whether this is because individual representations are extended or because of multiplicity.

This issue is relevant to Barlow's (1972) classic Perception paper. Barlow takes as his object grandmother, following Letvin (Gross, 2002) and discusses the plausibility of a 'grandmother cell' in the sense of a single cell that fires with 100% sensitivity and specificity for grandmother. This bears a relation to Sherrington's (1940) pontifical cell but not to that of Leibniz or James. Barlow suggested that grandmother was probably not important enough to have her own cell and that, more likely, grandmother would be encoded by the activity of perhaps a thousand 'cardinal' cells, each representing an aspect of grandmother such as a mouth or nose, any of which might presumably contribute to encoding other faces in other combinations. These elements of the percept are then seen as combining rather in *the way words combine in a sentence* (an analogy also used by Marr, 1982). Note that Barlow is not proposing a redundancy-for-safety strategy with information distributed in a 'holographic' way to several cells, each with a sensitivity and specificity of less than 100%. He is giving each cell a separate and specific job.

The odd thing here is that Barlow appears to be describing the activity of cells upstream of a site of recognition of grandmother. If each cell is responding to signals which together encode a feature not entirely specific and sensitive for grandmother then grandmother can only be recognized, and social responses activated, by a downstream group of cells receiving inputs from these thousand cells, *some of which downstream cells will fire and some not*. It would be these downstream cells whose inputs would encode all grandmother's features and it would therefore be their domains that we could (perhaps) expect to support a 'percept' of granny in the sense of manifestation of all of grandmother's features, whether or not they fired. And it would be the pattern of firing and non-firing of these latter cells that would 'represent' (in the denoting sense) to domains in the rest of the brain the presence, but not the pattern of features, of this individual. Whether or not within this latter group of cells there are cells with 100% sensitivity and specificity for grandmother is a different issue that need not bear on the search for the domains supporting the representations known as percepts.

More recently, Quian Quiroga and Kreiman (2010), has discussed the interpretation of experiments on individual cell responses to faces (Quian Quiroga et al., 2005) in relation to the grandmother cell concept. In this case the grandmother cell is rejected on redundancy grounds. While emphasizing the complexity of the grandmother cell concept, discussion seems to bypass the crucial issue of the distinction between the site of experience of a pattern such as a face and the site of recognition of such a pattern. Nevertheless, it seems to support the idea that inputs carrying information about a pattern such as a face will be received by not one, but many cellular computational units.

The above discussion emphasizes a number of issues relating to this crucial question. It seems that representations (in the broadest sense) of a given referent in the brain must be multiple and diverse. At each level many cells will be involved in representing. Representation and recognition are not likely to be commensurate. So far the discussion has been in terms of individual cells despite the general assumption in the literature that representations in brains each involve many cells. The grounds for such an assumption need be revisited in the light of the preceding arguments.

A RETURN TO THE NEURON DOCTRINE

As already noted, to be useful, the concept of representation-as-percept, has to imply a step in a causal chain with content encoded in the input to some domain. It is also difficult to see how a representation can have a meaning, or interpretation, to a domain, unless its content is encoded in the co-temporal input of a pattern of active signals and null signals to the domain. Representations like this do not occur in computers. Stored data in a computer can represent something meaningful to a human user accessing it via a screen but no representation based on a pattern of co-temporal input occurs *to* anything within the machine beyond the four trivial input options for an electronic gate of on/on, on/off, off/on and off/off. Moreover, we do not require that anything in a computer interprets, or attributes meaning to, input signals. It might be argued that a sequence of signals passing through a gate might constitute a representation. However, since each signal contributes to a separate computation this is problematic. The sequence of incoming signals is not subjected as a whole to a computation, other than as arbitrarily defined by a programmer. Within the machine any temporal 'chunking' of serial signals into 'representations' adds nothing to the causal account and at the gate in question no chunking should be apparent.

Within brains there *are* units that receive complex patterns co-temporally: neurons. Moreover, they are the only units that receive patterns relevant to percepts as far as we know. Barlow's 1000 cardinal cells are not a unit receiving a pattern of features of grandmother. Each has a separate input encoding one feature. For all 1000 features to contribute co-temporally to a representation 1000 cardinal cells must send all 1000 active or null signals to converge on at least one downstream neuron, which is within the range of neuronal inputs. The neuron doctrine, as was probably evident to Leibniz, entails the simple but surprising conclusion that representations qua percepts in brains can only be in individual neurons (Edwards, 2005; Sevush, 2006, 2016). There may be very large numbers of such representations, all encoding the same sensory data, distributed over a wide area, but each percept must be tied to the receiving unit that is the neuron.

This conclusion immediately resolves the paradox of localization and distribution of representation in the brain, since it implies that *local* representations can be present over a *widely distributed area*. This situation is familiar in the distribution of a newspaper, which is widespread but can only convey news if all the words of a news story are present in each copy read by an individual. To suggest that a single perceptual representation could be available to several cells is equivalent to saying that news can be understood by a group of people each of which receives one word from the paper.

The conclusion also resolves the question of precisely where in the brain are the representations that determine our actions. The answer is that they may be all over the brain. Even the question of where in the brain are the representations that determine considered verbalized behavior may have the same answer, although it seems reasonable to attach some special significance to representations in cells with multimodal inputs that would allow both the visual and auditory features and the concept of a dog to contribute to a 'percept' of a dog.

Putting representations in individual cells may appear implausible. However, it is unclear why a representation in a single cell should be more implausible than one involving many cells. The implausibility may be more salient simply because any proposal for a specific location for such representations brings into focus our lack of understanding of the rules of interpretation. This may be no bad thing. Ironically, the charge of implausibility tends to come from those who argue for functional rather than structural analysis and yet the conclusion is based on the 'functional' property of having input (and capacity) rather than structure. The conclusion might be branded over-reductive but one of its key features is that it makes explicit the boundary between reductive analysis and the non-reductive relation of 'interpretation,' rather than invoking an ill-defined internal no-man's-land where both are claimed to apply at different 'levels.'

Another attraction of the idea that representations are *to* the single computational (rule based input–output) units that are neurons is that it implies that the brain does not perform single operations on 'atomic' (structureless) symbols, but rather it performs operations on 'molecular' representations. That is to say that the basic data units that the brain operates on are irreducibly complex, with many degrees of freedom. This begins to address the puzzle of how the manipulation of symbols can be associated with an experience of complex patterns that reflect the complexity of their referents. It also provides a reason why, as appears to be increasingly recognized, syntax and semantics cannot be totally dissociated when considering meaning (Hinzen, 2006).

MULTIPLE REPRESENTATIONS OF MULTIPLE TYPES

The concept of multiplicity of representations of sensory data in the brain should not be unexpected if we consider the parallel and hierarchical nature of computation. There may be a lingering presumption that representations, qua percepts, ought to be single – belonging to a single 'me,' but this is not logically required. There is also a lingering discomfort with the idea that our actions may be guided by representations distinct from those we report as our percepts. Perhaps the best known 'redundancy' of representations is that implied by the dual path hypothesis for visual perception of Goodale and Milner (1992). The dissociation of percept and action described by Króliczak et al. (2006) for the hollow face illusion, presents the counterintuitive idea that the brain builds more than one spatial representation, which might seem redundant or extravagant in use of resources. This has the interesting implication that we consider the building of spatial representations qua percepts labor-intensive.

Figure 1 illustrates an approach to representation in the brain that suggests that this concern may be misplaced. It makes explicit the idea that 'representation' has two different meanings. One sense of representation (R) is an instance of a pattern, as in a picture or map, that acts as *origin* for a representation in the other sense (r) of an instance of representing *to* something via its *input*. At every stage of neural computation we can expect a representation-as-input to lead to an output that can act as representation-as-origin for the next stage. At every stage banks of cells will be involved but whereas such a bank of cells will hold a single representation-as-origin it will hold as many representations-as-inputs as there are cells in the bank. Perhaps surprisingly, although building a representation-as-origin is likely to be labor-intensive, much larger numbers of representations-as-inputs, which we could expect to correspond to percepts, would appear to come free of charge.

We are used to the idea that the nervous system generates motor output from sensory input at several levels of complexity. There are spinal reflexes, brainstem reflexes, automatic but co-ordinated responses involving cerebellum, routine purposive actions and deliberated actions. All of these can be expected to be associated with different levels of representation-as-origin and representations-as-input so we should not be surprised by the idea of multiple spatial representations even in terms of representations-as-origin. Perhaps more interestingly, as indicated on the right side of **Figure 1**, hierarchies of representation-as-origin give the opportunity for representations-as-input downstream to 'pick-'n'-mix' inputs from more than one level of this hierarchy. Thus there is nothing very surprising about the idea that the representations that guide our rapid actions appear to overlap in content in most but not all situations with those that form the basis of our percepts.

CONCLUSION

Mainstream neuroscience prides itself in being rigorously physicalist, in the sense of adhering to the basic precepts of natural science and general principles of causality. A consideration of representations in such a rigorous causal framework leads to the conclusion that all representations in the brain, including those that may form the basis of percepts, must ultimately be considered in terms of how they are cashed out in the inputs to individual neurons. These representations as inputs will occur at multiple levels of sensory processing and will be multiple at all levels, including levels associated with pattern recognition, denotation and reportable percepts. Such a model is counterintuitive but resolves certain important problems relating to the distributed nature of representation and may provide clues to the basis of meaning and language.

AUTHOR CONTRIBUTIONS

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

REFERENCES

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

Barlow, H. (1994). "The neuron doctrine in perception," in *The Cognitive Neurosciences*, Vol. 26, ed. M. Gazzaniga (Cambridge: MIT Press), 415–435.

Block, N. (1996). "What is functionalism?," in *The Encyclopedia of Philosophy Supplement*, ed. D. M. Borchert (London: Macmillan).

Branco, T., and Häusser, M. (2011). Synaptic integration gradients in single cortical pyramidal cell dendrites. *Neuron* 69, 885–892. doi: 10.1016/j.neuron.2011.02.006

Churchland, P. (1992). *A Neurocomputational Perspective*. Cambridge, MA: MIT Press.

Crane, T. (2014). "Unconscious belief and conscious thought, in *Apects of Psychologism*. (Cambridge, MA: Harvard University Press).

Dennett, D. (1988). "Quining qualia," in *Consciousness in Modern Science*, eds A. Marcel and E. Bisiach (Oxford: Oxford University Press).

Dretske, F. (1986). "Misrepresentation," in *Belief, Form, Content and Function*, ed. R. Brogdan (Oxford: Oxford University Press).

Edwards, J. C. (2005). Is consciousness only a property of individual cells? *J. Conscious. Stud.* 12, 60–76.

Fodor, J. A. (1975). *The Language of Thought*. New York, NY: Thomas Crowell.

Fodor, J. A. (1985). Fodor's guide to mental representation. *Mind* 94, 76–100. doi: 10.1093/mind/XCIV.373.76

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

Gold, I., and Stoljar, D. (1999). A neuron doctrine in the philosophy of neuroscience. *Behav. Brain Sci.* 22, 585–642.

Goodale, M. A., and Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends Neurosci.* 15, 20–25. doi: 10.1016/0166-2236(92)90344-8

Gross, C. G. (2002). Genealogy of the "grandmother cell". *Neuroscientist* 8, 512–518. doi: 10.1177/107385802237175

Hinzen, W. (2006). *Mind Design and Minimal Syntax*. Oxford: Oxford University Press.

Hommel, B. (2009). Action control according to TEC (theory of event coding). *Psychol. Res.* 73, 512–526. doi: 10.1007/s00426-009-0234-2

Hubel, D. H., and Wiesel, T. (2005). *Brain and visual perception*. Oxford: Oxford University Press.

Ishikawa, T., Shimuta, M., and Häusser, M. (2015). Multimodal sensory integration in single cerebellar granule cells in vivo. *Elife* 4:e12916. doi: 10.7554/eLife.12916

James, W. (1890/1983). *The Principles of Psychology*. Cambridge, MA: Harvard University Press.

Kosslyn, S. M. (1994). *Image and Brain: The Resolution of the Imagery Debate*. Cambridge, MA: MIT Press.

Króliczak, G., Heard, P., Goodale, M. A., and Gregory, R. L. (2006). Dissociation of perception and action unmasked by the hollow-face illusion. *Brain Res.* 1080, 9–16. doi: 10.1016/j.brainres.2005.01.107

Ladyman, J., and Ross, D. (2007). *Every Thing Must Go*. Oxford: Oxford University Press.

Male, D., Brostoff, J., Roth, D., and Roitt, I. M. (2012). *Immunology*, 8th Edn. Philadelphia, PA: Elsevier.

Marcus, G. (2001). *The Algebraic Mind*. Cambridge, MA: MIT Press.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York, NY: Freeman.

Orpwood, R. (2007). Neurological mechanisms underlying qualia. *J. Integr. Neurosci.* 6, 523–540. doi: 10.1142/S0219635207001696

Pribram, K. H. (1991). *Brain and Perception*. Upper Saddle River, NJ: Lawrence Erlbaum.

Quian Quiroga, R., and Kreiman, G. (2010). Postscript: about grandmother cells and Jennifer Aniston neurons. *Psychol. Rev.* 117, 297–299.

Quian Quiroga, R., Reddy, L., Kreiman, G., Koch, C., and Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature* 435, 1102–1107. doi: 10.1038/nature03687

Rosenberg, G. (2004). *A Place for Consciousness in Nature*. Oxford: Oxford University Press.

Ryder, D., Kingsbury, J., and Williford, K. (2012). *Millikan and Her Critics*. Chichester: John Wiley & Sons.

Seager, W., and Bourget, D. (2007). "Representationalism about consciousness," in *The Cambridge Handbook of Consciousness*, eds P. D. Zelazo, M. Moskovitch, and E. Thompson (Cambridge: Cambridge University Press).

Sevush, S. (2006). Single-neuron theory of consciousness. *J. Theor. Biol.* 238, 704–725. doi: 10.1016/j.jtbi.2005.06.018

Sevush, S. (2016). *Single-Neuron Theory: Closing in on the Neural Correlate of Consciousness*, Chap. 8. Basingstoke: Palgrave-MacMillan.

Sherrington, C. (1940). *Man on His Nature*. Cambridge: Cambridge University Press.

Simmons, W. K., and Barsalou, W. (2003). The similarity-in-topography principle: reconciling theories of conceptual deficits. *Cogn. Neuropsychol.* 20, 451–486. doi: 10.1080/02643290342000032

Smith, S. L., Smith, I. T., Branco, T., and Häusser, M. (2013). Dendritic spikes enhance stimulus selectivity in cortical neurons in vivo. *Nature* 503, 115–120. doi: 10.1038/nature12600

Stich, S. (1992). What is a theory of mental representation? *Mind* 101, 243–261.

Trehub, A. (1991). *The Cognitive Brain*. Cambridge, MA: MIT Press.

von der Malsburg, C. (1981). "The correlation theory of brain function," in *MPI Biophysical Chemistry, Internal Report 81–2. Reprinted in Models of Neural Networks II (1994)*, eds E. Domany, J. L. van Hemmen, and K. Schulten (Berlin: Springer).

Woolhouse, R. S., and Franks, R. (1998). *G.W. Leibniz, Philosophical Texts*. Oxford: Oxford University Press.

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

frontiers in Psychology

HYPOTHESIS AND THEORY
published: 09 August 2017
doi: 10.3389/fpsyg.2017.01216

Complexity Level Analysis Revisited: What Can 30 Years of Hindsight Tell Us about How the Brain Might Represent Visual Information?

John K. Tsotsos *

Department of Electrical Engineering and Computer Science, York University, Toronto, ON, Canada

Much has been written about how the biological brain might represent and process visual information, and how this might inspire and inform machine vision systems. Indeed, tremendous progress has been made, and especially during the last decade in the latter area. However, a key question seems too often, if not mostly, be ignored. This question is simply: do proposed solutions scale with the reality of the brain's resources? This scaling question applies equally to brain and to machine solutions. A number of papers have examined the inherent computational difficulty of visual information processing using theoretical and empirical methods. The main goal of this activity had three components: to understand the deep nature of the computational problem of visual information processing; to discover how well the computational difficulty of vision matches to the fixed resources of biological seeing systems; and, to abstract from the matching exercise the key principles that lead to the observed characteristics of biological visual performance. This set of components was termed *complexity level analysis* in Tsotsos (1987) and was proposed as an important complement to Marr's three levels of analysis. This paper revisits that work with the advantage that decades of hindsight can provide.

Keywords: vision, attention, complexity, pyramid representations, selective tuning model

OPEN ACCESS

Edited by:
Tarek Richard Besold,
University of Bremen, Germany

Reviewed by:
Sashank Varma,
University of Minnesota, United States
Johan Kwisthout,
Radboud University Nijmegen, Netherlands
Ulrike Stege,
University of Victoria, Canada

***Correspondence:**
John K. Tsotsos
tsotsos@cse.yorku.ca

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *15 August 2016*
Accepted: *03 July 2017*
Published: *09 August 2017*

Citation:
Tsotsos JK (2017) Complexity Level Analysis Revisited: What Can 30 Years of Hindsight Tell Us about How the Brain Might Represent Visual Information?. Front. Psychol. 8:1216. doi: 10.3389/fpsyg.2017.01216

INTRODUCTION

This paper has two main parts. In the first, there is a brief recapitulation of 30 years of research[1] that addresses the question: do proposed solutions to how the brain processes visual information match the reality of the brain's resources? The main goal of this activity had three components: to understand the deep nature of the computational problem of visual information processing; to discover how well the computational difficulty of vision matches to the fixed resources of biological seeing systems; and, to abstract from the matching exercise the key principles that lead to the observed characteristics of biological visual performance. The second part of the paper uses the results of that analysis and extends them to specifically connect to how the brain represents visual information. We begin by motivating the analysis as presented three decades ago.

[1]There is a distinct focus on our own work throughout this paper simply because the goal of this presentation is to examine that old work and how its conclusions have stood the test of time. This is not to say that no other work has appeared since nor that all other work is unimportant. Far from it! However, most other developments along complexity theoretic lines do not line up with the main thread of this paper, namely, what can this analysis tell us about representations in the brain.

A universally acclaimed landmark in the development of computational theories of intelligence is the presentation of the three levels of analysis defined by Marr (1982). Marr presents the three levels, now quoted, at which any machine carrying out an information-processing task must be understood:

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

This prescription has been used effectively ever since not only in vision modeling but throughout computational neuroscience and cognitive science. Unfortunately, Marr, not being a computer scientist, missed an important issue. He did not realize that it is not difficult to pose perfectly sensible computational solutions that are physically unrealizable. As argued in Tsotsos (1990) and elsewhere, there are a large number of perfectly well-defined computational problems whose general solution is provably intractable—unrealizable on available physical resources or requiring time longer than the age of the universe[2]. Even worse, there are well-defined problems that are undecidable, meaning there provably exists no algorithm to determine the result[3]. As argued in Tsotsos (1993, 2011), such results that seem impossible do not negate their main impact: our brains seem to deal with all the problems they face remarkably well so it can only be the case that the formal definitions of the problems that lead to such intractable or impossible results cannot be the ones that our brains are actually solving.

This matching process as an idea has its roots in earlier works. Uhr (1972, 1975) describes "recognition cones" as a representation for perception. Although his papers are clear in their inspiration from neural systems, Uhr only hinted at their resource implications. Feldman and Ballard (1982), however, explicitly linked computational complexity to neural processes saying "Contemporary computer science has sharpened our notions of what is 'computable' to include bounds on time, storage, and other resources. It does not seem unreasonable to require that computational models in cognitive science be at least plausible in their postulated resource requirements." They go on to examine the resources of time and numbers of processors, and more, leading to a key conclusion that complex behaviors can be carried out in fewer than 100 (neural processing) time steps. The overall import of their paper was to stress the need for a careful matching of problem to resources in cognitive theories. *Resource-complexity matching is a source of critical constraints on the viability of theories*, especially those that attempt to provide a mechanistic theory as opposed to a descriptive one (see Brown, 2014).

Even though these arguments were very strong, they took the form of 'counting arguments' and a formalization could perhaps make them even stronger. An attempt to formalize those points was made beginning with Tsotsos (1987). We examined the inherent computational difficulty of visual information processing from formal and empirical perspectives[4]. The methods used have their roots in the theoretical sub-domain of computer science known as computational complexity. Computational complexity has the goal of discovering formal characterizations of the difficulty of achieving solutions to computational problems[5] in terms of the size and nature of the input. The difficulty of achieving solutions has direct impact on resources, such as computational power, memory capacity and processing time, as Feldman and Ballard (1982) also pointed out.

For this reason, a fourth level, the complexity level, was introduced in Tsotsos (1987, 1990), intended to ensure the logic of the strategy for solving the problem is actually realizable within its available resources:

- *Complexity analysis:* What is the computational complexity of the problem being addressed? How does it match with the resources used for its realization? If the problem is intractable and/or there are insufficient resources available for a realization of its solution, how can the problem be reframed to enable a solution?

This paper revisits the conclusions reached by the resulting series of papers with the advantage of decades of hindsight. Interestingly, a wide spectrum of predictions regarding the brain's visual processes that resulted from that analysis has enjoyed subsequent experimental support (see Tsotsos, 2011 for details). We begin with a brief overview of the main conclusions and assertions that complexity level analysis provided.

COMPLEXITY LEVEL ANALYSIS

In Tsotsos (1989, 2011), a number of mathematical proofs were presented that formalize the difficulty of perhaps the most

[2]Details on this assertion are beyond the scope of this paper. The interested reader can find a very accessible discussion in Stockmeyer and Chandra (1988), while those wishing a deeper treatment should see classic texts such as Garey and Johnson (1979), Papadimitriou (2003).

[3]Decidability is discussed in Davis (1958, 1965). Proof of *decidability* is sufficient to guarantee that a problem can be modeled computationally. It requires that the problem be formulated as a decision problem and that a Turing Machine is defined to provide solution. This formulation for the full generality of vision does not currently exist. If no sub-problem of vision can be found to be decidable, then it might be that perception as a whole is undecidable and thus cannot be computationally modeled. However, many decidable vision problems are mentioned throughout this paper so that is not the case.

[4]It is not within the scope of this paper to detail the full sequence of papers on the topic, so they are simply cited here so that the interested reader can examine them separately: Tsotsos (1987, 1988a,b, 1989, 1990, 1991, 1992, 1993, 1995a, 2011), Ye and Tsotsos (1996), Ye and Tsotsos (2001), Parodi et al. (1998) Andreopoulos and Tsotsos (2013).

[5]A *problem* is distinct from an *algorithm*. A problem is a general statement about something to be solved (Marr's computational level, Marr, 1982) whereas an algorithm is a proposed solution (Marr's representational and algorithmic level). One can address computational complexity at both levels: the inherent difficulty of a problem in its general form as well as the difficulty of a particular algorithm. Problem complexity applies to all possible solutions and any realization of them while algorithm complexity applies only to the specific algorithm analyzed. Here, we address only the former.

elemental of visual operations—essentially a sub-element of all visual operations—namely, visual matching[6]. Visual matching is the task of determining if some arbitrary image, a goal image, is a subset of some other image, the test image. In this definition, no knowledge of the target is allowed to influence the solution—the problem is thus termed *unbounded* in those papers. A function was assumed to exist that would quickly determine if a particular match was found, and it was not permitted to reverse engineer that function in order to guide the search. In other words, the solution was constrained to be one requiring a strictly data-driven approach[7]. The main proof, replicated by Rensink (1989) using a different approach, showed that this problem potentially had exponential time complexity in the number of image pixels, largely because in the worst case, it is unknown which image subset is the one that represents that goal image (think of an arbitrary sky full of stars—which subset of stars forms a hexagon?). The more important part of this is that it was proved that no single solution exists that is optimal for all possible problem instances. Due to the particular manner in which the proof was executed, the problem lends itself to a number of non-exponential, but not necessarily exact or optimal, solutions, as pointed out by Kube (1991)[8]. Following a more detailed examination, it was shown that although these non-exponential solutions are indeed valid, they do not really help because they all rely on solution elements that have no biological counterpart and have execution times that do not reflect human performance (Tsotsos, 1991)[9]. Note that this is likely true also for the other problems cited throughout this paper; they may also have known non-exponential solutions and realizable solutions for small enough or special case instances. A puzzling situation thus results: can we or can we not rely on the theoretical work as a guide? Our everyday experience with our own visual systems exhibits no such intractability. The only conclusion therefore is that the brain is not solving the problem as formalized for those proofs: the human brain is solving a different version of visual matching. This is admittedly a non-standard use of complexity theory because it disallows solutions that are not biologically realizable or plausible[10]. It does however show that the prevailing thoughts of the time (i.e., 1980's and somewhat beyond) that vision can be formulated as a purely bottom-up (i.e., stimulus-driven) process needed to be re-considered. To preview the endgame of this paper, that reformulation is one that allows differing levels of solution precision and different expenditures of processing time for different subsets of problem instances.

At this point in this presentation, it seems important to emphasize that the proofs mentioned in the previous paragraph do indeed point to sensible conclusions because there are many other researchers who have reached similar conclusions, i.e., that their problems are likely intractable, for a variety of visual and non-visual problems that are associated with human intelligent behavior. Selected examples of other works focusing on vision and neural networks and thus relevant for this paper include: polyhedral scene line-labeling (Kirousis and Papadimitriou, 1988); loading shallow architectures (neural network learning with finite depth networks) (Judd, 1988); relaxation procedures for constraint satisfaction networks (Kasif, 1990); finding a single, valid interpretation of a scene with occlusion (Cooper, 1998); unbounded stimulus-behavior search (Tsotsos, 1995a); and 3D sensor planning for visual search (Ye and Tsotsos, 1996).

The impact of computational complexity has also been pursued by many researchers in artificial intelligence and cognitive science (too many to properly mention here, however, see van Rooij, 2008, for a nice review). To round out this section, the important paper focusing on algorithm complexity, as opposed to problem complexity addressed by the previously cited authors, in vision by Grimson (1990) must be highlighted. Biologists also contributed with consistent and complementary conclusions (Thorpe and Imbert, 1989; Lennie, 2003, and others).

So how to proceed with the complexity level analysis? The whole point was to ensure that solutions are tractable within the constraints of biological processing structures. The strategy we chose which first appeared in Tsotsos (1987) is to simply start with the obvious, brute-force, worst-case complexity for the visual problem first described in this section's opening paragraph, termed Visual Match in Tsotsos (1989) and Comparison in Macmillan and Creelman (2005) (which is not provable as a bound on the time complexity in any way) and see how it might be altered to fit within a brain[11]. It's as if we were

[6]If we look at the perceptual task definitions provided by Macmillan and Creelman (2005), we see that all psychophysical judgments are of one stimulus relative to another — the basic process is comparison. The most basic task is termed discrimination, the ability to tell two stimuli apart. The fact that it is a sub-element of all visual tasks means that the difficulty of any visual task is at least as great as that of this sub-element. Interestingly, this is a decidable perceptual problem and is an instance of the Comparing Turing Machine (Yasuhara, 1971). Further discussion is found in Tsotsos (2011).

[7]Although it is admittedly unusal to include this restriction, it makes sense if one wishes to follow the Marr approach to vision, i.e., that visual processing included no top-down or knowledge-based guidance. Marr (1982; p 96) restricted his approach to be applicable for the first 160ms of processing by the brain and for stimuli where target and background have a clear psychophysical boundary. Our original motivation was to show that this approach would not suffice for all stimuli; this was successfully accomplished.

[8]In general, it is true that for problems that are proven to have such complexity characteristics, it only means that sufficiently large problem instances may not be realizable and that perhaps small ones, or particular subsets or special cases of the overall problem, may be perfectly realizable. The point of the complexity proof is to characterize a general solution that applies for all possible instances. For vision, this is a tall order. The space of all possible images is impossibly large. Pavlidis (2014) derives possible characterizations of this space. He claims that a very conservative lower bound to the number of all possible human-discernible images is 10^{25} and may be as large as 10^{400}. The practical import is that any solution that one proposes must apply to this full set.

[9]Kube (1991) pointed out that the Knapsack problem, which forms the foundation of the proof, is known to have efficient solutions under certain circumstances. Tsotsos (1991) surveys those efficient solutions and notes that they are not easily matched to, let alone relevant for, biologically plausible architectures and processes. It is beyond the scope to give further details on this here but the sequence of commentaries in Tsotsos (1990, 1991), Kube (1991) provide more detail.

[10]Traub (1991) also struggles with this issue. He suggests that a theory of complexity of scientific problems is needed such that formulations capture the essence of the science and that they be tractable.

[11]This is essentially the same process as seen in Judd (1988), van Rooij et al. (2012), van Rooij and Wareham (2012), and others, where they effectively used intractability results to guide a search for methods and problem re-formulations that would lead to realizable solutions. However, a major difference is the need to further constrain that search to be consistent with neuroanatomical and neurophysiological knowledge.

tasked, in some imaginary world, to design the first ever visual system from scratch. Tsotsos (2011) gives this simple-minded worst-case complexity as $O(P^2 2^P 2^M)$[12]. P represents the number of image elements (pixels, photoreceptors), M is the number of features represented (e.g., color, shape, texture, etc.); these are the starting elements from which we need to design vision. Recall that the problem is termed 'unbounded' since there is no bounding information arising from task or world knowledge that limits the search—as designers of the first ever visual system, it might not yet be apparent to us that we need task or world knowledge! In other words, we begin with the Marr approach (see footnote 7). Any image subset can be the correct one, and thus the powerset of image elements gives the worst-case scenario, and processing proceeds in a purely data-directed manner. The three elements of the complexity function arise in the following manner: P^2-the worst-case cost of computing the matching functions; 2^P-the worst-case number of image subsets in an image of P pixels; 2^M-the worst-case number of feature subsets associated with each pixel.

In Artificial Intelligence, a central concept is that of Rational Action. Rational Action, carried out by a rational agent, maximizes goal achievement given the agent's current knowledge, the agent's ability to acquire new knowledge, and the current computational and time resources available to the agent (Russell et al., 2003). In everyday behavior, we humans only rarely attempt to optimize solutions, but rather, just need to get something done (when drinking from a glass, we do not optimize the path to minimize energy or distance; rather, we simply want to get the glass to our mouth). In other words, we mostly resort to solutions that may not be optimal in any way but that are *good enough* for the current needs. Often, these are heuristic solutions that simply accomplish our goals[13]. One of these heuristics is to seek a *Satisficing* solution. Satisficing is a strategy that entails searching through the available alternatives until an acceptability threshold is met. This differs from *optimal decision-making*, an approach that attempts to find the best feasible alternative. The term *satisficing*, (a combination of *satisfy* and *suffice*), was introduced by Herb Simon in 1956. Satisficing can take more than one form. If one is faced with a problem and has the luxury of time, then one can spend as much time as one likes to find an acceptable solution among all the possible ones. One the other hand, if time is limited, perhaps strictly limited by the need to act before something else occurs, then a different sort of search would occur, one that would find a *just in time* solution, the best one within the time limit. If time is extremely tight, then an almost *reflexive* response is needed, perhaps the first one that comes to mind. Clearly, external tasks and situations as well as internal motivations play an important role in determining the right sort of approach to employ. Different from this strategy is the one where subsets of the full problem are defined where optimal procedures apply without infeasible characteristics. Here, the first step is to determine when such a problem is presented. Then, the most appropriate solution can be deployed. A rational agent, then, attempts to achieve its current goal, given its current constraints, by applying such selection methods to choose among its many possible solution paths. This points to the need for some kind of executive to control the process (one review for executive function in the brain, of the many available, can be found in Funahashi, 2001).

Knowledge of the intractability of visual processing in the general case—that is, that no single solution can be found that is optimal and realizable for all instances—forces a reframing of the original problem. The space of all problem instances can be partitioned into sub-spaces where each may be solvable by a different method. Some of those methods—whether satisficing, optimal, just in time, reflexive or other type—may lead to fast realizations (for example, if there is a special case problem subset that leads to non-exponential algorithm[14]), others slow ones, and some perhaps no realization. Given that a fixed processing resource such as the brain is to be employed, the need to apply a variety of different solution strategies in a situation dependent manner implies that resources must be *dynamically tunable*[15]. In order to support such a decision process, representations of visual, task, and world information and more must be available to support the reasoning involved that an executive controller performs (a sketch of how this might occur appears in Tsotsos and Womelsdorf, 2016).

The second stage of complexity level analysis looks for ways of matching the available resources with the computational difficulty of the problem to be solved. For vision, and specifically for human vision, those resource constraints include numbers of neurons, synapses, neural transmission times, behavioral response times, and so on. As Garey and Johnson (1979) point out, using the main variables of the problem definition as a guide is useful; variables that appear in exponents are the most important to try and reduce. Only the conclusion of this exercise will be given here since the details have appeared in several past papers (see Tsotsos, 2011 for overview). The key activity is to reduce the worst-case time complexity expression so that it can lead to an algorithm that is matched to the size and behavior of the human brain. The main conclusions are:

1. Use a pyramid representation to reduce the number of image locations searched. A pyramid is a layered representation, each layer with decreasing spatial resolution and with bidirectional connections between locations in adjacent layers (Jolion and Rosenfeld, 1994 provide review). Introduced by Uhr (1972), they permit an image to be abstracted so that a smaller number of locations at the top level may be the only ones over which some algorithm needs to search. At

[12]The notation *O(-)*, known as Big-O notation, signifies the order of the time complexity function, that is, its dominating terms asymptotically.

[13]Garey and Johnson (1979) detail a variety of strategies and heuristics for dealing with intractable problems theoretically and these are as applicable here as for theoretical computer science problems.

[14]One additional possibility is that of a fixed parameter-tractable algorithm, that is, an algorithm that is exponential only in the size of a fixed parameter while polynomial in the size of the input (see Downey and Fellows, 1999; van Rooij and Wareham, 2007 for more).

[15]This is of course, not without a cost. Tuning takes time to affect the processing, and processing itself may also then take longer. That different visual tasks take different amounts of processing time is well documented and is related to dynamic tuning in Tsotsos et al. (2008), Tsotsos (2011). See also **Figure 5** and caption.

least, they may provide the starting point for a coarse-to-fine search strategy from top to bottom of the pyramid. such a representation would reduce the size of the variable *P*. **Figure 1** shows a hypothetical pyramid of 3 layers. The number of locations represented in the lowest layer (layer 1) is *p1*; $p1 > p2 > p3$. In most pyramid definitions, the value at each location in each layer is determined by a computation based on a subset of the other layer values. Each element is not only connected to others in the adjacent layers but may also be connected to elements within the same layer. Such a representation has much in common with the hierarchical organization of early visual cortex as revealed by the work of Hubel and Wiesel (1962, 1965).

2. The objects and events of the visual world are mostly spatially and temporally confined to some region; however, we can also recognize scattered items as well (such as star constellations, or collections of animals as flocks or herds, group motion say as in a rugby play, etc.). Spatio-temporally localized receptive fields reduce the number of possible receptive fields from $O(2^P)$ to $O(P^{1.5})$ (this assumes contiguous receptive fields of all possible sizes centered at all locations in the image array and is derived in Tsotsos, 1987). **Figure 1** not only shows a three-layer pyramid but also a typical element (neuron) within the middle layer and an illustration of the breadth of its connections within the pyramid showing that connectivity is limited in feedforward, feedback and lateral directions.
3. Selection of a single or group of receptive fields to consider can further reduce the $P^{1.5}$ term to some value $P' < P^{1.5}$. This may be not only a selection of location, but also a selection of a local region or size. Such selection of region of interest is the most common use of attention in models (Tsotsos and Rothenstein, 2011; Tsotsos et al., 2015).
4. For some given task, feature selectivity to relevant features can further reduce the M term to some value M', where $2^{M'} < 2^M$, that is, the subset M' of all possible features actually present in the image or important for the task at hand. $M \ll P$ in any case so its presence in the exponent poses much less of a problem. This implies that features are best organized into separate representations, one for each feature, permitting a processing mechanism to involve only the required features into a computation and leaving the irrelevant ones outside the computation. Such separate representations likely lead into separate processing pathways as features are abstracted. Human vision has the characteristic of performing differently depending on the feature complexity of stimuli, as has been shown many times since Duncan and Humphreys (1989). Their experiments showed that in visual search tasks, difficulty increases with increased similarity of targets (that is, feature overlap and thus the ability to remove irrelevant features from the computation) to non-targets and decreased similarity between non-targets, producing a continuum of search efficiency. This is yet another form of a restrictive attentive process, that may be termed priming in this instance.

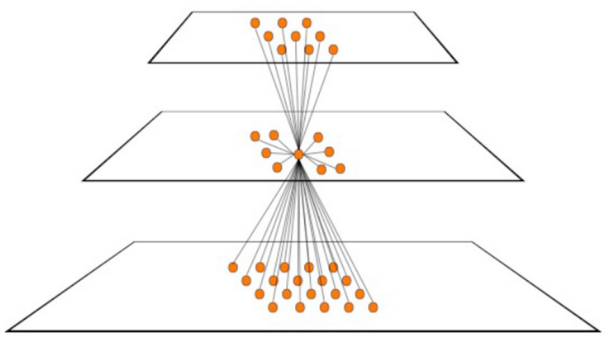


FIGURE 1 | A hypothetical 3-layer pyramid representation. The number of locations represented in the bottom layer (layer 1) is p_1; $p_1 > p_2 > p_3$. A typical element of each layer is shown in the center of the middle layer (layer 2). The figure shows how that element is connected to its immediate neighbors in the layer, as well as to elements in the lower and higher layers. All connections are potentially bidirectional. The figure shows the converging pattern of feedforward connections from layer 1 to 2, the diverging pattern of feedforward connections from layer 2 to 3, the converging pattern of feedback connections from layer 3 to 2 and the diverging pattern of feedback connections from layer 2 to 1. Each element of each layer features this pattern of connectivity.

These[16] achieve our goal, that is, to reduce the exponential complexity function to a much lower complexity expression, $O(2^{M'} P'^{3.5})$. It is important to note that attentional selection to either select a single candidate or to restrict consideration to a small set of candidates forces a serialization of the problem solution. If the chosen candidate is correct, the algorithm of course terminates. However, if it is not, the next candidate must be selected for consideration. A related situation arises for stimuli that are not spatially localized (such as the examples of a star constellation or flock of birds given earlier) and in such cases, full image comparisons or more complex methods (such as piecing together results from the available sub-image matches) would be required, again perhaps necessitating a serial search. No single solution will handle all problem instances; different strategies can be applied in succession until success is achieved, each with a successively higher processing cost. This characteristic is unavoidable and representations must support the process.

This leads to the final stage of complexity level analysis, which is to determine what impact arises from the previous stages that provide the foundations for developing a theory of human vision. This impact is summarized here:

- Pyramidal abstraction affects the problem through the loss of location information and signal combination. It affects the problem solution by sometimes enabling shorter search processes, commonly known as coarse-to-fine search.
- Spatiotemporally localized receptive fields force the system to look at features across a receptive field instead of finer grain combinations and thus arbitrary combinations of locations must be handled by some other strategy.
- Attentional processes permit selection and restriction within the input data to control the overall size of input to be considered.

[16]It should be noted that the original formulation included consideration of the set of world models *N* whose search efficiency can be logarithmically improved by hierarchical organization (Tsotsos, 1987). This is omitted here since it does not alter that nature of the problem.

What this demonstrates is that although the analysis began considering solutions for the full space of problem instances, the need to fit a solution within the brain's resources forced a shrinking of that full space into something smaller. In other words, the restriction that Marr placed on his approach—that is, a clear figure-ground boundary—manifests itself as a restriction on the set of problem instances. Unfortunately, it is not easy to characterize this subspace. However, there is a possible taxonomy of visual tasks that can help. **Figure 2** shows this taxonomy; there is no claim that it is complete. What it does point out is that the visual task most current AI systems address (such as Fukushima, 1988; LeCun and Bengio, 1995; Riesenhuber and Poggio, 1999; Krizhevsky et al., 2012), namely categorization, comprises only a small part of the taxonomy. It must be stressed that this taxonomy of tasks is not the same as a depiction of the space of problem instances. Each task has its own set of possible instances (and there may be overlap). For example, within categorization, there are instances that are easy (clear figure-ground boundary is seen) and instances that are difficult (without a clear figure-ground delineation).

To this point, the possibility of task influence on how a vision problem might be approached has not been discussed. The reason is that in his formulation, Marr discounted its use entirely and our approach was originally motivated by his perspective. However, increasingly, cognitive psychology and neuroscience has demonstrated that task influence plays a major role (see Carrasco, 2011; Tsotsos, 2011; Herzog and Clarke, 2014). In fact, accompanying the intractability proof in Tsotsos (1989) was a second theorem that showed that simple task knowledge can *bound* the search; it provides limits on the search space making it linear, rather than exponential, in the number of image elements (Wolfe, 1998 provides a relevant visual search review). The task knowledge can be as specific as target size or as generic as statistical regularities (as Parodi et al., 1998, illustrate empirically). This is a form of attentional priming (in advance of task execution) which limits what is processed in the location, feature and object domains. In **Figure 2**, task knowledge is critical for all the MG tasks as part of their basic definition, but also for the AG tasks since it bounds any search processes that might be employed in their solution. In effect, therefore, the original problem of Visual Match has been significantly reframed into a set of more specific problems as **Figure 2** shows, with different constraints on the solution for each and together extending the temporal range of visual tasks far beyond Marr's 160 ms. This is consistent with van Rooij et al. (2012) who proposed computational-level theory revision as a way of dealing with intractability.

Thus, in addition to the three bullet points presented above regarding impact of the analysis, we add two more:

- The use of task or world knowledge can have profound impact on the computational complexity of a visual problem and should be employed whenever available (of course, there must be a default processing state when none is available),
- The discussion on different decision-making strategies and the complex taxonomy of visual tasks of **Figure 2** strongly motivates the need for an executive control process that would dynamically decide on how to best approach and solve visual tasks as they are presented.

THE PROBLEMS WITH PYRAMIDS

Although pyramids played a strong role in reducing complexity, they do cause new problems with how information might flow within them. Some were first described in Tsotsos et al. (1995). **Table 1** provides a characterization of each (more

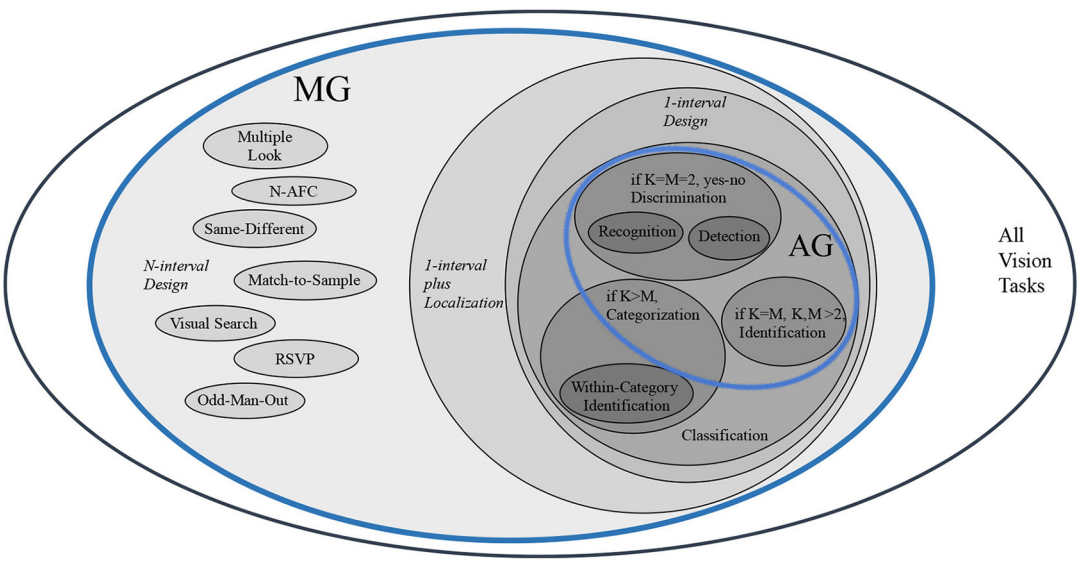


FIGURE 2 | A taxonomy of visual tasks (adapted from Tsotsos, 2011 and task naming based on Macmillan and Creelman, 2005). Within each taxonomy element, there are both easy and difficult instances. AG, At-a-Glance tasks are those that can be solved using only a single feed-forward pass through the brain's visual processing machinery; MG, More-than-a-Glance tasks are those that require more processing than a single feed-forward pass through the brain's visual processing machinery; K, the number of possible images; *M*, the number of object categories of interest.

details can be found in Tsotsos, 2011) and the reader is encouraged to refer to **Figure 3** while reading the table entries. These are all consequences of the basic connectivity pattern of **Figure 1**.

The consideration of representational issues, such as the problem with information flow in a pyramid is not common in the modeling literature (but see Anderson and Van Essen's Shifter circuits, 1987, that were strongly motivated by information routing issues). For the most part, the information flow problems require dynamic solutions that change from moment to moment depending on task and input. Models that ignore these routing characteristics are not only incomplete but lose out on incorporating the constraints that arise.

LATTICE OF PYRAMIDS

The pyramid representation as described so far fits very naturally into the hierarchical view of Hubel and Wiesel (1965, 1968). However, it is insufficient. Felleman and Van Essen (1991) give a set of criteria for determining hierarchical relationships among the visual areas in the cortex. These are:

> *"each area must be placed above all areas from which it receives ascending connections and/or sends descending connections. Likewise, it must be placed below all areas from which it receives descending connections and/or sends ascending connections. Finally, if an area has lateral connections, these must be with other areas at the same hierarchical level."*

This characterization of connectivity resembles that of a general lattice, as shown in **Figure 4B** (see Birkoff, 1967, for a mathematical discussion on the properties of lattice structures). In contrast to the pyramid of **Figure 4A**, i.e., exactly the representation found in convolutional neural networks (CNN-see LeCun and Bengio, 1995; Riesenhuber and Poggio, 1999; Krizhevsky et al., 2012), **Figure 4B** highlights the fact that there may be more than one pathway from input, as is well-documented in visual cortex. Tsotsos (2011) marries the concept of the pyramid with that of the lattice to define the P-Lattice, or lattice of pyramids in order to fully accommodate the criteria laid out by Felleman and Van Essen.

Each element or layer of the pyramid will be referred to as a *sheet*—an array of retinotopically organized neurons of common tuning profile. Each sheet may be connected to more than one other sheet in a feed-forward, recurrent or lateral manner. The main constraint is that no matter which path is taken from lower to higher level, each sheet at a lower level has the same or larger number of elements compared to any higher-level sheet on its

TABLE 1 | A summary description of the main information flow problems resulting from pyramid representations.

Problem	Data flow	Basic characteristic
Blurring **Figure 3A**	↑	Feedforward neural connections have a diverging pattern, a one-to-many mapping, so that spatial precision is not preserved.
Crosstalk **Figure 3B**	↑	Two spatially separated stimuli each root a feedforward diverging cone of connections which may intersect thus presenting neurons within the intersection with a conflicted (corrupted with respect to the stimulus of interest) signal.
Context **Figure 3C**	↑	The receptive field of a neuron—a many-to-one mapping—in the higher layers of the pyramid can be potentially large enough to include not only a stimulus of interest but a significant local spatial context which may confound the stimulus interpretation.
Multiple foci **Figure 3D**	↑↓	If more than one neuron at the output layer is considered, the ability to tease their meanings apart depends on the spatial separation of the receptive fields (the inverted version of the crosstalk problem). In the forward flow direction, contexts due to each overlap to some degree, thus neural responses at the top cannot be considered independent. In the top-down direction, there is a complication when solving the routing problem (see part 3F) which although seemingly trivial for this simple example, would be quite difficult for scenes with many stimuli, such as natural scenes.
Boundary **Figure 3E**	↑	In a hierarchy of spatial convolutions, at each layer, a kernel half-width at the edge of the visual field is left unprocessed because the kernel does not have full data for its convolution. This is compounded layer by layer because the half-widths are additive layer to layer. The result is that a sizeable boundary region at the top layer is left undefined (a true information loss) and thus the number of locations that represent veridical results of neural selectivity from the preceding layer is smaller and restricted to the central portion of the visual field. Solutions, such as used in current CNN's were first described in van der Wal and Burt (1992); they have no biological counterpart. See Tsotsos (2011) and Tsotsos et al. (1995, 2016) for a theory on how the brain deals with the boundary problem.
Routing **Figure 3F**	↑↓	Because of the above problems, a difficulty arises in the search for the neural pathway that connects a stimulus to the neurons that best represent it. If the search is bottom-up—from stimulus to highest layer neuron—then the search is constrained to the feed-forward cone outlined by the dotted lines. If the decisions are based on locally maximal neural responses (such as max pooling), then there is nothing to prevent a bottom-up search losing its way, due to the diverging feedforward connectivity, and missing the globally maximum response at the top layer. It is clear that to be successful, the correct path must always go through the overlap regions shown in dark ovals. But nothing guarantees that the local maximum must lie within those overlap regions. If the search is top-down—from the globally maximum responding neuron to the stimulus—the search is constrained by the dashed lines. Only top-down search is guaranteed to correctly connect the best responding neuron at the top with its stimulus because the search is constrained by the connectivity pattern of the source neuron which necessarily contains the goal stimulus.

Other such problems, not described here are the Sampling, Lateral Spread, Spatial Spread, Spatial Interpolation, and Convergent Recurrence problems and the interested reader can find these in Tsotsos (2011).

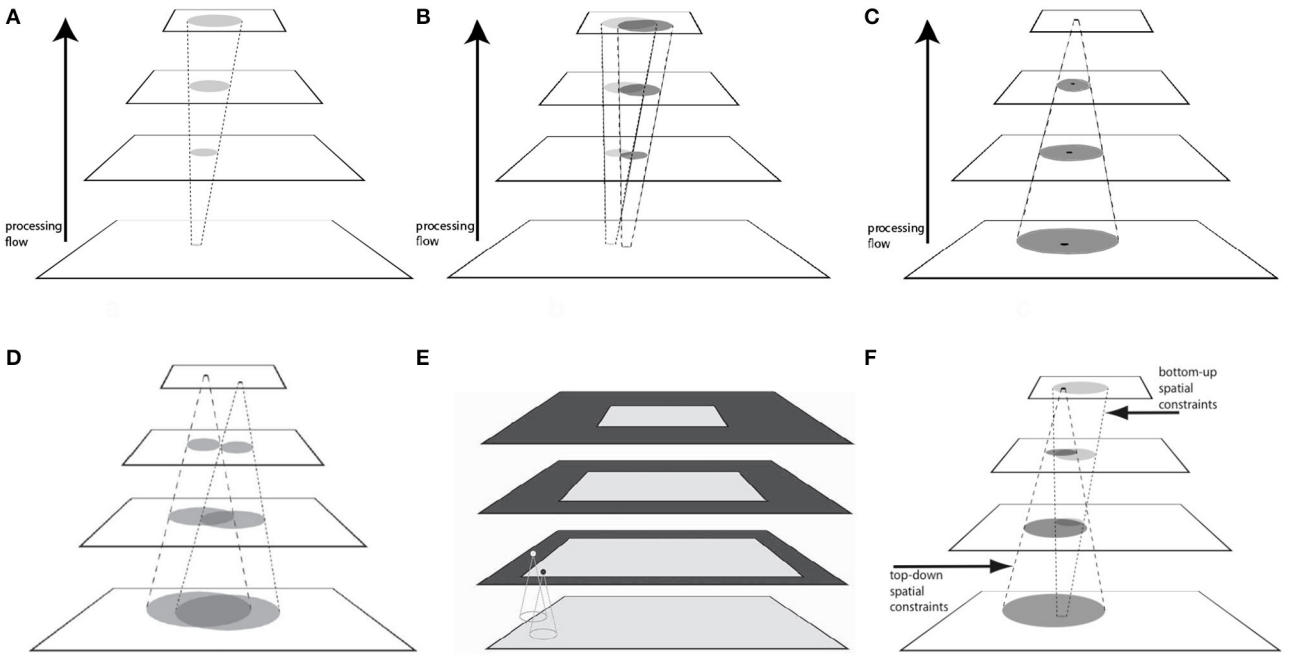


FIGURE 3 | The breadth of problems inherent in pyramid representations. **(A)** The Blurring Problem. An input element in the lowest layer will affect, via its feed-forward connections, a diverging pattern of locations in the higher layers of the pyramid. **(B)** The Crosstalk Problem. Two input stimuli activate feed-forward projections that overlap, with the regions of overlap containing neurons that are affected by both. Those might exhibit unexpected responses with respect to their tuning profiles. **(C)** The Context Problem. A stimulus (black dot) within the receptive field of a top layer neuron, showing its spatial context defined by that receptive field. **(D)**. The Multiple Foci Problem. Regions of overlap show the extent of interference if two (or more) output nodes are considered simultaneously. **(E)** The Boundary Problem. The two units depicted in the second layer from the bottom illustrate how the extent of the black unit's receptive field is entirely within the input layer while only half of the receptive field of the gray unit is within the input layer. The bottom layer represents the retina; the next layer of the pyramid (say area V1) represents the spatial dimension of the viewing field in a manner that gives more cortical area to central regions than peripheral. The boundary problem forces more and more of the periphery to be unrepresented in higher layers of the pyramid. **(F)** The Routing Problem. Interacting top-down and bottom-up spatial search constraints are shown with the areas of overlap representing the viable search regions for best neural pathway. (Reproduced from Tsotsos, 2011).

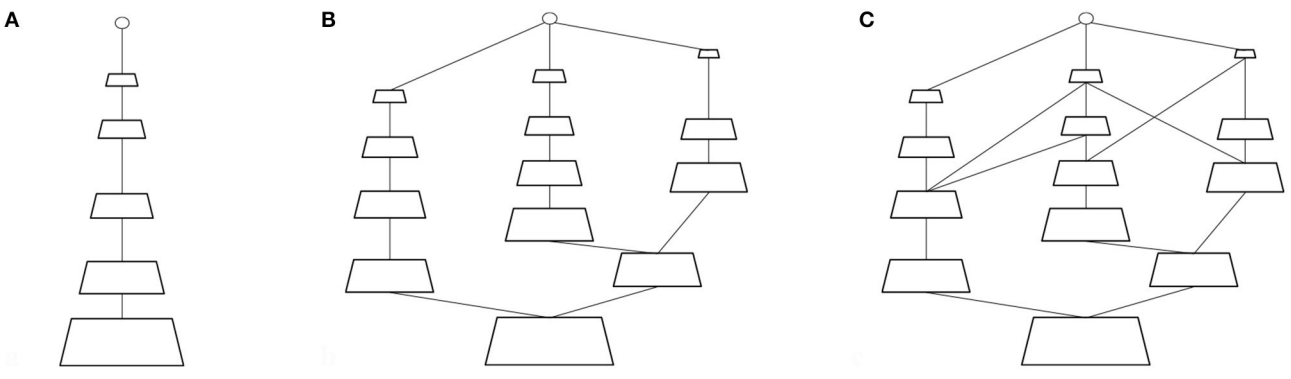


FIGURE 4 | From pyramids to P-lattices. **(A)** A simple pyramid representation. **(B)** A lattice of three pyramids. **(C)** A lattice of pyramids showing complex connectivity.

path. Both **Figures 4B,C** are P-Lattices; the **Figure 4C** shows a more complex version of **Figure 4B** in order to illustrate the full nature of the representation. The formalization will not be further described, but is developed in Tsotsos (2011). It should be apparent that the P-Lattice representation is much more faithful to the organization of different processing areas in the brain than the standard CNN.

The P-Lattice concept also lends itself very naturally to thinking about an organization that includes not only a part-whole relationship as is common for pyramids, but also a specialization relationship. Different features may be separated out into different sheets, and those may then be specialized differently along each pathway of the P-Lattice.

SELECTIVE TUNING

As a result of the complexity level analysis, a series of papers outlined the development of a model for how the main conclusions in the previous sections might impact a visual processing hierarchy (Tsotsos, 1988b, 1990, 1995b, 2011; Tsotsos et al., 1995, 2001; Rothenstein and Tsotsos, 2014). This model, named Selective Tuning (ST) was intended to provide a mechanistic explanation for how not only attentive selection and restriction might occur, but also, how the visual system deals with the many problems of information flow described in the previous section. To this end, ST incorporated pyramid representations, spatiotemporally limited receptive fields, separable feature representations, dynamic tuning and attentive selection. In order to deal with the Context Problem, ST employs a suppressive mechanism, recurrent localization, to inhibit portions of a receptive field deemed 'ground' while attending to 'figure' (see Tsotsos et al., 1995; Tsotsos, 2011 for details). Thus, suppression must be added to selection and restriction to form the full suite of attentional mechanisms. ST also offers an explanation for a wide variety of attentional phenomena; it is among the oldest and most studied models of attention. ST, beginning with the earliest papers, made a number of predictions about visual attention at both neural and behavioral levels, which, starting in the late 1990's, have seen broad and strong experimental support[17] (reviewed in Tsotsos, 2011; also in Hopf et al., 2010; Carrasco, 2011 and more).

Figure 5 illustrates the main features of the model showing how there are many aspects to attentive processing, and which are executed determined by the nature of the task of the moment. It shows the different stages of processing of the visual hierarchy needed for different visual tasks. The five components of the figure represent processing stages ordered in time, from left to right. The stages may be described as **Figure 5A**: pre-stimulus (shown as blank to portray a visual hierarchy ready for a new stimulus); **Figure 5B**. top-down priming for task; **Figure 5C**: feedforward stimulus processing and figure selection; **Figure 5D**: recurrent localization and local suppression, if the task requires it; **Figure 5E**: secondary feedforward processing. This illustrates the main cost associated with dynamic tuning, namely, time. Each hierarchy traversal may be primed for different function. Different visual tasks require different processing times depending on passes through the hierarchy. A smaller additional cost would be the process of actual tuning. Different visual tasks require different sets of these basic elements, sometimes with repeated elements and this shows how *dynamic tuning* can be realized.

To summarize, ST features several major elements not present in other models of attention: (1) the recurrent localization process; (2) the integration of multiple attentional processes within a single framework; (3) both local and global attentional operations; (4) the realization that not all vision occurs within the 150 ms time frame and that different kinds of visual tasks require different processes and thus take different durations to complete; (5) the capacity to dynamically tune the visual processing hierarchy depending on task; and (6) the use of inhibitory mechanisms rather than enhancement in order to achieve attentive effects (enhancement is a side-effect of suppression of competing stimuli).

[17]These predictions - all asserted before any supporting experimental data - include, for example, the suppressive surround in spatially attended stimuli, a suppressive surround in the attended feature dimension, the latency of attentional neural modulation having a top-down pattern, that neural modulation due to attention is present throughout the visual hierarchy, that neural baseline firing increases for an attended location and decreases elsewhere, and more.

NATURE OF SIGNAL INTERFERENCE IN THE P-LATTICE

The impressive successes of deep learning approaches to vision system development may lead one to think that vision is a solved problem, and that all one needs is a fast-enough computer and enough training data[18]. The complexity level analysis does indeed tell us something of interest here: that with enough computational capacity, *some* vision problems can be solved. Recall that the role of image size in the complexity function; this dictates the primary barrier without attentive selection. Proponents of deep learning widely acknowledge that the advent of GPU's and faster processors contributed to the recent successes. This is not the same as saying the vision problem has been made tractable: all it means is that with enough GPU power, the size of image—that is, the value of P that can be realized in the complexity expression—is now a reasonable number for practical applications. Importantly, it cannot be as large as the size of a human retina. We also note that although those approaches do indeed receive some motivation from biological vision, that motivation is almost entirely based on knowledge of the late 1960's. The methods validate the concepts of spatially limited receptive field size, convolution processing and hierarchical processing levels, but not much more. The representations typically used in deep learning are also not easily related to neural representations nor their methods for decoding those representations. None of this of course should detract from their practical success. The point here is simply that there is a great deal more work to be done with respect to understanding how biological systems deal with visual problems.

Let us return to the representation problem. Pyramid representations help with reducing complexity but as shown above, add new complications that can, as a group, be considered as signal interference. In other words, all incoming signals are represented in all layers of a pyramid (this is true for central regions, but not for peripheral—see **Figure 2E**), as they are in all layers of a modern CNN too. But they are not easily discriminable due to the interference that the context, boundary, blurring problems impose. It is important to examine interference more deeply.

The Context Problem is due to many-to-one neural mapping, the Blurring Problem due to one-to-many neural mapping and the Boundary Problem due to the realities of convolution processes. Of these, only the Boundary Problem leads to actual information loss and specifically in the periphery; the rest lead to

[18]Amnon Sha'shua, for example, asserted this in his keynote lecture the 2016 IEEE Computer Vision and Pattern Recognition conference, Las Vegas NV (Sha'shua, 2016). Elon Musk also claimed autonomous driving is solved, for which vision is a key technology, in Eadicicco (2016).

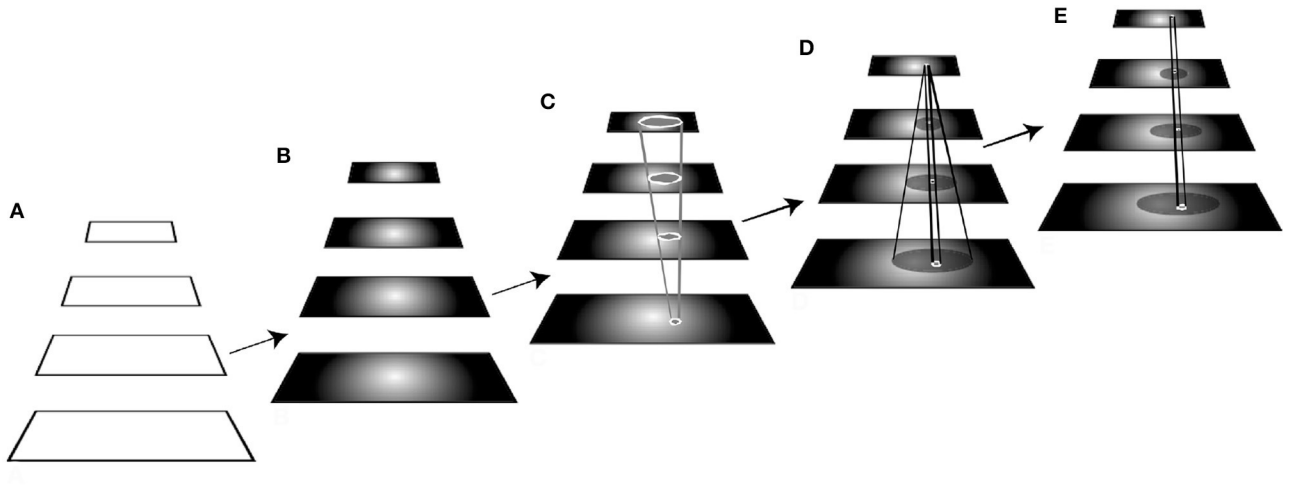


FIGURE 5 | The different stages of processing of the visual hierarchy needed for different visual tasks. The five components of the figure represent processing stages ordered in time, from left to right. **(A)** In the first stage, the network is portrayed as "blank," that is, without stimulus or top-down influences, as it might be prior to the start of an experiment, for example. **(B)** The second stage shows the network affected by a top-down pass tuning the network with any priming information to set up its expectation for a stimulus to appear, when such information is available. Here, the network is set up to expect a stimulus that is centrally located and is imposed via a global suppression of non-task-relevant locations and/or features. **(C)** At this point, the stimulus appears and is processed by the tuned network during a single feedforward pass. If the task is sufficiently simple, such as a detection or categorization tasks with sufficiently simple stimuli so that figure can be selected from ground, processing is complete. **(D)** If the required task for this stimulus cannot be satisfied by the first feedforward pass, such as for a within-category identification or the need for an eye movement response, the recurrent localization algorithm is deployed that traverses the network in a top-down manner, identifying the selected components while suppressing their spatial surrounds locally. **(E)** A subsequent feedforward pass then permits a re-analysis of the attended stimulus with interfering signals reduced or eliminated. It also permits a continuation of the cycle in a repeating fashion, such as would be needed for visual search. This illustrates the main cost associated with dynamic tuning, namely, time. Different visual tasks require different processing times depending on passes through the hierarchy. A smaller additional cost would be the process of actual tuning. (Reproduced from Tsotsos and Kruijne, 2014).

signal interference via combination. Every signal continues to be represented during the feedforward traversal of an input signal, except that it becomes increasingly intertwined and amalgamated with nearby signals, dictated by receptive field sizes. Modern theories prescribe computational decoding procedures that are able to take this muddled representation as input and decode it to extract meaning. For example, Hung et al. (2005) used a classifier-based readout technique (linear SVM) to interpret the neural coding of selectivity and invariance at the IT population level. The activity of small neuronal populations over very short time intervals (as small as 12.5 ms) contained accurate and robust information about both object "identity" and "category." Coarse information about position and scale could be read out over three positions. Isik et al. (2014) used neural decoding analysis (also known as multivariate pattern analysis, or readout) to understand the timing of invariant object recognition in humans. Neural decoding analysis applies a machine learning classifier to assess what information about the input stimulus is present in the recorded neural data. They found that size—and position-invariant visual information appear around 125 and 150 ms, respectively, and both develop in stages, with invariance to smaller transformations arising before invariance to larger transformations. They claimed that this supports a feed-forward hierarchical model of invariant object recognition where invariance increases at each successive visual area along the ventral stream. This is in contrast to work by Zhang et al. (2011) who show how a classifier can be trained on data from isolated-object trials and then make predictions about which objects were shown on either different isolated-object trials or on trials in which three objects are shown. They concluded that by focusing on how information is represented by populations of neurons, competitive effects that occur when two stimuli are presented within a neuron's RF, and global gain-like effects that occur when a single stimulus is presented within a neuron's RF, can both be viewed as restoring patterns of neural activity for object identity and position information, respectively. The competitive interactions Zhang et al. refer to are attentive mechanisms whose intent is to reduce interference, which was the goal of their study. The difference between the last two papers is due to the different stimuli used, the latter requiring attention and the former not. We can conclude that although coarse location information is likely easily extracted after a single feedforward pass for detection tasks, more complex visual tasks that require image details of precise features of location likely are not. The Multiple Foci problem of **Figure 3D** illustrates this nicely; spacing within the visual field dictates the degree of interference.

Let's continue to examine this neural interference. It is well-known and studied that the size of visual receptive fields generally increases with higher levels (or greater abstraction) of processing within the visual hierarchy of the brain. There is a further dependency not only on abstraction level but also eccentricity, or distance of the receptive field from the center of gaze. Kay et al. (2013) provide illuminating plots of receptive field sizes

in many visual areas of human cortex as a function of retinal eccentricity, reproduced in **Figure 6**. It is clear that the receptive field size increases with eccentricity within each visual map. Second, the receptive field size differs between maps, with the smallest pRFs in V1, and much larger pRFs in ventral (hV4, VO-1/2) and lateral (LO-1/2, TO-1/2) maps, showing a progression from least to most abstract in terms of processing. It is important to note—as the complexity level analysis pointed out earlier—that receptive fields are space-limited, i.e., there seem to be no fully connected layers where all receptive fields are connected to all others. There is a well-defined feedforward as well as feedback connectivity pattern (mostly symmetric) so that each element of a representation affects a clear feedforward diverging cone of elements in the next representation, is fed by a clear converging cone of elements from the earlier representation and these connections are bidirectional (this is exactly what **Figure 1** illustrates). A re-plotting of the elements of **Figure 6** leads to an explicit view in **Figure 7** of the spatial extent of feedforward convergence. Superimposing the receptive field maps, V3 onto V1, V4 onto V1 and a hypothetical LO1 receptive field (using values from **Figure 6** at 20° eccentricity) shows clearly that degree of signal convergence onto single neurons with higher levels of visual processing in cortex. These figures are a concrete demonstration of the Blurring and Context Problems of **Figure 3**. How can the visual system function at all under such circumstances? Most models do not consider how such eccentricity-dependent receptive field size variations might be usefully incorporated.

First, it might be the case that there are many more target representations at higher levels than previously thought, something hinted at by the very recent results of Glasser et al. (2016). That is, the breadth of the P-Lattice representation in the brain may be significant. Perhaps these might be specializations as suggested earlier, thus removing some of the interference that way. Second, lateral interactions within representations could assist in well-known ways by enhancing contrast, contrast in

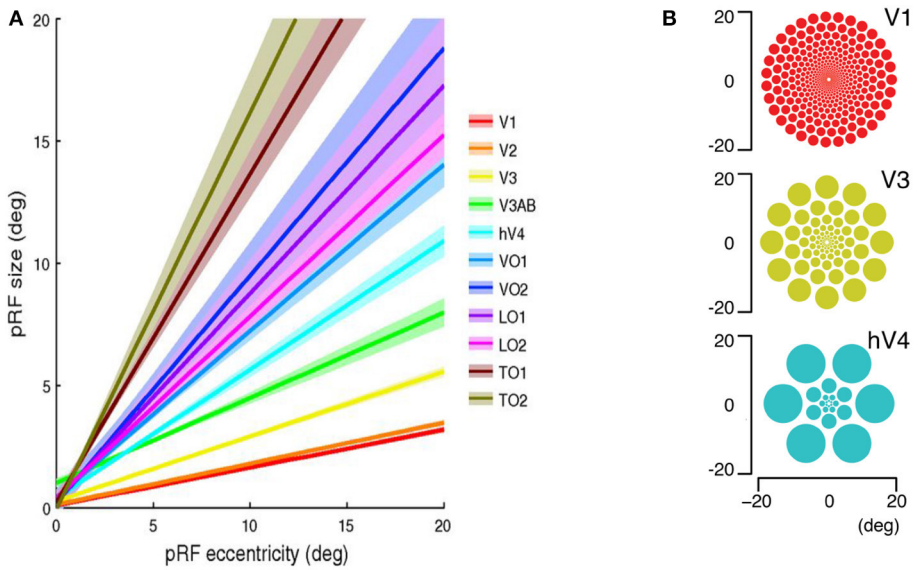


FIGURE 6 | Regularities in human population receptive field properties measured with functional MRI. **(A)** Population receptive field size as a function of eccentricity in several human retinotopic maps. Two clear trends are evident. First, the population receptive field size increases with eccentricity within each map. Second, the population receptive field size differs between maps, with the smallest pRFs in V1, and much larger pRFs in ventral (hV4, VO-1/2) and lateral (LO-1/2, TO-1/2) maps **(B)** The spatial array of pRFs using the parameters in the left panel. The radius of each circle is the apparent receptive field size at the appropriate eccentricity. [a-from Kay et al., 2013, Reproduced with permission of the publisher; b-Reproduced with permission of J. Winawer (https://archive.nyu.edu/handle/2451/33887)].

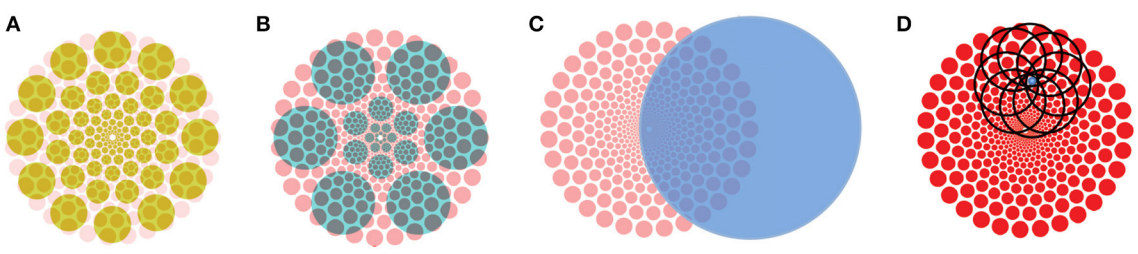


FIGURE 7 | Superposed spatial arrays of receptive fields using plots of **Figure 6**. **(A)** V3 over V1. **(B)** V4 over V1. **(C)** LO1 over V1 with the LO1 receptive field centerd over 20° eccentricity to match the data from **Figure 6**. **(D)** The black circles represent the feedforward divergence of outputs from a single V1 neuron at the V4 level.

this case not being restricted to luminance but to contrast in any featural or conceptual space. But this contrast enhancement cannot be total because local decisions may be wrong (Marr's, 1982 principle of least commitment; Herzog and Clarke, 2014).

It is not hard to believe that a classifier can indeed be trained to extract location for simple (Marr-like) images with small numbers of separated stimuli as Hung et al. report. But such a situation is not representative of real vision. Something more is needed for natural images and for tasks where more precision is required than simple coarse position. There are really two choices: 1-provide mechanisms that dynamically ameliorate the interference before interpretation; or, 2-provide mechanisms to correctly interpret corrupted representations. The methods just described are of the latter type. We chose to explore the former possibility. A key feature of the Selective Tuning model of visual attention is the use of a recurrent localization process that imposes a suppressive surround around the attended stimulus as shown in **Figure 5D** (Carrasco, 2011; Tsotsos, 2011) to deal with the Context and Routing problems. This would require a top-down pass through the processing hierarchy after the initial feedforward pass, consistent with the behavioral timing observed for such tasks. The requirement for an additional top-down pass for localization is not inconsistent with the claims of Isik et al. (2014). In ST, it is the recurrent localization process that replaces the role of the classifier, and in contrast to current classifiers presents a biologically plausible mechanism (supported experimentally, e.g., Boehler et al., 2009, 2011; Hopf et al., 2010).

Signal interference within a pyramid representation is a reality that seems insufficiently addressed in general. To be sure, the majority of experimental work, whether neural or behavioral, focus on foveal or near-foveal stimuli and as the plots of **Figure 6** show, the interference impact is not so great. Further, most experimenters use relatively simple stimuli, spaced apart and with little conflicting context. As the diagrams of **Figure 3** show, the distance between stimuli matters for the Blurring, Crosstalk, and Context Problems and it is experimentally possible to minimize the effect, thus making it appear as if the problem does not exist. As a result, experimental work does not fully address the problem in order to determine if and how it might cause interference or how the brain might deal with it. New experimental paradigms seem required.

ATTENTIVE PROCESSING AND ADAPTIVE BEAMFORMING

The most common way in which attention has found its way into theories and models of visual processing or other human sensory or cognitive abilities is as a mechanism to defeat capacity limits. This is also true for computational systems. The most prevalent mechanism is that of selecting a region of interest in some modality of the sensory input or in some conceptual space, such as a task-relevant sub-domain of interest. In a behaving agent, eye movements are most often considered the primary indicator of a shift in attention. Nevertheless, as Tsotsos (2011) argues and as any review of visual attention (such as Carrasco, 2011) amply illustrates, attention is a much broader capability with, sadly, no real consensus on how it might be characterized. One possibility for such a broad characterization appeared in Tsotsos (2011) where is was proposed that attention is a set of mechanisms that tune and control the search processes inherent in perception and cognition, with the major types of mechanisms being Selection, Suppression, and Restriction. Within each type are several specific mechanisms as shown in **Figure 8**.

Earlier, as a result of the complexity level analysis, it was asserted that the original vision problem is reframed by partitioning the space of problem instances into sub-spaces where each might be solvable by a different method instead of having a single, optimal, algorithm for all problem instances. The resource limits—which are fixed and common for all sub-problems in the case of the brain—guide the choices. A key element of the process is to have a method that, when confronted with a visual problem instance, can quickly determine which solution method to apply. And this is where attention is critical. A sufficiently flexible attentive process can start from the general and thus largest possible problem definition, and then focus in and scale down the problem to more manageable sub-problems. Combining all of these seemingly disparate tools, as shown in **Figure 8**, within a single formulation seems a daunting task, but this is what the Selective Tuning model of attention attempts to do (Tsotsos, 1988b, 1990, 2011; Tsotsos et al., 1995).

Interestingly, a related combination of disparate tools has not only been attempted previously, but has developed into a well-understood and very widely use technology, namely adaptive beamforming. Beamforming is a signal processing technique used in sensor arrays for directional signal transmission or reception (Van Veen and Buckley, 1988). Electromagnetic waves are additive and if more than one wave co-exists in space and time, this additive property causes each waveform to interfere with the others. Beamforming attempts to minimize this interference. This is achieved by controlling how elements combine so that some signals experience constructive interference while others experience destructive interference. Beamforming can be used at both the transmitting and receiving ends in order to achieve spatial selectivity. Beamforming can be used for radio or sound waves and has found numerous applications in radar, sonar, seismology, wireless communications, radio astronomy, acoustics, and biomedicine. An adaptive beamformer dynamically adjusts in order to maximize or minimize a desired parameter, such as signal-to-interference-plus-noise ratio. Dynamically adjusting phase and magnitude will cause the antenna gain pattern to change and provides for directional sensitivity without physically moving an array of receivers or transmitters.

The essence of beamforming seems precisely what attention seeks to accomplish: to pick out the relevant signal from among all the irrelevant ones. This connection between attention and beamforming has been made previously in the auditory domain (see Kidd et al., 2015, for a recent effort) in order to provide solutions to the well-known Cocktail Party problem. There are components of constructive and destructive interference within the attentional mechanisms of ST, and more, but it would be beyond the scope here to further explore the relationship. However, it is clear that any representations of visual information processing must

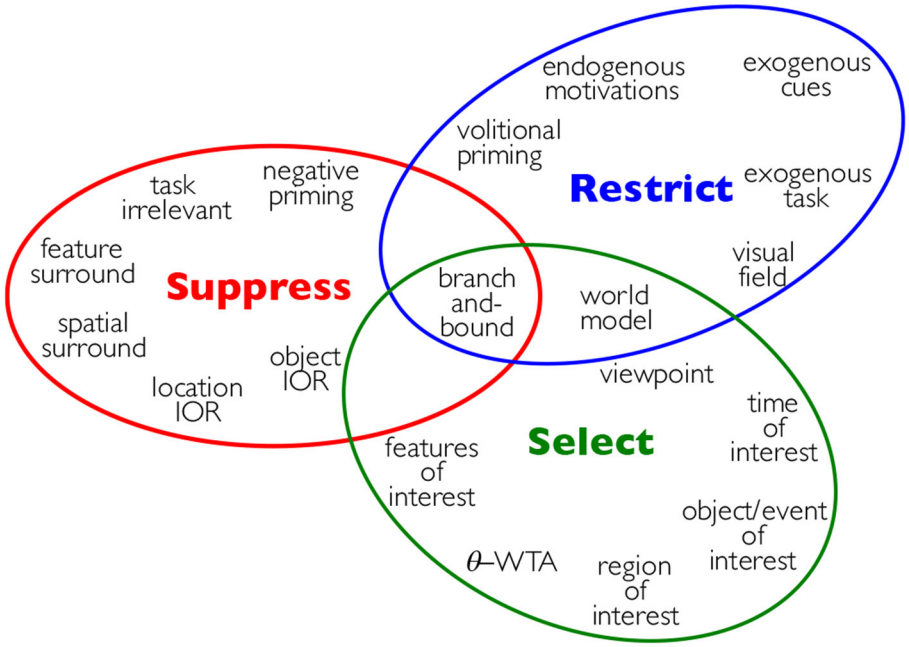


FIGURE 8 | Attention is a set of mechanisms that tune and control the search processes inherent in perception and cognition, with the major types of mechanisms being Selection, Suppression, and Restriction. See Tsotsos (2011) for details on each of the sub-mechanisms.

support these mechanisms. Adaptive beamforming—or perhaps more appropriately *attentive beamforming*—might present an appropriate analogy for formalization of dynamic visual attention processes.

CONCLUSIONS

The hallmark of human vision is its generality. The same brain and same visual system allow one to play tennis, drive a car, perform surgery, view photo albums, read a book, gaze into your loved one's eyes, go online shopping, solve 1,000-piece jigsaw puzzles, find your lost keys, chase after your young daughter when she appears in danger, and so much more. The reality is that incredible as the AI successes so far have been, it is humbling to acknowledge how far there is still to go. Recent AI systems even sometimes outperform humans so it is difficult to determine how well they might provide an explanation for human intelligence. With respect to an explanation for human intelligence, it is as important to ensure that model systems behave correctly as humans and with the same response times, as it is to ensure model systems fail as humans do. The successes have all been uni-taskers (they have a single, narrowly defined function)—the human visual system is a multi-tasker, and the tasks one can teach that system seem unbounded. And it is an infeasible solution to simply create a brain that includes a large set of uni-taskers.

Representation has been central to AI since its inception and it is only recently that it seems supplanted by the success of the machine learning approach. Unfortunately, the representations that learning systems create—except possibly for limited aspects of early vision—seem inscrutable. It might be that in order to make progress, there remains a need to better understand the kinds of representations and their transformations as they may be occurring in the brain, a sentiment appearing decades ago. Zucker (1981) stressed the importance of representation. He pointed out that computational models have two essential components—representational languages for describing information, and mechanisms that manipulate those representations, and: "*One of the strongest arguments for having explicit abstract representations is the fact that they provide explanatory terms for otherwise difficult (if not impossible) notions.*"

Our presentation has focused on the constraints that complexity level analysis presents for the representations and for the visual processes that operate on them in the brain (or in machines). It is clear that the main claim, namely, that resource-complexity matching is a source of critical constraints on the viability of theories, remains intact. The 30 years that have passed since their first introduction in this context have given us the luxury of seeing how they stood the test of time. None of the conclusions were in common use back then and some indeed were firmly believed to be incorrect[19]. Throughout, we have argued for a very specific view on representation and their processing, whose features include:

[19]For example, the prediction of spatial surround suppression due to attention, first described in Tsotsos (1988b), was in fact "proved" infeasible in the brain by Crick and Koch (1990; p. 959) but now is widely confirmed (see review by Carrasco, 2011). See also the various peer commentaries published along with Tsotsos (1990).

- an overall organization of visual areas into a lattice of pyramids,
- spatiotemporally limited receptive fields,
- specialized pathways based on visual features,
- a suite of attentional mechanisms that dynamically suppress, select and restrict processing to control the input space and to ameliorate the signal interference problem, and,
- the use of task or world knowledge can have profound impact on the computational complexity of a visual problem and should be employed whenever available,
- a partitioning of the space of visual tasks into a taxonomy of sub-tasks, each with its own specific characteristics and requiring differing methods all realized on that same processing substrate,
- the different decision-making strategies and the complex taxonomy of visual tasks strongly motivates the need for an executive control process that would dynamically decide on how to best approach and solve visual tasks as they are presented.

Moreover, the intractability results of our own work and of all other authors cited here, and more, show the futility of pursuing single criterion algorithms of any kind (for example, Friston's (2010) free-energy principle). Much is already in line with current knowledge of the brain, many of these features have found their way into the successful systems of the present, but much still requires further study. There is no suggestion that complexity level analysis can replace any other type of analysis. However, it is a critical component of theory development and provides an important source of constraint that models cannot do without.

AUTHOR CONTRIBUTIONS

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

FUNDING

This research was supported by several sources for which the author is grateful: Air Force Office of Scientific Research (FA9550-14-1-0393), Office of Naval Research (N00178-15-P-4873), the Canada Research Chairs Program (950-219525), and the Natural Sciences and Engineering Research Council of Canada (RGPIN-2016-05352).

REFERENCES

Anderson, C., and Van Essen, D. (1987). Shifter circuits: a computational strategy for dynamic aspects of visual processing. *Proc. Nat. Acad. Sci.* 84, 6297–6301. doi: 10.1073/pnas.84.17.6297

Andreopoulos, A., and Tsotsos, J. K. (2013). A computational learning theory of active object recognition under uncertainty. *Int. J. Comp. Vis.* 101, 95–142. doi: 10.1007/s11263-012-0551-6

Birkoff, G. (1967). *Lattice Theory, 3rd Edn*. Providence, RI: American Mathematical Society.

Boehler, C. N., Tsotsos, J. K., Schoenfeld, M., Heinze, H.-J., and Hopf, J.-M. (2009). The center-surround profile of the focus of attention arises from recurrent processing in visual cortex. *Cereb. Cortex* 19, 982–991. doi: 10.1093/cercor/bhn139

Boehler, C. N., Tsotsos, J. K., Schoenfeld, M., Heinze, H.-J., Hopf, J.-M. (2011). Neural mechanisms of surround attenuation and distractor competition in visual search. *J. Neurosci.* 31, p5213–p5224. doi: 10.1523/JNEUROSCI.6406-10.2011

Brown, J. W. (2014). The tale of the neuroscientists and the computer: why mechanistic theory matters. *Front. Neurosci.* 8:349. doi: 10.3389/fnins.2014.00349

Carrasco, M. (2011). Visual attention: the past 25 years. *Vis. Res.* 51, 1484–1525. doi: 10.1016/j.visres.2011.04.012

Cooper, M. C. (1998). The tractability of segmentation and scene analysis. *Int. J. Comput. Vis.* 30, 27–42. doi: 10.1023/A:1008013412628

Crick, F., and Koch, C. (1990). "Some reflections on visual awareness," in *Cold Spring Harbor Symposia on Quantitative Biology*, eds J. Watson and J. A. Witkowski (Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press), 953–962.

Davis, M. (1958). *Computability and Unsolvability*. New York, NY: McGraw-Hill.

Davis, M. (1965). *The Undecidable*. New York, NY: Hewlett Raven Press.

Downey, R. G., and Fellows, M. R. (1999). *Parameterized Complexity*. New York, NY: Springer.

Duncan, J., and Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychol. Rev.* 96:433. doi: 10.1037/0033-295X.96.3.433

Eadicicco, L. (2016). *Elon Musk Just Made These 5 Bold Claims About the Future, Time*. Available online at: http://time.com/4354864/elon-musk-mars-driverless-cars-apple-tesla-spacex (Accessed June 2, 2016).

Feldman, J., and Ballard, D. (1982). Connectionist models and their properties. *Cogn. Sci.* 6, 205–254. doi: 10.1207/s15516709cog0603_1

Felleman, D., and Van Essen, D. (1991). Distributed hierarchical processing in the primate visual cortex. *Cereb. Cortex* 1, 1–47. doi: 10.1093/cercor/1.1.1

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138. doi: 10.1038/nrn2787

Fukushima, K. (1988). Neocognitron: a hierarchical neural network capable of visual pattern recognition. *Neural Netw.* 1, 119–130. doi: 10.1016/0893-6080(88)90014-7

Funahashi, S. (2001). Neuronal mechanisms of executive control by the prefrontal cortex. *Neurosci. Res.* 39, 147–165. doi: 10.1016/S0168-0102(00)00224-8

Garey, M., and Johnson, D. (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness*. San Francisco, CA: Freeman.

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., et al. (2016). A multi-modal parcellation of human cerebral cortex, *Nature* 536, 171–178. doi: 10.1038/nature18933

Grimson, W. E. L. (1990). The combinatorics of object recognition in cluttered environments using constrained search. *Artif. Intel.* 44, 121–165. doi: 10.1016/0004-3702(90)90100-E

Herzog, M. H., and Clarke, A. M. (2014). Why vision is not both hierarchical and feedforward. *Front. Comput. Neurosci.* 8:135. doi: 10.3389/fncom.2014.00135

Hopf, J.-M., Boehler, N., Schoenfeld, M., Heinze, H.-J., Tsotsos, J. K. (2010). The spatial profile of the focus of attention in visual search: insights from MEG recordings. *Vision Res.* 50, 1312–1320. doi: 10.1016/j.visres.2010.01.015

Hubel, D., and Wiesel, T. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* 160, 106–154. doi: 10.1113/jphysiol.1962.sp006837

Hubel, D., and Wiesel, T. (1965). Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *J. Neurophysiol.* 28, 229–289.

Hubel, D. H., and Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *J. Physiol.* 195, 215–243. doi: 10.1113/jphysiol.1968.sp008455

Hung, C. P., Kreiman, G., Poggio, T., and DiCarlo, J. J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science* 310, 863–866. doi: 10.1126/science.1117593

Isik, L., Meyers, E. M., Leibo, J. Z., and Poggio, T. (2014). The dynamics of invariant object recognition in the human visual system. *J. Neurophysiol.* 111, 91–102. doi: 10.1152/jn.00394.2013

Jolion, J.-M., and Rosenfeld, A. (1994). *A Pyramid Framework for Early Vision*. Dordrecht: Kluwer.

Judd, S. (1988). On the complexity of loading shallow neural networks. *J. Complex.* 4, 177–192. doi: 10.1016/0885-064X(88)90019-2

Kasif, S. (1990). On the parallel complexity of discrete relaxation in constraint satisfaction networks. *Artif. Intell.* 45, 275–286. doi: 10.1016/0004-3702(90)90009-O

Kay, K. N., Winawer, J., Mezer, A., and Wandell, B. A. (2013). Compressive spatial summation in human visual cortex. *J. Neurophysiol.* 110, 481–494. doi: 10.1152/jn.00105.2013

Kidd, G., Mason, C. R., Best, V., and Swaminathan, J. (2015). Benefits of acoustic beamforming for solving the cocktail party problem. *Trends Hear.* 19:2331216515593385. doi: 10.1177/2331216515593385

Kirousis, L., and Papadimitriou, C. (1988). The complexity of recognizing polyhedral scenes. *J. Comp. Sys. Sci.* 37, 14–38. doi: 10.1016/0022-0000(88)90043-8

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "Imagenet classification with deep convolutional neural networks" in *Advances in Neural Information Processing Systems*, Vol. 25 (Lake Tahoe, NV: Neural Information Processing Systems Conferences), 1097–1105. Available online at: papers.nips.cc

Kube, P. R. (1991). Unbounded visual search is not both biologically plausible and NP-complete. *Behav. Brain Sci.* 14, 768–770. doi: 10.1017/s0140525x00072472

LeCun, Y., and Bengio, Y. (1995). "Convolutional networks for images, speech, and time series," in *The Handbook of Brain Theory and Neural Networks*, ed M. A. Arbib (Cambridge, MA: MIT Press), 276–279.

Lennie, P. (2003). The cost of cortical computation. *Curr. Biol.* 13, 493–497. doi: 10.1016/S0960-9822(03)00135-0

Macmillan, N. A., and Creelman, C. D. (2005). *Detection Theory: A User's Guide*. New York, NY: Lawrence Erlbaum Associates.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York, NY: Henry Holt and Co.

Papadimitriou, C. H. (2003). *Computational Complexity*. Chichester: John Wiley and Sons Ltd.

Parodi, P., Lanciwicki, R., Vijh, A., and Tsotsos, J. K. (1998). Empirically-derived estimates of the complexity of labeling line drawings of polyhedral scenes. *Artif. Intell.* 105, 47–75. doi: 10.1016/S0004-3702(98)00077-0

Pavlidis, T. (2014). The challenge of general machine vision. *Signal Image Video Proc.* 8, 191–195. doi: 10.1007/s11760-013-0549-8

Rensink, R. (1989). *A New Proof of the NP-Completeness of Visual Match.* Technical Report Department of Computer Science, University of British Columbia, 89–22.

Riesenhuber, M., and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nat. Neurosci.* 2, 1019–1025. doi: 10.1038/14819

Rothenstein, A. L., and Tsotsos, J. K., (2014). Attentional modulation and selection – an integrated approach, public library of science *PLoS ONE* 9:e99681. doi: 10.1371/journal.pone.0099681

Russell, S. J., Norvig, P., Canny, J. F., Malik, J. M., and Edwards, D. D. (2003). *Artificial Intelligence: A Modern Approach*. Upper Saddle River, NJ: Prentice hall.

Sha'shua, A. (2016). *Autonomous Driving, Computer Vision and Machine Learning, IEEE CVPR 2016*. Available online at: https://www.youtube.com/watch?v$=$n8T7A3wqH3Q

Stockmeyer, L., and Chandra, A. (1988). Intrinsically difficult problems, *Sci. Am. Trends Comput.* 1, 88–97.

Thorpe, S., and Imbert, M. (1989). "Biological constraints on connectionist modelling," in *Connectionism in Perspective*, eds R. Pfeifer, Z. Schreter, F. Fogelman-Souli é, and L. Steels eds (Amsterdam: Elsevier), 63–93.

Traub, J. F. (1991). "What is scientifically knowable," in *Twenty-Fifth Anniversary Symposium, School of Computer Science* (Reading, MA: Carnegie-Mellon University, Addison-Wesley), 489–503.

Tsotsos, J. K. (1987). "A 'complexity level' analysis of vision," in *Proceedings of the 1st International Conference on Computer Vision* (London; Washington, DC: IEEE Computer Society Press), 346–355.

Tsotsos, J. K. (1988a). A "complexity level" analysis of immediate vision. *Int. J. Comput. Vision* 2, 303–320. doi: 10.1007/BF00133569

Tsotsos, J. K. (1988b). "How does human vision beat the computational complexity of visual perception?," in *Computational Processes in Human Vision: An Interdisciplinary Perspective*, ed Z. Pylyshyn (Norwood, NJ: Ablex Press), 286–338.

Tsotsos, J. K. (1989). "The complexity of perceptual search tasks," in *Proceedings of the 11th International Joint Conference on Artificial Intelligence*, ed N. Sridharan (Detroit, MI), 1571–1577.

Tsotsos, J. K. (1990). Analyzing vision at the complexity level. *Behav. Brain Sci.* 13, 423–445. doi: 10.1017/S0140525X00079577

Tsotsos, J. K. (1991). Is complexity theory appropriate for analyzing biological systems? *Behav. Brain Sci.* 14, 770–773. doi: 10.1017/S0140525X00072484

Tsotsos, J. K. (1992). On the relative complexity of passive vs. active visual search. *Int. J. Comput. Vis.* 7, 127–141. doi: 10.1007/BF00128132

Tsotsos, J. K. (1993). "The role of computational complexity in understanding perception," in *Foundations of Perceptual Theory*, ed S. Masin (Amsterdam: North-Holland Press), 261–296.

Tsotsos, J. K. (1995a). Behaviorist intelligence and the scaling problem. *Artif. Intell.* 75, 135–160. doi: 10.1016/0004-3702(94)00019-W

Tsotsos, J. K. (1995b). "Towards a computational model of visual attention," in *Early Vision and Beyond*, eds T. Papathomas, C. Chubb, A. Gorea, and E. Kowler (Cambridge, MA: MIT Press/Bradford Books), 207–218.

Tsotsos, J. K. (2011). *A Computational Perspective on Visual Attention*. Cambridge, MA: MIT Press.

Tsotsos, J. K., Culhane, S., and Cutzu, F. (2001). "From theoretical foundations to a hierarchical circuit for selective attention," in *Visual Attention and Cortical Circuits*, eds J. Braun, C. Koch, and J. Davis (Cambridge, MA: MIT Press), 285–306.

Tsotsos, J. K., Culhane, S., Wai, W., Lai, Y., Davis, N., and Nuflo, F. (1995). Modeling visual attention via selective tuning. *Artif. Intell.* 78, 507–547. doi: 10.1016/0004-3702(95)00025-9

Tsotsos, J. K., Eckstein, M. P., and Landy, M. S. (2015). Computational models of visual attention. *Vision Res.* 116(Pt B):93. doi: 10.1016/j.visres.2015.09.007

Tsotsos, J. K., and Kruijne, W. (2014). Cognitive programs: software for attention's executive. *Front. Psychol.* 5:1260. doi: 10.3389/fpsyg.2014.01260

Tsotsos, J., Kotseruba, I., and Wloka, C. (2016). A focus on selection for fixation. *J. Eye Mov. Res.* 9, 1–34.

Tsotsos, J. K., Rodriguez-Sanchez, A., Rothenstein, A., and Simine, E. (2008). Different binding strategies for the different stages of visual recognition. *Brain Res.* 1225, 119–132. doi: 10.1016/j.brainres.2008.05.038

Tsotsos, J. K., and Rothenstein, A. (2011). Computational models of visual attention. *Scholarpedia* 6:6201. doi: 10.4249/scholarpedia.6201

Tsotsos, J. K., and Womelsdorf, T. (2016). "Visual tasks lead to unique sequences of cyclic attentional signals," in *F1000Research 2016*, 5:2467. Available online at: https://f1000research.com/posters/5-2467

Uhr, L. (1972). Layered "recognition cone" networks that preprocess, classify, and describe. *IEEE Trans. Comput.* 100, 758–768. doi: 10.1109/T-C.1972.223579

Uhr, L. (1975). *Recognition Cones that Perceive and Describe Scenes that Move and Change Over Time.* TR-235, Computer Sciences Department, University of Wisconsin-Madison.

van der Wal, G., and Burt, P. (1992). A VLSI pyramid chip for multiresolution image analysis. *Int. J. Comput. Vis.* 8, 177–190. doi: 10.1007/BF00055150

van Rooij, I. (2008). The tractable cognition thesis. *Cogn. Sci.* 32, 939–984. doi: 10.1080/03640210801897856

van Rooij, I., and Wareham, T. (2007). Parameterized complexity in cognitive modeling: foundations, applications and opportunities. *Comput. J.* 51, 385–404. doi: 10.1093/comjnl/bxm034

van Rooij, I., and Wareham, T. (2012). Intractability and approximation of optimization theories of cognition. *J. Math. Psychol.* 56, 232–247. doi: 10.1016/j.jmp.2012.05.002

van Rooij, I., Wright, C. D., and Wareham, T. (2012). Intractability and the use of heuristics in psychological explanations. *Synthese* 187, 471–487. doi: 10.1007/s11229-010-9847-7

Van Veen, B. D., and Buckley, K. M. (1988). Beamforming: a versatile approach to spatial filtering. *IEEE ASSP Magazine* 5, 4. doi: 10.1109/53.665

Wolfe, J. (1998). "Visual search," in *Attention*, ed H. Pashler (London: University College London), 13–74.

Yasuhara, A. (1971). *Recursive Function Theory and Logic*. New York, NY: Academic Press.

Ye, Y., and Tsotsos, J. K. (1996). "3D sensor planning: its formulation and complexity," in *Proceedings 4th International Symposium on Artificial Intelligence and Mathematics*, eds H. Kautz and B. Selman (Fort Lauderdale, FL).

Ye, Y., and Tsotsos, J. K. (2001). A complexity level analysis of the sensor planning task for object search. *Comput. Intell.* 17, 605–620. doi: 10.1111/0824-7935.00166

Zhang, Y., Meyers, E. M., Bichot, N. P., Serre, T., Poggio, T. A., and Desimone, R. (2011). Object decoding with attention in inferior temporal cortex. *Proc. Natl. Acad. Sci. U.S.A.* 108, 8850–8855. doi: 10.1073/pnas.1100999108

Zucker, S. W. (1981). "Computer vision and human perception: an essay on the discovery of constraints," in *Proceedings 7th International Conference on Artificial Intelligence,* eds P. Hayes and R. Schank (Vancouver, BC), 1102–1116.

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

frontiers in Psychology

ORIGINAL RESEARCH
published: 02 February 2017
doi: 10.3389/fpsyg.2017.00099

A Spiking Neuron Model of Word Associations for the Remote Associates Test

Ivana Kajić[1,2]*, *Jan Gosmann*[2], *Terrence C. Stewart*[2], *Thomas Wennekers*[1] and *Chris Eliasmith*[2]

[1] School of Computing, Electronics and Mathematics, University of Plymouth, Plymouth, UK, [2] Centre for Theoretical Neuroscience, University of Waterloo, Waterloo, ON, Canada

Generating associations is important for cognitive tasks including language acquisition and creative problem solving. It remains an open question how the brain represents and processes associations. The Remote Associates Test (RAT) is a task, originally used in creativity research, that is heavily dependent on generating associations in a search for the solutions to individual RAT problems. In this work we present a model that solves the test. Compared to earlier modeling work on the RAT, our hybrid (i.e., non-developmental) model is implemented in a spiking neural network by means of the Neural Engineering Framework (NEF), demonstrating that it is possible for spiking neurons to be organized to store the employed representations and to manipulate them. In particular, the model shows that distributed representations can support sophisticated linguistic processing. The model was validated on human behavioral data including the typical length of response sequences and similarity relationships in produced responses. These data suggest two cognitive processes that are involved in solving the RAT: one process generates potential responses and a second process filters the responses.

Keywords: semantic search, vector representations, semantic spaces, Neural Engineering Framework (NEF), spiking neurons, remote associates test

OPEN ACCESS

Edited by:
Tarek Richard Besold,
University of Bremen, Germany

Reviewed by:
Frank Van Der Velde,
University of Twente, Netherlands
Xavier Hinaut,
Inria Bordeaux - Sud-Ouest Research Centre, France

***Correspondence:**
Ivana Kajić
i2kajic@uwaterloo.ca

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *13 August 2016*
Accepted: *16 January 2017*
Published: *02 February 2017*

Citation:
Kajić I, Gosmann J, Stewart TC, Wennekers T and Eliasmith C (2017) A Spiking Neuron Model of Word Associations for the Remote Associates Test. Front. Psychol. 8:99. doi: 10.3389/fpsyg.2017.00099

1. INTRODUCTION

Language acquisition is highly dependent on the ability to create associations (Elman et al., 1997; Rogers and McClelland, 2004), as they are a central means of expanding both vocabulary and syntax (Brown and Berko, 1960; Hills, 2013). As well, associations allow infants to learn about previously unseen objects or concepts in terms of semantic similarities and semantic distinctions (Mandler and McDonough, 1993). While acquisition of language occurs in the earliest stages of human growth and development, starting with utterances of simple words and sentences, language skills continue to develop over the lifetime. Because associative mechanisms play such a crucial role in language and human cognition more generally, it is important to understand how the brain might represent, store, and deploy them.

The representation of linguistic content is an actively researched topic in various disciplines. For example, in Natural Language Processing (NLP) researchers work on optimal representations for the extraction of information from large corpora of text, as well as algorithms for text comprehension and production. Technology companies such as Facebook and Google are actively researching how to make machines better at understanding human language to improve their services and the efficiency of interactions between machines and humans

(Mikolov et al., 2013; Bordes et al., 2014; Weston et al., 2014; Hermann et al., 2015). One of the primary goals in NLP is to reach high performance on practical problems. Because this goal is generally adopted without regard for psychological or biological plausibility, it is unclear how such approaches can provide useful insights into how the brain solves the same problems.

Potentially more promising contributions to understanding the representation of lexical content and its meaning in the brain come from neuroimaging studies (Binder et al., 2009). Several studies have used fMRI data to construct semantic maps spanning broad areas of the cerebral cortex (Huth et al., 2012, 2016). Also, direct brain stimulation in the frontal cortex, left perisylvian cortex, and posterior temporal cortex of epileptic patients has identified regions essential for language production and comprehension (Ojemann et al., 1989).

While such studies provide us with a high-level perspective on possible brain regions involved in the processing of language, they do not shed light on how words and word associations might be represented by individual neurons and small networks. Improving our understanding of these lower-level mechanisms is a daunting task due to difficulty of locating, accessing, and recording from the brain regions responsible for such processes. In addition, direct recordings of neurons are invasive and can seldom be done on healthy humans.

Here, we opt to use a modeling approach to circumvent these problems while still gaining insight into representational structures and mechanisms that may potentially be used by the brain. We chose a linguistic task, the Remote Associates Test (RAT), to verify that the chosen representations of words and associations allow the model to perform in correspondence with human behavioral data. The model is hybrid insofar as it does not simulate the developmental process underlying the neural behavior. Rather, we use an analytical approach to derive the neural connectivity and then use simulated spiking neurons to produce the search process in the RAT. This makes it a non-developmental neural model, and we believe this is an important step toward the ultimate goal of having a complete neural account of the entire process that results in RAT behavior.

The choice of a particular neuron model also represents an important decision in the process of constructing the model. While there is a wide variety of neural models, we have chosen the leaky integrate-and-fire (LIF) neuron model due to its favorable trade-off between computational efficiency, analytical tractability, and its ability to capture some of the basic features of neuronal dynamics observed in biological systems (Izhikevich, 2007). In particular, synaptic dynamics and noise from fluctuations introduced by spiking impose constraints that a theoretical approach used to simulate neural systems needs to account for. The LIF neuron model is a spiking neuron model as it imitates the spiking behavior observed in biological neurons.

In biological neurons, electrically charged ions are exchanged across the cell membrane and an influx of positive ions into the cell can cause the neuron to trigger an action potential (also known as a spike). A spike can be registered by another, receiving neuron, if it has a synaptic connection with the neuron emitting a spike. In our modeling approach, spiking neurons are also connected by synapses so that the arrival of a spike at the side of a receiving neuron causes a post-synaptic current. The relevant neuron and synapse model parameters such as the membrane and synaptic time constants, and the shape of the post-synaptic currents conform to empirically measured value ranges and properties. These constraints are ensuring that the modeled system approximates the biological system and provides an account of the internal mechanisms underlying the investigated behavior.

1.1. The Remote Associates Test (RAT)

The RAT was developed in the early 1960s (Mednick, 1962) to study an individual's ability to think creatively. A creative thought or idea can often be described as novel and unusual (Boden, 2003). In the RAT subjects are presented with three cue words and have to find a solution word related to all cues within a time limit. An aspect of creativity is thought to be captured by subjects generating solution words that are only remotely associated with the problem cues, requiring subjects to relate familiar words in a novel way. For example, given a cue triplet *fish*, *mine*, and *rush*, thinking about common associations of each of the triplets such as *water*, *coal*, and *hour* is not helpful. Instead, *gold*, a less frequent associate of each of the words, is the correct solution as it can be meaningfully combined with each of the cues. The associative relationship between the cues and the solution in the RAT can vary: it can be a compound word such that each cue and the solution form a new word (e.g., *firefly*); it can be semantically related (e.g., *water* and *ice*); or it can form an expression (e.g., *mind game*). Mednick (1962) proposed that creative individuals are more likely to think of unstereotypical words that are solutions in the RAT. He attributed this to their flat associative hierarchy, in which the probability of coming up with an association is not very different for typical and untypical associations. In contrast, individuals scoring lower on the RAT would produce stereotypical associates with higher probability than untypical associates, which Mednick (1962) described as the characteristic of a steep associative hierarchy.

Performance on the test is expressed as the number of correctly solved items within a time limit, which is typically somewhere between a few seconds and a few minutes. Longer intervals correlate with higher solution rates (Bowden and Jung-Beeman, 2003), and it is assumed that longer solving periods allow for deliberate search processes, while shorter solving times are more likely to reflect sudden and involuntary insight solutions (Kounious and Beeman, 2014). Analyses of responses people give when attempting to solve a RAT problem have shown particular search patterns that differentiate search in the RAT from other related search processes (Raaijmakers and Shiffrin, 1981; Hills et al., 2012). Specifically, the RAT search process retrieves words that are strongly related to one of the three problem cues, shows occasional switching between the cues (Smith et al., 2013; Davelaar, 2015), and involves a local search strategy (Smith et al., 2013; Smith and Vul, 2015).

Performance on the RAT has been characterized by experimental, theoretical, and computational studies (Gupta et al., 2012; Kenett et al., 2014; Klein and Badia, 2015; Olteteanu and Falomir, 2015). Mednick's proposal about flat associative hierarchies of high-scoring individuals has been supported

experimentally by studies showing that indviduals who score higher on the RAT tend to avoid high-frequency answers on both incorrect and correct trials (Gupta et al., 2012; Kenett et al., 2014). This observation was further supported using NLP approaches that achieve better-than-human performance on the RAT (Klein and Badia, 2015; Olteteanu and Falomir, 2015). The properties of individual subjects' semantic networks correlates with their performance on the RAT (Kenett et al., 2014; Monaghan et al., 2014). Specifically, individuals who score high on a battery of creativity tests have semantic networks with small-world properties (Kenett et al., 2014). The connectivity in such networks is sparse, as they are characterized by short average path lengths between words, and strong local clustering. However, even though every node in the network is only sparsely connected, it takes just a few associations to reach any other node in the network. This kind of topology would assist in the solution of the RAT because quick, efficient searches can cover much of the semantic network.

1.2. Neural Representation

The question of word representation is central to all models concerned with linguistic tasks, including the RAT. Borrowing from the early theories of semantic memory in cognitive psychology (Collins and Quillian, 1969; Collins and Loftus, 1975), it is reasonable to approach the RAT by creating a semantic network where individual words are represented as nodes connected via edges indicating associations. Then, the process of finding the solution involves either a random or directed search in the network. Indeed, several models have used such representations to demonstrate performance on par with human performance (Bourgin et al., 2014; Monaghan et al., 2014; Kajić and Wennekers, 2015).

In terms of neurally plausible representations, these models would most closely correspond to the localist theory of representation (Bowers, 2009). Localist representations imply that a single neuron or a small group of neurons carries meaning. While this approach is often considered problematic in that it implies the existance of so-called "grandmother cells" (where there are particular neurons dedicated to representing the concept "grandmother"), some support for this type of representation can be seen in studies recording from single-cells which show high degrees of specificity in their response to external stimuli (Hubel and Wiesel, 1968; Moser et al., 2008; Quian Quiroga, 2012). In contrast to localist representations, distributed representations (McClelland and Rumelhart, 1987; Rogers and McClelland, 2004) are characterized by the assumption that a concept is represented by a population of neurons, where each individual neuron participates in the representation of multiple concepts. More recently, it has been argued that the kind of data used to support localist representations is often exhibited by distributed models (Stewart and Eliasmith, 2011; Eliasmith, 2013, p. 98–99, 369–370). Importantly, as will be described in more detail below, the method for distributed representation of concepts used in this paper suggests that each neuron within a distributed representation has a preferred state. This means that some neurons might be highly specific while others will have broad responses in our biologically informed distributed representation (Stewart et al., 2011a).

1.3. Modeling the Remote Associates Test

Despite arguments and evidence that distributed representations are used in many parts of the brain, there is no agreed upon approach to characterizing the representation of cognitive or linguistic structures using such representations. In particular, it is an open question of how such representations support word associations and how they might be employed in tasks requiring associative processing. We suggest answers to these questions by building a model that bridges from individual spiking neurons to the behavioral level and validating it on the RAT task.

To construct the model, we used the Neural Engineering Framework (NEF; Eliasmith and Anderson, 2003) described in the following section. It allows us to derive the required neural network to implement the necessary representations and transformations for performing the RAT. We describe the specific model in Section 2.3 and evaluation methods in Section 2.4. The quantitative and qualitative results are presented in Section 3, followed by a discussion and concluding remarks.

2. MATERIALS AND METHODS

The hybrid model presented in this paper was constructed with the methods of the NEF (Eliasmith and Anderson, 2003). The NEF specifies how a wide variety of functions can be implemented in biological neurons. It has been successfully used to model a diverse set of neural systems including those controlling behaviors such as eye position control, directed arm movements, and lamprey locomotion (Eliasmith and Anderson, 2003). It has also been successfully applied to the modeling of higher cognitive tasks such as serial working memory and action selection, and was the basis for the construction of the first detailed brain model capable of performing multiple tasks, called Spaun (Eliasmith et al., 2012). In this section we introduce the essentials of the NEF required to represent words with neural populations and to manipulate these representations. Using these basic methods, we describe the organization of a neural network to realize the cognitive processes in RAT memory search. We conclude by describing the semantic analysis methods used to validate the model.

2.1. Neural Engineering Framework (NEF)

We first describe how a group of neurons encodes a vector-valued stimulus $\boldsymbol{x}$, which lays the foundation for the representation of single words. Neurons have preferred stimuli, that is, they will respond more strongly to some stimuli than to other stimuli. For example, neurons in the striate cortex show selective responses to vertical bars of different orientations (Hubel and Wiesel, 1968) and neurons known as place cells in the hippocampus selectively exhibit specific firing patterns when an animal is present in a particular location in an environment (Moser et al., 2008). This stimulus preference can be expressed by assigning a preferred direction vector $\boldsymbol{e}_i$ to each neuron i. The inner product $\boldsymbol{e}_i^\top \boldsymbol{x}$ expresses how strongly a neuron will respond to a given stimulus; it increases as the stimulus vector aligns with the preferred

direction. This value can be thought of as being proportional to the amount of current flowing into a neuron, leading to the equation

$$a_i(t) = a_i\big(\boldsymbol{x}(t)\big) = G_i\Big[\alpha_i \boldsymbol{e}_i^\top \boldsymbol{x}(t) + J_i^{\text{bias}}\Big] \tag{1}$$

which gives the neuron activity $a_i(t)$ at time t for a time-dependent stimulus $\boldsymbol{x}(t)$. Here we convert the inner product into an input current to a neuron by means of a gain factor α_i and a bias current J_i^{bias}, used to capture observed neural responses also known as neural tuning curves. The spiking activity a_i of a neuron is given by applying a neuron non-linearity G_i to the input current.

While a wide variety of neuron non-linearities can be used with the NEF, here we use the LIF neuron model, which captures important properties related to neuronal excitability observed in biological neurons (Koch, 2004, Chapter 14). The incoming currents are accumulated as membrane voltage until a firing threshold is reached. At that point, the neuron emits a spike and the membrane voltage is reset to its resting value for a refractory period during which the neuron is unable to produce spikes. Without incoming currents, the membrane voltage will slowly decay to a resting potential due to leak currents. The left panel of **Figure 1** shows an example of how individual neurons in a set of seven LIF neurons respond to inputs in the range $x \in [0, 1]$. In this one-dimensional space, all preferred directions are either -1 or 1. For this example specifically, we assigned preferred directions of 1 to all neurons, as indicated by the increasing firing rate with increase of x. This captures the effect where stronger environmental stimuli (larger values of x) elicit stronger neural responses.

Given the firing in a group of neurons, how do we reconstruct the represented value $\boldsymbol{x}$? With LIF neurons, $a_i(t)$ is a spike train, i.e., $a_i(t)$ is 0 at all times t that no spike occurred and peaks at the spike times. However, biologically, each spike causes a post-synaptic current, which can be modeled as an exponential filter of the form $h(t) = \frac{1}{\tau}\exp(-t/\tau)$. This function can be combined with a linear decoding to provide a weighted linear filter that estimates the original vector $\boldsymbol{x}$. That is:

$$\hat{\boldsymbol{x}}(t) = \sum_i a_i(t) * \big[\boldsymbol{d}_i h(t)\big]. \tag{2}$$

The weights $\boldsymbol{d}_i$ are obtained by a global least-squares optimization of the error $E = \sum_k \|\boldsymbol{x}_k - \hat{\boldsymbol{x}}_k\|^2$ of the reconstructed stimulus across k sample points and all neurons in the group. The decoding process is visualized in the top right panel of **Figure 1**. The decoding weights scale the tuning curves (left panel) and the represented value is estimated with a sum over the scaled tuning curves.

Representing and reconstructing values is not sufficient for functionally interesting neural networks. Information needs to be transmitted and manipulated between groups of neurons. To do this, we need to find the synaptic connection weights that will perform this transformation. These can be computed from the decoding weights $\boldsymbol{d}_i$ of the pre-synaptic neurons that reconstruct an estimate of the represented value $\hat{\boldsymbol{x}}$. In addition, the input current to a post-synaptic neuron j depends on its preferred direction $\boldsymbol{e}_j$ and gain α_j. Because the quantities $\boldsymbol{d}_i$, $\boldsymbol{e}_j$, and α_j do

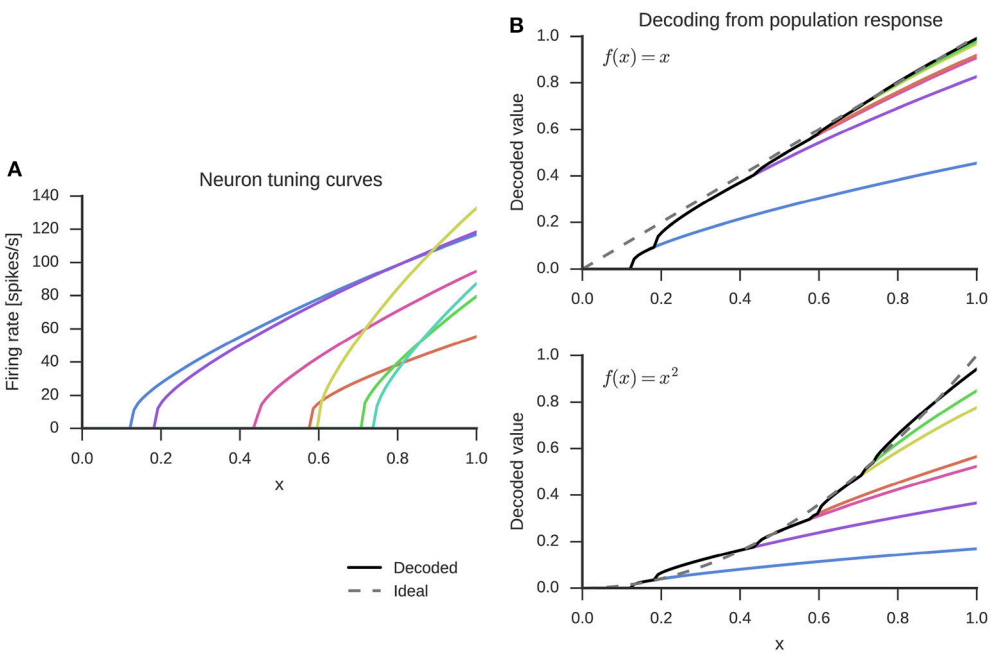


FIGURE 1 | Randomly generated tuning curves for seven neurons (left) and linear combination of these to decode the represented value x **(top right)** or decode a function, here x^2 **(bottom right)**. The dashed gray line is the ideal output and the black solid line the decoded value from all seven neurons.

not change over time[1], they can be multiplied together to provide standard neural network connection weights as follows

$$W_{ji} = \alpha_j \boldsymbol{e}_j^\top \boldsymbol{d}_i \tag{3}$$

where W_{ji} comprise the synaptic weight matrix $\boldsymbol{W}$. This is the optimal synaptic connection weight matrix for transmitting information from one neural group to another (Eliasmith and Anderson, 2003).

Finally, in addition to finding synaptic weight matrices that simply pass information from one group of neurons to the next, the NEF also allows us to find weight matrices that will compute functions $f(\boldsymbol{x})$ with neurons. This is done by using alternate decoding weights $\boldsymbol{d}_i^f$. Again, these can be determined from a least-squares optimization, but with a different error function $E^f = \sum_k \left\| f_k(\boldsymbol{x}) - \hat{f}_k(\boldsymbol{x}) \right\|^2$. The decoding with such alternative weights for the example of $f(x) = x^2$ is shown in the bottom right panel of **Figure 1**. The optimization for finding decoding weights is done separately for each function, but over all neurons within a group of neurons at once.

To summarize, the NEF allows us to state how a time-varying, vector-valued stimulus is encoded in neural populations, how the value represented in a neural population can be decoded, and how to connect neural populations to compute functions using those represented values. All connection weights are determined in an offline optimization without the need for an online process.

2.2. Representing Words and Associations with the NEF

To model the word search process in the RAT, words and associations among them need to be represented. We centrally adopt a representation where the activity of several neurons contributes to a representation of multiple words. In the NEF, this is achieved by using vectors to represent words, which we have elsewhere referred to as *Semantic Pointers* (Eliasmith, 2013)[2]. With the random distribution of preferred direction vectors $\boldsymbol{e}_i$, each neuron will be involved in the representation of multiple words and the representation is distributed across the neurons.

Representing words as vectors has a long tradition within NLP, for example Latent Semantic Analysis (Deerwester et al., 1990; Landauer and Dumais, 1997) and word2vec (Mikolov et al., 2013) are just two prominent approaches of generating word vectors from text corpora. Approaches like LSA and word2vec usually try to encode semantic structure or associations into the similarity or distance between the vectors. However, given two associated words A and B, this makes it difficult to decide which of these words is represented under the noisy conditions of plausible spiking neural representations. The word vector for A might become more similar to B than to A due to the noise. Thus, we avoid this kind of representation and use nearly orthogonal vectors. Specifically, we generate random unit-vectors with the constraint that no pair of such vectors exceeds a similarity of 0.1 as measured by the dot product. To fulfill the similarity constraint, a sufficient vector dimensionality has to be used. For the $N = 5018$ words used in the model, we set the dimensionality to $D = 2048$. This is considerably below the number of words because the number of almost orthogonal vectors, that can be fit into a vector space, grows exponentially with the number of dimensions (Wyner, 1967).

Such vector-based word representations have been successfully used to implement a variety of cognitive tasks such as the Tower of Hanoi task (Stewart and Eliasmith, 2011), inferential word categorization (Blouw et al., 2015), and Raven's Advanced Progressive Matrices (Rasmussen and Eliasmith, 2014). These representations have been shown in simulation to be robust to neural damage (Stewart et al., 2011a) and are consistent with the type of distributed representation found throughout sensory and motor cortex (Georgopoulos et al., 1986).

We next turn to the methods we used to compute the connection matrix between groups of neurons representing associations. These methods refer to algebraic operations and we do not consider them to be a part of the model. Instead, we use them to compute a matrix $\tilde{\boldsymbol{A}}$, which is implemented in connection weights among groups of neurons. The matrix $\tilde{\boldsymbol{A}}$ is used to describe associations between words and it transforms a word vector $\boldsymbol{w}$ to a linear combination of its associates. This matrix can be derived from an association matrix $\boldsymbol{A}$ where A_{ij} gives the associative strength from word i to word j. To do so we need to define the $N \times D$ matrix $\boldsymbol{V}$ that collects all the word vectors, i.e., row i of $\boldsymbol{V}$ is the vector representing word i. Then we can state $\tilde{\boldsymbol{A}} = \boldsymbol{V}^\top \boldsymbol{A}^\top \boldsymbol{V}$. Applied to a vector $\boldsymbol{w}$, this will first correlate $\boldsymbol{w}$ with all the word vectors ($\boldsymbol{Vw}$) to yield an N-dimensional vector indicating the similarity with each word; then $\boldsymbol{A}^\top$ is used to retrieve the corresponding associations before $\boldsymbol{V}^\top$ projects those associations back into a D-dimensional vector. As all of this collapses into a single linear transformation matrix $\tilde{\boldsymbol{A}}$, the retrieval of associations can be easily implemented with the NEF in the connection weights between two groups of neurons, computing the function $\boldsymbol{y} = \tilde{\boldsymbol{A}}\boldsymbol{x}$.

The model assumes that this set of connection weights is given. That is, we do not aim to explain the underlying developmental process, or speculate on whether particular mechanisms are innate or acquired developmentally. We would expect that the learning of associations and word representations occurs separately from the search process. Prior work (Bekolay et al., 2013; Voelker, 2015) has demonstrated the learning of NEF connection weights with spiking neurons in a biologically plausible manner, but we leave the investigation of these processes in this context to future work.

2.3. Model Description

We describe the core parts of the model most relevant to the RAT here, omitting implementational details not relevant to the main model function. A complete description can be found in the Supplementary Material. We used the Nengo neural network

[1]This assumes no synaptic weight changes, e.g., through learning, are happening. These could also be handled in a biologically realistic manner by the NEF (Bekolay et al., 2013), but are out of the scope of this paper.

[2]Semantic Pointers for words are more sophisticated than the representations used here, as they can encode structured representations as well. However, those structured representations are also vectors encoded as described here, so the present model is consistent with using semantic pointers for cognitive representation more generally.

simulator (Bekolay et al., 2014) for the implementation of the model. The model source code can be found at https://github.com/ctn-archive/kajic-frontiers2016.

All components of the model can be grouped into two main parts (see **Figure 2**):

- A *cue selection network* that randomly selects one of the three input cues as the primary cue. This selection is repeated in certain intervals to allow the primary cue to switch.
- A *response network* that selects an association as a response based on the current primary cue and previous responses.

While all three cues are being provided as input to the model, only one of them at a time will be regarded as the primary cue to generate associations. This is consistent with the way humans generate responses (Smith et al., 2013). To select a single cue each input cue is fed through a group of gating neurons that project to the neurons representing the primary cue. Inhibiting a set of gating neurons will prevent the transmission of the corresponding cue. To select a single cue, a simple winner-take-all (WTA) mechanism, seeded by white noise, is used. To get the required inhibitory gating signal, the WTA output has to be inverted. This is done with additional groups of neurons biased to represent 1. The WTA output will inhibit one of these groups to deactivate its inhibition of the gating neurons.

In the response network a single associated word is selected by a clean-up memory (Stewart et al., 2011b) with an added WTA mechanism. The input vector is correlated with the individual word vectors and each correlation value is represented by a group of neurons. In this way, the preferred direction vectors are not randomly distributed as in other parts of the model, but the preferred direction of every neuron and every group of neurons corresponds to one of the words. Furthermore, these groups of neurons threshold the represented value at 0.1. Each group is connected with every other group with lateral inhibitory connections, and to itself with a self-excitatory connection. This allows the group with the strongest input to remain active, while inhibiting all other groups. Another feedback connection is used to capture the evidence on the locality of search (Smith et al., 2013). This connection implements a transformation $\tilde{\boldsymbol{A}}$, so that the associates of the current response are fed as additional input to the WTA network. In **Figure 2**, all of these recurrent connections are denoted by a single feedback connection on the WTA in the response network.

The response inhibition plays a crucial role in allowing the next word to appear in the search process. It is realized as a neural group acting as a leaky integrator. Without external input, the represented vector will slowly decay to the zero vector. A recurrent connection feeding the output of the neural group back to itself prevents the otherwise very quick decay. External input to the integrator will slowly shift the represented value toward the input vector. This input is provided from the WTA network, while at the same time the response inhibition is used to inhibit the WTA network. Thus, the active word in the WTA network will be subject to increasing inhibition until finally a new word is selected. This switch will typically happen before the vector represented by the response inhibition shifted completely to the input vector. Because of that, the response inhibition will represent an additive mixture of the sequence of the last couple words and prevents those from reappearing in the search process in short succession.

The list of free model parameters and their values which produce the described model behavior is provided in **Table 1**. The values have been determined manually by observing which ranges produce the desired word-selection behavior.

To generate the association transformation matrix $\tilde{\boldsymbol{A}}$, we used the Free Association Norms dataset (FAN; Nelson et al., 2004). This dataset was constructed from a series of free association experiments conducted with over 6000 participants over the course of a few decades. In the experiments, participants were presented with a cue word and asked to write down the first word they thought of. In this way, a distribution of associates was created for every cue word by norming the frequency of response with the number of participants performing the task. The FAN data have been shown to provide a good match with human performance on the RAT when using it for solving RAT problems with a 2 s time limit (Kajić et al.,

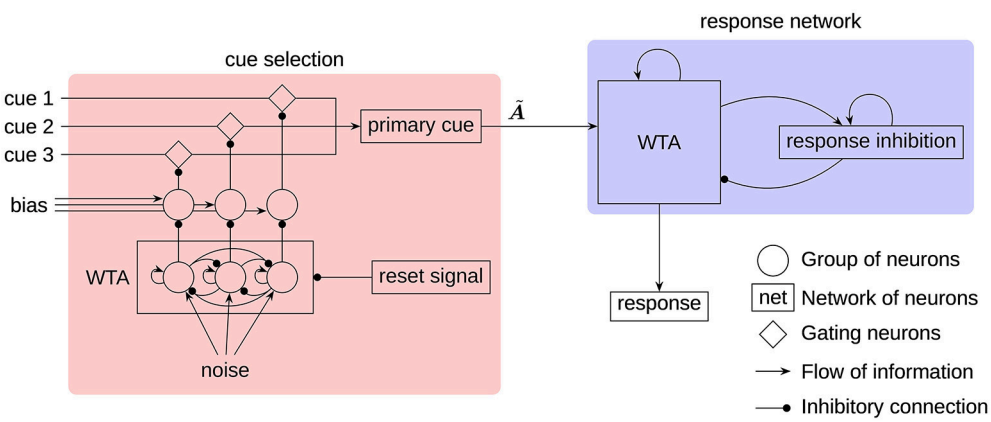


FIGURE 2 | The architecture of the RAT model. All neural groups, gating neurons, and networks consist of spiking neurons. The cues, noise, and bias are provided as external input to the model. The $\tilde{\boldsymbol{A}}$ label indicates the transformation, implemented on the connection to produce the associates of the primary cue.

TABLE 1 | Free parameters of the RAT model.

Parameter	Value	Description
d	2048	Number of dimensions per word vector
th	0.6–0.8	Percentage of randomly removed associations in the association matrix (varies with simulation)
assoc_th	0.05	WTA cut-off input threshold
cue_strength	0.1	Input strength of individual cues to WTA network
primary_cue_strength	0.7	Input strength of primary cue to WTA network
wta_feedback_strength	0.5	Input strength of associates of current response to WTA network
noise_std	0.01	Standard deviation of the zero-centered Gaussian noise in the cue selection network
integrator_feedback	0.95	Strength of recurrent connection on response inhibition

2016). Here, we use a binary association matrix $\boldsymbol{A}$ by assigning 1 to all non-zero association strengths and 0 otherwise. This disregards the associative strengths and only considers the (non-) existence of associative links. Preliminary simulations indicated that this approach gives a better match to the human data than weights proportional to the frequency of associations. To model individual differences in associative networks and adjust solution probabilities to match human data, we randomly remove between 60% and 80% of associations in the matrix by setting them to zero. This range has been determined empirically.

Not all potential responses produced by the response network qualify as a valid response to a RAT problem. Some words might be the result of an implicit priming effect, where a previous response primed a word which is not related to any of the cues. Also, it is reasonable to assume that participants in the experiment have typed only a subset of words that they thought of. To account for these effects, we implement a filtering procedure that regards only certain words as responses to a RAT problem. For every generated word, a similarity measure to the problem cues is calculated and, if it is below a threshold, the word is dismissed. The similarity is the sum of association strengths between every cue and the word.

Prior work (Kajić et al., 2016) allowed us to focus on a single source of association data for the generation of potential responses. However, we have no reason to assume that the same data is optimal for the filtering procedure. As such, we explore two association matrices and their binary variants for filtering purposes: FAN and the Google Books Ngram Viewer dataset (version 2 from July 2012, Michel et al., 2011, here referred to as Ngrams). We have previously shown both datasets to be suitable for modeling the RAT (Kajić et al., 2016). Although, the two sources of data contain similar information, there are interesting differences: approximately 6.5 million association pairs exist in the Ngram matrix and not in the FAN matrix. Conversely, only about 26,000 associations exist in the FAN but do not exist in the Ngram matrix.

Unlike the Ngram matrix, the FAN matrix contains non-reciprocal association strengths because associations are not bi-directional. Ninety-four percent participants in the free association experiment responded with the word *right* when given a cue *left*. However, the cue word *right* has much lower association to the word *left*, as only 41% participants responded with *left* and 39% participants responded with *wrong*. We used the sum of the FAN matrix with its transpose to obtain a symmetric association matrix.

While FAN data provides empirically derived association information through experiments with humans, a co-occurrence matrix for the Ngram data set has been derived by counting frequencies of n-grams across 5 million books published up to 2008. This is the second matrix we use. Here, we focus on 2-gram (bi-grams) only for words which exist in the FAN database. The Ngram matrix was constructed by iterating over all combinations of associative word pairs w_1 and w_2 and summing up occurrences of the 2-gram (w_1, w_2) and the 1-gram $w_1 w_2$ in the corpus.

Apart from using matrices that contain association strengths (for FAN) and co-occurrence frequencies (for Ngrams), we also explore whether just the existence of an association is sufficient to obtain the distribution of responses similar to the distribution of human responses. This is easily achieved by setting every non-zero entry in the matrix to one and gives the binary matrices bFAN and bNgram.

2.4. Model Evaluation

To evaluate the model, we use a set of 25 RAT problems and compare the model responses to the human responses from Smith et al. (2013). For each of the 25 problems, we ran 56 simulations with different random number seeds to ensure the independence of the results from the initial conditions, such as the choice of neurons and word vectors. For the analysis of responses, we adapt a set of analysis tools from Smith et al. (2013), which was originally developed to analyze human responses and characterize memory search in the RAT. The same analysis tools are used for human responses and model responses. While the experimental details about the data collection and detailed descriptions of analysis methods are available in the original publication, we present a brief overview of the data and a description of the adapted methods.

The data set contains responses from 56 participants, which were given 2 min to solve each RAT problem. Every participant was given 25 problems and was instructed to type every word which came to their mind, as they were solving the problem. Participants indicated when they thought they had provided the correct solution word with a key press. Thus, every trial consists of a sequence of responses from one participant to one RAT problem, ideally ending with the correct solution. Here, the analysis of responses has been performed over 1396 human trials and 1400 model trials. For each RAT problem, we ran 56 simulations, corresponding to the number of human participants. In 169 trials, human participants marked an incorrect response as correct and we excluded those from qualitative analyses, as they could have skewed analyses comparing how participants approached the final answer on incorrect and correct trials.

For every trial we did a series of pre-processing steps, as per Smith et al. (2013). Word pairs with words not available in the Free Norms or words identical to one of the cues were

excluded from the analysis. Responses repeated twice in a row were merged into a single response. Then, we assigned a 300-dimensional word vector to every word, including problem cues, the solution, and human responses. Those vectors were based on the Word Association Space (WAS; Steyvers et al., 2004), constructed by reducing the dimensionality of an association matrix. This matrix was the WAS $S^{(2)}$ measure based on the FAN, which includes not only direct association strengths between two words $\boldsymbol{w}_i$ and $\boldsymbol{w}_j$, but also links across one intermediary word, i.e., associations from $\boldsymbol{w}_i$ to $\boldsymbol{w}_k$ to $\boldsymbol{w}_j$. The similarity between words was measured as the cosine angle between the assigned word vectors. To conclude the pre-processing, every response was assigned the word vector with the highest similarity as the primary cue vector.

Metrics were calculated on the pre-processed data to evaluate the model. First, we determined the *average response similarity* for within and across cluster response pairs of adjacent responses. Clusters were defined on the primary cue of the responses; adjacent responses with the same primary cue are considered to be part of the same cluster. This was done to test for bunching of responses around cues by comparing the similarity between word pairs in each cluster. The assumption is validated with a *permutation test for average response similarity* by assigning cues from another trial and checking for conservation of similarity trends. The average response similarity within clusters is also computed in a cleaned data set, where all missing entries were dropped, which yielded new response pairs. Second, the *probability of switching primary cues* is computed as the number of response pairs with the different cues divided by the total number of response pairs. This value needs to be compared against a baseline probability based on the frequency each cue was selected under an independence assumption. This baseline calculation is required because certain cues might be selected more or less often than pure chance would predict. Third, the *similarity between adjacent and non-adjacent responses* within a cluster is computed to test for the direct influence of the previous response on the next one. The same is done for the responses with different primary cues, which occur right at the cluster breaks. Fourth, we tested whether the similarity to the final response increases as participants approach the final answer (either correct or incorrect).

3. RESULTS

In this section model responses are presented and compared to human responses using the methods described. Quantitative comparisons refer to the statistics of responses in terms of the number of correct solutions and the average number of responses for each RAT problem. The qualitative analysis addresses semantic properties of responses. Semantic analysis is based on the WAS space as described in Section 2.4. The aim of the qualitative analysis is to investigate whether response search trends, observed in human responses, match with those produced by the model. In particular, this refers to bunching of responses around problem cues, local search strategy, and clustering patterns.

3.1. Quantitative Comparison

The model solved on average 43% of the problems, showing a moderate correlation (Pearson correlation coefficient $r = 0.49$, $p < 0.05$) with humans who on average solved 42% problems. The left panel of **Figure 3** shows the accuracy on the 25 problems averaged, respectively, over all model simulations and over all human subjects. By applying the two-sided exact binomial test we find that for 14 out of 25 problems there is a statistically significant difference ($p < 0.05$) between the human and model responses[3]. These results are expected given that there are some problems which are easier for humans, and others that are easier for the model. On two problems—*dust, cereal, fish*; and *speak, money, street*—the model accuracy was more than 35 percentage points greater than the human accuracy on the same problems. On the other hand, there was one problem, *safety, cushion, point* where the human score was more than 35% points higher than the model score. However, **Table 2** indicates that, while the accuracy of this model matches well to the human performance, this model produces a much longer sequence of outputs than observed in the human subjects (40.20 vs. 7.78).

To deal with this discrepancy, we consider that there is some filter applied between the output of the model and the actual reported responses (in other words, the subjects do not actually write down all the words that come to mind while performing the task). As described in Section 2.3, this means that only a subset of all words produced by the model will be regarded as a set of responses to a RAT problem. In particular, a word that has a connection strength to all three cues below a threshold will be discarded. Thresholds have been determined as the lowest connection strength between the sets of three cues and solution for all problems. In this way, filtering will ensure that all solution words pass the filter. As a result, the accuracy and the correlation with the human accuracies are independent of the filtering method. **Table 2** summarizes the statistics for the raw data and various filtering methods. We compared the average number of responses per trial, the shortest and longest response sequence, and the match between distributions of the number of responses.

Overall, the Ngram matrix and the binary Ngram matrix yield distributions that best match to human data ($r = 0.95$ and $r = 0.93$, respectively). The threshold for the binary Ngram matrix has been set to 3, so that a word will pass the filter if it is an associate of all three problem cues. Reducing the threshold to two decreases the correlation with the distribution to $r = 0.36$ ($p < 0.05$) and increases the average number of responses per problem to 17.73. **Figure 4** displays the distributions for all filters plotted against the distribution of human responses. The Ngram derived matrices are more aggressive in filtering the responses compared to the FAN derived matrices. The former preserve $\approx$ 20% of words produced by the model, while the latter did so for $\approx$ 40% of the responses. Although, the Ngram matrix and the binary Ngram matrix yield comparably good matches

[3]It should be noted that if the number of problems is increased sufficiently, then there will always be a statistically significant difference for all conditions. For this reason, we take this test as a means of identifying problems where model responses deviate the most from human responses rather than as a measure of the quality of the model.

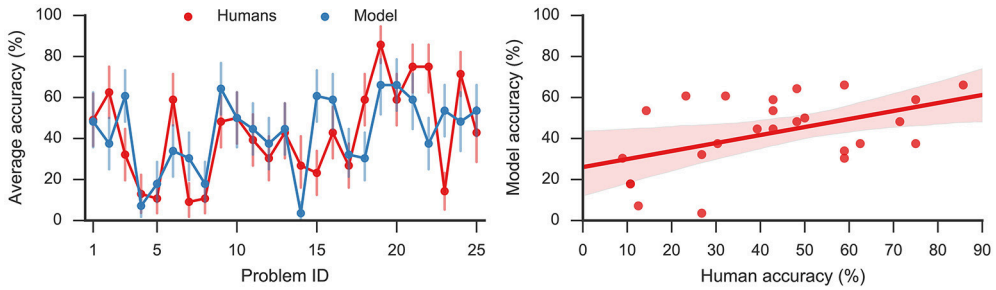


FIGURE 3 | (Left): Average accuracy on 25 RAT problems for model responses and human responses. Error bars denote 95% bootstrap confidence intervals. **(Right):** Linear regression (Pearson correlation coefficient $r = 0.49$, $p < 0.05$) with 95% bootstrap confidence intervals.

TABLE 2 | Quantitative analysis of raw and filtered model responses.

Analysis	Humans	Raw	Filtering method			
			FAN	bFAN	Ngram	bNgram
Filtering threshold			0.006	1	0.006	3
Shortest response sequence	1	2	1	1	1	1
Longest response sequence	49	46	39	40	27	33
Mean response sequence length	7.78	40.20	16.99	17.47	8.33	8.44
- Correlation with human data (*r*)		−0.30*	0.54***	0.51***	0.95***	0.93***

*Association matrices used for the filtering are: FAN, Free Association Norms; bFAN, binary FAN; Ngram; bNgram, binary Ngram. Values significant at $p < 0.05$ are marked with *, significant at $p < 0.001$ with ***.*

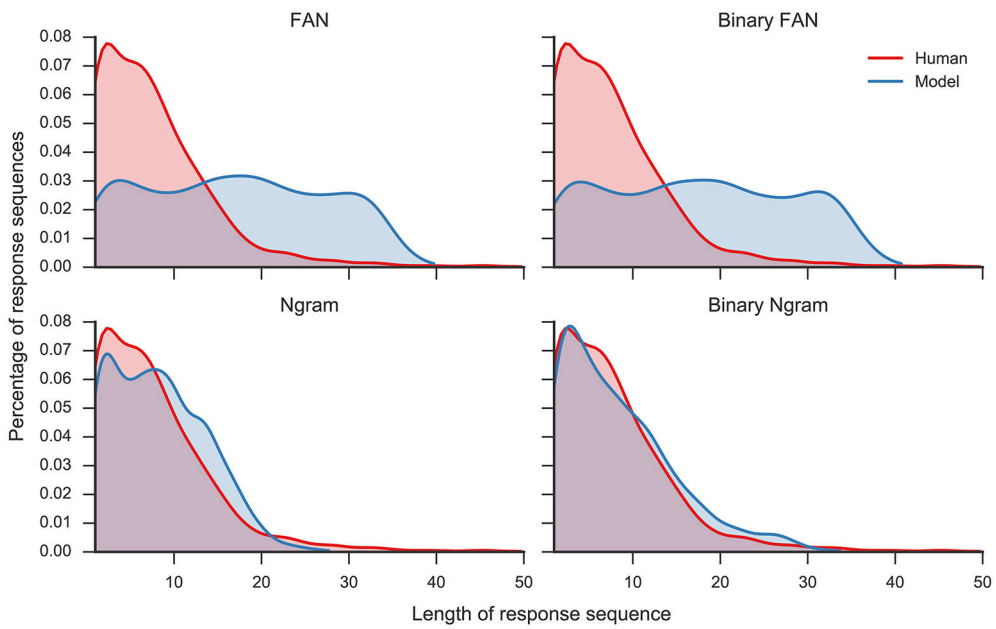


FIGURE 4 | Distribution of number of responses per trial for human responses and model responses plotted as Kernel Density Estimates (Gaussian kernel, bandwidth determined with Scott's rule, Scott, 1979). Four different model distributions were produced with four different filters (see text for details): Free Association Norms (FAN), binary FAN, Ngram, and binary Ngram.

with response distributions, in the following analyses we use the binary Ngram matrix which provided a slightly better match for some of the qualitative analyses.

To further investigate the effects of this filter, we tried applying it to the human data. Similarly to the model data it also reduced the dataset considerably, leaving <10% of overall responses.

3.2. Qualitative Comparison

We analyze responses obtained by applying the filter to raw model outputs in terms of their semantic similarity. The analysis compares similarity between two groups of response pairs, where groups refer to primary cue assignment of response pairs (same cue vs. different cue) and their proximity in a sequence of responses (adjacent vs. non-adjacent word pairs). Such analyses on human responses (Smith et al., 2013; Davelaar, 2015) showed that responses humans give tend to bunch around one of the three problem cues, and that different cues can be selected while search for the solution unfolds. Also, responses show sequential dependence, where the next response is dependent on the previous one. We use the set of analysis methods described in Section 2.4 to explore whether model responses exhibit such similarity patterns.

All analysis results are summarized in **Table 3**. To test for bunching of responses around problem cues, we explore the similarity of response pairs with a common primary cue. The similarity is greater for word pairs with the same cue compared to word pairs with different cues [0.141 vs. 0.054; two-sided t-test $t_{(9915)} = 20.4$]. This trend is preserved when we use the permutation test, which randomly assigns cues from a different trial [0.142 vs. 0.054; $t_{(4729)} = 13.7$]. Evidence for sequential dependence of word responses has been found by comparing similarities for word pairs within the same cluster; adjacent word pairs within the same cluster are more similar than pairs which are further apart [0.141 vs. 0.076; $t_{(13652)} = 17.8$]. Additional evidence for sequential search arises from greater similarity between adjacent word pairs with different primary cues compared to non-adjacent word pairs with different primary cues [0.054 vs. 0.011; $t_{(12,819)} = 22.4$]. We found that when the model produced a response, it produced another response with the same primary cue in 54.4% of cases. As done in the previous studies (Smith et al., 2013; Bourgin et al., 2014), we also analyzed the change in similarity between the final response (either correct or incorrect) and each one of the ten words prior to the final response. We identified a positive slope in similarity rates as responses were approaching the final answer.

TABLE 3 | Performance on the RAT and similarity patterns in the response search.

Analysis	Humans	Model
Average problem accuracy	42%	43%
-Correlation with human data (r)		0.49*
Shortest response sequence	1	1
Longest response sequence	49	33
Average number of responses per trial	7.78	8.44
-Correlation with human data (r)		0.93***
AVERAGE RESPONSE SIMILARITY		
-Within vs. across cue clusters	0.189 vs. 0.041	0.141 vs. 0.054
	CI: [0.134, 0.162]	CI: [0.079, 0.095]
-Permutation test	0.182 vs. 0.040	0.142 vs. 0.054
	CI: [0.124, 0.160]	CI: [0.077, 0.100]
-Within vs. across cue clusters (cleaned responses)	0.180 vs. 0.039	0.141 vs. 0.054
	CI: [0.128, 0.154]	CI: [0.079, 0.095]
Baseline vs. actual percentage of response pairs with the same primary cue (two-sided exact binomial test)	33.3 vs. 37.1%***	34.2 vs. 54.4%***
AVERAGE SIMILARITY BETWEEN ADJACENT AND NON-ADJACENT RESPONSES		
-With different primary cues (across cluster)	0.041 vs. 0.016	0.054 vs. 0.011
	CI: [0.063, 0.098]	CI: [0.038, 0.047]
-With same primary cues (within cluster)	0.189 vs. 0.108	0.141 vs. 0.076
	CI: [0.063, 0.098]	CI: [0.057, 0.072]

*Stated 95% confidence intervals are computed on the difference of reported mean values. Values significant at $p < 0.05$ are marked with *, significant at $p < 0.001$ with ***.*

3.3. Neural Outputs

We now turn to the neural responses generated by the model. Consequently, most observations in this section can be regarded as qualitative comparisons to spiking patterns observed in cortical neurons.

Figure 5 shows the spiking activity in three parts of the model during one simulation run. In the shown time frame, the primary cue starts as *widow*, but changes to *bite* about halfway through. This change is induced by the rising reset signal inhibiting the cue selection and causing a reselection of the primary cue. During the active period of either cue, the response neurons sequentially represent different words associated to the cue. Note, while four associations are shown for either cue, the number of responses generated during each active phase of a primary cue differs.

The spike raster plots (**Figure 5**) and firing rate estimates in **Figure 6** reveal interesting neuron tuning properties. We observe neurons that appear to be selective to cue words: some neurons only fire for *widow* (**Figure 6A**), while others only fire for *bite* (**Figure 6B**) in the shown time span. However, it is important to note that we did not test the response of these neurons to all possible cues and there might be other words which also elicit their response. Notwithstanding, such selective and explicit response behavior is consistent with observations from single-neuron recordings in medial temporal cortex in humans (Földiák, 2009; Quian Quiroga, 2012). We also observe neurons that fire for both cues, but with different firing rates. This word-dependent change in firing rate is more prominent for some neurons (**Figure 6C**), while it is more subtle for others (**Figure 6D**). The response population also includes neurons that are primarily active when a word is being represented, but not otherwise (**Figure 6E**).

From a single neuron perspective, of particular interest is the reset signal. Here, the neurons produce a clear bursting pattern during the onset of the reset signal. Such behavior is often thought to need an explanation in terms of complex neuron models that intrinsically burst (Izhikevich, 2007), which is not a charactersitic of LIF neurons. Nevertheless, we observe a bursting behavior because of the recurrent network dynamics producing the reset signal.

The presented neural network model, constrained by biological properties like membrane and synaptic time constants, shows a reasonable match to behavioral data. With that in mind, we believe that proposed mechanisms, such as word selection and word inhibition realized in spiking neurons, demonstrate the biological plausibility of this approach. Future work can address a

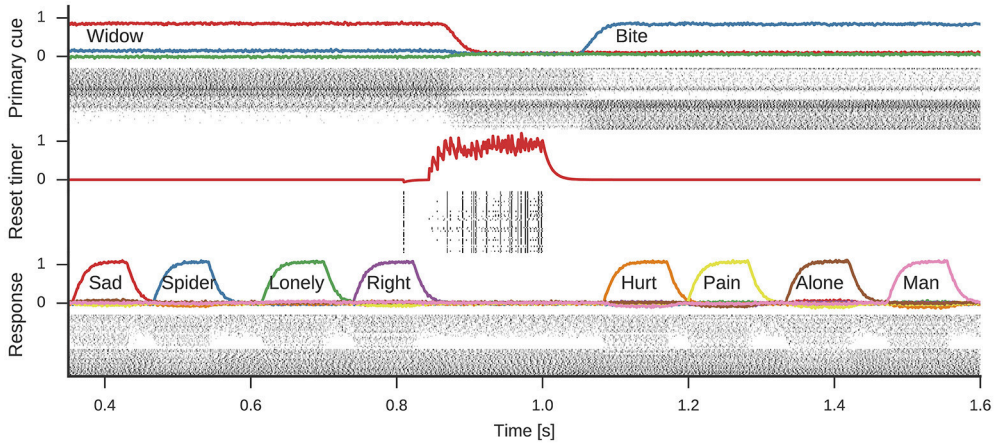


FIGURE 5 | Spikes and decoded values for three neural groups in the model. Data shown are an excerpt from a longer single simulation run. From top to bottom data for neurons representing the *primary cue*, the cue selection *reset signal*, and the *response* neurons are shown. Line plots for the primary cue and response show the similarity of the decoded vector with word vectors and are annotated with the corresponding words. The reset signal line plot shows the decoded scalar value. These line plots are interleaved with corresponding spike raster plots showing a subset of the neurons partaking in the representations.

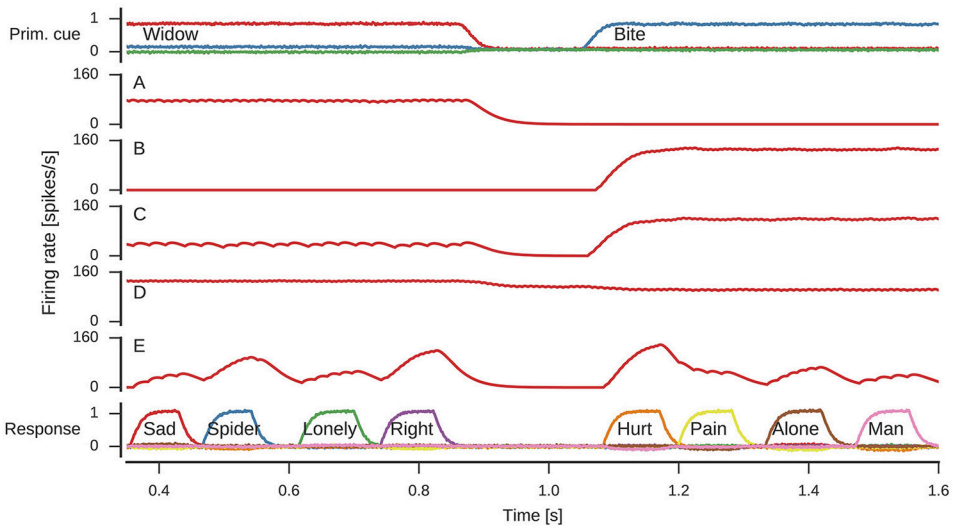


FIGURE 6 | Firing rates of individual neurons. Spike trains where filtered with $h(t) = [\alpha^2 t \exp(-\alpha t)]_+$ to obtain firing rate estimates. **(A)** Neuron responding to *widow*. **(B)** Neuron responding to *bite*. **(C)** Neuron responding to both *widow* and *bite* to a varying degree. **(D)** Neuron responding to both *widow* and *bite* with a more subtle difference. **(E)** Neuron responding to varying degrees whenever a response is produced.

stronger claim about the connection between measurable neural signals and the proposed mechanisms by using the model to generate fMRI predictions from the physiological properties of spiking neurons and their dendritic activity, as done in previous work (Eliasmith, 2013, Chapter 5).

4. DISCUSSION

We proposed a spiking neural network model that solves the Remote Associate Test, a task commonly used in creativity research. The model shows a significant correlation with human accuracy on the test, and its responses replicated similarity patterns observed in human responses (Smith et al., 2013; Davelaar, 2015). At the same time it implements possible biological mechanisms that generate these behavioral patterns, thus connecting multiple scales of the cognitive system.

The existing body of modeling studies have contributed to the general understanding of the RAT, including factors that influence the difficulty. Specifically, word frequency has been investigated as important in determining solvability of a RAT problem; an aspect that was already discussed when the test was developed (Mednick, 1962). Based on the frequency of a word or an expression in text corpora, it is possible to determine whether a problem will be easy or hard for humans (Gupta et al., 2012; Olteteanu and Falomir, 2015). Our model has reproduced the pattern of RAT item difficulty by showing a correlation with human accuracies on the 25 problems

from Smith et al. (2013). Individual differences in associative networks known to influence the performance on the test (Kenett et al., 2014) were modeled by randomly dropping a fraction of associations from the association matrix. Moving beyond the accuracy measure, we also looked at the quantitative and qualitative characteristics of response sequences. In terms of quantitative statistics, we analyzed the distribution of response sequence lengths which showed a strong correlation with the human data. We also observed a good match of the model and human data with respect to qualitative properties, such as bunching of responses around a single cue, cue switching, and sequential search. Such statistical similarity patterns were also successfully reproduced with probabilistic approaches in Bourgin et al. (2014), but without reference to cognitive processes underlying the search. Our model extends current findings by proposing biologically plausible network components involved in the search. In addition, we demonstrated how the representations in the model can display both specificity (commonly attributed to localist representation) and broad tuning (commonly attributed to distributed representation) depending on how single neuron activity is analyzed.

Previous studies identified the FAN as a viable source of associative data to model the RAT (Gupta et al., 2012; Bourgin et al., 2014; Kajić et al., 2016) and provided the motivation to use it in this model. Steyvers and Tenenbaum (2005) have shown that the FAN exhibits small-world properties. That means that its shortest paths between nodes are short on average, the clustering coefficients are high, and the connectivity is sparse. In our model, however, we removed associative links to model individual differences. It is left to future research to explore how this changes the properties of the associative network. For example, it might be possible that the small-world property gets disrupted leading to a lower performance on the RAT (Kenett et al., 2014).

Besides the generation of potential responses, we identified that it is important to filter out some of these responses to match human data. Interestingly, the Ngram data proved to be better suited for this task than the FAN. This leads to the hypothesis that humans use both sorts of information at different stages in the search process. But the cause could also be that most solutions, in the set of 25 problems, created compound words or word phrases with the cues, which is a property reflected to a larger degree in the co-occurrence data of the Ngrams. Nevertheless, it remains interesting that the Ngram data does not seem to be used for the generation of potential responses (Kajić et al., 2016).

While the current model offers a first unified account of the RAT search process in terms of both psychological and biological mechanisms, significant possible improvements remain for future work. First, switching of the primary cue is induced in quite regular intervals in our model. While we cannot exclude the possibility that this is the case in the actual cognitive process, we expect the actual process to be more complex. It would be interesting to explore how changing this part of the model can improve the match to human data, especially regarding the percentage of response pairs with the same primary cue. Second, the filtering of potential responses could be further investigated by exploring methods which discard less of the human and model responses, providing a closer match with the plausible cognitive mechanism. Furthermore, biologically plausible filtering with neurons should be implemented to extend the plausibility of the mechanisms of the complete model. While we have a proof-of-concept implementation of the filtering methods in spiking neurons, it is not yet complete. Third, current analysis methods filter out repeated responses, but these might give additional information on the search process and considering their occurrence patterns would allow us to refine the response inhibition network. Finally, the current model does not explain how humans learn word associations, or how the process of learning relates to changes in connection weights that store the relevant information. Since the acquisition of linguistic structure happens early in childhood and continues to develop throughout adulthood (Elman et al., 1997), a full account of word representation in the brain would also need to address learning at multiple time-scales, as well as mechanisms which enable such learning.

5. CONCLUSION

The RAT model proposed here specifies both cognitive processes and their neural implementation, which makes it unique among models of the RAT task. The model was validated on empirical data and shows a good match to this data. In the process of matching this data we identified that two processes might be at work: the generation of potential answers and the filtering of the answers to provide reported responses. Furthermore, the model sheds light on how the task relevant information can be represented in biologically realistic spiking neurons.

AUTHOR CONTRIBUTIONS

Conceived the model and experiments: IK, JG, TS, TW, and CE. Implemented the model: IK, JG. Analyzed the data: IK, JG. Wrote the paper: IK, JG, TS.

FUNDING

This work has been supported by the Marie Curie Initial Training Network FP7-PEOPLE-2013-ITN (CogNovo, grant number 604764), the Canada Research Chairs program, the NSERC Discovery grant 261453, Air Force Office of Scientific Research grant FA8655–13–1–3084, CFI, and OIT.

ACKNOWLEDGMENTS

The authors would like to thank Kevin Smith for sharing the human data and helpful advice.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2017.00099/full#supplementary-material

REFERENCES

Bekolay, T., Bergstra, J., Hunsberger, E., DeWolf, T., Stewart, T. C., Rasmussen, D., et al. (2014). Nengo: a Python tool for building large-scale functional brain models. *Front. Neuroinform.* 7:48. doi: 10.3389/fninf.2013.00048

Bekolay, T., Kolbeck, C., and Eliasmith, C. (2013). "Simultaneous unsupervised and supervised learning of cognitive functions in biologically plausible spiking neural networks," in *35th Annual Conference of the Cognitive Science Society* (Austin, TX), 169–174.

Binder, J. R., Desai, R. H., Graves, W. W., and Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cereb. Cortex* 19, 2767–2796. doi: 10.1093/cercor/bhp055

Blouw, P., Solodkin, E., Thagard, P., and Eliasmith, C. (2015). Concepts as semantic pointers: a framework and computational model. *Cogn. Sci.* 40, 1128–1162. doi: 10.1111/cogs.12265

Boden, M. A. (2003). *The Creative Mind: Myths and Mechanisms, 2nd Edn.* Routledge. New York, NY: Basic Books.

Bordes, A., Chopra, S., and Weston, J. (2014). Question answering with subgraph embeddings. *CoRR abs/1406.3676*, *arXiv preprint*. Available online at: http://arxiv.org/abs/1406.3676

Bourgin, D. D., Abbott, J. T., Griffiths, T. L., Smith, K. A., and Vul, E. (2014). "Empirical evidence for Markov Chain Monte Carlo in memory search," in *Proceedings of the 36th Annual Meeting of the Cognitive Science Society* (Austin, TX), 224–229.

Bowden, E. M., and Jung-Beeman, M. (2003). Normative data for 144 compound remote associate problems. *Behav. Res. Methods Instrum. Comput.* 35, 634–639. doi: 10.3758/BF03195543

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

Brown, R., and Berko, J. (1960). Word association and the acquisition of grammar. *Child Dev.* 31, 1–14.

Collins, A. M., and Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychol. Rev.* 82, 407 – 428.

Collins, A. M., and Quillian, M. R. (1969). Retrieval time from semantic memory. *J. Verb. Learn. Verb. Behav.* 8, 240–247.

Davelaar, E. J. (2015). Semantic Search in the Remote Associates Test. *Top. Cogn. Sci.* 7, 494–512. doi: 10.1111/tops.12146

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *J. Am. Soc. Inform. Sci.* 41, 391–407.

Eliasmith, C. (2013). *How to Build a Brain: A Neural Architecture for Biological Cognition*. New York, NY: Oxford University Press.

Eliasmith, C., and Anderson, C. H. (2003). *Neural Engineering: Computation, Representation, and Dynamics in Neurobiological Systems*. Cambridge, MA: MIT Press.

Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., and Rasmussen, D. (2012). A large-scale model of the functioning brain. *Science* 338, 1202–1205. doi: 10.1126/science.1225266

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., and Plunkett, K. (1997). *Rethinking Innateness: A Connectionist Perspective on Development*. Cambridge, MA: MIT Press.

Földiák, P. (2009). Neural coding: non-local but explicit and conceptual. *Curr. Biol.* 19, R904–R906. doi: 10.1016/j.cub.2009.08.020

Georgopoulos, A. P., Schwartz, A. B., and Kettner, R. E. (1986). Neuronal population coding of movement direction. *Science* 233, 1416–1419.

Gupta, N., Jang, Y., Mednick, S. C., and Huber, D. E. (2012). The road not taken: creative solutions require avoidance of high-frequency responses. *Psychol. Sci.* 23, 288–294. doi: 10.1177/0956797611429710

Hermann, K. M., Kociský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. (2015). "Teaching machines to read and comprehend," in *Advances in Neural Information Processing Systems*, eds C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (Montreal, QC: Curran Associates), 1693–1701.

Hills, T. (2013). The company that words keep: comparing the statistical structure of child- versus adult-directed language. *J. Child Lang.* 40, 586–604. doi: 10.1017/S0305000912000165

Hills, T. T., Todd, P. M., and Jones, M. N. (2012). Optimal foraging in semantic memory. *Psychol. Rev.* 119, 431–440. doi: 10.1037/a0027373

Hubel, D. H., and Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *J. Physiol.* 195, 215–243.

Huth, A., Nishimoto, S., Vu, A., and Gallant, J. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron* 76, 1210–1224. doi: 10.1016/j.neuron.2012.10.014

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., and Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* 532, 453–458. doi: 10.1038/nature17637

Izhikevich, E. M. (2007). *Dynamical Systems in Neuroscience*. Cambridge, MA: MIT Press.

Kajić, I., Gosmann, J., Stewart, T. C., Wennekers, T., and Eliasmith, C. (2016). "Towards a cognitively realistic representation of word associations," in *38th Annual Conference of the Cognitive Science Society* (Austin, TX: Cognitive Science Society), 2183–2188.

Kajić, I., and Wennekers, T. (2015). "Neural network model of semantic processing in the Remote Associates Test," in *Proceedings of the NIPS Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches Co-located with the 29th Annual Conference on Neural Information Processing Systems (NIPS 2015)* (Montreal, QC), 73–81.

Kenett, Y. N., Anaki, D., and Faust, M. (2014). Investigating the structure of semantic networks in low and high creative persons. *Front. Hum. Neurosci.* 8:407. doi: 10.3389/fnhum.2014.00407

Klein, A., and Badia, T. (2015). The usual and the unusual: solving Remote Associates Test tasks using simple statistical natural language processing based on language use. *J. Creat. Behav.* 49, 13–37. doi: 10.1002/jocb.57

Koch, C. (2004). *Biophysics of Computation: Information Processing in Single Neurons*. Computational Neuroscience Series. New York, NY: Oxford University Press.

Kounious, J., and Beeman, M. (2014). The cognitive neuroscience of insight. *Annu. Rev. Psychol.* 65, 71–93. doi: 10.1146/annurev-psych-010213-115154

Landauer, T. K., and Dumais, S. T. (1997). A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol. Rev.* 104, 211–240.

Mandler, J. M., and McDonough, L. (1993). Concept formation in infancy. *Cogn. Dev.* 8, 291–318.

McClelland, J. L., and Rumelhart, D. E. (1987). *Parallel Distributed Processing, Vol. 2*. Cambridge, MA: MIT Press.

Mednick, S. A. (1962). The associative basis of the creative process. *Psychol. Rev.* 69, 220–232.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., et al. (2011). Quantitative analysis of culture using millions of digitized books. *Science* 331, 176–182. doi: 10.1126/science.1199644

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems 26*, eds C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger (Lake Tahoe, NV: Curran Associates, Inc), 3111–3119.

Monaghan, P., Ormerod, T., and Sio, U. N. (2014). "Interactive activation networks for modelling problem solving," in *Computational Models of Cognitive Processes: Proceedings of the 13th Neural Computation and Psychology Workshop* (Singapore: World Scientific), 185–195.

Moser, E. I., Kropff, E., and Moser, M.-B. (2008). Place cells, grid cells, and the brain's spatial representation system. *Annu. Rev. Neurosci.* 31, 69–89. doi: 10.1146/annurev.neuro.31.061307.090723

Nelson, D. L., McEvoy, C. L., and Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behav. Res. Methods Instrum. Comput.* 36, 402–407. doi: 10.3758/BF03195588

Ojemann, G., Ojemann, J., Lettich, E., and Berger, M. (1989). Cortical language localization in left, dominant hemisphere: an electrical stimulation mapping investigation in 117 patients. *J. Neurosurg.* 71, 316–326.

Olteteanu, A.-M., and Falomir, Z. (2015). comRAT-C: a computational compound Remote Associates Test solver based on language data and

its comparison to human performance. *Pattern Recognit. Lett.* 67, 81–90. doi: 10.1016/j.patrec.2015.05.015

Quian Quiroga, R. (2012). Concept cells: the building blocks of declarative memory functions. *Nat. Rev. Neurosci.* 13, 587–597. doi: 10.1038/nrn3251

Raaijmakers, J. G., and Shiffrin, R. M. (1981). Search of associative memory. *Psychol. Rev.* 88, 93–134.

Rasmussen, D., and Eliasmith, C. (2014). A spiking neural model applied to the study of human performance and cognitive decline on Raven's Advanced Progressive Matrices. *Intelligence* 42, 53–82. doi: 10.1016/j.intell.2013.10.003

Rogers, T., and McClelland, J. (2004). *Semantic Cognition: A Parallel Distributed Processing Approach*. Cambridge, MA: MIT Press.

Scott, D. W. (1979). On optimal and data-based histograms. *Biometrika* 66, 605–610.

Smith, K. A., Huber, D. E., and Vul, E. (2013). Multiply-constrained semantic search in the Remote Associates Test. *Cognition* 128, 64–75. doi: 10.1016/j.cognition.2013.03.001

Smith, K. A., and Vul, E. (2015). The role of sequential dependence in creative semantic search. *Top. Cogn. Sci.* 7, 543–546. doi: 10.1111/tops.12152

Stewart, T. C., Bekolay, T., and Eliasmith, C. (2011a). Neural representations of compositional structures: representing and manipulating vector spaces with spiking neurons. *Connect. Sci.* 22, 145–153. doi: 10.1080/09540091.2011.571761

Stewart, T. C., and Eliasmith, C. (2011). "Neural cognitive modelling: a biologically constrained spiking neuron model of the Tower of Hanoi task," in *33rd Annual Conference of the Cognitive Science Society* (Austin, TX: Cognitive Science Society), 656–661.

Stewart, T. C., Tang, Y., and Eliasmith, C. (2011b). A biologically realistic cleanup memory: autoassociation in spiking neurons. *Cogn. Syst. Res.* 12, 84–92. doi: 10.1016/j.cogsys.2010.06.006

Steyvers, M., Shiffrin, R. M., and Nelson, D. L. (2004). "Word association spaces for predicting semantic similarity effects in episodic memory," in *Experimental Cognitive Psychology and its Applications: Festschrift in Honor of Lyle Bourne, Walter Kintsch, and Thomas Landauer*, ed A. F. Healy (Washington, DC: American Psychological Association), 237–249.

Steyvers, M., and Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. *Cogn. Sci.* 29, 41–78. doi: 10.1207/s15516709cog2901_3

Voelker, A. R. (2015). *A Solution to the Dynamics of the Prescribed Error Sensitivity Learning Rule*. Technical report, Centre for Theoretical Neuroscience, Waterloo, ON.

Weston, J., Chopra, S., and Bordes, A. (2014). Memory networks. *CoRR, abs/1410.3916, arXiv preprint*. Available online at: http://arxiv.org/abs/1410.3916

Wyner, A. D. (1967). Random packings and coverings of the unit n-sphere. *Bell Syst. Tech. J.* 46, 2111–2118.

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

frontiers in Psychology

ORIGINAL RESEARCH
published: 21 February 2017
doi: 10.3389/fpsyg.2017.00215

Semi-Supervised Learning of Cartesian Factors: A Top-Down Model of the Entorhinal Hippocampal Complex

András Lőrincz and András Sárkány*

Neural Information Processing Group, Faculty of Informatics, Eötvös Loránd University, Budapest, Hungary

The existence of place cells (PCs), grid cells (GCs), border cells (BCs), and head direction cells (HCs) as well as the dependencies between them have been enigmatic. We make an effort to explain their nature by introducing the concept of Cartesian Factors. These factors have specific properties: (i) they assume and complement each other, like direction and position and (ii) they have localized discrete representations *with* predictive attractors enabling implicit metric-like computations. In our model, HCs make the distributed and local representation of direction. Predictive attractor dynamics on that network forms the Cartesian Factor *"direction."* We embed these HCs *and* idiothetic visual information into a semi-supervised sparse autoencoding comparator structure that compresses its inputs and learns PCs, the distributed local and direction independent (allothetic) representation of the Cartesian Factor of global space. We use a supervised, information compressing predictive algorithm and form direction sensitive (oriented) GCs from the learned PCs by means of an attractor-like algorithm. Since the algorithm can continue the grid structure beyond the region of the PCs, i.e., beyond its learning domain, thus the GCs and the PCs *together* form our metric-like Cartesian Factors of space. We also stipulate that the same algorithm can produce BCs. Our algorithm applies (a) a bag representation that models the "what system" and (b) magnitude ordered place cell activities that model either the integrate-and-fire mechanism, or theta phase precession, or both. We relate the components of the algorithm to the entorhinal-hippocampal complex and to its working. The algorithm requires both spatial and lifetime sparsification that may gain support from the two-stage memory formation of this complex.

Keywords: Cartesian factors, entorhinal hippocampal complex, integrate-and-fire neurons, head direction cells, place cells, grid cells, border cells

OPEN ACCESS

Edited by:
Tarek Richard Besold, University of Bremen, Germany

Reviewed by:
Florian Röhrbein, Technische Universität München, Germany
Terrence C. Stewart, University of Waterloo, Canada

****Correspondence:***
András Lőrincz
lorincz@inf.elte.hu

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *31 July 2016*
Accepted: *03 February 2017*
Published: *21 February 2017*

Citation:
Lőrincz A and Sárkány A (2017) Semi-Supervised Learning of Cartesian Factors: A Top-Down Model of the Entorhinal Hippocampal Complex. Front. Psychol. 8:215. doi: 10.3389/fpsyg.2017.00215

1. INTRODUCTION

The fact that we are able to describe autobiographic events, can discover rules, in spite of the many dimensional inputs, such as the retina (millions of photoreceptors), the ear (cca. 15,000 inner plus outer hair cells), the large number of chemoreceptors as well as proprioceptive, mechanoreceptive, thermoceptive and nociceptive sensory receptors is puzzling, since the number of sensors enters the exponent of the size of the state space. This number is gigantic even if the basis of exponent is only 2, but it is typically much larger. How is it possible to remember for anything in such a huge space?

An illuminating and classical observation has been made by Kohonen (1982): the brain develops low dimensional representations, sometimes in the form of topographic maps manifested by retinotopy in the visual system, tonotopy in the auditory system, somatotopy in the somatosensory system, and so on. Kohonen considered these maps as some kind of *implicit* metric of the sensed space, being visual, auditory, or body related. The dimensionality of these maps is low, unlike the number of the sensors that give rise to those maps. Similarly low-dimensional representations of *the external space* appear in the entorhinal-hippocampal complex (EHC), although the topography is sometimes missing. The derivation of the abstract and low-dimensional representation of space from the actual and high dimensional sensory information is critical for goal oriented behavior as noted in the context of reinforcement learning, see, e.g., Kearns and Koller (1999), Boutilier et al. (2000), and Szita and Lőrincz (2009). In this context, one is directed to the EHC. The importance of this complex was discovered many years ago by Scoville and Milner (1957). Now, it is widely accepted that the EHC is responsible for episodic memory, see, e.g., Squire and Zola (1998) and Moscovitch et al. (2016) for an earlier review and for a recent one, respectively. In their paper, Buzsáki and Moser (2013) propose that (i) planning has evolved from navigation in the physical world, (ii) that navigation in real and mental space are fundamentally the same, and (iii) they underline the hypothesis that the EHC supports navigation and memory formation.

We believe that one of the functional tasks of this complex is the learning of low-dimensional Cartesian Factors that we define as follows. We say that (i) a low-dimensional representation discretizes a low dimensional variable, if discretization means that individual neurons [e.g., *place cells* (PCs) discovered more than 40 years ago (O'Keefe and Dostrovsky, 1971; O'Keefe and Nadel, 1978) represent local regions of their space (the so called place fields for PCs)] and thus the representation of the variable is distributed, (ii) the variable could be used as a coordinate in controlling and cognitive tasks, and (iii) an attractor network can predict by means of the local representation and, in turn, it can work as an implicit metric. As a further specification, we distinguish two factor types. Components of the first kind may exist even if other ones are not present, whereas components of Cartesian Factors do assume each other; no Cartesian Factor may exist without the others although many of them can be latent. We detail this below:

- Type I factors make no (or minor) assumptions about each other. Non-negative matrix factorization (NMF), for example, originates from chemistry: it is used in mass spectrometry and radiology among other fields, where absorbing or radiating components can sum up. In a given environment and for a given detector system, the observation of different isotopes depends on the environment and the detector, but they do not influence each other's spectrum except that—to a good approximation—they sum up. Another example is slow vs. faster or fast features (Franzius et al., 2007; Schönfeld and Wiskott, 2015). Such Type I factors are called features in most cases; they can be independent, one of them may not have to imply the presence of others. In other words, if one of the NMF or slow feature components is present, others can be missing.
- Type II factors assume each other. For example, texture, shape, weight, material components belong to the same object and any object possess all of these features. Some of them can be relevant when considering the value of a tool in a task. Another example is the information about the position of an object in space that can be given by the spatial coordinates and its pose. The speed of the object is another component, being necessary for the characterization of its state in certain tasks.

Latent Type II factors can serve cognition by decreasing the description and thus the state space. Keeping the example of the space, path planning requires the discretization of space and information about the neighboring relations of the PCs, i.e., the neighbor graph. Then an algorithm can find the shortest path on the graph. This path planning procedure doesn't require directional information; it works in a reduced dimensional space. We are concerned with such complementing and dimension reducing factors that may alleviate cognition in different ways in different tasks.

We assume that there is at least one Type II factor that can be sensed directly and this factor is represented in a topographic manner: it has some kind of (implicit) metric. This factor plays the role of a *semi-supervisor* in the learning of the complementing Type II factor(s). We also assume that the complementing factor is also low dimensional. Allothetic representation of the space is one example of such factors and it is the complementing factor of the allothetic representation of direction. *Head direction cells* (HCs) (see e.g., the work of Taube, 2007 and the references therein) make the discretized allothetic head direction representation. An attractor network can predict the activity pattern of the representation during rotation making it an (implicit) metric-like representation of direction. In turn, the set of HCs make a *Cartesian Factor*. We will consider how a metric-like representation may emerge from neurally plausible dynamics and the PC representation via predictive methods.

We note that according to Winter et al. (2015), in rodents, HCs are needed for the development of PCs, which are localized (i.e., discretized and distributed) allothetic representation of space; Type II factor according to our concepts.

There are neurons that respond along trigonal grids. These are the so called *grid cells* (GCs) (Fyhn et al., 2004; Moser et al., 2014). Results of Bonnevie et al. (2013) indicate that the presence of GCs is conditioned on both the presence of PCs *and* on the availability of HCs. For a recent review of the grid cells and the place cells see, e.g., the collection edited by Derdikman and Knierim (2014) as well as the references cited therein.

Our contributions are as follows.

1. We present a unified model of the EHC. We put forth the idea that this complex tries to solve the problem of nonlinear dimensionality reduction via Type II factors. These reduced dimensions function are like Cartesian coordinates if attractor networks enable them to form an implicit metric. Such Cartesian Factors can be reasoned with like symbolic variables. Consequently, we see the continuation of the grid as *learnable manipulation* at the symbolic level called *mind*

travel by Sanders et al. (2015). The grounding of the symbolic manipulation beyond the known domain seems as a necessity for acting according to Harnad (1990). A simple example is *homing behavior*; the transformation of goals in allothetic PCs to idiothetic action series.

2. The model is a learning model, which is capable of explaining (a) peculiar findings on the inter-dependencies of PCs and GCs, including (b) the corruptions that occur upon lesioning of different components and (c) the order of learning as described in the recent review paper of Rowland et al. (2016).
3. Direction sensitive GCs are developed from PCs and HCs by means of a predictive and compressing supervised algorithm working on *magnitude ordered neural activities*. We argue that either (a) integrate-and-fire characteristics or (b) theta phase precession can give rise to magnitude ordering in the time domain. We apply two simple linear algorithm on the ordered representation; we use *pseudoinverse computation* and *partial least squares* (PLS) regression. We show that PLS regression produces orientation sensitive, close to hexagonal grids in an incommensurate squared environment. We demonstrate that magnitude ordered predictive grid representation can be continued beyond the experienced environment.
4. We show that the predictive mechanism that gives rise to direction sensitive GCs can support the learning of Border Cells (BCs).
5. Our autoencoder model exploits sparsification and has the following constraints: we find that *lifetime sparsification*, i.e., sparsification over a larger number of inputs is necessary for efficient learning. Lifetime sparsification is not possible in real time, when individual input based sparsification, called *spatial sparsification* is needed. We propose that the two types of sparsification may be (one of) the underlying reason(s) of the two-stage memory formation in the EHC loop (Buzsáki, 1989).

Cartesian Factors have been introduced in two previous conference papers (Lőrincz et al., 2016; Lőrincz, 2016). The definition presented here is more precise and more elaborate: *Cartesian Factors complement each other and assume metric-like representations*. PCs have been developed in those publications and we review the results here. The extension of the model with orientation sensitive grid cells appears here for the first time alike to the proposal that magnitude ordered representation can serve the learning. Both integrate-and-fire behavior and theta phase precession are neurally plausible mechanism for magnitude ordering. In the first case, the spike representing the highest magnitude input comes first. In the second case, highest firing rates occur in the middle of the theta cycles. The combined model of direction sensitive GCs, PCs, and BCs is presented here for the first time.

In the following sections, we review background information and list some of the models of place cell and grid formation (Section 2). We describe the algorithmic components of our model in Section 3. More details of the algorithms are provided in the Appendix. The results section (Section 4) presents PC and directional sensitive GC representations. Results are discussed from the point of view of neuroscience in Section 5. We also consider symbolic representation, symbol manipulation and the symbol grounding problem in this section. We argue that all components—i.e., Cartesian Factors, place cell forming algorithms, oriented grid learning computational methods, and border cell formation—may fit the features of the EHC. Conclusions are drawn in Section 6.

2. BACKGROUND

2.1. Review of Related Findings in the EHC

The set of PCs, also called the *cognitive map*, the orientation independent representation of space, was discovered more than 40 years ago (O'Keefe and Dostrovsky, 1971; O'Keefe and Nadel, 1978). Since then we have learned many features of these cells, which are present in the CA3 and CA1 subfields of the hippocampus. Theta frequency oscillations (5–10 Hz) in the rodent hippocampal system create theta sequences: (i) place cells fire in temporal order, (ii) the sequences cover past, present and future, and (iii) time compression can be as much as a factor of 10 (Skaggs and McNaughton, 1996). Such temporal series centered on the present are the so called (theta) phase precession of PCs. The CA3 subfield has a recurrent collateral structure that, during sharp wave ripple (SPW-R, 140–200 Hz) complexes, replays temporal series experienced during exploratory behavior, when theta oscillations occur. Time series compression in SPW-R is around twenty fold and forty fold, before and after learning, respectively as shown by Lee and Wilson (2002). Memory trace formation seems to require to stages, the theta-concurrent exploratory activity and the population burst during SPW-R following the explorations (Buzsáki, 1989; Chrobak and Buzsáki, 1994) and according to the widely accepted view, the EHC formed memories include episodic ones (Moscovitch et al., 2016). The hippocampal formation is needed for dead reckoning (path integration) (Whishaw et al., 2001).

Grid cells have been found in the medial entorhinal cortex (MEC). It turns out that MEC lesion can abolish phase precession (Schlesiger et al., 2015; Wang et al., 2015), but the lesion only corrupts hippocampal place cells, it can't fully eliminate them (Hales et al., 2014). On the other hand, grid cells require hippocampal input (Bonnevie et al., 2013). The excellent review of Sanders et al. (2015) about place cells, grid cells, and phase precession includes a novel model about the two halves, i.e., about the past and the future. They claim that different mechanisms operate during the two halves.

Another important feature is that both the grid representation in the entorhinal cortex and the place cell representation of the hippocampus depend strongly on the vestibular information. There are indications put forth by Winter and Taube (2014) that head direction cells may not be critical for place cell formation since those can be controlled by environmental cues, like visual landmarks. However, it was shown by Winter et al. (2015) that the disruption of head direction cells can impair grid cell signals and are crucial for the formation of the allothetic representation including both place cells and grid cells . They also reported that theta waves are spared upon the same manipulation.

We shall argue that several findings follow from the constraints of developing the Cartesian Factor abstraction and the related metric-like representations.

2.2. Related Models

The number of place cell models is considerable, we list only a few of them. The interested reader is directed to the recent publication of Schultheiss et al. (2015) that reviews both mechanistic bottom-up models and top-down models.

Neural representation of trajectories traveled and the connectivity structure developed during such paths have been suggested as the method for place cell formation by Redish and Touretzky (1998). Incoming information includes external cues and internally generated signals. They are fused to develop place cells in the paper of Arleo and Gerstner (2000). Place cells were derived by Solstad et al. (2006) from linear combinations of entorhinal grid cells (Fyhn et al., 2004) and vice versa, neuronal level model can produce grid cell firing from place cell activities as shown by Burgess and O'Keefe (2011). Time plays the key role in the slow feature analysis model of place cells put forth by Franzius et al. (2007) and Schönfeld and Wiskott (2015). Time plays the opposite role in the independent component analysis based autoencoding place cell models (Lőrincz and Buzsáki, 2000; Lőrincz and Szirtes, 2009). In these works, time appears in a so called novelty detection (time differentiation) step.

We think that all of these models, i.e., navigation based models, models based on interaction between representations, models that search for components that change slowly in time, and models that consider novelty detection may have their merits in the development of low-dimensional representation of Cartesian Factors, since the development os such representations—as it has been mentioned earlier—are crucial for reinforcement learning of goal oriented behavior. For example, navigation in partially observed environments, like the Morris maze or when in dark, can be supported by temporal integration. As another example, novelty detection may support the separation of a rotating platform from remote, non-rotating cues studied by the Stuchlik group (Stuchlik and Bures, 2002; Stuchlik et al., 2013). Further, the relevance of learning of low-dimensional task oriented representations can't be underestimated since state space and thus learning time decreases tremendously if the dimension is decreased.

It seems straightforward to us that information both from the environment and from self-motion should be combined for an efficient and precise neural representation of self motion in the external space (Evans et al., 2016) and that different signals and latent variables can be advantageous under different conditions and may support each other. The case is similar to object recognition, when the different mechanisms, such as stereoscopic information, structure from motion, shape from shading, texture gradient, and occlusion contours among others work together in order to disambiguate the "blooming, buzzing confusion" of the visual information in different conditions, see, e.g., the work of Todd (2004) and the references in that paper.

Due to the critical nature of the vestibular input, our goal is to derive place cells under the assumption that only this component of the Cartesian representation, namely the egocentric direction relative to an allothetic coordinate system is available and we ask if the allothetic representation of space can be derived by using only (i) directional information and (ii) the egocentric, i.e., idiothetic visual information.

3. REVIEW OF THE ALGORITHMS

3.1. The Logic of the Algorithmic Components

The logic is as follows:

(i) We start with an autoencoding network and meet the comparator hypothesis of Vinogradova (2001).
(ii) Firing in the hippocampus is very sparse, see, e.g., the work of Quiroga et al. (2008), and we apply sparse models.
(iii) We find limitations and include lifetime sparsity beyond the spatial one. It is supported by the two-stage formation of memory traces.
(iv) We derive the dynamics of the grid structure by predicting in the simplest form: input–output pairs are formed by past and future experiences, respectively. The predicted values can be fed back, the input can be shifted by them and thus, prediction can be continued into the future. We compare linear models; the pseudoinverse computation and partial least square regression.
(v) Prediction concerns the actual firing pattern instead of the individual neurons that fire and components are ordered by their magnitudes: the largest magnitude signal makes the first component of the input and so on in decreasing order. This feature may appear naturally in integrate-and-fire mechanisms.
(vi) We assume view invariant observations of the objects. We use indices: a visible object activates an index. This is like the recognition of the presence of the object ("what") without the knowledge about its position ("where"). This "what" representation resembles to the so called "bag model" (Harris, 1954; Csurka et al., 2004).

Below, we elaborate on these algorithms and then we present our results.

3.2. Autoencoder

An autoencoder is the self-supervised version of the Multilayer Perceptron (MLP) and may have *deep* versions (Hinton and Salakhutdinov, 2006; Vincent et al., 2010). For the sake of general formulation, the deep version is described below although our numerical studies in this respect are limited.

Consider a series of non-linear mappings (layers) of the form:

$$\boldsymbol{H} = f_N\left(\cdots f_2\left(f_1(\boldsymbol{X}\boldsymbol{W}_1)\boldsymbol{W}_2\right)\cdots \boldsymbol{W}_N\right), \tag{1}$$

where $\boldsymbol{X} \in \mathbb{R}^{I\times J}$ is the matrix of I inputs of size J, $\boldsymbol{W}_n \in \mathbb{R}^{Q_{n-1},Q_n}$ are parameters with $Q_0 = J$, and f_n are non-linear almost everywhere differentiable element-wise functions ($n = 1,\ldots,N$; $N \in \mathbb{N}$). Then $\boldsymbol{H} \in \mathbb{R}^{I\times Q}$ is called the feature map ($Q_N = Q$). Typically, one takes two such mappings with reversed sizes—an encoder and a decoder—and composes them. Thereupon one can

define a so-called reconstruction error between the encoder input $\boldsymbol{X}$ and the decoder output $\widehat{\boldsymbol{X}} \in \mathbb{R}^{I\times J}$, normally the ℓ_2 or Frobenius norm of the difference, i.e.,

$$\frac{1}{2}\|\boldsymbol{X}-\widehat{\boldsymbol{X}}\|_F^2 = \frac{1}{2}\sum_{i=1,\ldots,I}\sum_{j=1,\ldots,J}(\boldsymbol{X}_{i,j}-\widehat{\boldsymbol{X}}_{i,j})^2$$

and try to find a local minimum of it in terms of parameters $\boldsymbol{W}_n$ after random initialization, by taking advantage of a step-size adaptive mini-batch subgradient descent method (Duchi et al., 2011; Zeiler, 2012; Kingma and Ba, 2014). The non-linearity can be chosen to be the rectified linear function $f_n(x) = x \cdot \mathbb{I}\ (x > 0)$ for $x \in \mathbb{R}$ (Nair and Hinton, 2010; Dahl et al., 2013) to avoid the vanishing gradient problem (Hochreiter, 1991; Hochreiter et al., 2001), where $\mathbb{I}$ designates the indicator function.

3.3. Spatial Sparsity and Lifetime Sparsity

Deep Autoencoders are often used as a pretraining scheme, see, e.g., the work of Erhan et al. (2010), or as a part of supervised algorithms as in the paper of Rasmus et al. (2015), but they lack the ability to learn a meaningful or simple data representation without prior knowledge (Sun et al., 2017). To obtain such a description, one might add regularizers or constraints to the objective function as did Grant and Boyd (2014) and Becker et al. (2011), or employ a greedy procedure like Tropp and Gilbert (2007) and Dai and Milenkovic (2009). It is well known that minimizing the sum of ℓ_2 norms of parameters $\boldsymbol{W}_n$ can reduce model complexity by yielding a dense feature map, and similarly, the ℓ_1 variant may result in a sparse version (Tibshirani, 1996; Ng, 2004).

An alternative possibility is to introduce constraints in the non-linear function f_n. For example, one may utilize a k-sparse representation by keeping solely the top k activation values in feature map $\boldsymbol{H}$, and letting the rest of the components zero as suggested by Makhzani and Frey (2013). This case, when features, i.e., the components of the representation, compete with each other is referred to as *spatial sparsity*.

Sparsification occurs on a different ground if indices of the representation on *many* inputs go up against each other. This case is called *lifetime sparsity*, see, e.g., the work of Makhzani and Frey (2015) and the references therein. Lifetime sparsification ensures that all indices may play a role, whereas spatial sparsification may render a large portion of the components of the representation quiet for all inputs. On the other hand, lifetime sparsification may not be used on any single input, the case needed for real time responses.

3.4. Predictive Partial Least Squares Regression

PLS regression started with the works of Kowalski et al. (1982) and Geladi and Kowalski (1986) back in the eighties. The PLS model assumes explanatory samples collected in matrix $\boldsymbol{R}$ made of t samples of l dimensions and a response matrix $\boldsymbol{Q}$ of m dimensions collected on the t observations. PLS combines features of principal component regression (PCR) and multiple linear regression (MLR): PCR finds maximum variance in $\boldsymbol{R}$, MLR is to maximize correlation between $\boldsymbol{R}$ and $\boldsymbol{Q}$. PLS regression tries to do both by maximizing covariance between them: first, it extracts a set of latent factors that explain the covariance between the explanatory and response variables and then the regression step predicts the values of the response variables.

In our case, explanatory variables and responses are connected by time: $\boldsymbol{R} = [\boldsymbol{r}(1), \ldots, \boldsymbol{r}(t)]$ and $\boldsymbol{Q} = \boldsymbol{R}(+) = [\boldsymbol{r}(2), \ldots, \boldsymbol{r}(t+1)]$ make the explanatory and the response variables, respectively. PLS regression takes the form

$$\boldsymbol{R} = \boldsymbol{T}\boldsymbol{P}^T + \boldsymbol{E} \tag{2}$$

$$\boldsymbol{R}(+) = \boldsymbol{U}\boldsymbol{Q}^T + \boldsymbol{F} \tag{3}$$

where $\boldsymbol{T}$ and $\boldsymbol{U}$ are matrices of dimensions $t \times n$, $\boldsymbol{P}$ and $\boldsymbol{Q}$ are the so called orthogonal loading matrices of dimensions $t \times n$ ($\boldsymbol{P}^T\boldsymbol{P} = \boldsymbol{Q}^T\boldsymbol{Q} = I$), and matrices $\boldsymbol{E}$ and $\boldsymbol{F}$ are the error terms drawn from independent and identically distributed random normal variables. It is also assumed that covariance between matrices $\boldsymbol{T}$ and $\boldsymbol{U}$ are maximal. In the computations, we used the Python package *sklearn* (Pedregosa et al., 2011).

3.5. Prediction via Pseudoinverse Computation

PLS regression is one option for predictions. Deep networks can be considerably more efficient. The simplest method, on the other hand, is possibly pseudoinverse computation that can be embedded into a Hebbian network structure as suggested by Lőrincz and Szirtes (2009) and in some of the references cited. Using the notations of the previous section, the pseudoinverse solution can be formulated as follows:

$$\boldsymbol{r}(\tau+1) = \boldsymbol{M}\left(\boldsymbol{r}(\tau)^T, \ldots, \boldsymbol{r}(\tau-t)^T\right)^T + \boldsymbol{e}(t) \tag{4}$$

where $\boldsymbol{e}(t)$ is the error term at time t. Equation (4) gives rise to the solution $\widehat{\boldsymbol{M}} \approx \boldsymbol{R}(+)\mathfrak{R}^+$ where $\mathfrak{R}^+$ denotes the Moore–Penrose right pseudoinverse of the matrix constructed from the matrix with the i^{th} column formed by $\left(\boldsymbol{r}(i)^T, \ldots; \boldsymbol{r}(i-t)^T\right)^T$ and $i > n$ is assumed.

3.5.1. Continued Prediction

For the pseudoinverse method, matrix $\widehat{\boldsymbol{M}}$ and the estimated predicted activities can be used for shifting the prediction further in time

$$\widehat{\boldsymbol{r}}(\tau+1) \approx \widehat{\boldsymbol{M}}\left(\boldsymbol{r}(\tau)^T, \ldots, \boldsymbol{r}(\tau-t)^T\right)^T \tag{5}$$

$$\widehat{\boldsymbol{r}}(\tau+2) \approx \widehat{\boldsymbol{M}}\left(\widehat{\boldsymbol{r}}(\tau+1)^T, \ldots, \boldsymbol{r}(\tau-t+1)^T\right)^T \tag{6}$$

and so on

and the case is similar for the PLS regression.

3.6. Magnitude Ordered Activities

PC activities themselves are bounded to the PCs themselves. This representation can't fulfill our purposes since PCs are locked to already observed bag representations and thus they are not able to support prediction outside of the explored field. As we shall see, sparse autoencoder on the bag representation produces

densely packed PCs that have high activities at the centers and lower activities off-center. In turn, between two place cell bumps there should be a hump and a metric-like representation can take advantage of this periodicity. If we order activities according to their magnitudes then largest activity will reach its (local) maximum at the center of a place cell, it will be smaller at other (neighboring) positions and will become large at another center. We will develop latent predictive factors of the magnitude ordered place cell activities. Indications that magnitude based ordering may be present in the neural substrate is elaborated in the discussion (Section 5.2).

3.7. The Bag Model

We assume a high level representation of the visual information that correspond to the so called bag model of machine learning. The Bag of Words representation, for example, represents a document by the words that occur in the document, without any syntactic information (Harris, 1954). Similarly, the Bag of Keypoints representation of an image (see e.g., Csurka et al., 2004 and the references therein) contains the visual descriptors of the image without any information about the position of those descriptors. Such representations are similar to the *what system* in visual processing as described first by Mishkin and Ungerleider (1982), elaborated later by Goodale and Milner (1992) and that may also be present in the representations of other modalities, see, e.g., the work of Schubotz et al. (2003).

Our inputs are represented by the indices of the objects present in the visual field. If the object is present, then the value at corresponding input component is set to 1. Otherwise, it is set to zero. This representation is independent from the position of the object within the visual field, being an invariant representation of the object, since the value of the representation does not change as a function of idiothetic direction and allothetic position as long as the object is within the visual field.

3.8. Algorithmic Formulation of Cartesian Factor Learning

We assume that a latent random variable Z (e.g., the discretized allothetic representation of the state, that is, the place cells) and an observed random variable Y (e.g., the head direction, that is, a compass) are continuous and together they can fully explain away—by means of saved memories—another observed binary random variable X (e.g., the egocentric view with pixel values either one or zero taken in the direction of the head, or the invariant bag representation with ones and zeros). The ranges of Z and Y are supposed to be discretized finite r- and one-dimensional intervals, respectively. For more details, see **Figure 1** and the Appendix.

3.9. Simulation environment and numerical details

3.9.1. The Arena

For our study, we generated a squared "arena" surrounded by $d = 150$ boxes (**Figure 1**). The "arena" had no obstacles. Boxes were placed pseudo-randomly: they did not overlap. The "arena" was discretized by an $M \times M = 36 \times 36$ grid. From each grid point and for every 20°, a 28° field of view was created (i.e., $L = \frac{360^\circ}{20^\circ} = 18$, overlap: 4° between regions), and the visibility—a binary value (0 for occlusion or out of the angle of view)—for each box was recorded, according to Equation (7); we constructed a total of $I = 37 \cdot 37 \cdot 18 = 24,642$ binary ($\boldsymbol{x}^{(m,l)}$) vectors.

3.9.2. Masks and Information on Closeness

These vectors were processed further. Beyond the actual viewing direction and viewing angle of 28°, we also input visual information in neighboring directions: we varied the non-zeroed (non-masked) part of the input from a single direction (28°) to all 18 directions (360°). Formally, for various experiments, we defined masks $\boldsymbol{V}_{i,\cdot}$ summing to $\nu = 1, 3, \ldots, 17, 18$, for which we carried out the concatenation method for each visible sectors separated by 20° degrees that we appended with all-zero vectors for the non-visible sectors (see, **Figure 2** below and Equation 8 in the Appendix).

3.9.3. Normalization and Lifetime Sparsity

In some experiments we normalized the inputs to unit ℓ_2 norm for each $d = 150$ dimensional components, provided that at least one of the components differed from zero and dropped the input if all the components were zeroes. This is the "normalized case." We used spatial sparsification with $k = 1$. We also used lifetime sparsification. The dimension Q of the feature map of the autoencoder was set to 30 and we used probabilities of $p = \frac{100}{Q}\% = 3.33\%$ and $p = 6.66\%$. The $p = 3.33\%$ means that any component was active once on the average in the sample, but either none of them, one of them, or more than one of them may have assumed non-zero values for an individual input. The all zero case was dropped and thus the average probability was somewhat higher than $p = 3.33\%$. The ratio of dropped inputs was smaller for probability $p = 6.66\%$.

Concerning the error of the autoencoder we had two options: (a) error of the full output and (b) error only on the visible components that belonged to the viewing angle as in Equation (9). This latter is called masked experiment. We experimented with 3 and 5 layer autoencoders, with the middle layer representing the latent variables. For the 5 layer case, the sizes of the hidden layers were spaced linearly between 2700 and 30 giving rise to layers of dimensions 2,700, 1,335, 30, 1,335, 2,700 from input to output, respectively.

3.9.4. Magnitude Ordering

For each point in the arena we ordered the activity vector's components according to their magnitudes, with the largest being the first. Although the dimension of the representation remains, the individual indices of the place cells disappear: one doesn't know, which place cell has the largest activity, which one is the second largest, and so on. Nonetheless the largest activity will change along straight paths since between two place cell dumps there is always a hump. The oscillation is the basis of learning. Magnitude ordered activities along straight paths may provide information about displacements along the path, since the differences of the magnitudes change. Exceptions correspond to different positions that have the same set of activity

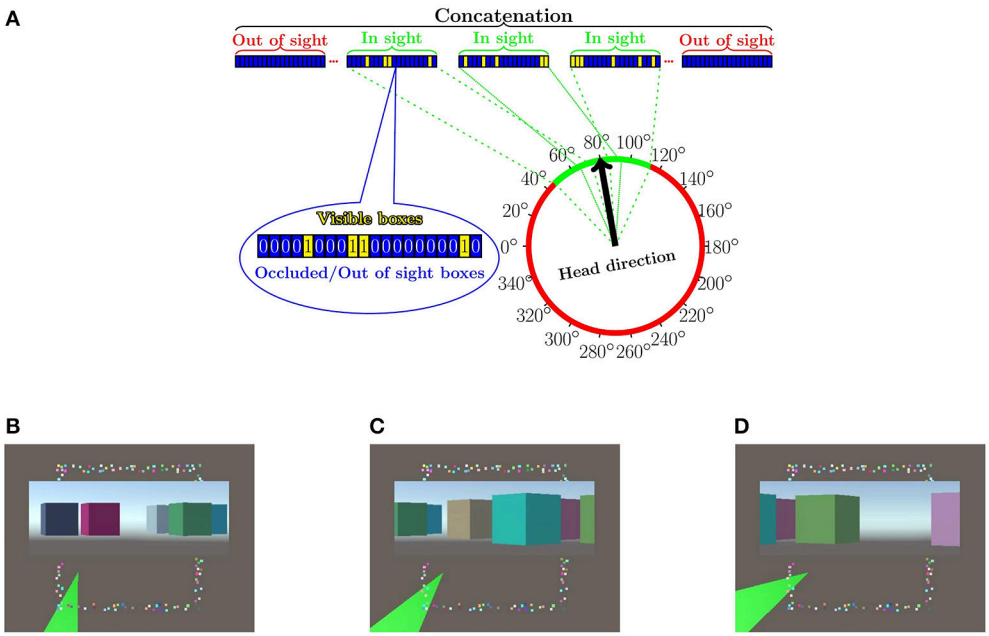


FIGURE 1 | Arrangements of the numerical experiments. (A) Input is concatenated from sub-vectors, which belong to different allothetic directions. A given index corresponds to the same box, the "remote visible cue," in all sub-vectors. The value of the a component of a sub-vector is 1 (0) if the box is visible (non-visible) in the corresponding direction (cf. bag representation, for more details, see text). More than one direction can be visible. The figure shows the case of three visible directions depicted by green color. Some boxes may be present in more than one visible direction, since they are large. **(B–D)** The "arena" from above with the different boxes around it plus some insets. Shaded green areas in **(B–D)**, show the visible portions within the field of view at a given position with a given head direction. Insets show the visual information for each portion to be transformed to 1 s and 0 s in the respective components of the sub-vectors. Components of out-of-view sub-vectors are set to zero. (Lőrincz et al., 2016 with permission).

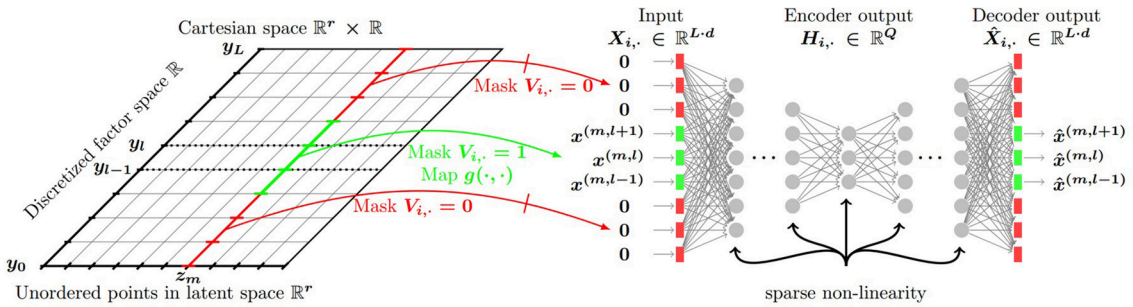


FIGURE 2 | General architecture. In the numerical experiments the notations correspond to the following quantities: Z latent positions, Y discretized "compass" values. Non-visible part of the input to the network is denoted by red, visible part is denoted by green. Visible part consists of 28° viewing angle in the actual direction and 28° viewing angle in neighboring directions separated by 20°. The number of neighbors was set to 2, 4, 16, 17 with 17 directions and the actual direction covering the whole 360°. For each viewing angle inputs represent boxes visible within the corresponding range. Values of the vector components representing the boxes are set to 1 if the range corresponds to the actual direction or if belongs to the set of neighbors. The full size of the input equals to the "No. of boxes × No. of viewing angle ranges." (Lőrincz et al., 2016 with permission).

magnitudes, which may occur for regular lattices and along lattice translation vectors.

3.9.5. Prediction along Straight Paths

We performed the prediction experiments on a place cell activity model trained by the autoencoder with a specific set of parameters: we used $p = 6.66\%$ lifetime sparsity with normalized input and masked loss function with a 220° viewing angle. The model was trained for 100 epochs. We discretized the arena to a 150×150 grid and collected place cell activities using the model from each of the $151 \times 151 = 22801$ grid points for all 18 directions.

We collected data in each direction separately. Distance between the steps equals the grid step distance of the discretized arena. In the learning phase we used $n = 60, 80, 100$ samples of the $m(= 30)$ magnitude ordered place cell activities from a n step

length straight path as inputs. For each step the sample of the closest grid point was taken. The m dimensional data sample of the $(n+1)^{st}$ step along the same path was used as supervisory predictive information. All sample paths where the necessary $n+1$ steps doesn't lead out from the arena were used during training.

With this method we can estimate the representation beyond the arena from an initial series of samples by using the predicted estimation for shifting the n consecutive samples and dropping the last one. The short distances between the steps aim to imitate gamma-wave sampling.

The software used in these studies can be downloaded from GitHub[1].

4. RESULTS

First, we review our recent results on place cells derived in Lőrincz et al. (2016) and in Lőrincz (2016) for the sake of argumentation and clarity. These results are reproduced in **Figures 3**, **4**, and in **Table 1**. Then we derive new features related to the place cells. This subsection is followed by the description of our new results on oriented grid cells. They, together, form the Cartesian Factor.

We note that uniformly distributed inputs and sparsification favors similarly sized sets of the input space, since latent units are competing for responses as we shall discuss it later. Competition gives rise to close packing. In 2D, the locally closest packing is the hexagonal structure and this arrangement is commensurate with the 2D space, so locally close packing can be continued and gives rise to a regular global structure, the triangular lattice. Our arena is, however, a square structure and has 90° symmetry, which is incommensurate with the hexagonal structure. In turn, we expect a close to hexagonal PC structure with reasonable amount of structural errors. Notably, self-supervised predictive compression gives rise to grids and emerging grids show improved hexagonal symmetry and tend to correct the errors of the place cells. Note that the larger the arena, the smaller the effect of the boundary is.

4.1. Cartesian Abstraction Yields Place Cells

The dependencies of the responses in the hidden representation vs. space and direction are shown in **Figures 3**, **4**, respectively. Linear responses of randomly selected latent units for different algorithms are depicted in **Figure 3**, illustrating the extent that the responses became localized even in the absence of competition after learning.

Figure 4 shows the direction (in)dependence of the responses. This figure has a special coding method: for each position and for each direction we computed the responses of all 30 neurons of the middle layer of the autoencoder and chose the one with the highest activity. In the ideal case a single neuron wins in all directions at a given position. Therefore, for each position we selected the neuron which won in the most directions (out of the 18) and assigned the number of its winnings to that position. Then we colored each position within the arena with a color between white, when the number is zero, i.e., none of the neurons is responding in any of the directions, and black, when the number is 18, i.e., the winner is the same neuron in all directions. Middle values between 0 and 18 are colored from light yellow to dark red in increasing order. **Figure 4** depicts results for different masks. The first column from the left is the case when only a single direction is not masked. Other columns from left to right correspond to cases when 3, 5, ... 18 directions are not masked.

One should ask (i) if the linear responses are local and activities far from the position of the peak activity are close to zero; (ii) if the number of dead latent units is small, (iii) if responses are direction independent, that is, if we could derive the discretization of space in allothetic coordinates. We found that spatial sparsity with the 3 layer network rendered the output of some or sometimes all hidden units to zero (**Table 1**). The same happened for the 5 layer network with dense $2nd$ and $4th$ layers and sparse $3rd$ layer. On the other hand, lifetime sparsity $p = 3.33\%$ with the 5 layer network produced excellent results. Lifetime sparsity $p = 6.66\%$ also produce high quality PCs. **Figure 4** shows that including the mask, direction-invariant activations start to develop at around about 100° (see the second and the third lines), whereas without the mask, similar activations appear at around 230°. For the sake of comparison, we also provide the ICA responses in **Figure 3**.

4.2. Place Cells Assume Close to Hexagonal Structure

Competition, as it was mentioned above, gives rise to hexagonal close packing in two dimensions, that is in a triangular lattice structure. In our experiments the symmetry is frustrated by the squared boundary of the "arena." The Delaunay triangulation of **Figure 5A** shows a number of distorted hexagons, some heptagons, pentagons and—closer to the edges of the "arena"—a few quadrilaterals, too. The more dark red the color, the smaller is the winning domain of the neuron. Sizes are more similar and shapes are more circle-like in the internal part of the "arena," whereas they are more distorted around the edges and at the corners. The size of the PCs are similar or larger at around the edges and the corners (**Figure 5C**). The paper written by Muller et al. (2002) reviews the different variables of sensory information that affect the sizes and the densities of PCs. We note that in the experiments, the bags are almost empty at the edges (in 180°) and in the corners (in 270°).

4.3. Predictive Methods Can Form Grid Cells from Place Cells

We use pseudoinverse and PLS regression methods to predict the next activity based on a series of previous ones. These methods work on magnitude ordered series and thus they are not associated with individual place cells. Magnitude ordered activities show oscillations along straight paths as shown in **Figure 6**. Such behavior suits prediction.

We show results for this two linear methods below.

[1]https://github.com/asarkany/ehcmodel

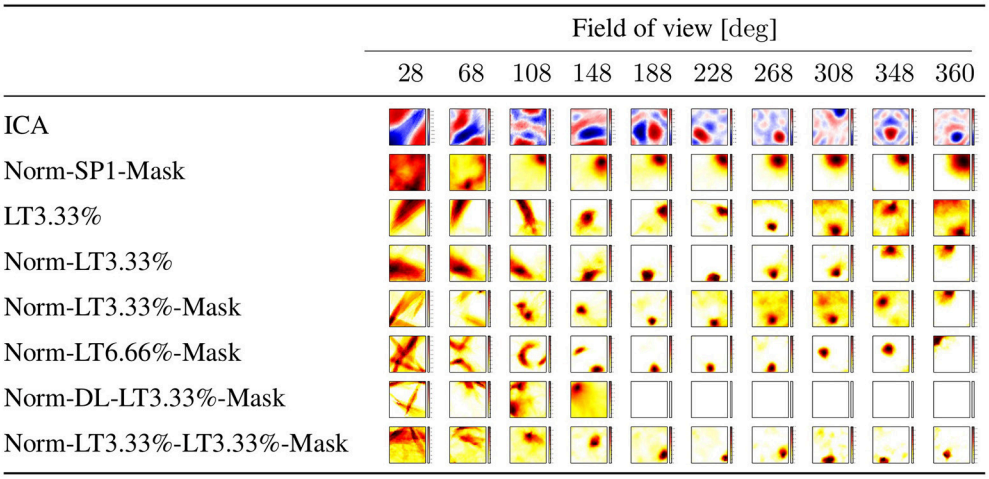


FIGURE 3 | Linear responses of individual latent units selected randomly: we chose neuron with index 2 from the latent layer. ICA: values may take positive and negative values. Other experiments: all units are ReLUs, except the output, which is linear. Color coding represents the sum of responses for all directions at a given point. SP1: spatial sparsity with $k = 1$, LT3.3%: lifetime sparsity = 3.3%, Norm: for each 150 components, the ℓ_2 norm of input is 1 if any of the components is non-zero, Mask: autoencoding error concerns only the visible part of the scene (i.e., the non-masked part of the input) DL: dense layer. "Norm-LT3.3%-LT3.3%-Mask" means normed input, masked error, 5 layers; the input layer, 3 layers with LT sparsity of 3.3% and the output layer. Columns correspond to masks of different angular extents separated by 20° and covering viewing angle of 28°, i.e., they overlap. Left column: a single viewing angle is non-masked. Other columns correspond to 3, 5, ... 17, 18 non-masked directions in increasing order to the right. (Lőrincz et al., 2016 with permission).

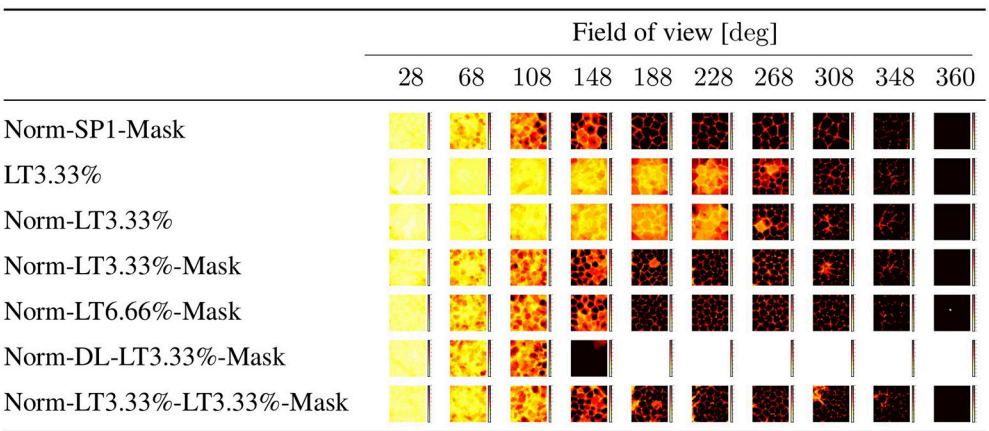


FIGURE 4 | Angle independence. Notations are the same as in **Figure 3**. The highest activity (winning) unit was selected for each input at each position in each direction. We counted the number of wins at each position for each unit and selected the largest number. Results are color coded. Black (18): there is a single winner for all angles at that position. White (0): no response at that point from any neuron in any direction. Values between 1 and 17: the darker the color the larger the direction independence for the best winner at that position. Rows represent different algorithmic components. SP1: spatial sparsity with $k = 1$, LT3.3%: lifetime sparsity = 3.3%, Norm: for each 150 components, the ℓ_2 norm of input is 1 if any of the components is non-zero, Mask: autoencoding error concerns only the visible part of the scene (i.e., the non-masked part of the input), DL: dense layer. "Norm-LT3.3%-LT3.3%-Mask" means normed input, masked error, 5 layers; the input layer, 3 layers with LT sparsity of 3.3% and the output layer. Columns correspond to masks of different angular extents separated by 20° and covering viewing angle of 28°, i.e., they overlap. Left column: a single viewing angle is non-masked. Other columns correspond to 3, 5, ... 17, 18 non-masked directions in increasing order to the right. (Lőrincz et al., 2016 with permission).

4.3.1. Prediction Outside of the "Arena"

Figures 7, **8** depict the results for the pseudoinverse method and for PLS regression, respectively

PLS regression is a better predictor than the pseudoinverse method. We show predictions starting from a straight line along different directions. Both methods produce results that depend on the position along the starting line. PLS also predicts periodic changes along the paths and this structure is close to hexagonal beyond the "arena": pentagons and heptagons or other non-hexagonal polygons are rare except around the edges of the predicted region (**Figure 9**). Predicted signal fades in most cases as prediction proceeds.

TABLE 1 | Dead neuron count: number of non-responsive computational units.

	Field of view [deg]									
	28	**68**	**108**	**148**	**188**	**228**	**268**	**308**	**348**	**360**
Norm-SP1-Mask	2	0	5	5	10	12	16	18	15	18
LT3.33%	0	0	0	0	0	2	2	6	8	9
Norm-LT3.33%	0	0	0	1	1	3	2	4	9	11
Norm-LT3.33%-Mask	0	0	0	0	0	0	1	2	7	11
Norm-LT6.66%-Mask	0	0	0	0	0	0	1	4	13	13
Norm-DL-LT3.33%-Mask	0	3	1	29	30	30	30	30	30	30
Norm-LT3.33%-LT3.33%-Mask	0	0	0	0	0	0	0	0	0	0

Figure 9 show predicted structures at angles 0° (**Figures 9A–C**) and in 340° (**Figures 9D–F**), respectively. Prediction takes past values of 60, 80, and 100 steps, respectively (see **Figure 9**). Outside the arena the number of predicted steps are in the order of 200. Note one step is very small compared to the PCs. If the size of the PCs is about the size of the rat, then the steps are about one twentieth of the rat's size.

For 0°, hexagonal structure is the best for 60 steps, but it fades quickly. Fading decreases for 80 steps, but the structure inherits the PC errors of the arena. This is more so for 100 steps. The case is somewhat different for predictions along 340°. In this case, fadings are similar. Visual inspection says that it is the smallest for the 80 step case. Hexagonal structure is relatively poor for 60 steps and is considerably better for 80 and 100 steps.

The figures demonstrate that close to hexagonal predictions can arise. The following notes are due here. The more the information from the past, the more the squared arena frustrates the hexagonal structure. Different directions approximate hexagonal structure differently, depending on the error structure within the squared arena. We also note that the ratio between length of the boundary and the size of the arena decreases the frustrating effect of boundary as the size of the arena increases.

From the point of view of model categories, the predictive network that uses its own output to complement (increment) its own input is an *attractor network*.

5. DISCUSSION

First, we review and discuss the general and specific features of our results. We also link them to the neural substrate and consider the computational potentials from the point of view of semantic memory, episodic memory, and reinforcement learning.

5.1. General Considerations

Our goal was to find hidden and abstract Cartesian Factor, that is, the discretization of the factor and the related attractor network that serves as an implicit representation of the related metric, provided that we have the complementing one. The method is general. We applied the approach as a model for the EHC. We assumed that we are having the head direction cells. From the point of view of the neuronal computations, attractor models working on set of cells are the most promising (see e.g., Skaggs et al., 1995; Redish et al., 1996 reviewed by Clark and Taube, 2012).

From the theoretical point of view, the abstraction that we want to develop is similar to geometrical abstractions or algebraic abstractions: they cannot be sensed directly, so they are latent. They are also Cartesian in the sense that they are like coordinates in an abstract space. In turn, they enable highly compressed descriptions. According to our assumptions, Cartesian Factors are low dimensional and only a few of them are needed for the mental solving of certain tasks and for the execution of decisions. Such elimination of variables is critical for reinforcement learning (Kearns and Koller, 1999; Boutilier et al., 2000; Szita and Lőrincz, 2009). The example in the context of navigation is path planning. Path planning can be accomplished in a discretized allothetic abstraction independently from idiothetic visual observations. This property lowers computational needs considerably. In turn, optimization of problem solving depends on the capability of forming low dimensional Cartesian Factors that are relevant for planning.

The concept of Cartesian Factors is closely related to Gestalt principles. Gestalt psychologists considered objects as perceived and as global constructs made of the constituting elements *within an environment*. Gestalt psychology has a number of concepts or laws on how to group things or events. Among these are the *Law of Proximity* and the *Law of Continuity*: according to Köhler (1929), "what moves together, belongs together" (see e.g., Paglieri, 2012 and the references therein). Self-motion, for example, allows the separation of the self from the rest of the environment and can be uncovered by temporal information. Such information drives the SFA procedure explored by Wiskott's group (Franzius et al., 2007; Schönfeld and Wiskott, 2015). They found that in realistic conditions and for large viewing angles, direction independent place cells can be formed by means of the temporal information. However, temporal information may be limited due to sudden environmental changes or occlusions. Furthermore, limiting the algorithm to temporal information limits the Gestalt principles to a few of them.

Another Gestalt principle is the *Law of Similarity*. This principle does not rely on temporal information and could be more adequate for general databases. Our algorithms implicitly exploit this principle through the concatenated input pieces that correspond to different viewing directions and may have identical, similar or very different information contents, subject

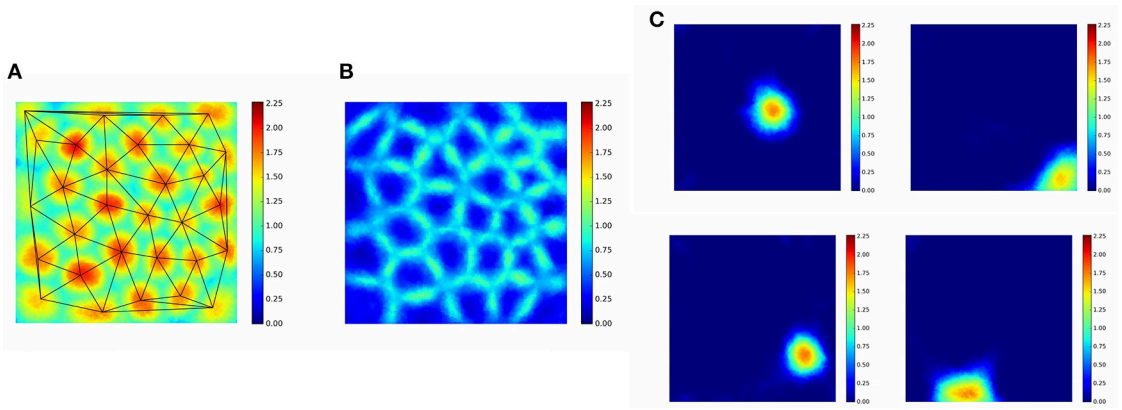


FIGURE 5 | PC positions make close to hexagonal structure constrained by the non-hexagonal form of the "arena." (A) Delaunay triangulation on the linear activities of the first (largest) component of the magnitude ordered representation. **(B)** Linear activities of the second(-largest) component of the magnitude ordered representation. **(C)** Individual PC activities. For more details, see text.

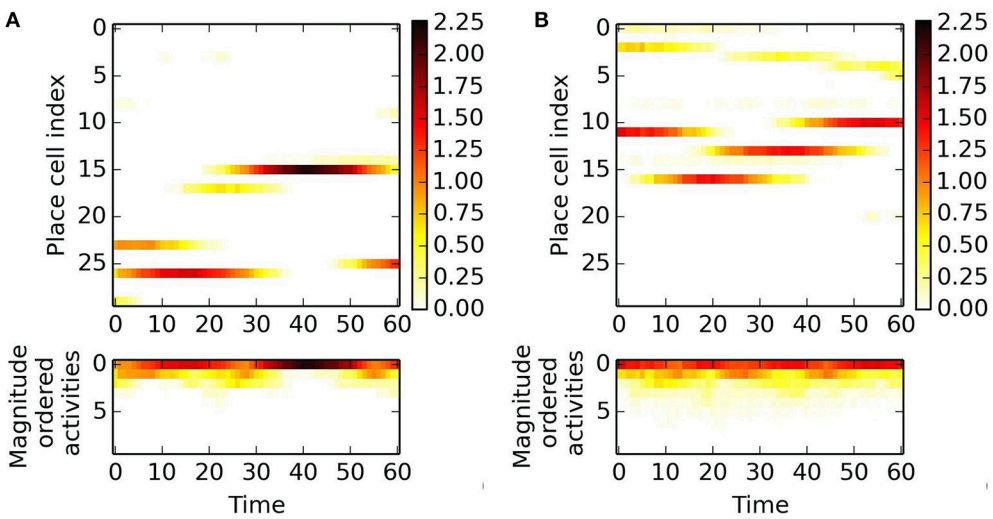


FIGURE 6 | Magnitude ordered examples at two different positions in two different directions. Activities are color coded. **(A)** 1st place and 1st direction. **Top:** activities of place cells along a 60 step paths, **bottom:** magnitude ordered activities. **(B)** Alike **(A)**, but for 2nd place and 2nd direction. Different place cells fire. About four place cells produce non-negligible outputs in both cases.

to the position and the orientation. In our work, we used head direction and idiothetic information. The idiothetic observation was in the form of a *bag model*. Bag models are widely used in natural language processing, called the *bag of words* (BoW) representation, and in image processing, called the *bag of keypoints* (BoK) representation in this case. It means that we have access to the components being present at a time, but not about their order in time or space. In other words, the bag model is similar to the *what system* of visual information processing, described first by Mishkin and Ungerleider (1982).

Considering the bag model from another point of view, any component in the bag requires an invariant representation. For BoW, stemming is the tool. BoK can be based, for example, on local scale invariant features introduced by Lowe (1999). Whereas stemming eliminates the details and becomes invariant of the syntax, scale invariant features incorporate scale and rotation variations in order to become invariant to transformations. The case of PCs is similar, their outputs are invariant to directional changes. In turn, our concept can be formulated as follows: we assume that beyond having a Cartesian Factor, (a) some "details," such as suffixes or scaling and rotations or orientation, can be measured, (b) the bag model has been built and the "suffixes" are either explicitly embedded into the complementing observations (i.e., into BoK) or neglected (i.e., from BoW), (c) the complementing observations hide a low dimensional space and thus it can be discretized with limited resources, and (d) this low dimensional space may have a related metric. In the case of documents, discretization may correspond to topics and the underlying structure is similar to a tree, since each topic may have subtopics. In the case of scale invariant features, the complementing space is the space of shapes and textures and it is very large. However, if the bag of environmental

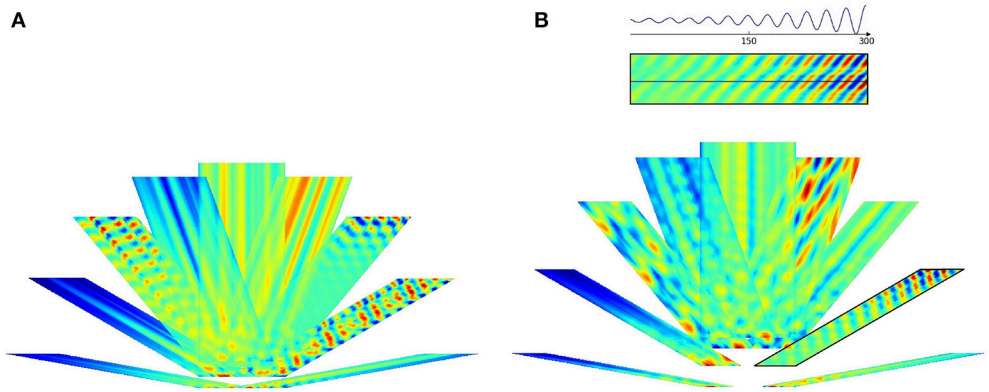


FIGURE 7 | Pseudoinverse predictions along straight paths beyond the "arena." Inset of **(B)** and the parallelogram framed with black line in **(B)** show how the figures were created: oscillating paths along a straight line are color coded and numerous parallel lines were computed. Different directions are cumulated into different parallelograms oriented according to the path traveled. Predictions start from the points of the lower edge of the "arena," have **(A)** 60 and **(B)** 100 time steps measured in different directions and proceed in that direction. Each direction had its trained predictive matrix that predicted activities from all neighboring oriented lines. Ideal activity pattern forms a triagonal lattice structure. The predictive matrix learned some characteristics of the displacements in the structure, but the prediction is poor.

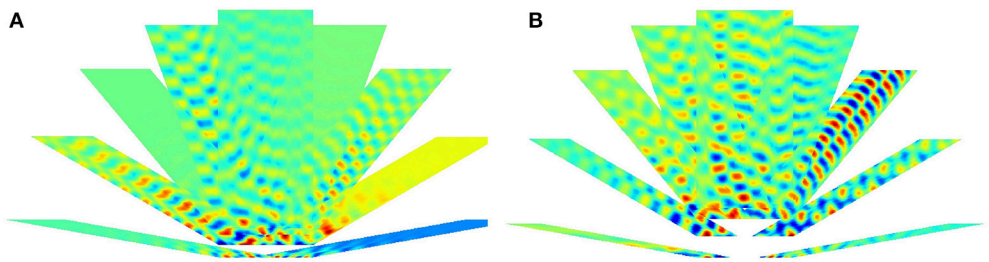


FIGURE 8 | PLS regression based predictions along straight paths beyond the "arena." The subfigures are created alike to those of **Figure 7**. Predictions start from the points of the lower edge of the "arena," have **(A)** 60 and **(B)** 100 time steps measured in different directions and proceed in that direction. Each direction had its trained predictive PLS algorithm. The predictions can represent displacement information along the parallel lines and periodic, approximately hexagonal structure appear in a number of directions.

visual cues can be formed as we did here, then it can support the discretization of the environment as we showed in our computer studies.

We should note that similarity based grouping is an alternative to temporal grouping and can be used if the latter is not available. For example, temporal grouping is impaired in akinetopsia, but the representation of the 3D world is not impaired. It seems reasonable to expect that temporal and similarity based algorithms *together* learn faster, perform more robustly and more precisely, e.g., if the task is forecasting.

The novelty of our contribution is the concept of Cartesian Factor. Such factors can be developed in many ways. Here, we put forth a similarity based algorithm, studied it, and suggest to unify it with other Gestalt principles. From the point of view of Gestalt theory, the novelty in this work is that we are looking for descriptors of the global context, that is, the environment itself. Compression takes place via sparse autoencoding, when encoding is based on the information that we apply via *masking* part of the input representation. Note that the input is in the form of a *bag representation*, which is a sufficient condition here.

We added temporal clues and developed predictive systems using pseudoinverse computations and PLS regression. Pseudoinverse computation seem to fit the structure of the superficial layers of the entorhinal cortex (Lőrincz and Szirtes, 2009) and the non-linear extensions are feasible. For pseudoinverse computation and for PLS, we found that PLS regression can provide more regular predictions. Furthermore, we found that the oriented hexagonal-like structures continued beyond the observed "arena" can keep the hexagonal regularity, sometimes to a better extent than the original set of PCs learned in a non-hexagonal environment. We suspect that the highly precise hexagonal grids (see e.g., the review written by Buzsáki and Moser, 2013 and the cited references therein) may emerge by including an interplay between the PCs and the oriented grids when orientation free grids are developed, since the trigonal grid is the common structure in the different directions.

5.2. Cell Types Developed

Using the bag model, we could develop place cells by covering viewing angles of about 100°. Further improvement can be expected if (i) deeper networks are applied and (ii) if temporal

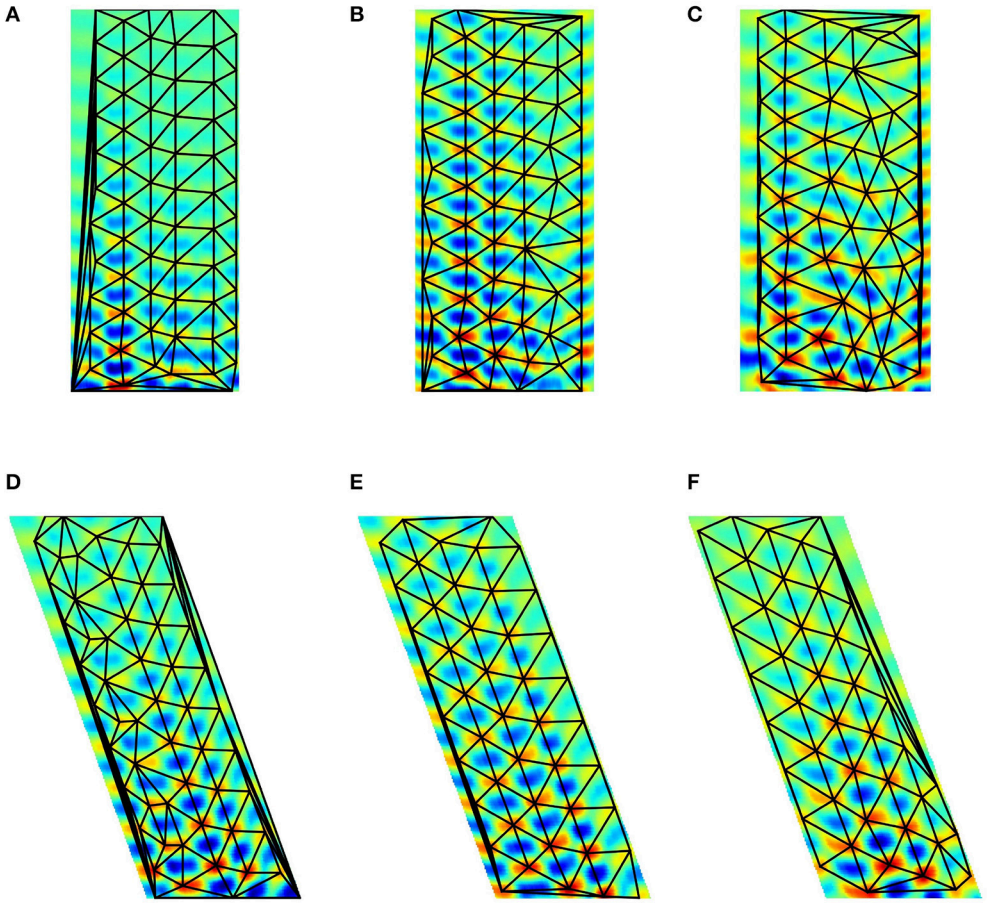


FIGURE 9 | Delaunay triangulation fitted to the predicted structure from models. (A–C) Trained on 0° paths **(D–F)** trained on 340° paths. Training paths are 60, 80, and 100 steps long for **(A)**+**(D)**, **(B)**+**(E)**, and **(C)**+**(F)** subfigures, respectively.

changes are included. We found in our simulations that sparsity should be kept for deeper networks at least for some of the layers. No experiments were conducted on pixel based visual information, a much higher dimensional representation that has pixel-wise nonlinearities. Such nonlinearities can be overcome in many ways, including temporal methods as demonstrated by Franzius et al. (2007) and Schönfeld and Wiskott (2015). An extension of our architecture to a hierarchy may also suffice.

While the first largest amplitude PC signal must belong to the closest cell, the second largest must belong to its nearest neighbor along the path. In turn, second largest amplitudes should uncover the Voronoi tessellation of the PCs as demonstrated in our computer experiments (**Figure 5B**).

From the algorithmic point of view, when a path proceeds toward the border of the "arena" and gets close to it, the second largest component becomes very small, since there is no cell beyond the border and the second nearest neighbor can be far at the sides. Assume that a cell responds to the ratio between the largest activity and the second largest one. This cell will show high activity when the path is directed toward the border and the position is close to the border, since the second largest activity belongs to a remote PC and is small. This cell would behave alike to *border cells* even in dark. We should note that according to the long held view, interneurons approximate arithmetic operations, such as *subtraction*, *division* or *shunting* of the excitation.

By means of PCs, we could develop oriented grid cells and could derive some precursors for border cells. Three simple and justifiable algorithmic operations were exploited, (i) the integrate-and-fire mechanism, (ii) features of the theta waves, and (iii) a self-supervisory compression in the form of pseudoinverse computation and PLS regression. Self-supervision means that actual signals supervise delayed signals during learning. Magnitude based ordering may occur in the neural substrate, e.g., if magnitudes are converted to time giving rise to time ordering. However, some kind of clock is needed for telling the zero instant of the ordering. Intriguingly, the phase of theta wave can play the role of such a clock. Indeed, during the first half of the theta cycle, cells that fire represent current position, whereas during the second half of the theta cycle temporally ordered (future) place cells fire (Sanders et al., 2015). These findings point to a more complex mechanism: cells that represent the past cant fire in the second half of the theta wave. We used a concatenation mechanism for prediction and, in turn, our model suggests a predictive learning mechanism that overbridges theta

cycles and exploits the activities of the second halves of the theta cycles.

Recent results from Ferrante et al. (2016b) show that different functional groups of pyramidal and inhibitory neurons are present in the entorhinal cortex. Such groups may satisfy our constraints that magnitude based ordering can support oriented grid cell formation via self-supervised prediction as well as border cell formation via shunting inhibition. Here is putative model for the latter. Consider the integrate-and-fire model. Spikes that come first excite the neuron and if delayed spikes that respond to the second largest activities are not capable for the ignition of shunting inhibition—e.g., if the animal is close to the border and no PC is in that direction—then the cell will fire and the cell will behave like a border cell. The head direction dependence is, however, more complex as reported by the original work of Solstad et al. (2008) calling for more detailed models based on sophisticated features, see e.g., the review of Kepecs and Fishell (2014) and the papers of Ferrante et al. (2016a), being outside of the scope of this paper.

5.3. Order of Learning in the Model

We used HCs for learning PCs without temporal information. We developed oriented grid cells from the PCs by means of temporal information and self-supervised compression. We showed that prediction becomes more regular (more hexagonal-like) if it is continued beyond the area represented by PCs. Temporal information on the second largest amplitudes gives rise to the Voronoi polygons on the set of PCs and may uncover border responses, e.g., by insufficient shunting inhibition. This algorithmic feature remains valid in dark, since it relies on the available set of PCs.

Other entorhinal cell types, such as speed cells and direction independent grid cells pose further challenges for our model. Speed cells described by Kropff et al. (2015), can be easily formed, since the firing rate of oriented grids is a monotone function of speed as found by Sargolini et al. (2006). For example, the max pooling operation, being well documented for the primary visual cortex (Movshon et al., 1978; Mechler and Ringach, 2002; Touryan et al., 2005), suits the needs. The idea can be traced back to the work of Fukushima (1980) and has gained attention from the point of view of (i) invariant representations (Serre et al., 2002), (ii) as a tool for efficient feature extraction, and (iii) reduction of the dimension of the representation (Huang et al., 2007). From the point of view of grid cells, a max pooling neuron outputs the largest activity and thus it loses orientation and displacement dependencies making the activity a monotone function of the speed.

The model of direction independent grid cells is more challenging, since there are additional constraints: firing should be continued (a) at any point, (b) including the absence of learned PCs, and (c) according to the displacement of the grid in any changes of the direction. A number of neurally plausible models based on different assumptions have been built see, e.g., the works of Burgess and O'Keefe (2011), Giocomo et al. (2011), and Kesner and Rolls (2015) and the cited references. The capability for planning, however, seems crucial as emphasized by Buzsáki and Moser (2013) and Sanders et al. (2015). It has been included into a detailed model by Sanders et al. (2015). Compared to these model, the Cartesian Factor principle is a high level description that aims to shed light onto the origin of the key algorithmic building blocks of the development of neural representations.

The Cartesian Factor principle suggests the following order of learning: (i) head direction cells, (ii) place cells, (iii) oriented grid cells, (iv) direction free grid cell representation by means of an interplay between place cells and grid cells. According to the recent paper from Rowland et al. (2016), there are two possible routes for grid cell formation: it is either species specific or spatial experience shapes the grid system. Our model proposes the latter option and fits the experimentally found order of learning reviewed in the cited paper.

We illustrated that the hexagonal like symmetry of the grid cells can be maintained in the absence of information form PCs. Planning and then traveling along loops, e.g., exploring and then homing, can serve the tuning of the grid cells. It may be worth noting that both grids and PCs change under slight distortion of the "arena" showing the coupling between these representations.

Along the same line of thoughts, our model is based on an autoencoder, which—by construction—is also a comparator (Lőrincz and Buzsáki, 2000) as suggested for the hippocampal function by Vinogradova (2001) and others, see the cited references. In the autoencoder, the input received is compared with the representation generated output. In case of mismatch, the adjustment of the representation may take place and the same error may drive Hebbian learning. Such error based optimization of the representation and learning were suggested by Lőrincz and Buzsáki (2000) and Chrobak et al. (2000) and elaborated by Lőrincz and Szirtes (2009).

Our sparse autoencoder hypothesis is supported by the fact that activity patterns are very sparse in the CA1 subfield of the hippocampus. We found in our numerical experiments that two stages are needed for the development of sparse representations, one for real time processing that uses spatial sparsity, and another one for off-line processing, when replayed inputs satisfy lifetime sparsity constraints. Such differences may show up in statistical evaluations of theta phase patterns and SPW-R patterns, with the former representing the actual path, whereas the latter may perform lifetime sparsification. However, behavioral relevance may modulate this simple picture.

5.4. Special Features of the Algorithms

The particular features of our algorithmic approach are as follows:

1. Sparse autoencoding requires two stage operation, one for real time and another one for learning. The latter should implement or approximate lifetime sparsity. Imperfect lifetime sparsity may give rise to silent neurons not responding to inputs. Homeostasis can counteract this process, enabling an adjustable reservoir of PCs for learning new information. Homeostatic maintenance of the activity may manifest itself through low spatial specificity. Such neurons have been found by Grosmark and Buzsáki (2016), but the picture seems more sophisticated.

2. Temporal ordering is necessary for the predictive compression in our model. This is the core step that sets the high-level grid representation free from external observations. Theta-waves or integrate-and-fire behavior, possibly both, are candidates for temporal ordering.
3. The bag model simplifies both the algorithm and representation; it decreases the dimensionality of the input and neglects many of the details. It keeps track of the components, but not their actual manifestations. The bag representation is analogous of the "what system" that has information about the objects present, but not about their positions, for example. From the point of view of component based representation, the bag model resembles to the "recognition by components" principle put forth by Biederman (1987) for visual inputs.
4. The model of Cartesian Factor formation needs neurons that can multiply and can produce conjunctive representations, e.g., between the visual cues and the head direction cells. Candidates for such computations include (i) the logical operations, such as the AND operation made possible by coincidence detection (for a recent review, see the work of Stuart and Spruston, 2015), (ii) the interplay between distal and proximal dendritic regions—when the proximal input enhances the propagation of the distal dendritic spikes—can also support a multiplicative function (Larkum et al., 2001; Jarsky et al., 2005). We note that the EHC has sophisticated interconnections between distant and proximal regions (Gigg, 2006). We exploited the multiplicative feature in our representation by using the product space and zero some of the inputs by (multiplicative) masking.

5.5. Relation to Meta-Level Cognition

Cartesian Factors select features of the world and a limited set of features may be sufficient for solving distinct problems. Path planning is an example. The grid like structure, its potentials for path planning and distance estimation as described in Huhn et al. (2009), for example, are high level descriptors of the world. They tell very little about the actual sensory information. The autoencoding principle can serve both functions that is (i) the manipulation at the meta-, or symbolic level, such as the computation of distances on the grid structure and (ii) the low level input-like representation via the estimations of the inputs or the inputs that follow. The autoencoding principle resolves the homunculus fallacy by saying that "making sense of the input" is the function of the representation that approximates the input (Lőrincz et al., 2002). We undersign the view that the estimation of the input occurs via hierarchical bag representations that neglect more and more details bottom-up and combine more and more (Cartesian) factors top-down. One may say that in the top-down generation of the estimated input, meta level description becomes semantically embedded by means of the contributing Cartesian Factors.

One can also treat episodic memory in the context of the autoencoding principle. The appearances or the disappearances of sparse codes by time can be seen as starting and ending points of events. Such description fits factored reinforcement learning (Szita et al., 2003). Taken together, our algorithms and the concept of Cartesian Factors can provide simple clues about the working mechanisms of the "cognitive map" in such a way that the computations avoid combinatorial explosions (Szita and Lőrincz, 2009) and thus escape the curse of dimensionality, explicated by Bellman (1958).

6. CONCLUSIONS

We put forth the novel concept of Cartesian Factors. The working was demonstrated by forming of place cells and grid cells, where we exploited the complementary information, the head direction cells. Our proposed cognitive mechanism does not work in the absence of such information. We note that upon destroying the vestibular system, which is critical for having head direction cells, no place cell is formed (Taube, 2007; Winter and Taube, 2014).

Our algorithm is a sparse autoencoding mechanism that can be deep, but should be sparse in the hidden layers according to the numerical studies. Our algorithm relies on the *bag model* that we related to the *what system*. The bag model works with a collection of input portions that represent the same quantity type, or object types, or episode types, such as idiothetic inputs collected at the same position but in different directions, or the different views of an object, or the different temporal variations starting from a given state and ending in an other one, respectively. The different mechanisms should support each other.

The particular feature of the Cartesian Factors is that a few of them may be sufficient for solving cognitive problems. An example is path planning on the "cognitive map" if neighbor relations are available. Elimination of directions from the path planning problem reduces the state space in the exponent. This is a very important advantage in decision making.

We used the discretized form of the Cartesian Factors to develop the (implicit) metric-like representation that can be continued beyond the experienced portion of the factor. The self-supervised predictive compression method was illustrated in oriented grid formation. We found that the predicted grids can be very regular and may compensate for the errors of the underling discretization of the factor. We used magnitude based ordering and suggested integrate-and-fire mechanism and theta wave based firing as candidate mechanisms for this learning stage. The attractive feature of magnitude ordering is that it detaches sensory information from the underlying (metrical) structure and enables extrapolation beyond the already observed part of the world.

The interplay between (a) the detachment of the direct sensory information, (b) the manipulation in the underlying space, and (c) the association of new sensory information to the extrapolated structure, in other words, the separation of grids from visual sensory information, the prediction on the grids can be seen as symbol learning, symbol manipulation, respectively. The association of grid cell activities to visual information, on the other hand, corresponds to symbol grounding in our framework and offers a solution to the grounding problem targeted first by Harnad (1990).

We found that the concept of Cartesian Factors approximates well the learning order and impairment related features of head

direction cells, place cells, and oriented grid cells. The concept also provides hints about border cells that can fire in the absence of visual information. We argued that border cells, direction free grid cells, and speed cells can emerge in the model via neurally plausible mechanisms, but they require further studies.

In sum, the concept of Cartesian Factors offers (a) a solution for the curse of dimensionality problem of reinforcement learning, (b) an explanation for a number of features of the EHC, such as sparse representation, distinct cell types, and the order of learning, (c) a framework for symbol formation, symbol manipulation, and symbol grounding processes, and (d) a mechanism for the learning of attractor models by means of magnitude ordering.

AUTHOR CONTRIBUTIONS

The main contributions of AL cover the basic concepts, including the idea of Cartesian Factors, the relations to the cognitive map, the grid structure, and other cell types, the connections to cognitive science, and factored reinforcement learning. Computer studies, including some discoveries from computation based modeling are the key contributions of AS. They contributed equally to the design of the work, the analysis, and the interpretation of data. The paper was written jointly with figures being produced mostly by AS, whereas writing is mostly due to AL.

FUNDING

This research was supported by the EIT Digital grant (Grant No. 16257).

ACKNOWLEDGMENTS

We are grateful to the reviewers for their helpful comments and suggestions.

REFERENCES

Arleo, A., and Gerstner, W. (2000). Spatial cognition and neuro-mimetic navigation: a model of hippocampal place cell activity. *Biol. Cybern.* 83, 287–299. doi: 10.1007/s004220000171

Becker, S. R., Candès, E. J., and Grant, M. C. (2011). Templates for convex cone problems with applications to sparse signal recovery. *Math. Program. Comput.* 3, 165–218. doi: 10.1007/s12532-011-0029-5

Bellman, R. (1958). *Combinatorial Processes and Dynamic Programming*. Technical report, DTIC Document.

Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., et al. (2010). "Theano: a CPU and GPU math expression compiler," in *Python Science Computer*, *Vol. 4*, 3–10. Available online at: https://conference.scipy.org/proceedings/scipy2010/pdfs/proceedings.pdf

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychol. Rev.* 94, 115–147. doi: 10.1037/0033-295X.94.2.115

Bonnevie, T., Dunn, B., Fyhn, M., Hafting, T., Derdikman, D., Kubie, J. L., et al. (2013). Grid cells require excitatory drive from the hippocampus. *Nat. Neurosci.* 16, 309–317. doi: 10.1038/nn.3311

Boutilier, C., Dearden, R., and Goldszmidt, M. (2000). Stochastic dynamic programming with factored representations. *Artif. Intell.* 121, 49–107. doi: 10.1016/S0004-3702(00)00033-3

Burgess, N., and O'Keefe, J. (2011). Models of place and grid cell firing and theta rhythmicity. *Curr. Opin. Neurobiol.* 21, 734–744. doi: 10.1016/j.conb.2011.07.002

Buzsáki, G. (1989). Two-stage model of memory trace formation: a role for noisyİ brain states. *Neuroscience* 31, 551–570. doi: 10.1016/0306-4522(89)90423-5

Buzsáki, G., and Moser, E. I. (2013). Memory, navigation and theta rhythm in the hippocampal-entorhinal system. *Nat. Neurosci.* 16, 130–138. doi: 10.1038/nn.3304

Chrobak, J. J., and Buzsáki, G. (1994). Selective activation of deep layer (V-VI) retrohippocampal cortical neurons during hippocampal sharp waves in the behaving rat. *J. Neurosci.* 14, 6160–6170.

Chrobak, J. J., Lőrincz, A., and Buzsáki, G. (2000). Physiological patterns in the hippocampo-entorhinal cortex system. *Hippocampus* 10, 457–465. doi: 10.1002/1098-1063(2000)10:4<457::AID-HIPO12>3.0.CO;2-Z

Clark, B. J., and Taube, J. S. (2012). Vestibular and attractor network basis of the head direction cell signal in subcortical circuits. *Front. Neural Circuits* 6:7. doi: 10.3389/fncir.2012.00007

Csurka, G., Dance, C., Fan, L., Willamowski, J., and Bray, C. (2004). "Visual categorization with bags of keypoints," in *Workshop on Statistical Learning in Computer Vision, ECCV*, *Vol. 1* (Prague), 1–21.

Dahl, G. E., Sainath, T. N., and Hinton, G. E. (2013). "Improving deep neural networks for LVCSR using rectified linear units and dropout," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (New York, NY: IEEE Press), 8609–8613.

Dai, W., and Milenkovic, O. (2009). Subspace pursuit for compressive sensing signal reconstruction. *IEEE Trans. Inform. Theory* 55, 2230–2249. doi: 10.1109/TIT.2009.2016006

Derdikman, D., and Knierim, J. J., (eds.) (2014). *Space, Time and Memory in the Hippocampal Formation*. Vienna: Springer.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.* 11, 12, 2121–2159.

Erhan, D., Bengio, Y., Courville, A., Manzagol, P.-A., Vincent, P., and Bengio, S. (2010). Why does unsupervised pre-training help deep learning? *J. Mach. Learn. Res.* 11, 625–660.

Evans, T., Bicanski, A., Bush, D., and Burgess, N. (2016). How environment and self-motion combine in neural representations of space. *J. Physiol.* 594, 6535–6546. doi: 10.1113/jp270666

Ferrante, M., Shay, C. F., Tsuno, Y., William Chapman, W., and Hasselmo, M. E. (2016a). Post-inhibitory rebound spikes in rat medial entorhinal layer II/III principal cells: *in vivo, in vitro*, and computational modeling characterization. *Cereb. Cortex*. doi: 10.1093/cercor/bhw058. [Epub ahead of print].

Ferrante, M., Tahvildari, B., Duque, A., Hadzipasic, M., Salkoff, D., Zagha, E. W., et al. (2016b). Distinct functional groups emerge from the intrinsic properties of molecularly identified entorhinal interneurons and principal cells. *Cereb. Cortex*. doi: 10.1093/cercor/bhw143. [Epub ahead of print].

Franzius, M., Sprekeler, H., and Wiskott, L. (2007). Slowness and sparseness lead to place, head-direction, and spatial-view cells. *PLoS Comput. Biol.* 3:e166. doi: 10.1371/journal.pcbi.0030166

Fukushima, K. (1980). Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.* 36, 193–202. doi: 10.1007/BF00344251

Fyhn, M., Molden, S., Witter, M. P., Moser, E. I., and Moser, M.-B. (2004). Spatial representation in the entorhinal cortex. *Science* 305, 1258–1264. doi: 10.1126/science.1099901

Geladi, P., and Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Anal. Chim. Acta* 185, 1–17. doi: 10.1016/0003-2670(86)80028-9

Gigg, J. (2006). Constraints on hippocampal processing imposed by the connectivity between ca1, subiculum and subicular targets. *Behav. Brain Res.* 174, 265–271. doi: 10.1016/j.bbr.2006.06.014

Giocomo, L. M., Moser, M.-B., and Moser, E. I. (2011). Computational models of grid cells. *Neuron* 71, 589–603. doi: 10.1016/j.neuron.2011.07.023

Goodale, M. A., and Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends Neurosci.* 15, 20–25. doi: 10.1016/0166-2236(92)90344-8

Grant, M., and Boyd, S. (2014). *CVX: Matlab Software for Disciplined Convex Programming, Version 2.1.* Available online at: http://cvxr.com/cvx

Grosmark, A. D., and Buzsáki, G. (2016). Diversity in neural firing dynamics supports both rigid and learned hippocampal sequences. *Science* 351, 1440–1443. doi: 10.1126/science.aad1935

Hales, J. B., Schlesiger, M. I., Leutgeb, J. K., Squire, L. R., Leutgeb, S., and Clark, R. E. (2014). Medial entorhinal cortex lesions only partially disrupt hippocampal place cells and hippocampus-dependent place memory. *Cell Rep.* 9, 893–901. doi: 10.1016/j.celrep.2014.10.009

Harnad, S. (1990). The symbol grounding problem. *Physica D* 42, 335–346. doi: 10.1016/0167-2789(90)90087-6

Harris, Z. S. (1954). Distributional structure. *Word* 10, 146–162. doi: 10.1080/00437956.1954.11659520

Hinton, G. E., and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science* 313, 504–507. doi: 10.1126/science.1127647

Hochreiter, S. (1991). *Untersuchungen zu Dynamischen Neuronalen Netzen*. Master's Thesis, Institut für Informatik, Technische Universität, München.

Hochreiter, S., Bengio, Y., and Frasconi, P. (2001). "Gradient flow in recurrent nets: the difficulty of learning long-term dependencies," in *Field Guide to Dynamical Recurrent Networks*, eds J. Kolen and S. Kremer (New York, NY: IEEE Press).

Huang, F. J., Boureau, Y.-L., and LeCun, Y. (2007). "Unsupervised learning of invariant feature hierarchies with applications to object recognition," in *2007 IEEE Conference on Computer Vision and Pattern Recognition* (New York, NY: IEEE Press), 1–8.

Huhn, Z., Somogyvári, Z., Kiss, T., and Érdi, P. (2009). Distance coding strategies based on the entorhinal grid cell system. *Neural Netw.* 22, 536–543. doi: 10.1016/j.neunet.2009.06.029

Jarsky, T., Roxin, A., Kath, W. L., and Spruston, N. (2005). Conditional dendritic spike propagation following distal synaptic activation of hippocampal ca1 pyramidal neurons. *Nat. Neurosci.* 8, 1667–1676. doi: 10.1038/nn1599

Kearns, M., and Koller, D. (1999). "Efficient reinforcement learning in factored MDPs," in *International Joint Conference on Artificial Intelligence, vol. 16* (San Francisco, CA: Morgan Kaufmann Publishers Inc.), 740–747.

Kepecs, A., and Fishell, G. (2014). Interneuron cell types are fit to function. *Nature* 505, 318–326. doi: 10.1038/nature12983

Kesner, R. P., and Rolls, E. T. (2015). A computational theory of hippocampal function, and tests of the theory: new developments. *Neurosci. Biobehav. Rev.* 48, 92–147. doi: 10.1016/j.neubiorev.2014.11.009

Kingma, D., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv:1412.6980*.

Köhler, W. (1929). *Gestalt Psychology. [Psychologische Probleme 1933]*. New York, NY: Horace Liveright.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biol. Cybern.* 43, 59–69. doi: 10.1007/BF00337288

Kowalski, B., Gerlach, R., and Wold, H. (1982). "Chemical systems under indirect observation," in *Systems Under Indirect Observation*, eds K. Jöreskog and H. Wold (Amsterdam: North-Holland), 191–209.

Kropff, E., Carmichael, J. E., Moser, M.-B., and Moser, E. I. (2015). Speed cells in the medial entorhinal cortex. *Nature* 523, 419–424. doi: 10.1038/nature14622

Larkum, M. E., Zhu, J. J., and Sakmann, B. (2001). Dendritic mechanisms underlying the coupling of the dendritic with the axonal action potential initiation zone of adult rat layer 5 pyramidal neurons. *J. Physiol.* 533, 447–466. doi: 10.1111/j.1469-7793.2001.0447a.x

Lee, A. K., and Wilson, M. A. (2002). Memory of sequential experience in the hippocampus during slow wave sleep. *Neuron* 36, 1183–1194. doi: 10.1016/S0896-6273(02)01096-6

Lőrincz, A. (2016). Cartesian abstraction can yield 'cognitive maps'. *Procedia Comput. Sci.* 88, 259–271.

Lőrincz, A., and Buzsáki, G. (2000). Two-phase computational model training long-term memories in the entorhinal-hippocampal region. *Ann. N. Y. Acad. Sci.* 911, 83–111. doi: 10.1111/j.1749-6632.2000.tb06721.x

Lőrincz, A., Sárkány, A., Milacski, Z. Á., and Tősér, Z. (2016). "Estimating cartesian compression via deep learning," in *International Conference on Artificial General Intelligence* (Berlin: Springer), 294–304. doi: 10.1007/978-3-319-41649-6_30

Lőrincz, A., Szatmáry, B., and Szirtes, G. (2002). The mystery of structure and function of sensory processing areas of the neocortex: a resolution. *J. Comput. Neurosci.* 13, 187–205. doi: 10.1023/A:1020262214821

Lőrincz, A., and Szirtes, G. (2009). Here and now: how time segments may become events in the hippocampus. *Neural Netw.* 22, 738–747. doi: 10.1016/j.neunet.2009.06.020

Lowe, D. G. (1999). "Object recognition from local scale-invariant features," in *Proceedings of the Seventh IEEE International Conference on Computer Vision, 1999, Vol. 2* (New York, NY: IEEE Press), 1150–1157. doi: 10.1109/ICCV.1999.790410

Makhzani, A., and Frey, B. (2013). k-sparse autoencoders. *arXiv:1312.5663*.

Makhzani, A., and Frey, B. J. (2015). "Winner-take-all autoencoders," in *Advances in Neural Information Processing Systems*, eds C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (Red Hook, NY: Curran Associates, Inc.), 2773–2781.

Mechler, F., and Ringach, D. L. (2002). On the classification of simple and complex cells. *Vis. Res.* 42, 1017–1033. doi: 10.1016/S0042-6989(02)00025-1

Mishkin, M., and Ungerleider, L. G. (1982). Contribution of striate inputs to the visuospatial functions of parieto-preoccipital cortex in monkeys. *Behav. Brain Res.* 6, 57–77. doi: 10.1016/0166-4328(82)90081-X

Moscovitch, M., Cabeza, R., Winocur, G., and Nadel, L. (2016). Episodic memory and beyond: the hippocampus and neocortex in transformation. *Ann. Rev. Psychol.* 67, 105–134. doi: 10.1146/annurev-psych-113011-143733

Moser, E. I., Roudi, Y., Witter, M. P., Kentros, C., Bonhoeffer, T., and Moser, M.-B. (2014). Grid cells and cortical representation. *Nat. Rev. Neurosci.* 15, 466–481. doi: 10.1038/nrn3766

Movshon, J. A., Thompson, I. D., and Tolhurst, D. J. (1978). Receptive field organization of complex cells in the cat's striate cortex. *J. Physiol.* 283:79. doi: 10.1113/jphysiol.1978.sp012489

Muller, R. U., Poucet, B., and Rivard, B. (2002). "Sensory determinants of hippocampal place cell firing fields," in *The Neural Basis of Navigation*, ed P. E. Sharp (Berlin: Springer), 1–22.

Nair, V., and Hinton, G. E. (2010). "Rectified linear units improve restricted Boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML)* (Haifa: Omnipress), 807–814.

Ng, A. Y. (2004). "Feature selection, ℓ_1 vs. ℓ_2 regularization, and rotational invariance," in *Proceedings of the 21st International Conference on Machine Learning (ICML)* (New York, NY: ACM), 78.

O'Keefe, J., and Dostrovsky, J. (1971). The hippocampus as a spatial map. preliminary evidence from unit activity in the freely-moving rat. *Brain Res.* 34, 171–175. doi: 10.1016/0006-8993(71)90358-1

O'Keefe, J., and Nadel, L. (1978). *The Hippocampus as a Cognitive Map*, *vol. 3*. Oxford: Clarendon Press.

Paglieri, F. (2012). *Consciousness in Interaction: The Role of the Natural and Social Context in Shaping Consciousness*, *Vol. 86*. Amsterdam: John Benjamins Publishing.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.

Quiroga, R. Q., Kreiman, G., Koch, C., and Fried, I. (2008). Sparse but not grandmother-cellcoding in the medial temporal lobe. *Trends Cogn. Sci.* 12, 87–91. doi: 10.1016/j.tics.2007.12.003

Rasmus, A., Berglund, M., Honkala, M., Valpola, H., and Raiko, T. (2015). "Semi-supervised learning with ladder networks," in *Advances in Neural Information Processing Systems*, eds C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (Red Hook, NY: Curran Associates, Inc.), 3532–3540.

Redish, A. D., Elga, A. N., and Touretzky, D. S. (1996). A coupled attractor model of the rodent head direction system. *Network* 7, 671–685. doi: 10.1088/0954-898X_7_4_004

Redish, A. D., and Touretzky, D. S. (1998). The role of the hippocampus in solving the morris water maze. *Neural Comput.* 10, 73–111. doi: 10.1162/089976698300017908

Rowland, D. C., Roudi, Y., Moser, M.-B., and Moser, E. I. (2016). Ten years of grid cells. *Ann. Rev. Neurosci.* 39, 19–40. doi: 10.1146/annurev-neuro-070815-013824

Sanders, H., Rennó-Costa, C., Idiart, M., and Lisman, J. (2015). Grid cells and place cells: an integrated view of their navigational and memory function. *Trends Neurosci.* 38, 763–775. doi: 10.1016/j.tins.2015.10.004

Sargolini, F., Fyhn, M., Hafting, T., McNaughton, B. L., Witter, M. P., Moser, M.-B., et al. (2006). Conjunctive representation of position, direction, and velocity in entorhinal cortex. *Science* 312, 758–762. doi: 10.1126/science.1125572

Schlesiger, M. I., Cannova, C. C., Boublil, B. L., Hales, J. B., Mankin, E. A., Brandon, M. P., et al. (2015). The medial entorhinal cortex is necessary for temporal organization of hippocampal neuronal activity. *Nat. Neurosci.* 18, 1123–1132. doi: 10.1038/nn.4056

Schönfeld, F., and Wiskott, L. (2015). Modeling place field activity with hierarchical slow feature analysis. *Front. Comput. Neurosci.* 9:51. doi: 10.3389/fncom.2015.00051

Schubotz, R. I., von Cramon, D. Y., and Lohmann, G. (2003). Auditory what, where, and when: a sensory somatotopy in lateral premotor cortex. *Neuroimage* 20, 173–185. doi: 10.1016/S1053-8119(03)00218-0

Schultheiss, N. W., Hinman, J. R., and Hasselmo, M. E. (2015). "Models and theoretical frameworks for hippocampal and entorhinal cortex function in memory and navigation," in *Analysis and Modeling of Coordinated Multi-neuronal Activity*, ed M. Tatsuno (Berlin: Springer), 247–268.

Scoville, W. B., and Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *J. Neurol. Neurosurg. Psychiatry* 20, 11–21. doi: 10.1136/jnnp.20.1.11

Serre, T., Riesenhuber, M., Louie, J., and Poggio, T. (2002). "On the role of object-specific features for real world object recognition in biological vision," in *International Workshop on Biologically Motivated Computer Vision* (Berlin: Springer), 387–397.

Skaggs, W. E., Knierim, J. J., Kudrimoti, H. S., and McNaughton, B. L. (1995). "A model of the neural basis of the rat's sense of direction," in *Advances in Neural Information Processing Systems 7*, eds G. Tesauro, D. S. Touretzky, and T. K. Leen (Cambridge, MA: MIT Press), 173–180.

Skaggs, W. E., and McNaughton, B. L. (1996). Theta phase precession in hippocampal. *Hippocampus* 6, 149–172.

Solstad, T., Boccara, C. N., Kropff, E., Moser, M.-B., and Moser, E. I. (2008). Representation of geometric borders in the entorhinal cortex. *Science* 322, 1865–1868. doi: 10.1126/science.1166466

Solstad, T., Moser, E. I., and Einevoll, G. T. (2006). From grid cells to place cells: a mathematical model. *Hippocampus* 16, 1026–1031. doi: 10.1002/hipo.20244

Squire, L. R., and Zola, S. M. (1998). Episodic memory, semantic memory, and amnesia. *Hippocampus* 8, 205–211.

Stuart, G. J., and Spruston, N. (2015). Dendritic integration: 60 years of progress. *Nat. Neurosci.* 18, 1713–1721. doi: 10.1038/nn.4157

Stuchlik, A., and Bures, J. (2002). Relative contribution of allothetic and idiothetic navigation to place avoidance on stable and rotating arenas in darkness. *Behav. Brain Res.* 128, 179–188. doi: 10.1016/S0166-4328(01)00314-X

Stuchlík, A., Petrásek, T., Prokopová, I., Holubová, K., Hatalová, H., Valeš, K., et al. (2013). Place avoidance tasks as tools in the behavioral neuroscience of learning and memory. *Physiol. Res.* 62(Suppl. 1), 1–19.

Sun, Y., Mao, H., Sang, Y., and Yi, Z. (2017). Explicit guiding auto-encoders for learning meaningful representation. *Neural Comput. Appl.* 28, 429–436. doi: 10.1007/s00521-015-2082-x

Szita, I., and Lőrincz, A. (2009). "Optimistic initialization and greediness lead to polynomial time learning in factored MDPs," in *Proceedings of the 26th International Conference Machine Learning* (New York, NY: ACM), 1001–1008.

Szita, I., Takács, B., and Lőrincz, A. (2003). ε-MDPs: learning in varying environments. *J. Mach. Learn. Res.* 3, 145–174.

Taube, J. S. (2007). The head direction signal: origins and sensory-motor integration. *Ann. Rev. Neurosci.* 30, 181–207. doi: 10.1146/annurev.neuro.29.051605.112854

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B (Methodol.).* 58, 267–288.

Todd, J. T. (2004). The visual perception of 3d shape. *Trends Cogn. Sci.* 8, 115–121. doi: 10.1016/j.tics.2004.01.006

Touryan, J., Felsen, G., and Dan, Y. (2005). Spatial structure of complex cell receptive fields measured with natural images. *Neuron* 45, 781–791. doi: 10.1016/j.neuron.2005.01.029

Tropp, J. A., and Gilbert, A. C. (2007). Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Trans. Inform. Theory* 53, 4655–4666. doi: 10.1109/TIT.2007.909108

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P.-A. (2010). Stacked denoising autoencoders. *J. Mach. Learn. Res.* 11, 3371–3408.

Vinogradova, O. S. (2001). Hippocampus as comparator: role of the two input and two output systems of the hippocampus in selection and registration of information. *Hippocampus* 11, 578–598. doi: 10.1002/hipo.1073

Wang, Y., Romani, S., Lustig, B., Leonardo, A., and Pastalkova, E. (2015). Theta sequences are essential for internally generated hippocampal firing fields. *Nat. Neurosci.* 18, 282–288. doi: 10.1038/nn.3904

Whishaw, I. Q., Hines, D. J., and Wallace, D. G. (2001). Dead reckoning (path integration) requires the hippocampal formation: evidence from spontaneous exploration and spatial learning tasks in light (allothetic) and dark (idiothetic) tests. *Behav. Brain Res.* 127, 49–69. doi: 10.1016/S0166-4328(01)00359-X

Winter, S. S., Clark, B. J., and Taube, J. S. (2015). Disruption of the head direction cell network impairs the parahippocampal grid cell signal. *Science* 347, 870–874. doi: 10.1126/science.1259591

Winter, S. S., and Taube, J. S. (2014). "Head direction cells: from generation to integration," in *Space, Time and Memory in the Hippocampal Formation*, eds D. Derdikman and J. J. Knierim (Berlin: Springer), 83–106.

Zeiler, M. D. (2012). Adadelta: an adaptive learning rate method. arXiv:1212.5701.

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

APPENDIX: DETAILS OF THE ALGORITHMIC FORMULATION OF CARTESIAN FACTOR LEARNING

Assume that a latent random variable Z and an observed random variable Y are continuous and together they fully explain away another observed binary random variable X. The ranges of Z and Y are supposed to be grid discretized finite r- and one-dimensional intervals, respectively. We denote the resulting grid points by $(z^{(m)}, y^{(l)}) \in \mathbb{R}^r \times \mathbb{R}; l = 0, \ldots, L, m = 1, \ldots, (M+1)^r$, $L, M, r \in \mathbb{N}$. The indices $m = 1, \ldots, (M+1)^r$ are supposed to be scrambled throughout training (i.e., we assume no topology between $z^{(m)}$). Then observation $\boldsymbol{x}^{(m,l)} \in \{0,1\}^d$ is generated by a highly non-linear function $g\colon \mathbb{R}^r \times \{1, \ldots, L\} \to \{0,1\}^d$ from grid point $z^{(m)}$ and grid interval $[y^{(l-1)}, y^{(l)})$ as

$$\boldsymbol{x}^{(m,l)} = g(\boldsymbol{z}^{(m)}, l) \tag{7}$$

for $m = 1, \ldots, (M+1)^r$; $l = 1, \ldots, L$. For each fixed m, one is given masks $\boldsymbol{V}_{i,\cdot} \in \{0,1\}^L$; $\sum_{l=1}^{L} \boldsymbol{V}_{i,l} = v \in \mathbb{N}$ indexing pairs of the form $(l, \boldsymbol{x}^{(m,l)})$, where $i = 1, \ldots, I$ is a global index. Provided such a sample from Y and X, we aim to approximate the discretized version of Z.

We formulated the above problem as a multilayer feedforward *lifetime sparse autoencoding* (Makhzani and Frey, 2015) procedure with input matrix $\boldsymbol{X} \in \{0,1\}^{I \times J}$ utilizing two novelties: concatenated input vectors and a masked loss function are motivated by the input structure. In order to construct the inputs $\boldsymbol{X}_{i,\cdot}; i = 1, \ldots, I$ of size $J = L \cdot d$, we coupled each v-tuple of $\boldsymbol{x}^{(m,l)}$ vectors for fixed m into a single block-vector using the $\boldsymbol{V}_{i,\cdot}$ values as follows:

$$\boldsymbol{X}_{i,\cdot} = \left[\boldsymbol{V}_{i,1} \cdot \boldsymbol{x}^{(m,1)}, \ \ldots, \boldsymbol{V}_{i,l} \cdot \boldsymbol{x}^{(m,l)}, \ \ldots, \boldsymbol{V}_{i,L} \cdot \boldsymbol{x}^{(m,L)}\right]. \tag{8}$$

Then, we used the ℓ_2 reconstruction error as the loss, but on a restricted set of elements, namely, on the v non-zero blocks for each input:

$$l(\boldsymbol{X}, \widehat{\boldsymbol{X}}, \boldsymbol{V}) \colon = \frac{1}{I} \sum_{\substack{i=1,\ldots,I \\ j=1,\ldots,J}} \boldsymbol{V}_{i,\lfloor \frac{j-1}{d} + 1 \rfloor} \cdot (\boldsymbol{X}_{i,j} - \widehat{\boldsymbol{X}}_{i,j})^2 \tag{9}$$

where $\widehat{\boldsymbol{X}}$ denotes the output of the decoder network. Finally, a sparse non-linearity was imposed on top of each encoder layer, which selected the k percent topmost activations across one component. We applied both lifetime (Makhzani and Frey, 2015) and spatial sparsification (Makhzani and Frey, 2013). Multilayer autoencoders with rectified linear units, $k = 1$ spatial sparsity, p%-sparse lifetime sparsity, and linear decoder output layer make the non-linear units of the network.

We implemented our method in the Python library Theano (Bergstra et al., 2010) based upon the SciPy2015 GitHub repository[2].

[2]https://github.com/kastnerkyle/SciPy2015

frontiers in Psychology

REVIEW
published: 22 November 2016
doi: 10.3389/fpsyg.2016.01820

Spaces in the Brain: From Neurons to Meanings

*Christian Balkenius * and Peter Gärdenfors*

Cognitive Science, Lund University, Lund, Sweden

Spaces in the brain can refer either to psychological spaces, which are derived from similarity judgments, or to neurocognitive spaces, which are based on the activities of neural structures. We want to show how psychological spaces naturally emerge from the underlying neural spaces by dimension reductions that preserve similarity structures and the relevant categorizations. Some neuronal representational formats that may generate the psychological spaces are presented, compared, and discussed in relation to the mathematical principles of monotonicity, continuity, and convexity. In particular, we discuss the spatial structures involved in the connections between perception and action, for example eye–hand coordination, and argue that spatial organization of information makes such mappings more efficient.

Keywords: chorus transform, conceptual spaces, eye–hand coordination, population coding, radial basis function, similarity, stimulus generalization

OPEN ACCESS

Edited by:
Tarek Richard Besold,
University of Bremen, Germany

Reviewed by:
Serge Thill,
University of Skövde, Sweden
Terrence C. Stewart,
University of Waterloo, Canada

****Correspondence:***
Christian Balkenius
christian.balkenius@lucs.lu.se

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *19 July 2016*
Accepted: *03 November 2016*
Published: *22 November 2016*

Citation:
Balkenius C and Gärdenfors P (2016) Spaces in the Brain: From Neurons to Meanings. Front. Psychol. 7:1820. doi: 10.3389/fpsyg.2016.01820

1. INTRODUCTION

Within psychology there is considerable evidence that many aspects of human perception and categorization can be modeled by assuming an underlying spatial structure (Shepard, 1987; Gärdenfors, 2000). A paradigmatic example is the color space (Vos, 2006; Renoult et al., 2015), but also, for example, the emotion space (Russell, 1980; Mehrabian, 1996) and musical space (Longuet-Higgins, 1976; Shepard, 1982; Large, 2010) have been extensively studied. Within cognitive linguistics, such spaces are also assumed to be carriers of meaning. For example, Gärdenfors (2000, 2014) has proposed that the semantic structures underlying major word classes such as nouns, adjectives, verbs and prepositions can be analyzed in terms of "conceptual spaces."

For some of the psychological spaces, there exist models that connect neural structures to perception. For example, it is rather well understood how the different types of cones and rods in the human retina result in the psychological color space (see Renoult et al., 2015 for a review). The mammalian brain sometimes represents space in topographic structures. A clear example is the three layers in the superior colliculus for visual, auditory and tactile sensory inputs (Stein and Meredith, 1993). Another example of a topographic representation is the mapping from pitch to position in the cochlea and the tonotopic maps of auditory cortex (Morel et al., 1993; Bendor and Wang, 2005).

For most psychological spaces, however, the corresponding neural representations are not known. Our aim in this article is to investigate the hypothesis that also other representing mechanisms in the brain can be modeled in terms of spatial structures, even if they are not directly mapped onto topographic maps. We present some neuronal representational formats that may generate the psychological spaces. We want to show how psychological spaces naturally emerge from the underlying neural spaces by dimension reduction that preserve similarity structures and

thereby preserve relevant categorizations. In this sense, the psychological and the neural spaces correspond to two different levels of representation.

Furthermore, we argue that spatial representations are fundamental to perception since they naturally support similarity judgments. In a spatial representation, two stimuli are similar to each other if they are close in the space (Hutchinson and Lockhead, 1977; Gärdenfors, 2000). Spatial representations also help generalization since a novel stimulus will be represented close to other similar stimuli in the space, and will thus be likely to belong to the same category or afford the same actions.

One of the main tasks of the brain is to mediate between perception and action (Churchland, 1986; Jeannerod, 1988; Stein and Meredith, 1993; Milner and Goodale, 1995). We argue that this task is supported by spatial representations. When both the sensory input and the motor output use a spatial representation, the task of mapping from perception to action becomes one of mapping between two spaces. To be efficient, spatial representations need to obey some general qualitative constraints on such a mapping. We focus on continuity, monotonicity, and convexity.

In the following section we present some basic psychological spaces and possible connections with neural representations. In Section 3, the role of similarity in psychological spaces, in particular in relation to categorization is presented and conceptual spaces are introduced as modeling tools. Section 4 is devoted to arguing that spatial coding is implicit in neural representations, in particular in population coding. In Section 5, we show how spatial structures are used in mappings between perception and action. Some computational mechanisms, in particular the chorus transform, are discussed in Section 6.

2. BASIC PSYCHOLOGICAL SPACES

We share many psychological spaces with other animals. In this section, we briefly present some of the most basic spaces and outline the representational formats. First and foremost, most animal species have some representations of the external physical space. Even in insects such as bees and ants, one can find advanced systems for navigation (Gallistel, 1990; Shettleworth, 2009). However, the neuro-computational mechanisms that are used vary considerably between species. Mammals have a spatial representation system based on place cells in the hippocampus that are tuned to specific locations in the environment such that the cell responds every time the animal is in a particular location (O'Keefe and Nadel, 1978). This system is complemented by the grid cells in the entorhinal cortex that show more regular firing patterns that are repeated at evenly spaced locations in the environment (Moser et al., 2008). Taken together, the responses of these cells represent a location in space. This code is redundant in the information theoretical sense since many more neurons are used than would be strictly necessary to represent a point in three-dimensional space. One reason for this is that a redundant coding is less sensitive to noise, but it also supports the spatial computations made by the brain as we will see in Section 4.

A second example is the emotion space that is shared with many animal species. Mammals, birds, and other species show clear indications of at least fear, anger and pleasure and there are evolutionarily old brain structures that regulate these emotions and their expressions. For the psychological space of human emotions, there exist a number of models. Many of these models can be seen as extensions of Russell's (1980) two-dimensional circumplex (**Figure 1A**). Here, the emotions are organized along two orthogonal dimensions. The first dimension is valency, going from pleasure to displeasure; the second is the arousal-sleep dimension. Russell shows that the meaning of most emotions words can be mapped on a circumplex spanned by these two dimensions. Other models of psychological emotion space sometimes include a third dimension, for example a "dominance" dimension that expresses the controlling nature of the emotion (Mehrabian, 1996). For example while both fear and anger are unpleasant emotions, anger is a dominant emotion, while fear is non-dominant.

In relation to the topic of this paper, a central question concerns what are the neurophysiological correlates of the psychological emotion space. A recent hypothesis is the three-dimensional emotion cube based on neuromodulators proposed by Lövheim (2012), where the axes correspond to the level of serotonin, dopamine and noradrenaline respectively. By combining high or low values on each of the dimensions, eight basic emotions can be generated. For example, "fear" corresponds to high dopamine, low serotonin and noradrenaline, while "joy" corresponds to high noradrenaline, high serotonin and dopamine (see **Figure 1B**). The mapping between the representation in terms of neurotransmitters and the psychological emotion space remains to be empirically evaluated, but Lövheim's model presents an interesting connection between brain mechanism and the psychological emotion space. Unlike the coding of physical space, this representation has a direct relation between the underlying physiological variables, the transmitter substances, and the psychological emotion space.

A third example of a psychological space that is shared between many species is the color space. The human psychological color space can be described by three dimensions: The first dimension is hue, which is represented by the familiar color circle. The second dimension of color is saturation, which ranges from gray (zero color intensity) to increasingly greater intensities. The third dimension is brightness, which varies from white to black and is thus a linear dimension with end points. There are several models of this human psychological space that differ in some detail concerning the geometric structure, but they are all three dimensional (Vos, 2006).

In other animal species, the psychological color space has only been investigated, via discrimination tasks, for a limited number of species (Renoult et al., 2015). However, the dimensionality of the space varies from one-dimensional (black-white) two-dimensional (in most mammals), three-dimensional (e.g., in primates), to four-dimensional (in some birds and fish). For example, some birds with a four-dimensional space can distinguish between a pure green color and a mixture of blue and yellow, something that most humans cannot (Jordan and Mollon, 1993; Stoddard and Prum, 2008).

The next question then becomes how these various psychological color spaces can be grounded in the

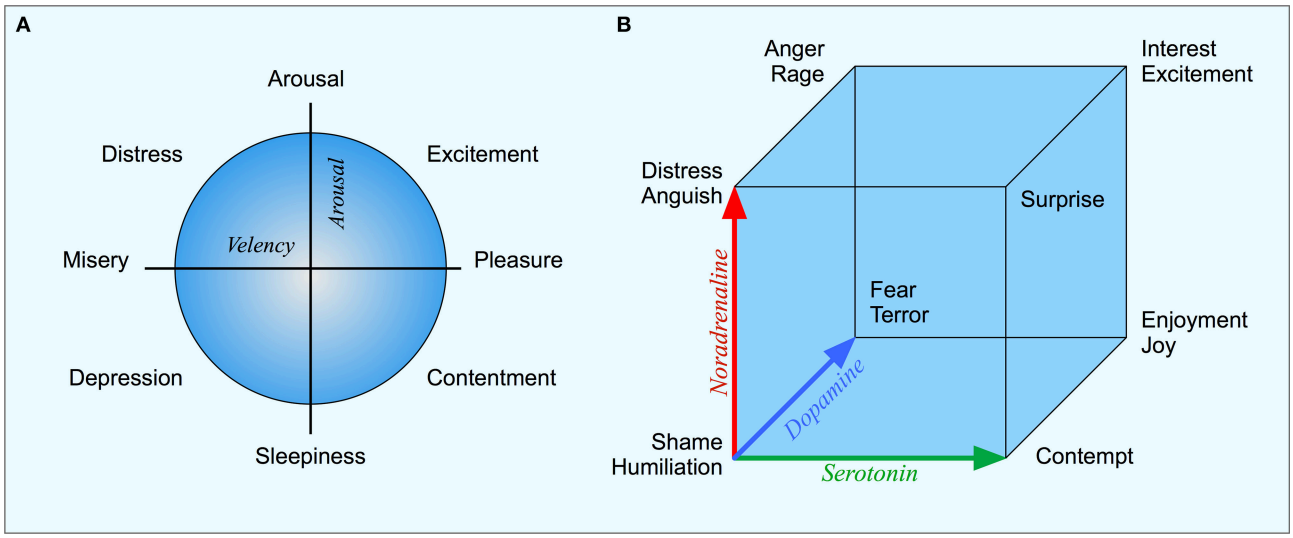


FIGURE 1 | (A) Russell's circumplex with the two basic dimensions of valency and arousal and different emotions arranged in a circular structure. **(B)** Lövheim's emotion cube where the three axes represent the levels of dopamine, noradrenaline, and serotonin respectively.

neurophysiology of the vision systems of different species. The retinas of tri-chromats such as humans have three types of cones that generate color perception: short wavelength (blue), medium wavelength (green) and long wavelength cones (red). Tetra-chromats typically have an additional type of cone that is sensitive to ultra-violet light (Endler and Mielke, 2005). Although every photoreceptor is tuned to a particular wavelength of light, its response intertwines its light intensity with spectral content (Hering, 1964). A change in photoreceptor response can be the results of a change in light intensity as well as a change in color. It is only when the responses of receptors with different tuning are combined that the brain can distinguish between brightness, saturation, and hue.

There exist different theories regarding the connection between the signals from the cones and the rods and the perceived color. One is the opponent-process theory that claims that for tri-chromats there are three opponent channels: red vs. green, blue vs. yellow, and black vs. white. The perceived color is then determined from the differences between the responses of the cones (Hering, 1964). The theory has received support also in several animal species with known tri-chromacy, for examples in primates, fish and bees (see Svaetichin, 1955; De Valois et al., 1958; Backhaus, 1993). For tetra-chromats, a similar theory has been proposed (Endler and Mielke, 2005; Stoddard and Prum, 2008).

It is interesting to note that even though both the receptor space and the psychological color space are both of low dimension, they are not the same. For example, the subjective experience of a color circle has no correspondence in sensory physiology. For humans, the color coded at the receptor level is a cube while the psychological space has the shape of a double cone. None of these spaces are a direct representation of the physical light spectrum.

3. MODELS OF PSYCHOLOGICAL SPACES

3.1. Similarity as a Central Factor

Perhaps the most important cognitive function of the brain is to provide a mapping from perception to action (Milner and Goodale, 1995). In the case of simple reflex mechanism, the mapping is more or less fixed and automatic. In most cases, however, the mapping has to be learned (Schouenborg, 2004) and it is a function not only of the current perception, but also of memory and context (Bouton, 1993). It is central that such a mapping can be learnable in an efficient way. A general economic principle for cognition is that similar perceptions should lead to similar actions. Therefore, similarity should be a fundamental notion when modeling the mapping from perception to action.

In the behavioristic tradition, connections between stimuli and responses were investigated. This research lead to the principle of stimulus generalization that says that, after conditioning, when the subject is presented with a stimulus that is similar to the conditioned stimulus, it will evoke a similar response (Hanson, 1957, 1959). For example, work by Shepard (1957) was seminal in showing that stimulus generalization can be explained in terms of similarity between stimuli. Within this tradition, it was seldom studied what made a stimulus similar to another. What was meant by similarity was taken for granted or induced by varying a physical variable (Nosofsky and Zaki, 2002).

If we leave the behavioristic tradition and turn to more cognitively oriented models, a general assumption is that the connection between stimuli and responses is mediated by a categorization process and that it is the outcome of the categorization that determines the action to be taken. For example, stimuli are categorized as food or non-food, which then determines whether an act of eating will take place.

There exist several psychological category learning models based on similarity. Some are based on forming prototypes of

categories (Rosch, 1975; Gärdenfors, 2000). One way of using prototypes to generate concepts is by Voronoi tessellations (see next subsection) that are calculated by placing any stimulus in the same category as the nearest prototype (Gärdenfors, 2000). Other category learning methods are based on learning a number of exemplars for the different concepts in a domain. (Nosofsky, 1988; Nosofsky and Zaki, 2002). Then a new stimulius is categorized as the same as its nearest neighbor among the exemplar. This is also a technique that is commonly used for pattern recognition in an engineering context (Cover and Hart, 1967).

A general problem for such categorization models is that only for special types of stimuli it is known how the underlying similarity structure can be described. For most stimuli, the modeling will have to be based on hypotheses. The idea that similar perceptions should lead to similar action can, however, be formulated in terms of some general principles that a mapping from perceptions to actions should fulfill. In mathematical terms, the principles can be described as monotonicity, continuity and convexity. Monotonicity means that an increase in a perceptual variable should correspond to an increase in an action variable. For example, if an object B is perceived as being further away than object A, then the agent must reach further to grasp B than to grasp A. Continuity means that small changes in a perceptual variable should correspond to an small change in an action variable. Again eye–hand coordination provides an example: When reaching for an object, the agent makes small adjustments to hand movements in order to adjust for small perceptual discrepancies between hand and object. Convexity means that closed regions of a perceptual space are mapped onto a closed region of action space. To continue with the reaching example, this requirement entails that if object C is located between objects A and B, then the motor signals to reach C should also lie between the motor signals to reach A and to reach B. Even if these three requirements do not determine the mapping from perceptions to actions, they provide strong constraints on such a mapping. The important thing to notice is that once perceptions and actions are spatially represented, a continuous mapping from perception to action typically also fulfills the criteria of monotonicity and convexity. These properties are also important from the perspective of control theory, for example when a robot needs to interpolate between learned movements in novel situations (Schaal and Atkeson, 2010).

Furthermore, when an agent is learning, for example, to coordinate the information from the eyes with the actions of the hands, the fact that the mapping satisfies these conditions potentially makes the learning procedure considerably more efficient. Even with little training, it would be possible to interpolate between already trained mappings from eye to hand and to test a movement that likely is close to the correct one. Under ideal conditions, it is sufficient to have learned how to reach three points on a plane to be able to reach any position on that plane. Other points can be reached by interpolating (or extrapolating) from the movements that reaches each of these three points. Although such interpolation does not necessarily lead to a perfect behavior, it is a good starting point and as more movements are tested the mapping will quickly converge on the correct one.

3.2. Conceptual Spaces

A modeling problem is how psychological and neurological spaces can best be represented. Gärdenfors, 2000 proposes that categories can be modeled as convex regions of a conceptual space. A psychological conceptual space consists of a number of domains such as space, time, color, weight, size, and shape, where each domain is endowed with a particular topology or geometry. Convexity may seem a strong assumption, but it is a remarkably regular property of many perceptually grounded categories, for example, color, taste, and vowels. Although a main argument for convexity is that it facilitates the learnability of categories (Gärdenfors, 2000), it is also crucial for assuring the effectiveness of communication (Warglien and Gärdenfors, 2013). In this article, we focus on the role of convexity in mappings from perception spaces to action space.

There are interesting comparisons to make between analyzing categories as convex regions and the prototype theory developed by Rosch and her collaborators (Rosch, 1975; Mervis and Rosch, 1981; Lakoff, 1987). When categories are defined as convex regions in a conceptual space, prototype effects are to be expected. Given a convex region, one can describe positions in that region as being more or less central. Conversely, if prototype theory is adopted, then the representation of categories as convex regions is to be expected. Assume that some conceptual space is given, for example, the color space; and that the intention is to decompose it into a number of categories, in this case, color categories. If one starts from a set of categories prototypes—say, the focal colors—then these prototypes should be the central points in the categories they represent. The information about prototypes can then be used to generate convex regions by stipulating that any point within the space belongs to the same categories as the *closest* prototype. This rule will generate a certain decomposition of the space: a so-called Voronoi tessellation (see **Figure 2**). The illustration of the tessellation is two-dimensional, but Voronoi tessellations can be extended to any arbitrary number of dimensions. An important feature of Voronoi tessellations is that they always generate a convex partitioning of the space.

The prototype structure of concepts is also central for modeling meanings of words. Gärdenfors (2000, 2014) develops a semantic theory where elements from the main word classes are mapped onto convex regions of domain or convex sets of vectors over a domain. This way of representing word meanings can explain many features of how children learn their first language. Again, the low-dimensional structure of the domains are essential for rapid learning of new word meanings (Gärdenfors, 2000).

4. SPATIAL CODING IS IMPLICIT IN NEURAL REPRESENTATIONS

We next turn to a more general account of how space may be neurally represented. We suggest that a spatial coding is implicit in most neural mechanisms, and that concepts of distances and betweenness are readily applicable to such codes.

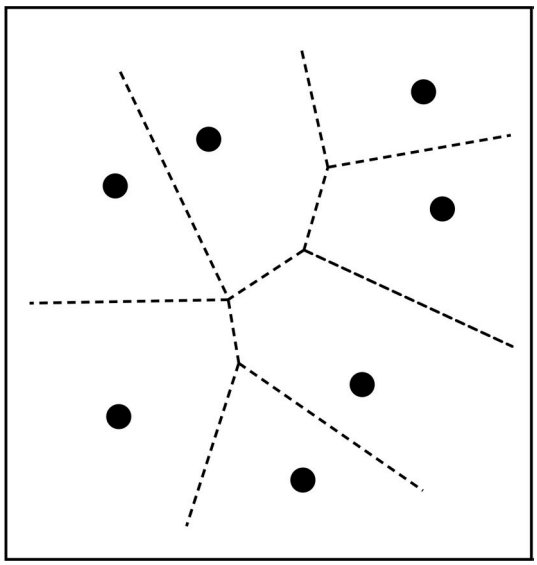


FIGURE 2 | Voronoi tessellation of the plane into convex categories. Each point represents a prototype and the lines show the borders between the categories.

As a first example, we look at the neurons in motor cortex. These neurons code for the direction of movement using a *population code* where each individual neuron is tuned to movement in a particular direction (Georgopoulos et al., 1988) and modulated by distance (Fu et al., 1993). In a population code, a stimulus or a motor command is coded by the joint activities of a set of neurons. Before the movement, the response of each cell is proportional to the angle between the direction vector represented by that cell and the direction of the following movement. Cells with vectors close to the movement direction will respond more than cells that code for different movement directions.

The set of neurons can be seen as a basis for a highly redundant high-dimensional coding of a low-dimensional vector space for movement direction. The responses of all neurons taken together represent a population vector that can be computed by adding together the direction vectors of each individual neuron weighted by its response magnitude (**Figures 3A,B**). The population vector is thus the low-dimensional "decoding" of the high-dimensional population code.

The similarity between two population codes can be calculated by considering the population codes as vector in the high-dimensional space. The similarity is defined by the cosine of the angle between these vectors. This similarity measure varies between 0 and 1, where 1 indicates identical population codes, and a value of 0 indicates two maximally dissimilar codes. This is different from calculating the similarity between the population vectors that lie in the low-dimensional space. A fundamental aspect of the population coding is that population codes that are similar using this measure in the high-dimensional neural space will produce population vectors that are also similar in the low-dimensional movement space.

Population codes are not only used for motor coding but are also used for perceptual tasks. In their seminal study of population coding of human faces in the anterior inferotemporal cortex (AIT) and anterior superior temporal polysensory area (STP) of macaque monkeys, Young and Yamane (1992) showed that the recorded responses of these brain regions contain information about the identity (AIT) and possibly familiarity (STP) of the faces the monkey is viewing. The responses of a large number of neurons to different faces were recorded. Using multidimensional scaling, they mapped the recordings onto a lower dimensional space. The dimensions of this space are not visible when looking at a single neuron that responds preferably to a single stimulus and gradually decreases its response as the stimulus moves away from the preferred one. However, by looking at the low-dimensional code, they were able to show that two dimensions explained most of the variation in the population code for each of the two brain regions.

This implies that the macaque brain implicitly uses a low-dimensional space to code different faces. Although a high-dimensional population code is used, most of the information is contained in a small number of dimensions. Each face is coded in a unique location in this space, and faces coded close to each other in the space share visual characteristics such as the amount of hair and the general shape of the face. The distance between points in this low-dimensional space represents the similarity between the coded faces (**Figure 3C**) and may correspond to the psychological face space of the monkey.

For both examples of population coding described above, the underlying space appears to be two-dimensional, but this is clearly an artifact of the experimental details. In Georgopoulos' experiments, the monkey moved its arm in two dimensions, the vectors found are consequently also two-dimensional, but we must assume that the same principle holds for movement in three dimensions and possibly also for more complex movements that are extended in time and involves more degrees of freedom (Graziano et al., 2002). The only difference in this case is that a larger number of dimensions are necessary. Similarly, in the experiment by Young and Yamane, two dimensions were sufficient to capture most of the variation necessary to distinguish the different faces, but presumably, the monkey could have access to more dimensions had it been necessary to differentiate between the faces. The exact number of dimensions in neural representation is not important as long as a low-dimensional reduction of the space covers most of the information.

There are two ways to view the coding in the brain—one at a detailed level, the other at an aggregated level. The first is to look at each neuron individually. By systematically testing different stimuli, it is possible to find the stimulus that each neuron maximally responds to (Tanaka, 2003). In this case, the neuron is considered a detector tuned to that preferred stimulus. The preferred stimulus can be seen as the *prototype* for that neuron (Edelman and Shahbazi, 2012), and the more similar a stimulus is to that prototype, the stronger the neuron will react.

The other approach is to look at the whole population of neurons and view the activity pattern as a point in a high dimensional space. In this case, the response of each neuron is

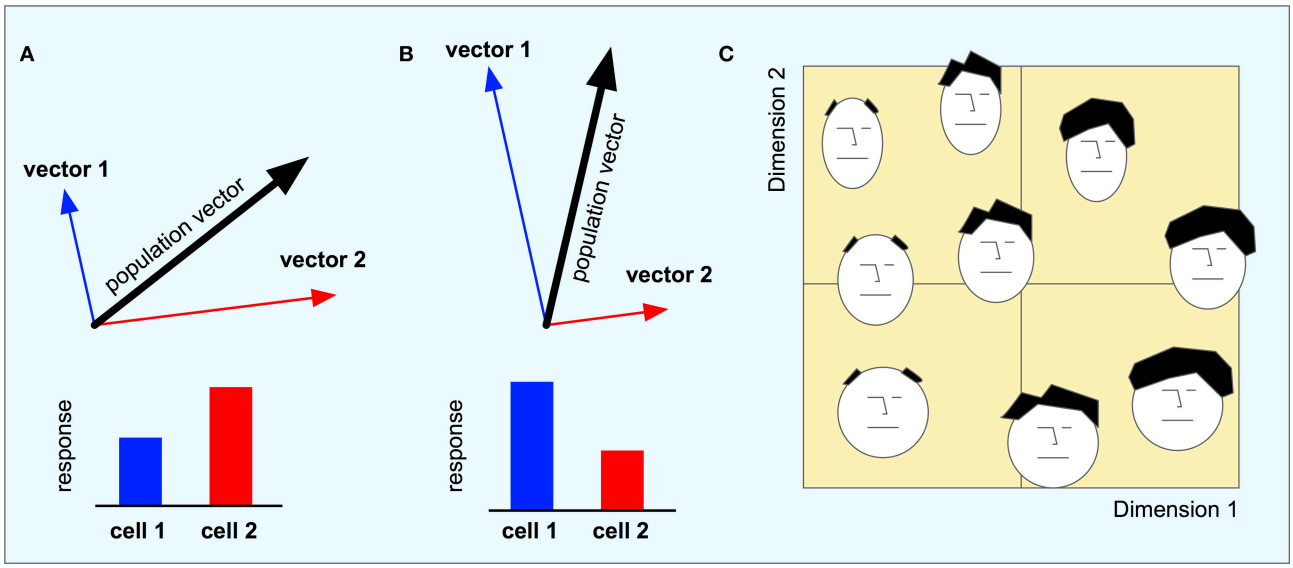


FIGURE 3 | (A,B) The direction vectors from a minimal population of two cells are combined into a population vector. Each cell codes for a particular movement direction and the responses of the two cells weigh together the two vectors into a combined population vector that corresponds to the subsequent movement direction. When the response of cell 2 is higher than that of cell 1, the population vector will point in a direction that is closer to vector 2 than that of vector 1. When cell 1 has the higher response, the population vector will be more aligned with vector 1. **(C)** A hypothetical face space of the type found by Young and Yamane (1992). Different faces are arranged along two dimensions where faces that are similar to each other are located close to each other in the space.

seen as a basis function and every stimulus is coded as a blend of these *basis functions*[1]. In this case, the responses of individual neurons are not necessarily meaningful on their own. Although these two views may look contrasting, they are actually two sides of the same coin and are both equally valid.

Although a population code consists of the activity of multiple neurons that are not necessarily located close to each other on the cortical surface, Erlhagen and Schöner (2002) have suggested that neurons that make up a stable activity pattern may be linked by mutual excitation in such a way that they functionally can be considered a point in a higher dimensional topographic space. This is a central component of the Dynamic Field Theory that studies the temporal dynamics of such activity patterns.

We have here looked at how low dimensional spaces are implicitly coded in a high dimensional population code, but the brain also constructs lower dimensional codes explicitly throughout the sensory system. This is often modeled as successive steps of dimensionality reduction, or compression, in hierarchical networks (e.g., Serre et al., 2007). In the semantic pointer architecture (Eliasmith, 2013), relatively low dimensional codes that are constructed in this way are used to define a "semantic space" where different concepts can be represented. The high dimensional representation at lower levels in the hierarchy can be partially reconstructed from the low-dimensional semantic pointer. Furthermore, the architecture allows for recursive binding through the operation of circular convolution. Unlike earlier methods using tensor operations (Smolensky, 1990), circular convolution does not increase the dimensionality of the representation and can be performed in several successive steps to produce deep embeddings (Blouw et al., 2015). Many other forms of binding mechanisms are discussed by van der Velde and De Kamps (2006). Common to all are that the individual constituents can have the spatial structure described above.

5. THE USE OF SPACES AS MAPPINGS BETWEEN PERCEPTION AND ACTION

We next turn to neuro-cognitive models that include both the sensory and the motor side. Specifically, we want to show that sensory-motor mappings can be described as mappings between points in low-dimensional spaces, Here we only consider basic examples of sensory-motor mappings, but the principles we present are general.

A direct form of sensory to motor mapping is used when we keep our head stationary and let the eyes saccade to an object. The location of the object is captured in eye-centered coordinates and it is necessary to convert these into the appropriate motor commands to move the eyes to that location. This sensory-motor transformation can be seen as a mapping between two representational spaces, one for the object location and one for the movements of the eyes.

For a saccade, the mapping is relatively simple since every location on the retina could in principle be mapped to a unique motor command (Salinas and Abbott, 1995). When a target is detected on the retina, a motor command would be produced that would point the eye in the correct direction. There would thus be one eye direction vector for each retinal position. Here,

[1]Basis functions are elementary function that can be linearly combined to produce more complex functions in a particular function space.

the desired gaze direction is a function of the target location on the retina and the function satisfies the three conditions of monotonicity, continuity and convexity. As the target moves further away from the center of the eye, the required movement is larger and the mapping is thus monotone. A small change in the target position on the retina, requires a small change of the corresponding movement. The mapping is thus continuous. It also follows that the convexity criteria is met since the movements are mapped out in an orderly fashion over the retina. The movement to a target that is projected between two arbitrary points on the retina lies somewhere in-between the two movements required for the two points.

Deneve and Pouget (2003) suggested that mappings from sensory to motor systems could be performed using basis function maps. Such maps use basis functions to represent all possible stimuli in such a way that linear combinations of basis functions can compute any motor command. More specifically, they propose that the neurons of the supplementary eye field of the parietal cortex form a set of basis units. Each basis unit corresponds to a single prototype in the sensory space. The output from the unit codes the distance from the input to that prototype. The task for the subjects in their experiment was to saccade to the left or the right part of an object that appeared at an arbitrary location and orientation on the retina. This task is interesting since it requires that the whole object is identified before it is possible to localize its sides and thus it appears to require object-centered representations and as such a sequence of coordinate transformation would be necessary. However, Deneve and Pouget showed that this task can be performed as a single mapping by a three layer network where the middle layer consists of basis function units (**Figure 4**). The basis units work together so that inputs that match several basis units will produce an output that is a combination of the outputs from each of the individual basis units. A finite set of basis units can thus cover the whole input and output spaces. The responses of all these basis units together constitute a population code (Pouget and Snyder, 2000).

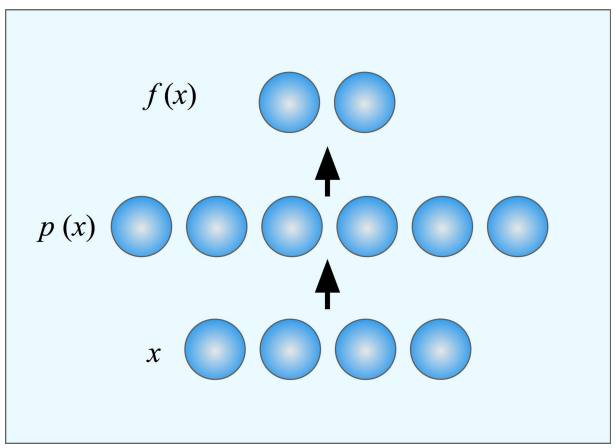


FIGURE 4 | A three-layer neural network. The input x is mapped onto a population code $p(x)$ in a hidden layer where each neuron (or basis unit) is tuned to different position in the input space (its prototype). The output function $f(x)$ is computed by weighting together the responses of the units in the hidden layer. Different functions can be computed by weighting the outputs differently. Learning in the network consists of finding the appropriate weights for the desired function.

Since we know that the input can be described by a small number of variables (here the position and orientation of the object combined with the instruction to look at the left or right part of it), it is clear that the responses of the basis function units implicitly codes for this low-dimensional space. Similarly, the output is a point in a two dimensional space consisting of the possible targets for the saccade. We can thus interpret the operation of the network as a mapping from a four-dimensional to a two-dimensional space, although the computations are made implicitly in a high-dimensional space as a linear combination of basis unit responses.

Similar models have been proposed to explain the sensory-motor transformations necessary to reach for a visually identified object (Zipser and Andersen, 1988). For example, to point at a visual target the brain needs to take into account the position of the target on the retinas of the two eyes, the orientation of the head and eyes and the posture of the body. To compute the location of the target relative to the hand, the target must first be identified on the retina and then it is necessary to compensate for the location of the eyes relative to the hand and the rest of body. This can be viewed as a sequence of coordinate transformations, but it is also possible that, like in Deneve and Pouget's (2003) model, the target location could be found in a single step by mapping from a space coding retinal position and the positions of all the relevant joints. In either case, these computations can be made as mappings between population codes in different layers of a network (Zipser and Andersen, 1988; Eliasmith and Anderson, 2003).

The relative roles of retinal target position and joint angles can be seen in an experiment by Henriques et al. (2003). The experiment showed that reaching is easier when we look directly at the target compared to when the target is off-gaze. This indicates that the orientation of the eyes has a larger influence on the movement than the retinal position of the target and supports the idea that joint position are used in computations of spatial locations. The result is probably a consequence of the fact that we most often look directly at an object we try to reach.

In the brain, the mapping from the retinal position and eye direction to the external target location that controls reaching movements is believed to take place in the posterior parietal cortex (PPC) (Jeannerod, 1997). Zipser and Andersen (1988) looked at the responses of the neurons in area 7a of PPC and trained an artificial neural network on the mapping from eye direction and retinal position to head-centered coordinates. The network consisted of three layers where units in the first layer code for retinal position and eye direction. The activity of the output layer indicated the head-centered location of the target. The model produced similar response properties as the real neurons of PPC. The neurons in the hidden layer became tuned to retinal position but they are also modulated by eye position. Like the saccade control described above, this learned mapping fulfills the criteria of monotonicity, continuity and convexity.

The type of population coding that is found in area 7a, and that also emerges in the hidden layer of the model, is often called a *gain field* (Zipser and Andersen, 1988; Buneo and Andersen, 2006). Like other types of population coding, different neurons take care of different parts of the mapping and the final result is obtained by weighing together the contributions of each neuron (**Figures 3**, **4**). However, a gain field is characterized by the fact that the neurons are primarily organized along only some of the input dimensions. For example, Zipser and Andersen (1988) found that neurons coding for target position were retinotopically tuned, but responded differently depending on the eye positions.

Salinas and Abbott (1995) also addressed the question of how the brain can transfer information from sensory to motor system using population codes. They investigated the coordinate transformations in visually guided reaching and proposed a model that uses a Hebbian learning mechanism to learn the sensory-motor mapping. Unlike Zipser and Andersen, they assume that the input space is already covered by a large set of prototypes coded by a set of basis units. They view the problem of eye–hand coordination as a form of function approximation where the problem is to find the appropriate weights for the outputs of each basis unit to obtain the desired mapping (**Figure 4**). They show how these weights can be learned using Hebbian learning and the general model they present can be used to describe arbitrary mappings between spaces.

We now turn to the slightly more complicated situation where reaching is followed by grasping an object. Here, we not only need to locate the target, we also need to shape the hand in the correct way both before reaching the object and subsequently to grasp it. Despite the added complexity, this too can be seen as a mapping between two spaces. In the case that only visual information is used, the input space codes for the location and the shape of the object while the output space minimally contains the movement direction for reaching, the parameters to preshape the hand and finally the force vectors to perform the grasping movement.

Although little is known about how shapes are represented in the brain, work in mathematics (Kendall, 1984) and computer graphics (Blanz and Vetter, 1999; Kilian et al., 2007) show that it is possible to design shape spaces where different shapes can be synthesized from combinations of basic shapes in a way that is similar to how basis units work together to represent a point in a space using a population code. For a known rigid object however, it is sufficient to code the orientation of the object. This can be done in a three dimensional space of the rotation angles that describes the orientation of the object[2]. The orientation can be represented in a way similar to position by a set of basis units that together code for all possible orientations of the object. Here the final mapping is between a six-dimensional space representing position and orientation to a space that represents the critical parameters of the reaching movement.

There are a number of spaces that could potentially be involved in eye–hand coordination. Depending on task constraints, the brain is thought to use both egocentric and allocentric representations of space (Crawford et al., 2004) and there is evidence that neurons in PPC code for targets in relation to both gaze (Batista et al., 1999) and hand (Buneo and Andersen, 2006). Investigating spatial representations for reaching in the superior parietal lobule (SPL) of PPC, Buneo and Andersen (2006) found evidence for representations of targets in both eye-centered coordinates and of the difference in position between the hand and target.

Figure 5 summarizes some of the coordinate systems involved in eye–hand coordination. The target object can be represented in either allocentric or one of the egocentric spaces. For reaching, an egocentric frame of reference is more suitable but as we have seen, there are several egocentric spaces corresponding at least to the eye and the hand, but it is likely that many more exist and presumably the brain is able to map freely between them. To grasp an object, its representation must be mapped on the space that contains the possible motor actions. The dimensionality of this space is high enough to contain all possible grasp movements, but still of limited dimensionality. For the brain to learn these mapping in an efficient way, it is necessary that where possible, these mappings fulfill the three conditions of monotonicity, continuity and convexity. We submit that population codes are used to make this possible.

6. MECHANISMS

Population codes of spaces as described above are instances of a coding scheme where each input is coded by the distance to a number of prototypes. The optimal stimulus for each neuron, or basis unit, in the population can be considered the prototype for that unit. One such form of population coding is given by the *chorus transform* proposed by Edelman (1999) who calls it a

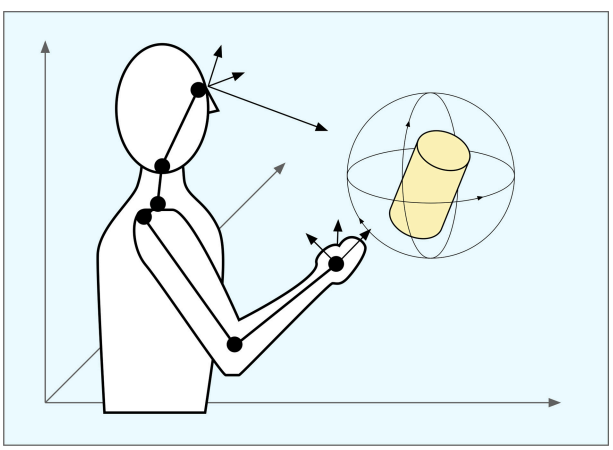


FIGURE 5 | Coordinate systems in eye–hand coordination. The position of an object can be coded in in relation to allocentric space, the eye or the hand and potentially also other parts of the body. To reach and grasp the object, the brain must map representations of the object from a sensory space to a motor space in order to control the arm and hand in an appropriate way.

[2]Euler angles are generally problematic as orientation representations, in particular because they do not fulfill the three criteria of monotonicity, continuity and convexity. However, there exist other representations, such as quarternions, that do.

"chorus of prototypes." In the simplest case, the response of each unit is the similarity to the prototype measured by some suitable metric. **Figure 6** shows an example of a chorus transform. The input to the transform is an image of a face. This face is compared to each of five face prototypes and the resulting transform is the set of similarity measurements. The chorus transform thus describes a type of population coding of the input.

An important property of the chorus transform is that it preserves Voronoi tessellations of the input space (Edelman, 1999, p. 268). It also approximately preserves the inter-point distances in the original space. This means that category boundaries in the input space are mostly preserved in the output space (Edelman and Shahbazi, 2012). This has several critical consequences for both neural coding and mappings of spatial representations:

Stimulus generalization. The spatial representation naturally supports *generalization* since novel stimuli will be coded by the similarity to known stimuli and the coding will gradually change if the stimulus gradually changes. There is thus a continuous mapping from stimulus properties to the representation of the stimulus.

Discrimination. Since discrimination borders are typically Voronoi borders and these are mostly preserved by the coding, it means that discrimination borders in the input space are preserved in the coding space.

Categorization. For the same reason, categories induced by the Voronoi tessellation are preserved in the population coding.

When looking at mapping between two spaces coded by a population of units we note that these properties of the chorus transform imply that a linear mapping from such a representation also have these properties. This has consequences for sensory-motor mapping between spaces:

Interpolation. Since similar stimuli are coded by similar population codes, similar stimuli will be mapped to similar motor outputs. This is equivalent to the continuity criteria introduced above.

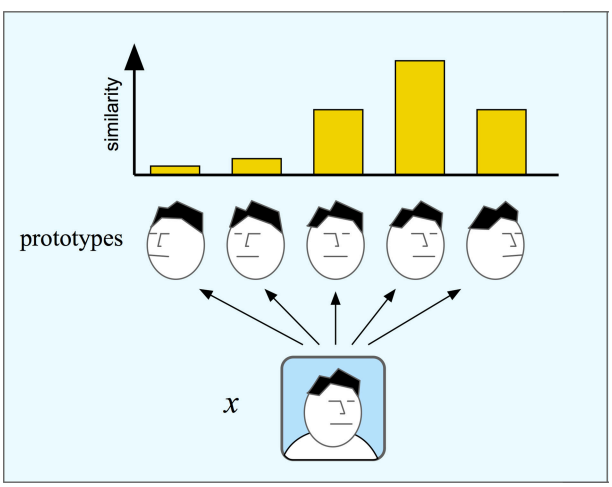


FIGURE 6 | In the chorus transform (and in RBF networks) an input is mapped to its similarity to a number or prototypes. Here, an image x is compared to five face prototypes with different orientation. Each comparison produces one component of the chorus transform that will here consist of five similarity measures.

Sensory-motor categories. When object categories are represented as Voronoi borders in the stimulus space, different stimulus categories can be mapped to different motor actions by a single mapping. When the input crosses the Voronoi border between two categories, so will the output, and in the same way a small part of the input space represents a particular category, a corresponding part of the output space can represent actions suitable for that category. Furthermore, since Voronoi borders are preserved in the mapping it follows that the convexity criteria is also fulfilled by these types of mappings.

The most commonly used computational architecture that uses the chorus transform is the *radial basis function* (RBF) network (Moody and Darken, 1989). This artificial neural network consists of three layers (**Figure 4**). The middle layer consists of units that are tuned to different stimulus prototypes and their response is maximal when the input is identical to the prototype. The prototypes can be selected in different ways. One possibility is to use one prototype for each exemplar that has been encountered. Alternatively, the prototypes can be selected by trying to cover the input space by suitably spaced prototypes. Finally, the prototypes can be found by learning. Once selected, each unit in the middle layer contributes to the output depending on how well the input matches its prototype. RBF-networks have been used in numerous applications for both categorization and function approximation tasks and can easily learn complex relations between their input and output.

A type of RBF-network that is of special interest is normalizing radial basis function networks (Bugmann, 1998). This model differs from the standard model in that the output is normalized. This is an operation that has been suggested to be implemented by lateral inhibition and it is ubiquitous in the brain (Carandini and Heeger, 2012). The importance of the normalization stage is that it makes the output of the RBF-network consist of a convex combination of the outputs of the individual units, where each output is weighed by how close its prototype is to the input. This property guarantees that the output will be a continuous function of the input that quickly converges on the correct mapping during learning. Like other multilayer feed-forward networks, RBF-networks are universal approximators, which means that they can learn any mapping between finite dimensional spaces with any desired accuracy as long as there are a sufficient number of hidden units.

Salinas and Abbott (1995) investigated how the number of units influenced the accuracy of the coding and decoding of different magnitudes. More recently, Eliasmith and Anderson (2003) presented general mathematical recipes for how low dimensional quantities can be coded and decoded in the brain using population codes as well as suggestions for how such quantities can be combined in different ways to implement different arithmetic operations and mappings between spaces. The same type of reasoning can be applied to many different tasks including other sensorimotor transformations, learning and short-term memory (Pouget and Snyder, 2000). This shows that population codes are a general way to code quantities

in one- or multi-dimensional spaces and to perform arbitrary operations on them. This further supports that the machinery available to the brain is ideally suited for processing spatial representations.

It is interesting to note that the basis unit coding has many similarities to approaches in control theory, such as locally weighted learning (Atkeson et al., 1997). In fact, Stulp and Sigaud (2015) have shown that many models and algorithms working according to these principles use exactly the same underlying model as the three-layer network described above. This lends support to the idea that there is something fundamental about these types of mechanisms where functions are computed using units that each react to different parts of the input space and the output is subsequently calculated as a combination of the outputs from those individual units.

To summarize, we have proposed that mappings between spaces consist of two steps. The first is a comparison between the input and a number of prototypes and the second is the weighting of the output from each prototype unit by its similarity to the input. The chorus transform provides a good model for the usefulness of population codes, both as a way to represent points in psychological spaces and as a mechanism for mapping between such spaces. RBF-networks constitute the canonical way to model learning of such mappings, but many other models are possible.

7. CONCLUSION

This article has treated two levels of spaces in the brain—psychological and neurocognitive. The psychological spaces, for example the color space, can be studied in psychophysical experiments, in particular with the aid of discrimination or similarity judgments. These spaces can often be represented in a small number of dimensions and we have shown how conceptual spaces can be used to model categorization processes. Neurocognitive representations are implemented implicitly using populations coding where different neurons process different regions of the spaces and allow for efficient mappings between spaces. Furthermore, spatial coding naturally supports generalization from learned examples by interpolation and extrapolation. We have also argued that the psychological spaces naturally emerge from the neural codings.

Although there exist examples of topographic representations in the brain, the spatial representations are typically not topographically organized. This is not even the case for the representations of physical space in the hippocampus. Instead, a population code is used to implicitly represent the spaces. The main advantage of this is that it allows the brain to potentially learn any functional mapping and not only those that can be represented by mappings between two-dimensional spaces.

The main function of spatial representations is to make the mapping from perception to action more efficient. Many models of computations with population codes use explicit representations of perceptual and motor variables. This is useful when investigating what the model is doing, but does not mean that we should expect to find such neurons in the brain, where there is no need to decode the population codes until the final stage when they are used to produce movements. To reveal the low-dimensional spatial representations and to make them match the psychological results, it is necessary to decode the population codes in a low-dimensional space but such a decoding is never explicitly required by the brain itself.

By analyzing the neural representations and reducing them to low-dimensional representations, we have argued that they to a large extent can explain the structure of the psychological spaces. Moreover, we have shown how spatial representations are useful as a basis for categorization and sensory-motor mappings and how they can be implicitly coded by populations of neurons. This suggests that spatial representations can be found everywhere in the brain.

AUTHOR CONTRIBUTIONS

CB and PG contributed equally to the reported research and the writing of the article.

REFERENCES

Atkeson, C. G., Moore, A. W., and Schaal, S. (1997). Locally weighted learning for control. *Artif. Intell. Rev.* 11, 75–11. doi: 10.1023/A:1006511328852

Backhaus, W. (1993). Color vision and color choice behavior of the honey bee. *Apidologie* 24, 309–309. doi: 10.1051/apido:19930310

Batista, A. P., Buneo, C. A., Snyder, L. H., and Andersen, R. A. (1999). Reach plans in eye-centered coordinates. *Science* 285, 257–260. doi: 10.1126/science.285.5425.257

Bendor, D., and Wang, X. (2005). The neuronal representation of pitch in primate auditory cortex. *Nature* 436, 1161–1165. doi: 10.1038/nature03867

Blanz, V., and Vetter, T. (1999). "A morphable model for the synthesis of 3D faces," in *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, (Boston, MA: ACM Press/Addison-Wesley Publishing Co.), 187–194. doi: 10.1145/311535.311556

Blouw, P., Solodkin, E., Thagard, P., and Eliasmith, C. (2015). Concepts as semantic pointers: a framework and computational model. *Cogn. Sci.* 40, 1128–1162. doi: 10.1111/cogs.12265

Bouton, M. E. (1993). Context, time, and memory retrieval in the interference paradigms of Pavlovian learning. *Psychol. Bull.* 114:80. doi: 10.1037/0033-2909.114.1.80

Bugmann, G. (1998). Normalized Gaussian radial basis function networks. *Neurocomputing* 20, 97–110. doi: 10.1016/S0925-2312(98)00027-7

Buneo, C. A., and Andersen, R. A. (2006). The posterior parietal cortex: sensorimotor interface for the planning and online control of visually guided movements. *Neuropsychologia* 44, 2594–2606. doi: 10.1016/j.neuropsychologia.2005.10.011

Carandini, M., and Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nat. Rev. Neurosci.* 13, 51–62. doi: 10.1038/nrn3136

Churchland, P. M. (1986). Cognitive neurobiology: a computational hypothesis for laminar cortex. *Biol. Philos.* 1, 25–51. doi: 10.1007/BF00127088

Cover, T. M., and Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* 13, 21–27. doi: 10.1109/TIT.1967.1053964

Crawford, J. D., Medendorp, W. P., and Marotta, J. J. (2004). Spatial transformations for eye–hand coordination. *J. Neurophysiol.* 92, 10–19. doi: 10.1152/jn.00117.2004

Deneve, S., and Pouget, A. (2003). Basis functions for object-centered representations. *Neuron* 37, 347–359. doi: 10.1016/S0896-6273(02)01184-4

De Valois, R. L., Smith, C. J., Karoly, A. J., and Kitai, S. T. (1958). Electrical responses of primate visual system. I. Different layers of macaque lateral geniculate nucleus. *J. Comp. Physiol. Psychol.* 51, 662–668.

Edelman, S. (1999). *Representation and Recognition in Vision*. Cambridge, MA: MIT Press.

Edelman, S., and Shahbazi, R. (2012). Renewing the respect for similarity. *Front. Comput. Neurosci.* 6:45. doi: 10.3389/fncom.2012.00045

Eliasmith, C., and Anderson, C. H. (2003). *Neural Engineering: Computation, Representation, and Dynamics in Neurobiological Systems*. Cambridge, MA: MIT Press.

Eliasmith, C. (2013). *How to Build a Brain: A Neural Architecture for Biological Cognition*. Oxford: Oxford University Press.

Endler, J. A., and Mielke, P. W. (2005). Comparing entire colour patterns as birds see them. *Biol. J. Linnean Soc.* 86, 405–431. doi: 10.1111/j.1095-8312.2005.00540.x

Fu, Q. G., Suarez, J. I., and Ebner, T. J. (1993). Neuronal specification of direction and distance during reaching movements in the superior precentral premotor area and primary motor cortex of monkeys. *J. Neurophysiol.* 70, 2097–2116.

Gallistel, C. R. (1990). *The Organization of Learning*. Cambridge, MA: MIT Press.

Gärdenfors, P. (2000). *Conceptual Spaces: The Geometry of Thought*. Cambridge, MA: MIT Press.

Gärdenfors, P. (2014). *Geometry of Meaning: Semantics Based on Conceptual Spaces*. Cambridge, MA: MIT Press.

Georgopoulos, A. P., Kettner, R. E., and Schwartz, A. B. (1988). Primate motor cortex and free arm movements to visual targets in three-dimensional space. II. Coding of the direction of movement by a neuronal population. *J. Neurosci.* 8, 2928–2937.

Graziano, M. S. A., Taylor, C. S. R., Moore, T., and Cooke, D. F. (2002). The cortical control of movement revisited. *Neuron* 36, 349–362. doi: 10.1016/S0896-6273(02)01003-6

Hanson, H. M. (1957). Discrimination training effect on stimulus generalization gradient for spectrum stimuli. *Science* 125, 888–889. doi: 10.1126/science.125.3253.888

Hanson, H. M. (1959). Effects of discrimination training on stimulus generalization. *J. Exp. Psychol.* 58, 321. doi: 10.1037/h0042606

Henriques, D. Y. P., Medendorp, W. P., Gielen, C. C. A. M., and Crawford, J. D. (2003). Geometric computations underlying eye-hand coordination: orientations of the two eyes and the head. *Exp. Brain Res.* 152, 70–78. doi: 10.1007/s00221-003-1523-4

Hering, E. (1964). *Outlines of a Theory of the Light Sense*. Cambridge, MA: Harvard University Press.

Hutchinson, J., and Lockhead, G. R. (1977). Similarity as distance: a structural principle for semantic memory. *J. Exp. Psychol. Hum. Learn. Mem.* 3, 660–678. doi: 10.1037/0278-7393.3.6.660

Jeannerod, M. (1988). *The Neural and Behavioural Organization of Goal-Directed Movements*. Oxford: Clarendon Press/Oxford University Press.

Jeannerod, M. (1997). *The Cognitive Neuroscience of Action*. Oxford: Blackwell Publishing.

Jordan, G., and Mollon, J. D. (1993). A study of women heterozygous for colour deficiencies. *Vis. Res.* 33, 1495–1508. doi: 10.1016/0042-6989(93)90143-K

Kendall, D. G. (1984). Shape manifolds, procrustean metrics, and complex projective spaces. *Bull. Lond. Math. Soc.* 16, 81–121. doi: 10.1112/blms/16.2.81

Kilian, M., Mitra, N. J., and Pottmann, H. (2007). Geometric modeling in shape space. *ACM Trans. Graph.* 26:64. doi: 10.1145/1276377.1276457

Lakoff, G. (1987). *Women, Fire, and Dangerous Things*. Chicago, IL: The University of Chicago Press.

Large, E. (2010). "Neurodynamics of music," in *Music Perception*, eds M. Riess Jones, R. R. Fay, and A. N. Popper (Berlin: Springer), 201–231.

Longuet-Higgins, H. C. (1976). Perception of melodies. *Nature* 263, 646–653. doi: 10.1038/263646a0

Lövheim H. (2012). A new three-dimensional model for emotions and monoamine neurotransmitters. *Med. Hypoth.* 78, 341–348. doi: 10.1016/j.mehy.2011.11.016

Milner, A. D., and Goodale, M. A. (1995). *The Visual Brain in Action*. Oxford: Oxford University Press.

Mehrabian, A. (1996). Pleasure-arousal-dominance: a general framework for describing and measuring individual differences in temperament. *Curr. Psychol.* 14, 261–292. doi: 10.1007/bf02686918

Mervis, C., and Rosch, E. (1981). Categorization of natural objects. *Annu. Rev. Psychol.* 32, 89–115. doi: 10.1146/annurev.ps.32.020181.000513

Moody, J., and Darken, C. J. (1989). Fast learning in networks of locally tuned processing units. *Neural Comput.* 1, 281–294. doi: 10.1162/neco.1989.1.2.281

Morel, A., Garraghty, P. E., and Kaas, J. H. (1993). Tonotopic organization, architectonic fields, and connections of auditory cortex in macaque monkeys. *J. Comp. Neurol.* 335, 437–459. doi: 10.1002/cne.903350312

Moser, E. I., Kropff, E., and Moser, M. B. (2008). Place cells, grid cells, and the brain's spatial representation system. *Neuroscience* 31, 69–89. doi: 10.1146/annurev.neuro.31.061307.090723

Nosofsky, R. M. (1988). Similarity, frequency, and category representations. *Mem. Cogn.* 14, 54–65. doi: 10.1037/0278-7393.14.1.54

Nosofsky, R. M., and Zaki, S. R. (2002). Exemplar and prototype models revisited: response strategies, selective attention, and stimulus generalization. *J. Exp. Psychol. Learn. Mem. Cogn.* 28, 924–940. doi: 10.1037/0278-7393.28.5.924

O'Keefe, J., and Nadel, L. (1978). *The Hippocampus as a Cognitive Map*. Oxford: Clarendon Press.

Pouget, A., and Snyder, L. H. (2000). Computational approaches to sensorimotor transformations. *Nat. Neurosci.* 3, 1192–1198. doi: 10.1038/81469

Renoult, J. P., Kelber, A., and Schaefer, H. M. (2015). Colour spaces in ecology and evolutionary biology. *Biol. Rev.* doi: 10.1111/brv.12230. [Epub ahead of print].

Rosch, E. (1975). Cognitive representations of semantic categories. *J. Exp. Psychol. Gen.* 104, 192–233. doi: 10.1037/0096-3445.104.3.192. [Epub ahead of print].

Russell, J. A. (1980). A circumplex model of affect. *J. Personal. Soc. Psychol.* 39, 1161–1178. doi: 10.1037/h0077714

Salinas, E., and Abbott, L. F. (1995). Transfer of coded information from sensory to motor networks. *J. Neurosci.* 15, 6461–6474.

Serre, T., Oliva, A., and Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proc. Natl. Acad. Sci. U.S.A.* 104, 6424–6429. doi: 10.1073/pnas.0700622104

Schaal, S., and Atkeson, C. G. (2010). Learning control in robotics. *IEEE Robot. Autom. Mag.* 17, 20–29. doi: 10.1109/MRA.2010.936957

Shepard, R. N. (1957). Stimulus and response generalization: a stochastic model relating generalization to distance in psychological space. *Psychometrica* 22, 325–345. doi: 10.1007/BF02288967

Shepard, R. N. (1982). Geometrical approximations to the structure of musical pitch. *Psychol. Rev.* 89, 305–333. doi: 10.1037/0033-295X.89.4.305

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

Shettleworth, S. J. (2009). *Cognition, Evolution, and Behavior*. Oxford: Oxford University Press.

Schouenborg, J. (2004). Learning in sensorimotor circuits. *Curr. Opin. Neurobiol.* 14, 693–697. doi: 10.1016/j.conb.2004.10.009

Erlhagen, W., and Schöner, G. (2002). Dynamic field theory of movement preparation. *Psychol. Rev.* 109, 545. doi: 10.1037/0033-295X.109.3.545

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

Stein, B. E., and Meredith, M. A. (1993). *The Merging of the Senses*. Cambridge, MA: The MIT Press.

Stoddard, C., and Prum, R. (2008). Evolution of avian plumage color in a tetrahedral color space: a phylogenetic analysis of New World buntings. *Am. Natural.* 171, 755–776. doi: 10.1086/587526

Stulp, F., and Sigaud, O. (2015). Many regression algorithms, one unified model: a review. *Neural Netw.* 69, 60–79. doi: 10.1016/j.neunet.2015.05.005

Svaetichin, G. (1955). Spectral response curves from single cones. *Acta Physiol. Scand.* 39(Suppl.), 17–46.

Tanaka, K. (2003). Columns for complex visual object features in the inferotemporal cortex: clustering of cells with similar but slightly different stimulus selectivities. *Cereb. Cortex* 13, 90–99. doi: 10.1093/cercor/13.1.90

van der Velde, F., and De Kamps, M. (2006). Neural blackboard architectures of combinatorial structures in cognition. *Behav. Brain Sci.* 29, 37–70. doi: 10.1017/s0140525x06009022

Vos, J. J. (2006). From lower to higher colour metrics: a historical account. *Clin. Exp. Optomet.* 89, 348–360. doi: 10.1111/j.1444-0938.2006.00091.x

Warglien, M., and Gärdenfors, P. (2013). Semantics, conceptual spaces, and the meeting of minds. *Synthese* 190, 2165–2193. doi: 10.1007/s11229-011-9963-z

Young, M. P., and Yamane, S. (1992). Sparse population coding of faces in the inferotemporal cortex. *Science* 256, 1327–1331. doi: 10.1126/science.1598577

Zipser, D., and Andersen, R. A. (1988). A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons. *Nature* 331, 679–684. doi: 10.1038/331679a0

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

frontiers in Psychology

HYPOTHESIS AND THEORY
published: 03 November 2016
doi: 10.3389/fpsyg.2016.01584

Information Compression, Multiple Alignment, and the Representation and Processing of Knowledge in the Brain

J. Gerard Wolff *

CognitionResearch.org, Menai Bridge, UK

The *SP theory of intelligence*, with its realization in the *SP computer model*, aims to simplify and integrate observations and concepts across artificial intelligence, mainstream computing, mathematics, and human perception and cognition, with information compression as a unifying theme. This paper describes how abstract structures and processes in the theory may be realized in terms of neurons, their interconnections, and the transmission of signals between neurons. This part of the SP theory—*SP-neural*—is a tentative and partial model for the representation and processing of knowledge in the brain. Empirical support for the SP theory—outlined in the paper—provides indirect support for SP-neural. In the abstract part of the SP theory (SP-abstract), all kinds of knowledge are represented with *patterns*, where a pattern is an array of atomic symbols in one or two dimensions. In SP-neural, the concept of a "pattern" is realized as an array of neurons called a *pattern assembly*, similar to Hebb's concept of a "cell assembly" but with important differences. Central to the processing of information in SP-abstract is information compression via the matching and unification of patterns (ICMUP) and, more specifically, information compression via the powerful concept of *multiple alignment*, borrowed and adapted from bioinformatics. Processes such as pattern recognition, reasoning and problem solving are achieved via the building of multiple alignments, while unsupervised learning is achieved by creating patterns from sensory information and also by creating patterns from multiple alignments in which there is a partial match between one pattern and another. It is envisaged that, in SP-neural, short-lived neural structures equivalent to multiple alignments will be created via an inter-play of excitatory and inhibitory neural signals. It is also envisaged that unsupervised learning will be achieved by the creation of pattern assemblies from sensory information and from the neural equivalents of multiple alignments, much as in the non-neural SP theory—and significantly different from the "Hebbian" kinds of learning which are widely used in the kinds of artificial neural network that are popular in computer science. The paper discusses several associated issues, with relevant empirical evidence.

Keywords: multiple alignment, cell assembly, information compression, unsupervised learning, artificial intelligence

OPEN ACCESS

Edited by:
Asim Roy,
Arizona State University, USA

Reviewed by:
Jonathan C. W. Edwards,
University College London, UK
Luis C. Lamb,
Federal University of Rio Grande do Sul, Brazil

***Correspondence:**
J. Gerard Wolff
jgw@cognitionresearch.org

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *18 May 2016*
Accepted: *29 September 2016*
Published: *03 November 2016*

Citation:
Wolff JG (2016) Information Compression, Multiple Alignment, and the Representation and Processing of Knowledge in the Brain. Front. Psychol. 7:1584. doi: 10.3389/fpsyg.2016.01584

1. INTRODUCTION

The *SP theory of intelligence*, and its realization in the *SP computer model*, is a unique attempt to simplify and integrate observations and concepts across artificial intelligence, mainstream computing, mathematics, and human perception and cognition. The name "SP" derives from the central importance in the theory of information compression, something that may be seen as a process of maximizing the *Simplicity* of a body of information, by removing information that is repeated, whilst retaining as much as possible of its non-repeated expressive *Power*. Also, the theory itself may be seen to compress empirical information by combining simplicity in the theory with wide-ranging explanatory and descriptive power.

This paper, which draws on Wolff (2006, chapter 11) with revisions and updates, describes how abstract structures and processes in the SP theory may be realized in terms of neurons, their interconnections, and the transmission of impulses between neurons. This part of the SP theory—called *SP-neural*—may be seen as a tentative and partial theory of the representation and processing of knowledge in the brain. As such, it may prove useful as a source of ideas for theoretical and empirical investigations in the future. For the sake of clarity, the abstract parts of the theory, excluding SP-neural, will be referred to as *SP-abstract*.

It is envisaged that SP-neural will be further developed in the form of a computer model. As with the existing computer model of SP-abstract (which, unless otherwise stated, will be referred to as "the SP computer model"), the development of the new computer model of SP-neural will help to guard against vagueness in the theory, it will serve as a means of testing ideas to see whether or not they work as anticipated, and it will be a means of demonstrating what the model can do, and validating it against empirical data.

The next section says something about the theoretical orientation of this research. Then SP-abstract will be described briefly as a foundation for the several sections that follow, which describe aspects of SP-neural and associated issues.

2. THEORETICAL ORIENTATION

Cosmologist John Barrow has written that "Science is, at root, just the search for compression in the world" (Barrow, 1992, p. 247), an idea which may be seen to be equivalent to Occam's Razor because, in accordance with the remarks above about the name "SP" and the theory itself, a good theory should combine conceptual *Simplicity* with descriptive or explanatory *Power*.

This works best when the range of phenomena to be described or explained is large. But this has not always been observed in practice: Newell (1973, p. 303) urged researchers in psychology to address "a genuine slab of human behavior"; and McCorduck (2004, pp. 417, 424) has described how research in artificial intelligence became fragmented into many narrow sub-fields.

In the light of these observations, and in the spirit of research on "unified theories of cognition" (Newell, 1990) and "artificial general intelligence[1]," the SP programme of research has attempted to simplify and integrate observations and concepts across a broad canvass, resisting the temptation to concentrate only on one or two narrow areas.

3. SP-ABSTRACT IN BRIEF

As a basis for the description of SP-neural, this section provides a brief informal account of SP-abstract. The theory is described most fully in Wolff (2006) and quite fully but more briefly in Wolff (2013). Details of other publications in the SP programme, most of them with download links, are shown on (http://www.cognitionresearch.org/sp.htm).

3.1. Origins and Foundations of the SP Theory

The origins of SP theory are mainly in a body of research by Attneave (1954) and Barlow (1959, 1969) and others suggesting that much of the workings of brains and nervous systems may be understood as compression of information, and my own research on language learning (summarized in Wolff, 1988) suggesting that, to a large extent, the learning of language may be understood in the same terms. There is more about the foundations of the theory in Wolff (2014d).

3.2. Elements of SP-Abstract

In SP-abstract, all kinds of knowledge are represented with *patterns*, where a pattern is an array of atomic *symbols* in one or two dimensions. At present, the SP computer model[2] works only with 1D patterns but it is envisaged that the model will be generalized to work with 2D patterns. In this connection, a "symbol" is simply a "mark" that can make a yes/no match with any other symbol—no other result is permitted.

In most of the examples shown in this paper, symbols are shown as alphanumeric characters or short strings of characters but, when the SP system is used to model biological structures and processes, such representations may be interpreted as low-level elements of perception such as formants or formant ratios in the case of speech or lines and junctions between lines in the case of vision (see also Section 4.2).

To help cut through mathematical complexities associated with information compression, the SP system—SP-abstract and its realization in the SP computer model—is founded on a simple, "primitive" idea: that information may be compressed by finding full or partial matches between patterns and merging or "unifying" the parts that are the same. This principle—"Information Compression via the Matching and Unification of Patterns" (ICMUP)—provides the foundation for a powerful concept of *multiple alignment*, borrowed and adapted from bioinformatics. The multiple alignment concept, outlined in Section 3.5, below, is itself central in the workings of SP-abstract

[1]See, for example, "Artificial General Intelligence", *Wikipedia*, http://bit.ly/1ZxCQPo, retrieved 2016-01-19.

[2]The current version of the SP computer model is SP71, the source code for which may be downloaded via a link near the bottom of www.cognitionresearch.org/sp.htm. This version of the computer model is very similar to SP70, described in Wolff (2006, Sections 3.9.2, 9.2).

and is the key to versatility and adaptability in the SP system. It has the potential to be as significant for the understanding of "intelligence" in a broad sense as is DNA for biological sciences.

3.3. SP Patterns, Multiple Alignment, and the Representation and Processing of Knowledge

In themselves, SP patterns are not very expressive. But in the multiple alignment framework (Section 3.5) they become a very versatile medium for the representation of diverse forms of knowledge. And the building of multiple alignments, together with processes for unsupervised learning (Sections 3.4, 3.7), has proved to be a powerful means of modeling diverse aspects of intelligence.

The two things together—SP patterns and multiple alignment—have the potential to be a "Universal Framework for the Representation and Processing of Diverse Kinds of Knowledge" (UFK), as discussed in Wolff (2014c, Section III).

An implication of these ideas is that there would not, for example, be any difference between the representation and processing of non-syntactic cognitive knowledge and the representation and processing of the syntactic forms of natural language. A framework that can accommodate both kinds of knowledge is likely to facilitate their seamless integration, as discussed in Section 3.8.2.

3.4. Early Stages of Learning

The SP theory is conceived as a brain-like system that receives *New* patterns via its "senses" and stores some or all of them, in compressed form, as *Old* patterns. In broad terms, this is how the system learns.

In the SP system, all learning is "unsupervised[3]," meaning that it does not depend on assistance by a "teacher," the grading of learning materials from simple to complex, or the provision of "negative" examples of concepts to be learned—meaning examples that are marked as "wrong" (*cf.* Gold, 1967). Notwithstanding the importance of schools and colleges, it appears that most human learning is unsupervised. Other kinds of learning, such as "supervised" learning (learning from labeled examples)[4], or "reinforcement" learning (learning with carrots and sticks)[5], may be seen as special cases of unsupervised learning (Wolff, 2014b, Section V).

At the beginning of processing by the system, when the repository of Old patterns is empty[6], New patterns are stored as they are received but with the addition of system-generated "ID" symbols at the beginning and end. For example, a New pattern like "`t h e b i g h o u s e`" would be stored as an Old pattern like "`A 1 t h e b i g h o u s e #A`." Here, the lower-case letters are atomic symbols that may represent actual letters but could represent basic elements of speech (such as formant ratios or formant transitions), or basic elements of vision (such as lines or corners), and likewise with other sensory data.

Later, when some Old patterns have been stored, the system may start to recognize full or partial matches between New and Old patterns. If a New pattern is exactly the same as an Old pattern (excluding the ID-symbols), then frequency measures for that pattern and its constituent symbols are incremented. These measures, which are continually updated at all stages of processing, have an important role to play in calculating probabilities of structures and inferences and in guiding the processes of building multiple alignments (Section 3.5) and unsupervised learning.

With partial matches, the system will form multiple alignments like the one shown in **Figure 1**, with a New pattern in row 0 and an Old pattern in row 1.

From a partial match like this, the system creates Old patterns from the parts that match each other and from the parts that don't. Each newly-created Old pattern will be given system-generated ID-symbols. The result in this case would be patterns like these: "`B 1 t h e #B`," "`C 1 h o u s e #C`," "`D 1 s m a l l #D`," and "`D 2 b i g #D`." In addition, the system forms an abstract pattern like this: "`E 1 B #B D #D C #C #E`" which records the sequence ["`B 1 t h e #B`," ("`D 1 s m a l l #D`" or "`D 2 b i g #D`"), "`C 1 h o u s e #C`"] in terms the ID-symbols of the constituent patterns.

Notice how "`s m a l l`" and "`b i g`" have both been given the ID-symbol "`D`" at their beginnings and the ID-symbol "`#D`" at their ends. These additions, coupled with the use of the same two ID-symbols in the abstract pattern "`E 1 B #B D #D C #C #E`" has the effect of assigning "`s m a l l`" and "`b i g`" to the same syntactic category, which looks like the beginnings of the "adjective" part of speech.

The overall result in this example is a collection of SP patterns that functions as a simple grammar to describe the phrases *the small house* and *the big house*.

In practice, the SP computer model may form many other multiple alignments, patterns and grammars which are much less tidy than the ones shown. But, as outlined in Sections 3.5, 3.7, the system is able to home in on structures that are "good" in terms of information compression.

As we shall (see Sections 3.5, 3.8.1, and 6), SP patterns, within the SP system, are remarkably versatile and expressive, with at least the power of context-sensitive grammars (Wolff, 2006, Chapter 5).

```
0     t h e       s m a l l h o u s e     0
      | | |                 | | | | |
1 A 1 t h e b i g           h o u s e #A 1
```

FIGURE 1 | A multiple alignment produced by the SP computer model showing a partial match between a New pattern (in row 0) and an Old pattern (in row 1).

[3]See "Unsupervised learning," *Wikipedia*, bit.ly/22nEPL2, retrieved 2016-03-17.
[4]See "Supervised learning," *Wikipedia*, bit.ly/1nR4ybK, retrieved 2016-03-17.
[5]See "Reinforcement learning," *Wikipedia*, bit.ly/1R0RoDv, retrieved 2016-03-17.
[6]Although it is likely that, contrary to what Noam Chomsky and others have suggested, a newborn child does *not* have any kind of detailed knowledge of the structure of natural language, it *is* likely he or she does have inborn knowledge such as how to suck milk from a breast. In this respect (and others), the SP theory, insofar it is seen as a model of human cognition, is not entirely accurate.

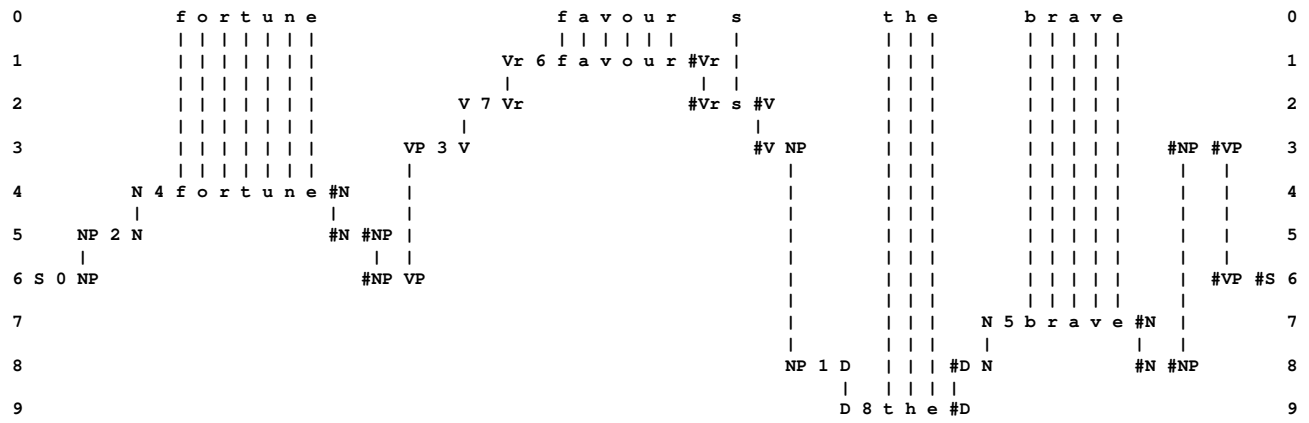


FIGURE 2 | The best multiple alignment produced by the SP computer model with a New pattern representing a sentence to be parsed and a repository of user-supplied Old patterns representing grammatical categories, including words. In the multiple alignment, the New pattern appears in row 0 and some of the Old patterns supplied to the system appear in rows 1 to 9, one pattern per row.

3.5. The Multiple Alignment Concept

The multiple alignment shown in **Figure 1** is unusually simple because it contains only two patterns. More commonly, the system forms "good" multiple alignments like the one shown in **Figure 2**, with one New pattern (in row 0) and several Old patterns (one in each of rows 1–9)[7]. As a matter of convention, the New pattern is always shown in row 0, but the order of the Old patterns across the other rows is not significant.

A multiple alignment like the one shown in **Figure 2** is built in stages, using heuristic search at each stage to weed out structures that are "bad" in terms of information compression and retaining those that are "good." Problems of computational complexity are reduced or eliminated by a scaling back of ambition: instead of searching for theoretically-ideal solutions, one merely searches for solutions that are "good enough."

In this example, multiple alignment achieves the effect of parsing the sentence into parts and sub-parts, such as a sentence ("S") defined by the pattern in row 6, one kind of noun phrase ("NP") defined by the pattern that appears in row 5, and another kind of noun phrase shown in row 8, a verb phrase ("VP") defined by the pattern in row 3, nouns ("N") defined by the patterns in rows 4 and 7, and so on. But there is much more than this to the multiple alignment concept as it has been developed in the SP programme. It turns out to be a remarkably versatile framework for the representation and processing of diverse kinds of knowledge—non-verbal patterns and pattern recognition, logical and probabilistic kinds of "rules" and several kinds of reasoning, and more (Sections 3.8.1, 6).

A point worth mentioning here is that, although the multiple concept is entirely non-hierarchical, it can model several kinds of hierarchy and heterarchy (Section 3.8.1), as illustrated by the parsing example in **Figure 2**. And such hierarchies or heterarchies may not always be "strict" because any pattern may be aligned with any other pattern and, within one multiple alignment, any pattern may be aligned with two or more other patterns.

3.6. Deriving a Code Pattern from a Multiple Alignment

From a multiple alignment like the one shown in **Figure 2**, the SP system may derive a *code pattern*—a compressed encoding of the sentence—as follows: scan the multiple alignment from left to right, identifying the ID-symbols that are *not* matched with any other symbol and create an SP pattern from the sequence of such symbols. In this case, the result is the pattern "`S 0 2 4 3 7 6 1 5 #S`." This code pattern has several existing or potential uses including:

- It provides a basis for calculating a "compression score" for the Old patterns in the multiple alignment, meaning their effectiveness as a means of compressing the New pattern. Compression scores like that have a role in sifting out one or more "good" grammars for any given set of New patterns.
- If the code pattern is treated as a New pattern then, with the same Old patterns as when the code pattern was produced, the SP system can recreate the original sentence, as described in Section 8.
- When SP-abstract is developed to take account of meanings as well as syntax, it is likely that each ID-symbol in the code pattern will take on a dual role: representing each syntactic form (word or other grammatical structure) and representing the meaning of the given syntactic form.
- It is envisaged that, with further development of the SP computer model, code patterns will enter into the learning process, as outlined in Section 3.7, next.

3.7. Later Stages of Learning

As we saw in Section 3.4, the earliest stage of learning in SP-neural—when the repository of Old patterns is empty or nearly so—is largely a matter of absorbing New information

[7]In this case, the SP computer model was supplied with an appropriate set of Old patterns. It did not learn them for itself.

directly with little modification except for the addition of system-generated ID-symbols. Later, when there are more Old patterns in store, the system begins to create Old patterns from partial matches between New and Old patterns. Part of this process is the creation of abstract patterns that describe sequences of lower-level patterns.

As the system begins to create abstract patterns, it will also begin to form multiple alignments like the one shown in **Figure 2**. And, as it begins to form multiple alignments like that, it will also begin to form code patterns, as described in Section 3.6.

At all stages of learning, but most prominent in the later stages, is a process of inferring one or more *grammars* that are "good" in terms of their ability to encode economically all the New patterns that have been presented to the system. Here, a "grammar" is simply a collection of SP patterns[8].

Inferring grammars that are good in terms of information compression is, like the building multiple alignments, a stage-by-stage process of heuristic search through the vast abstract space of alternatives, discarding "bad" alternatives at each stage, and retaining a few that are "good." As with the building of multiple alignments, the search aims to find solutions that are "good enough," and not necessarily perfect. These kinds of heuristic search may be performed by means of genetic algorithms, simulated annealing, and other heuristic techniques.

It is envisaged that the SP computer model will be developed so that, in this later phase of learning, learning processes will be applied to code patterns as well as to New patterns. It is anticipated that this may overcome two weaknesses in the SP computer model as it is now: that, while it forms abstract patterns at the highest level, it does not form abstract patterns at intermediate levels; and that it does not recognize discontinuous dependencies in knowledge (Wolff, 2013, Section 3.3).

In Wolff (2006, Chapter 9), there is a much fuller account of unsupervised learning in the SP computer model.

3.8. Evaluation of SP-Abstract

The SP theory in its abstract form may be evaluated in terms of "simplicity" and "power" of the theory itself (discussed in Section 3.8.1 next), in terms of its potential to promote simplification and integration of structures and functions in natural or artificial systems that conform to the theory (Section 3.8.2 below), and in comparison with other AI-related systems.

3.8.1. Simplicity and Power

In terms of the principles outlined in Section 2, the SP system, with multiple alignment center stage, scores well. One relatively simple framework has strengths and potential in the representation of several different kinds of knowledge, in several different aspects of AI, and it has several potential benefits and applications:

- *Representation and processing of diverse kinds of knowledge*. The SP system (SP-abstract) has strengths and potential in the representation and processing of: class hierarchies and heterarchies, part-whole hierarchies and heterarchies, networks and trees, relational knowledge, rules used in several kinds of reasoning, patterns with pattern recognition, images with the processing of images (Wolff, 2014a), structures in planning and problem solving, structures in three dimensions (Wolff, 2014a, Section 6), knowledge of sequential and parallel procedures (Wolff, 2014b, Section IV-H). It may also provide an interpretive framework for structures and processes in mathematics (Wolff, 2014d, Section 10).

  There is a fuller summary in Wolff (2014c, Section III-B) and much more detail in Wolff (2006, 2013).
- *Strengths and potential in AI*. The SP theory has things to say about several different aspects of AI, as described most fully in Wolff (2006) and more briefly in Wolff (2013). In addition to its capabilities in the parsing of natural language, described above, the SP system has strengths and potential in the production of natural language, the representation and processing of diverse kinds of semantic structures, the integration of syntax and semantics, fuzzy pattern recognition, recognition at multiple levels of abstraction, computer vision and modeling aspects of natural vision (Wolff, 2014a), information retrieval, planning, problem solving, and several kinds of reasoning (one-step "deductive" reasoning; abductive reasoning; reasoning with probabilistic decision networks and decision trees; reasoning with "rules"; nonmonotonic reasoning and reasoning with default values; reasoning in Bayesian networks with "explaining away"; causal diagnosis; reasoning which is not supported by evidence; and inheritance of attributes in an object-oriented class hierarchy or heterarchy). There is also potential for spatial reasoning (Wolff, 2014b, Section IV-F.1) and what-if reasoning (Wolff, 2014b, Section IV-F.2). The system also has strengths and potential in unsupervised learning (Wolff, 2006, Chapter 9).
- *Many potential benefits and applications*. Potential benefits and applications of the SP system include: helping to solve nine problems associated with big data (Wolff, 2014c); the development of intelligence in autonomous robots, with potential for gains in computational efficiency (Wolff, 2014b); the development of computer vision (Wolff, 2014a); it may serve as a versatile database management system, with intelligence (Wolff, 2007); it may serve as an aid in medical diagnosis (Wolff, 2006); and there are several other potential benefits and applications, some of which are described in Wolff (2014e).

In short, the SP theory, in accordance with Occam's Razor, demonstrates a favorable combination of simplicity and power across a broad canvass. As in other areas of science, this should increase our confidence in the validity and generality of the theory.

3.8.2. Simplification and Integration

Closely related to simplicity and power in the SP theory are two potential benefits arising from the use of one simple format (SP patterns) for all kinds of knowledge and one relatively simple framework (chiefly multiple alignment) for the processing of all kinds of knowledge:

[8]The term "grammar" has been adopted partly because of the origins of the SP system in research on the learning of natural language (Wolff, 1988) and partly because the term has come to be used in areas outside computational linguistics, such as pattern recognition.

- *Simplification*. Those two features (one simple format for knowledge and one simple framework for processing it) can mean substantial simplification of natural systems (brains) and artificial systems (computers) for processing information. The general idea is that one relatively simple system can serve many different functions. In natural systems, there is a potential advantage in terms of natural selection, and in artificial systems there are potential advantages in terms of costs.
- *Integration*. The same two features are likely to facilitate the seamless integration of diverse kinds of knowledge and diverse aspects of intelligence—pattern recognition, several kinds of reasoning, unsupervised learning, and so on—in any combination, in both natural and artificial systems. It appears that that kind of seamless integration is a key part of the versatility and adaptability of human intelligence and that it will be essential if we are to achieve human-like versatility and adaptability of intelligence in artificial systems.

With regard to the seamless integration of diverse kinds of knowledge, this is clearly needed in the understanding and production of natural language. To understand what someone is saying or writing, we obviously need to be able to connect words and syntactic structures with their non-syntactic meanings, and likewise, in reverse, when we write or speak to convey some meaning.

This has not yet been explored in any depth with the SP-abstract conceptual framework but preliminary trials with the SP computer model suggest that it is indeed possible to define syntactic-semantic structures in a set of SP patterns and then, with those patterns playing the role of Old patterns, to analyse a sample sentence and to derive its meanings (Wolff, 2006, Section 5.7, Figure 5.18), and, in a separate exercise with the same set of Old patterns, to derive the same sentence from a representation of its meanings (Wolff, 2006, Figure 5.19).

3.8.3. Distinctive Features and Advantages of the SP System Compared with Other AI-Related Systems

In several publications, such as Wolff (2006, 2007, 2014e), potential benefits and applications of the SP system have been described.

More recently, it has seemed appropriate to say what distinguishes the SP system from other AI-related systems and, more importantly, to describe advantages of the SP system compared AI-related alternatives. Those points have now been set out in some detail in *The SP theory of intelligence: its distinctive features and advantages* (Wolff, 2016). Of particular relevance to this paper are the several advantages of the SP system compared with systems for deep learning in artificial neural networks (Wolff, 2016, Section V).

Since many AI-related systems may also be seen as models of cognitive structures and processes in brains, Wolff (2016) may also be seen to demonstrate the relative strength of the SP system in modeling aspects of human perception and cognition.

In this connection, the SP system appears to have some advantages compared with concepts developed in research in "neural-symbolic computation," described in d'Avila Garcez et al. (2015), de Penning et al. (2011), d'Avila Garcez et al. (2009), Komendantskaya et al. (2007), and d'Avila Garcez (2007) amongst other publications. The main apparent advantages are:

- *The AI scope of the SP system*. The scope of SP-abstract in AI, meaning the range of AI-related capabilities where it has strengths and potential (summarized in Section 3.8.1), appears to be greater than the range of AI-related capabilities considered in research on neural-symbolic computation. There is potential for SP-neural to inherit that same wide scope.
- *Problems with deep learning in artificial neural networks, and potential SP solutions*. As mentioned above, the SP system has the potential to overcome several problems with deep learning in artificial neural networks (Wolff, 2016, Section V).

4. INTRODUCTION TO SP-NEURAL

As we have seen in Section 3, SP-abstract is a relatively simple system with descriptive and explanatory power across a wide range of observation and phenomena in artificial intelligence and related areas. How can such a system have anything useful to say about the extraordinary complexity of brains and nervous systems, both in their structure and in their workings?

An answer in brief is that SP-neural—a realization of SP-abstract in terms of neurons, their interconnections, and the transmission of impulses between neurons—may help us to interpret neural structures and processes in terms of the relatively simple concepts in SP-abstract. To the extent that this is successful, it may—like any good theory in any field—help us to understand empirical phenomena in our area of interest, it may help us to make predictions, and it may suggest lines of investigation.

It is anticipated that SP-neural will work in broadly the same way as SP-abstract, but the characteristics of neurons and their interconnections raise some issues that do not arise in SP-abstract and its realization in the SP computer model. These issues will be discussed at appropriate points in this and subsequent sections.

This section introduces SP-neural in outline, and sections that follow describe aspects of the theory in more detail, drawing where necessary on aspects of SP-abstract that have been omitted from or only sketched in Section 3.

4.1. Sensory Data and the Receptor Array

Figure 3, to be discussed in this and the following subsections, shows in outline how a portion of the multiple alignment shown in **Figure 2**, may be realized in SP-neural. with associated patterns and symbols.

In the figure, "sensory data" at the bottom means the visual, auditory or tactile data entering the system which, in this case, corresponds with the phrase "t h e b r a v e." In a more realistic illustration, the sensory data would be some kind of analog signal. Here, the letters are intended to suggest the kinds of low-level perceptual primitives outlined below.

It is envisaged that, with most sensory modalities, the receptor array would be located in the primary sensory cortex. Of course,

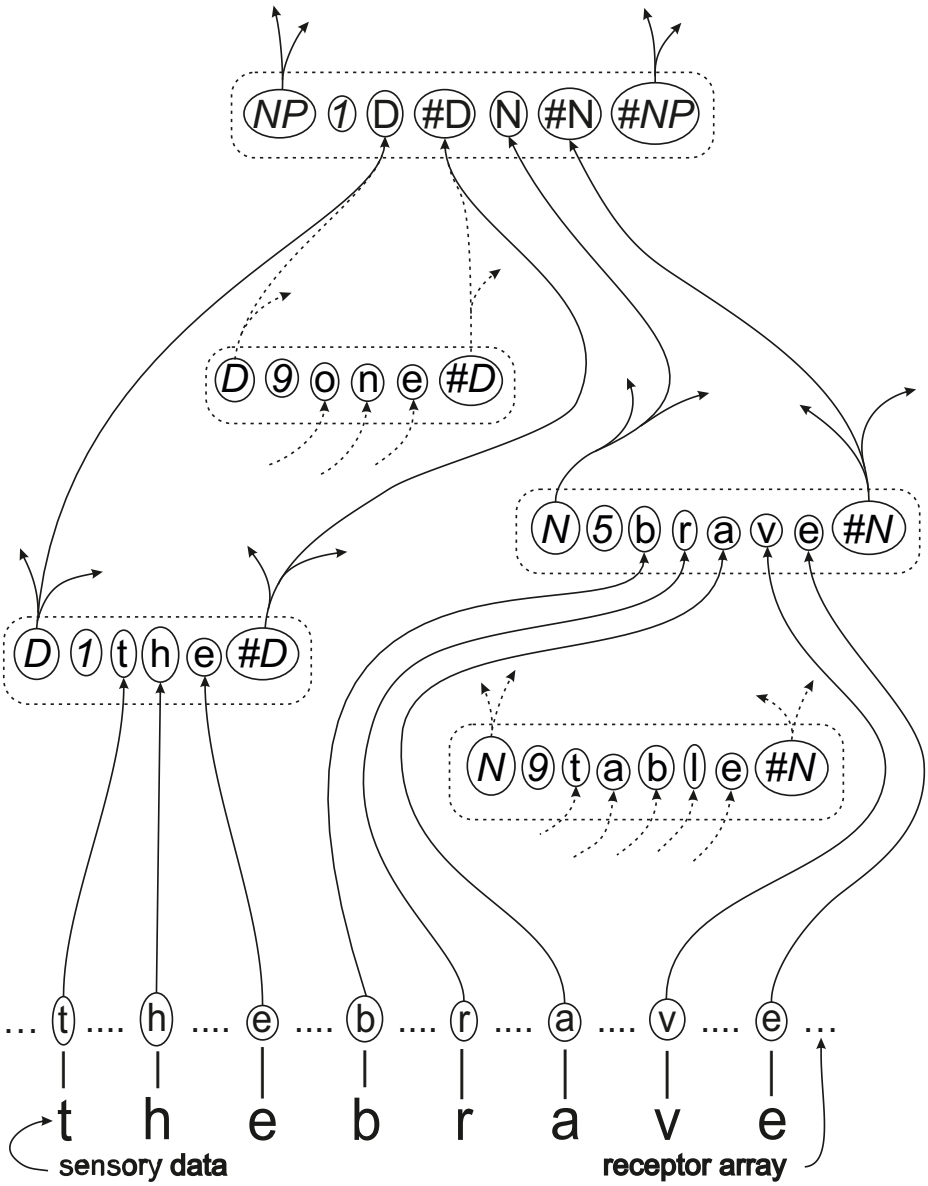


FIGURE 3 | A schematic outline of how part of the multiple alignment shown in Figure 2, with associated patterns and symbols, may be expressed in SP-neural as neurons and their inter-connections. The meanings of the conventions in the figure, and some complexities that are not shown in the figure, are explained in this main section and ones that follow.

a lot of processing goes on in the sense organs and elsewhere between the sense organs and the primary sensory cortices. But it seems that most of this early processing is concerned with the identification of the perceptual primitives just mentioned.

As with SP-abstract, it is anticipated that SP-neural will, at some stage, be generalized to accommodate patterns in two dimensions, such as visual images, and then the sensory data may be received in two dimensions, as in the human eye.

Between the sensory data and the *receptor array* (above it in the figure), there would be, first, cells that are specialized to receive particular kinds of input (auditory, visual, tactile etc.). These send signals to neurons that encode the sensory data as *neural symbols*, the neural equivalents of "symbols" in SP-abstract. In the receptor array, each letter enclosed in a solid ellipse represents a neural symbol, expressed as a single neuron or, more likely, a small cluster of neurons. As we shall see Section 5.1, the reality is more complex, at least in some cases.

In vision, neural symbols in the receptor array would represent such low-level features as lines, corners, colors, and the like, while in speech perception, they would represent such things as formants, formant ratios and transitions, plosive and fricative sounds, and so on. Whether or how the SP concepts can

be applied in the discovery or identification of features like these is an open question (Wolff, 2013, Section 3.3). For now, we shall assume that they can be identified and can be used in the creation and use of higher-level structures.

4.2. Pattern Assemblies

In the rest of **Figure 3**, each broken-line rectangle with rounded corners represents a *pattern assembly*—corresponding to a "pattern" in SP-abstract. The word "assembly" has been adopted within the expression "pattern assembly" because the concept is quite similar to Hebb's concept of a "cell assembly"—a cluster of neurons representing a concept or other coherent mental entity. Differences between Hebb's concept of a cell assembly and the SP concept of a pattern assembly are described in the Appendix.

Within each pattern assembly, as represented in the figure, each character or group of characters enclosed in a solid-line ellipse represents a *neural symbol* which, as already mentioned, corresponds to a "symbol" in SP-abstract. As with neural symbols in the receptor array, it is envisaged that each neural symbol would comprise a single neuron or, more likely, a small cluster of neurons.

It is supposed that, within each pattern assembly, there are lateral connections between neural symbols—but these are not shown in the figure.

It is envisaged that most pattern assemblies would represent knowledge that is learned and not inborn, and would be located mainly outside the primary sensory areas of the cortex, in other parts of the sensory cortices. Pattern assemblies that integrate two or more sensory modalities may be located in "association" areas of the cortex.

Research with fMRI recordings from volunteers (Huth et al., 2016) has revealed "semantic maps" that "show that semantic information is represented in rich patterns that are distributed across several broad regions of cortex. Furthermore, each of these regions contains many distinct areas that are selective for particular types of semantic information, such as people, numbers, visual properties, or places. We also found that these cortical maps are quite similar across people, even down to relatively small details"[9]. Of course, this research says nothing about whether or not the knowledge is represented with pattern assemblies and their interconnections. But it does apparently confirm that knowledge is stored in several regions of the cortex and throws light on how it is organized.

Although most parts of the mammalian cerebral cortex has six layers and many convolutions, it may be seen, topologically, as a sheet which is very much broader and wider than it is thick. Correspondingly, it is envisaged that 1D and 2D pattern assemblies will be largely "flat" structures, rather like writing or pictures on a sheet of paper. That said, it is quite possible, indeed likely, that pattern assemblies would take advantage of two or more layers of the cortex, not just one.

Incidentally, since 2D SP patterns may provide a basis for 3D models, as described in Wolff (2014a, Sections 6.1, 6.2), flat neural structures in the cortex may serve to represent 3D concepts.

4.3. Connections between Pattern Assemblies

In **Figure 3**, the solid or broken lines that connect with neural symbols represent axons, with arrows representing the direction of travel of neural impulses. Where two or more connections converge on a neural symbol, we may suppose that, contrary to the simplified way in which the convergence is shown in the figure, there would be a separate dendrite for each connection.

Axons represented with solid lines are ones that would be active when the multiple alignment in **Figure 2** is in the process of being identified. Broken-line connections show a few of the many other possible connections.

As mentioned in Section 4.2, it is envisaged that there would be one or more neural connections between neighboring neural symbols within each pattern assembly but these are not marked in the figure.

Compared with what is shown in the figure, it likely that, in reality, there would be more "levels" between basic neural symbols in the receptor array and ID-neural-symbols representing pattern assemblies for relatively complex entities like the words "one," "brave," "the," and "table," as shown in the figure.

In this connection, it is perhaps worth emphasizing that, as with the modeling of hierarchical structures in multiple alignments (Section 3.5), while pattern assemblies may form "strict" hierarchies, this is not an essential feature of the concept, and it is likely that many neural structures formed from pattern assemblies may be only loosely hierarchical or not hierarchical at all.

4.4. SP-Neural, Quantities of Knowledge, and the Size of the Brain

Given the foregoing account of how knowledge may be represented in the brain, a question that arises is "Are there enough neurons in the brain to store what a typical person knows?" This is a difficult question to answer with any precision but an attempt at an answer, described in Wolff (2006, Section 11.4.9), reaches the tentative conclusion that there are. In brief:

- Given that estimates of the size of the human brain range from 10^{10} up to 10^{11} neurons,[10] we may estimate, via calculations given in Wolff (2006, Section 11.4.9), that the "raw" storage capacity of the brain is between approximately 1000 and 10,000 MB.
- Given a conservative estimate that, using SP compression mechanisms, compression by a factor of 3 may be achieved across all kinds of knowledge, our estimates of the storage capacity of the brain will range from about 3000 MB up to about 30,000 MB.

[9]From the website of the Gallant Lab at UC Berkely, retrieved 2016-05-02, http://bit.ly/1WvvLhX. See also "Brain 'atlas' of words revealed," *BBC News*, 2016-04-27, bbc.in/1SGESLz.

[10]This is consistent with another estimate, not quoted in Wolff (2006, Section 11.4.9), that there may be as many as 86 billion neurons in the human brain (Herculano-Houzel, 2012).

- Assuming: (1) That the average person knows only a relatively small proportion of what is contained in the *Encyclopedia Britannica* (EB); (2) That the average person knows lots of "everyday" things that are *not* in the EB; (3) That the "everyday" things that we *do* know are roughly equal to the things in the EB that we *do not* know; Then (4), we may conclude that the size of the EB provides a rough estimate of the volume of information that the average person knows.
- The EB can be stored on two CDs in compressed form. Assuming that most of the space is filled, this equates to 1300 MB of compressed information or approximately 4000 MB of information in uncompressed form.
- This 4000 MB estimate of what the average person knows is the same order of magnitude as our range of estimates (3000 to 30,000 MB) of what the human brain can store.
- Even if the brain stores two or three copies of its compressed knowledge—to guard against the risk of losing it, or to speed up processing, or both—our estimate of what needs to be stored (lets say $4000 \times 3 = 12,000$ MB) is still within the 3000 to 30,000 MB range of estimates of what the brain can store.

4.5. Neural Processing

In broad terms, it is envisaged that, for a task like the parsing of natural language or pattern recognition:

1. SP-neural will work firstly by receiving sensory data and interpreting it as neural symbols in the receptor array—with excitation of the neural symbols that have been identified:
   - Excitatory signals would be sent from those excited neural symbols to pattern assemblies that can receive signals from them directly. In **Figure 3**, these would be all the pattern assemblies except the topmost pattern assembly.
   - Within each pattern assembly, excitatory signals will spread laterally via the connections between neighboring neural symbols.
   - Pattern assemblies would become excited, roughly in proportion to the number of excitatory signals they receive.
2. At this stage, there would be a process of selecting amongst pattern assemblies to identify one or two that are most excited.
3. From those pattern assemblies—more specifically, the neural ID-symbols at the beginnings and ends of those pattern assemblies—excitatory signals would be sent onwards to other pattern assemblies that may receive them. In **Figure 3**, this would be the topmost pattern assembly (that would be reached immediately after the first pass through stages 2 and 3).

   As in stage 1, the level of excitation of any pattern assembly would depend on the number of excitatory signals it receives, but building up from stage to stage so that the highest-level pattern assemblies are likely to be most excited.
4. Repeat stages 2 and 3 until there are no more pattern assemblies that can be sent excitatory signals.

The "winning" pattern assembly or pattern assemblies, together with the structures below them that have, directly or indirectly, sent excitatory signals to them, may be seen as neural analogs of multiple alignments (NAMAs), and we may guess that they provide the best interpretations of a given portion of the sensory data.

If the whole sentence, "`f o r t u n e f a v o u r s t h e b r a v e`," is processed by SP-neural with pattern assemblies that are analogs of the SP patterns provided for the example shown in **Figure 2**, we may anticipate that the overall result would be a pattern of neural excitation that is an analog of the multiple alignment shown in that figure.

When a neural symbol or pattern assembly has been "recognized" by participating in a winning (neural) multiple alignment, we may suppose that some biochemical or physiological aspect of that structure is increased as an at least approximate measure of the frequency of occurrence of the structure, in accordance with the way in which SP-abstract keeps track of the frequency of occurrence of symbols and patterns (Section 3.4).

Some further possibilities are discussed in Sections 5, 9.

5. SOME MORE DETAIL

The bare-bones description of SP-neural in Section 4 is probably inaccurate in some respects and is certainly too simple to work effectively. This section and the ones that follow describe some other features which are likely to figure in a mature version of SP-neural, drawing on relevant empirical evidence where it is available.

5.1. Encoding of Information in the Receptor Array

With regard to the encoding of information in the receptor array, it seems that the main possibilities are these:

1. *Explicit alternatives*. For the receptor array to work as described in Section 4, it should be possible to encode sensory inputs with an "alphabet" of alternative values at each location in the array, in much the same way that each binary digit (bit) in a conventional computer may be set to have the value 0 or 1, or how a typist may enter any one of an alphabet of characters at any one location on the page. At each location in the receptor array, each option may be provided in the form of a neuron or small cluster of neurons. Here, there seem to be two main options:
   a. *Horizontal distribution of alternatives*. The several alternatives may be distributed "horizontally," in a plane that is parallel to the surface of the cortex.
   b. *Vertical distribution of alternatives*. The several alternatives may be distributed "vertically" between the outer and inner surfaces of the cortex, and perpendicular to those surfaces.
2. *Implicit alternatives*. At each location there may be a neuron or small cluster of neurons that, via some kind of biochemical or neurophysiological process, may be "set" to any one of the alphabet of alternative values.
3. *Rate codes*. Something like the intensity of a stimulus may be encoded via "an interaction between [the] firing rates and the number of neurons [that are] activated by [the] stimulus." (Squire et al., 2013, p. 503).

4. *Temporal codes*. A stimulus that varies with time may be encoded via "the time-varying pattern of activity in small groups of receptors and central neurons." (Squire et al., 2013).

In support of option 1.a, there is evidence that neurons in the visual cortex (of cats) are arranged in columns perpendicular to the surface of the cortex, where, for example, all the neurons in a given column respond most strongly to a line at one particular angle in the field of view, that—within a "hypercolumn" containing several columns—the preferred angle increases progressively from column to column, and that there are many hypercolumns across the primary visual cortex (Barlow, 1982). "Hubel and Wiesel point out that the organization their results reveal means that each small region, about 1 mm^2 at the surface, contains a complete sequence of ocular dominance and a complete sequence of orientation preference." (Barlow, 1982, pp. 148–149).

Leaving out the results for ocular dominance, these observations are summarized schematically in **Figure 4**. In terms of this scheme, the way in which the receptor array is shown in **Figure 3**, is a considerable simplification—each neural symbol in the receptor array in that figure should really be replaced by a hypercolumn.

With something like the intensity of a stimulus, it seems that, at least in some cases: "... activity in one particular population of somatosensory neurons ... leads the CNS to interpret it as painful stimulus ...." (Squire et al., 2013, p. 503), while "An entirely separate population of neurons ... would signal light pressure." (Squire et al., 2013). Since it is likely that relevant receptors appear repeatedly across one's skin, this appears to be another example of option 1.a.

There seems to be little evidence of encoding via option 1.b. Indeed, since the concept of a cortical column is, in effect, defined by the fact that all the neurons in any one column have the same kind of receptive field, this seems to rule out the 1.b option (see also Section 5.2).

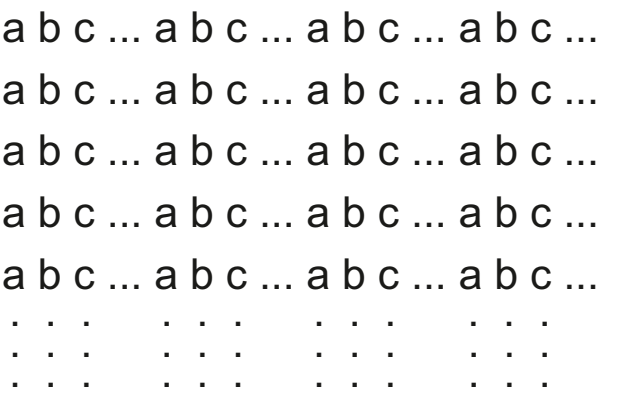


FIGURE 4 | Schematic representation of one hypercolumn in the receptor array in the cortex. Each letter represents a neural symbol that responds to a particular small pattern in the sensory data. The ellipsis, "...," in each row and each column represents other neural symbols that would be shown in a more comprehensive representation of the given hypercolumn. Each vertical sequence of letters, all of one kind such as "a" or "b," represents a simple column in the cortex.

But, with respect to option 2, it appears that in some cases, as noted above, the intensity of a stimulus may be encoded via the rates of firing of neurons, together with the numbers of neurons that are activated (option 3). And, since we can perceive and remember time-varying stimuli such as the stroking of a finger across one's skin, or the rising or falling pitch of a note, some kind of temporal encoding must be available (option 4).

Here, it must be acknowledged that options 3 and 4 appear superficially to be outside the scope of the SP theory, in view of the emphasis in many examples on discrete atomic symbols. But, as we know from the success of digital recording, or indeed digital computing, any continuum may be encoded digitally, in keeping with the digital nature of the SP theory. How the SP theory may be applied to the digital encoding and processing of continua has been discussed elsewhere in relation to vision (Wolff, 2014a) and the development of autonomous robots (Wolff, 2014b).

5.2. Why Are There Multiple Neurons with the Same Receptive Fields in Columns in the Cortex?

As we have seen (Section 5.1), some aspects of vision are mediated via columns of neurons in the primary visual cortex in which each column contains many neurons with receptive fields that are all the same, all of them responding, for example, to a line in the visual field with a particular orientation.

Why, at each of several locations across the visual cortex, should there be many neurons with the same receptive field, not just one? There seem to be two possible answers to this question (and they are not necessarily mutually exclusive):

- *Encoding of sensory patterns*. If, in the receptor array, we wish to encode two or more patterns such as "m e t" and "h e m," they need to be independent of each other, with repetition of the "e" neural symbol, otherwise there will be the possibly unwanted implication that such things as "m e m" or "h e t" are valid patterns.
- *Error-reducing redundancy*. At any given location in the receptor array, multiple instances of neurons representing a given neural symbol may help to guard against the problems that may arise if there is only one neuron at that location and if, for any reason, it becomes partially or fully disabled.

With regard to the first point, the receptor array may have a useful role to play, *inter alia*, as a short-term memory for many sensory patterns pending their longer-term storage (Section 11). In vision, for example, the receptor array may store many short glimpses of a scene, as outlined in Section 5.6, until such time as further processing may be applied to weld the many glimpses into a coherent structure (Wolff, 2014b) and to transfer that structure to longer-term memory.

5.3. The Labeled Line Principle

Section 4.5 suggests that normally, at some early stage in sensory processing, raw sensory data is encoded in terms of the excitation of neuronal symbols in a receptor array, then excited neural symbols send excitatory signals to appropriate neural symbols within pattern assemblies, and pattern assemblies that

are sufficiently excited send excitatory signals on to other pattern assemblies, and so on. As we shall see (Section 9), it is likely that, in this processing, there will also be a role for inhibitory processes.

At first sight, it may be thought that, in the same way that each location in the receptor array should provide an alphabet of alternative encodings (Section 5.1), the same should be true for the location of each neural symbol within each pattern assembly. But if a neural symbol in a pattern assembly (let's call it "NS1") receives signals only from neural symbols in the receptor array that represent a given feature, let us say, "a," then, in accordance with the "labeled line" principle (Squire et al., 2013, p. 503), NS1 also represents "a."

For most sensory modalities, this principle applies all the way from each sense organ, through the thalamus, to the corresponding part of the primary sensory cortex[11]. It seems reasonable to suppose that the same principle will apply onwards from each primary sensory cortex into non-primary sensory cortices and non-sensory association areas.

5.4. How the Ordering or 2D Arrangement of Neural Symbols May Be Respected

In SP-neural, as in SP-abstract and the SP computer model, the process of matching one pattern with another should respect the orderings of symbols. For example, "`A B C D`" matched with "`A B C D`" should be rated more highly in terms of information compression than, for example, "`A B C D`" matched with "`C A D B`[12]."

It appears that this problem may be solved by the adoption, within SP-neural, of the following feature of natural sensory systems:

> "Receptors within [the retina and skin surface] communicate with ganglion cells and those ganglion cells with central neurons in a strictly ordered fashion, such that relationships with neighbors are maintained throughout. This type of pattern, in which neurons positioned side by side in one region communicate with neurons positioned side-by-side in the next region, is called a *typographic pattern*." (Squire et al., 2013, p. 504) (emphasis in the original).

5.5. How to Accommodate the Variable Sizes of Sensory Patterns

A prominent feature of human visual perception is that we can recognize any given entity over a wide range of viewing distances, with correspondingly wide variations in the size, on the retina, of the image of that entity.

For any model of human visual perception that is based on a simplistic or naive process for the matching of patterns, this aspect of visual perception would be hard to reproduce or to explain. But the SP system is different: (1) Knowledge of entities that we may recognize are always stored in a compressed form; (2) The process of recognition is a process of compressing the incoming data; (3) The overall effect is that an image of a thing to be recognized can be matched with stored knowledge of that entity, regardless of the original size of the image.

As an example, consider how the concept of an equilateral triangle (as white space bounded by three black lines all of the same length) may be stored and how an image of such a triangle may be recognized. Regarding storage, there are three main redundancies in any image of that kind of triangle: (1) The white space in the middle may be seen as repeated instances of a symbol representing a white pixel; (2) Each of the three sides of the triangle may be seen as repeated instances of a symbol representing a black pixel; and (3) There is redundancy in that the three sides of the triangle are the same.

All three sources of redundancy may be encoded recursively as suggested in **Figure 5**[13], which shows a multiple alignment modeling the recognition of a one-dimensional analog of a triangle.

Column 0 shows information about the triangle to be recognized, comprising three "`corners`" and three sides of the triangle, each one represented by just two "`points`."

The pattern "`LN ln1 point LN #LN #LN`" in columns 1 and 2 is a self-referential and thus recursive definition of a line as a sequence of "`points`." It is self-referential because, within the body of the pattern, it contains a reference to itself via the symbols at the beginning and end of the pattern: "`LN #LN`." Because there is no limit to this recursion, it may represent a line containing any number of points. In a similar way, a second side is encoded via the same pattern in columns 6 and 7, and, again with the same pattern, the third line is encoded in columns 12 and 12.

In columns 4, 9 and 15 in the figure, the pattern "`SG sg1 CR #CR LN #LN #SG`" shows one of the three elements of a triangle as a corner ("`CR #CR`") followed by a line ("`LN #LN`"). And the recursion to encode multiple instances of that structure is in self-referential occurrences of the pattern "`TR tr1 SG #SG TR #TR #TR`" in columns 5, 10, and 22. Strictly speaking, the encoding is for a polygon, not a triangle, because there is nothing to stop the recursive repetition of "`SG sg1 CR #CR LN #LN #SG`." And, in terms of the problem, as described above, the representation is incomplete because there is nothing to show that the three sides of the triangle are the same.

These encodings account for the redundancy in the repetition of points along a line and also the redundancy in the repetition of three sides of a triangle. In a 2D version, they would also account for the redundancy in the white space within the body of the triangle, because they would allow most of the white space to be eliminated via shrinkage of the representation to the minimum needed to express the concept of a triangle.

[11] Thus, for example, "Even within one function, mappings of neurons [within the thalamus] are preserved so that there is separation of neurons providing touch information from the arm vs. from the leg and of neurons responding to low vs. high sound frequencies ...." (Squire et al., 2013, p. 507). Also, "Nuclei in the central pathways often contain multiple maps." but "The functional significance of multiple maps in general, however, remains to be clarified." (Squire et al., 2013).

[12] A possible exception is when one pattern is a mirror image or inversion of another, since Leonardo da Vinci, by repute, could read mirror writing as easily as ordinary writing, and it is now well established that people wearing inverting spectacles can learn quite quickly to see the world as if it was the right way up (Stratton, 1897).

[13] Compared with the multiple alignments shown in **Figures 1**, **2**, this multiple alignment has been rotated by 90°. The choice between these alternative presentations of multiple alignments depends entirely on what fits best on the page.

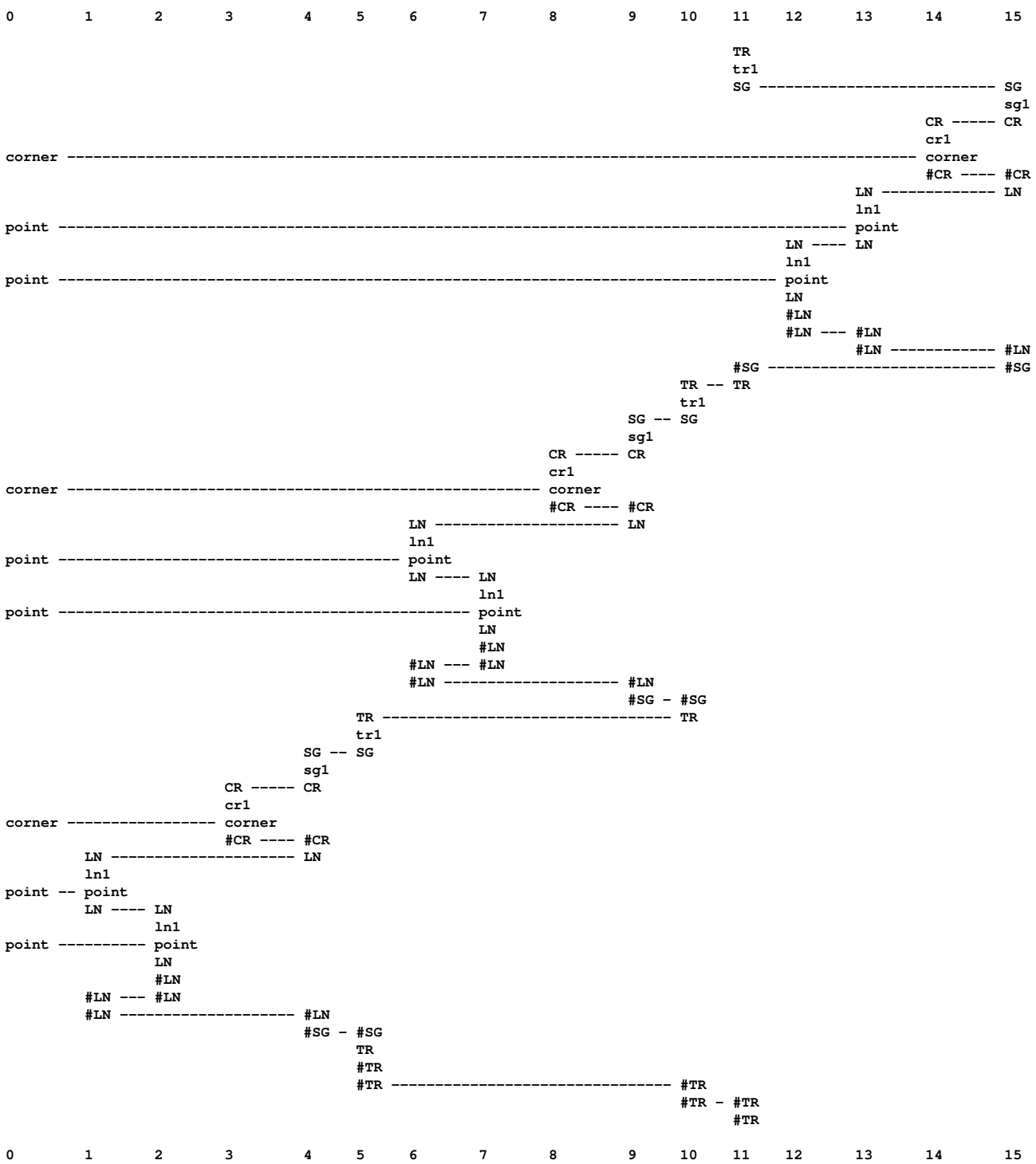


FIGURE 5 | A multiple alignment produced by the SP computer model showing how a one-dimensional analog of how an equilateral triangle may be perceived, as described in the text. Adapted from Wolff (2016, Figure 8), with permission.

5.6. We See Much Less than We Think We See

Most people with normal vision have a powerful sense that their eyes are a window on to a kind of cinema screen that shows what we are looking at with great detail from left to right and from top to bottom. But research shows otherwise:

- In the phenomenon of *inattentional blindness*, people may fail to notice salient things in their visual fields when they are

looking for something else, even if they are trained observers. In a recent demonstration (Drew et al., 2013), radiologists were asked to search for lung-nodules in chest x-rays but many of them (83%) failed to notice the image of a gorilla, 48 times the size of the average nodule, that was inserted into one of the radiographs.

- In the phenomenon of *change blindness*, people often fail to notice large changes to visual scenes. For example, if a conversation between two people—the investigator and the experimental subject—is interrupted by a door being carried between them, the experimental subject may fail to notice, when the door has gone by, that the person they are speaking to is different from the person they were speaking to before (Simons and Ambinder, 2005).
- Although each of our eyes has a blind spot[14], we don't notice it, even when we are viewing things with one eye (so that there is no possibility that the blind spot in one eye will be filled in via vision in the other eye). Apparently, our brains interpolate what is likely to be in the blind part of our visual field.

It seems that part of the reason for this failure to see things is that photoreceptors are concentrated at the fovea (Squire et al., 2013, p. 502), and cones are only found in that region (Squire et al., 2013), so that, with two eyes, we are, to a large extent, looking at the world through a keyhole composed of two circumscribed and largely overlapping views, one from each eye.

It seems that our sense that the world is displayed to us on a wide and deep cinema screen is partly because our perception of any given scene draws heavily on our memories of similar scenes and partly because we can piece together what will normally be a partial view of what we are looking at from many short glimpses through the "keyhole" as we move our gaze around the scene.

The SP theory provides an interpretation for these things as follows:

- The theory provides an account in some detail of how New (sensory) information may be related to Old (stored) information and how an interpretation of the New information may be built up via the creation of multiple alignments. When sensory information provides an incomplete description of some entity or scene (which is normally the case), we fill in the gaps from stored knowledge.
- The theory provides an account of how we can piece together a picture of something, or indeed a 3D model of something, from many small but partially-overlapping views, in much the same way that: (1) With digital photography, it is possible to create a panoramic picture from several partially-overlapping images; (2) The views in Google's Streetview are built up from many partially-overlapping pictures; (3) A 3D digital image of an object may be created from partially-overlapping images of the object, taken from viewpoints around it. These things are discussed in Wolff (2014a, Sections 5.4, 6.1).

With regard to the second point, it should perhaps be said that partial overlap between "keyhole" views is not an essential part of building up a big picture from smaller views. But if two or more views do overlap, it is useful if they can be stitched together, thus removing the overlap. And partial overlap may be helpful in establishing the relative positions of two or more views.

[14]See "Blind spot (vision)," *Wikipedia*, bit.ly/1oI0vyI, retrieved 2016-04-08.

5.7. A Resolution Problem and Its Possible Resolution

As we have seen (Section 5.1), each hypercolumn in the primary visual cortex of cats occupies about 1 mm^2 at the surface of the cortex, and it seems likely that each such hypercolumn provides a means of encoding one out of an alphabet of perceptual primitives, such as a line at a particular angle.

Assuming that this interpretation is correct, and if we view the primary visual cortex as if it was film in an old-style camera or the image sensor in a digital camera, it may seem that the encoding of perceptual primitives, with 1 mm^2 for each one, is remarkably crude. How could such a system—with the area of the primary visual cortex corresponding to the area of our field of view—create that powerful sense that, through our eyes, we see a detailed "cinema screen" view of the world (Section 5.6).

Part of the answer is probably that we see much less than we think we see (Section 5.6). But it seems likely that another part of the answer is to reject the assumption that the whole of the primary visual cortex corresponds to the area of our field of view. In the light of the remarks in Section 5.6, it seems more likely that, normally, in each of the previously-mentioned glimpses of a scene, all of the primary visual cortex or most of it is applied in the assimilation and processing of information capture by the fovea and, perhaps, parts of the retina that are near to the fovea.

In support of this idea: "*Cortical magnification* describes how many neurons in an area of the visual cortex are 'responsible' for processing a stimulus of a given size, as a function of visual field location. In the center of the visual field, corresponding to the fovea of the retina, a very large number of neurons process information from a small region of the visual field. If the same stimulus is seen in the periphery of the visual field (i.e., away from the center), it would be processed by a much smaller number of neurons. The reduction of the number of neurons per visual field area from foveal to peripheral representations is achieved in several steps along the visual pathway, starting already in the retina (Barghout-Stein, 1999)"[15].

With this view of visual processing, what appears superficially to be a rather course-grained recording and analysis of visual data, may actually be very much more detailed. As described in Section 5.6, it seems likely that our view of any scene is built up partly from memories and partly from many small snapshots or glimpses of the scene. And it seems like that each such snapshot or glimpse is processed using a relatively large neural resource.

5.8. Grandmother Cells, Localist and Distributed Representations

In terms of concepts that have been debated about how knowledge may be represented in the brain, the ID-neural-symbols for any pattern assembly are very much like the concept of a *grandmother cell*—a cell or small cluster of cells in one's brain that represents one's grandmother so that, if the cell or

[15]See "Cortical magnification," *Wikipedia*, bit.ly/1qJsQX1, emphasis in the original, retrieved 2016-04-14.

cells were to be lost, one would lose the ability to recognize one's grandmother[16].

It seems that the weight of observational and experimental evidence favors the belief that such cells do exist (Gross, 2002; Roy, 2013). This is consistent with the observation that people who have suffered a stroke or are suffering from dementia may lose the ability to recognize members of their close family.

Since SP-neural, like Hebb's (1949) theory of cell assemblies, proposes that concepts are represented by coherent groups of neurons in the brain, it is very much a "localist" type of theory. As such, it is quite distinct from "distributed" types of theory that propose that concepts are encoded in widely-distributed configurations of neurons, without any identifiable location or center.

However, just to confuse matters, SP-neural does *not* propose that all one's knowledge about one's grandmother would reside in a pattern assembly for that lady. Probably, any such pattern assembly would, in the manner of object-oriented design as discussed in Section 6 and illustrated in **Figure 6**, be connected to and inherit features from a pattern assembly representing grandmothers in general, and from more general pattern assemblies such as pattern assemblies for such concepts as "person" and "woman." And again, a pattern assembly for "person" would not be the sole repository of all one's knowledge about people. That pattern assembly would, in effect, contain "references" to pattern assemblies describing the parts of a person, their physiology, their social and political life, and so on.

Thus, while SP-neural is unambiguously localist, it proposes that knowledge of any entity or concept is likely to be encoded not merely in one pattern assembly for that entity or concept but also in many other pattern assemblies in many parts of the cortex, and perhaps elsewhere.

5.9. Positional Invariance

With something simple like a touch on the skin, or a pin prick, it is not too difficult to see how the sensation may be transmitted to the brain via any one of many relevant receptors located in many different areas of the skin. But with something more complex, like an image on the retina of a table, a house, or a tree, and so on, it is less straightforward to understand how we might recognize such a thing in any part of our visual field.

For each entity to be recognized, it seems necessary at first sight to provide connections, directly or indirectly, from every part of the receptor array to the relevant pattern assembly. In terms of the schematic representation shown in **Figure 3**, it would mean repeating the connections for "`t h e`" and "`b r a v e`" in each of many parts of the receptor array. Bearing in mind the very large number of different things we may recognize, the number of necessary connections would become very large, perhaps prohibitively so.

However, things may be considerably simplified via either or both of two provisions:

1. For reasons outlined in Section 5.6, it seems likely that, with vision, we build up our perception of a scene, partly from memories of similar scenes and partly via many relatively narrow "keyhole" views of what is in front of us. If that is correct, and if, as suggested in Section 5.7, most of the primary visual cortex is devoted to analysing information received via the fovea and, perhaps, via parts of the retina that are very close to the fovea, then the need to provide for any given pattern in many parts of the receptor array may be greatly reduced. Since, by moving our eyes, we may view any part of a scene, it is possible that any given entity would need only one or two sets of connections between the receptor array and the pattern assembly for that entity.
2. As noted in Section 4.3, it seems likely that, with regard to **Figure 3**, there would, in a more realistic example, be several levels of structure between neural symbols in the receptor array and relatively complex structures like words. At the first level above the receptor array there would be pattern assemblies for relatively small recurrent structures, and the variety of such structures would be relatively small. This should ease any possible problems in connecting the receptor array to pattern assemblies.

If it turns out that the number of necessary connections is indeed too large to be practical, or if there is empirical evidence against such numbers, then a possible alternative to what has been sketched in this paper is some kind of dynamic system for the making and breaking of connections between the receptor array and pattern assemblies. It seems likely that permanent or semi-permanent connections would be very much more efficient and the balance of probabilities seems to favor such a scheme.

In connection with positional invariance, it is relevant to note that "... lack of localization is quite common in higher-level neurons: receptive fields become larger as the features they represent become increasingly complex. Thus, for instance, neurons that respond to faces typically have receptive fields that cover most of the visual space. For these cells, large receptive fields have a distinct advantage: the preferred stimulus can be identified no matter where it is located on the retina." (Squire et al., 2013, p. 579). A tentative and partial explanation of this observation is that repetition of neurons that are sensitive to each of several categories of low-level feature—in the receptor array and as ID-neural-symbols for "low-level" pattern assemblies—is what allows positional invariance to develop at higher levels.

6. NON-SYNTACTIC KNOWLEDGE IN SP-NEURAL

As was emphasized in Section 3, the SP system (SP-abstract) has strengths and potential in the representation and processing of several different kinds of knowledge, not just the syntax of natural language. That versatility has been achieved using the mechanisms in SP-abstract that were outlined in that section. If those mechanisms can be modeled in SP-neural, it seems likely that the several kinds of knowledge that may be represented and processed in SP-abstract may also be represented and processed in SP-neural.

As an illustration, **Figure 6** shows a simple example of how, via multiple alignment, the SP computer model may recognize an unknown creature at several different levels of abstraction,

[16] See "Grandmother cell," *Wikipedia*, bit.ly/1UDulyV, retrieved 2016-08-26.

and **Figure 7** suggests how part of the multiple alignment, with associated patterns, may be realized in terms of pattern assemblies and their inter-connections.

Figure 6 shows the best multiple alignment found by the SP computer model with four symbols representing attributes of an unknown creature (shown in column 0) and a collection of Old patterns representing different creatures and classes of creature, some of which are shown in columns 1–4, one pattern per column. In a more detailed and realistic example, symbols like "`eats`," "`retractile-claws`," and "`breathes`," would be represented as patterns, each with its own structure.

From this multiple alignment, we can see that the unknown creature has been identified as an animal (column 4), as a mammal (column 3), as a cat (column 2) and as a specific cat, "Tibs" (column 1). It is just an accident of how the SP computer model has worked in this case that the order of the patterns across columns 1–4 of the multiple alignment corresponds with the level of abstraction of the classifications. In general, the order of patterns in columns above 0 is entirely arbitrary, with no significance.

Figure 7 shows how part of the multiple alignment from **Figure 6** may be realized in SP-neural. The figure contains pattern assemblies for "animal" and "mammal," corresponding to patterns from columns 4 and 3 of the multiple alignment. Notice that the left-right order of the pattern assemblies is different from the order of the patterns in the multiple alignment, in accordance with the remarks, above, about the workings of the SP computer model, and also because there is no reason to believe that pattern assemblies are represented in any particular order.

Neural connections amongst the things that have been mentioned so far are very much the same as alignments between neural symbols in **Figure 6**: "eats" on the left connects with "eats" in the "animal" pattern assembly; "furry" connects with "furry" in the "mammal" pattern assembly, and the "A" and "#A" connections for those two pattern assemblies correspond with the alignments of symbols in the multiple alignment. As in **Figure 3**, some neural connections are shown with broken lines to suggest that they would be relatively inactive during the neural processing which identifies one or more "good" NAMAs. And as before, it is envisaged that there would be one or more neural connections between each neural symbol and its immediate neighbors within each pattern assembly, but these are not marked in the figure.

The inclusion of a pattern assembly for "reptile" in **Figure 7**, with some of its neural connections, is intended to suggest some of the processing involved in identifying one or more winning NAMAs. In the same way that the pattern for "mammal" is

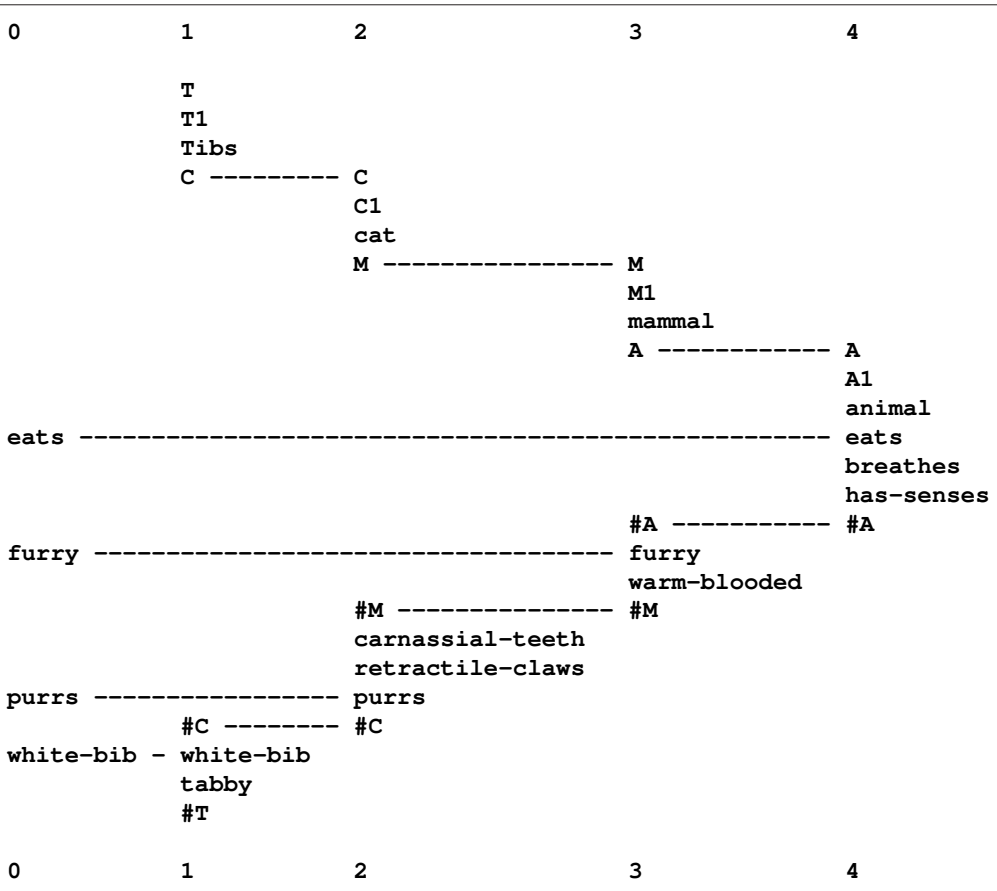


FIGURE 6 | The best multiple alignment found by the SP computer model with four one-symbol New patterns representing attributes of an unknown creature and a collection of Old patterns representing different creatures and classes of creature. Adapted from Figure 6.7 in Wolff (2006), with permission.

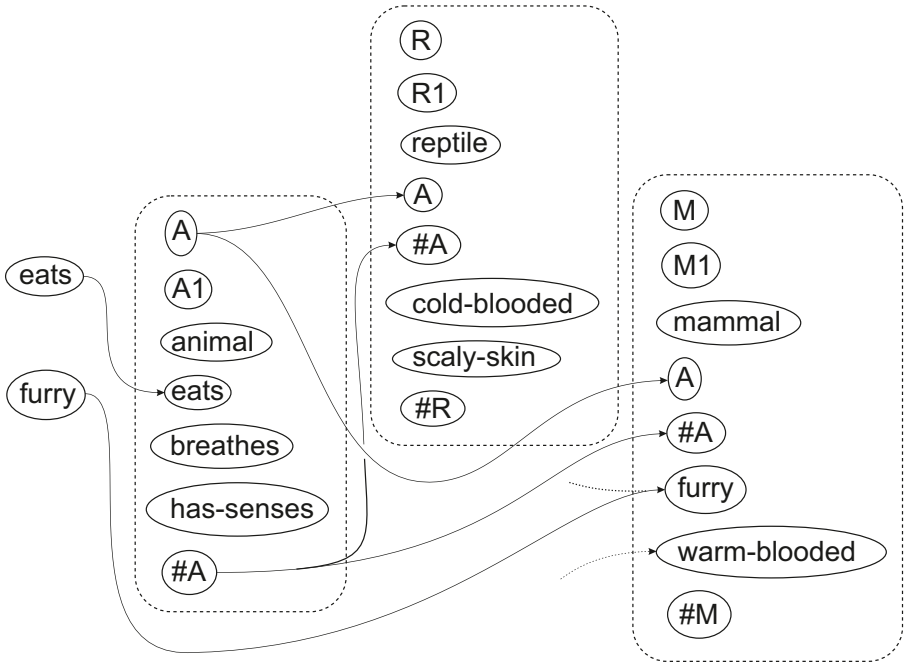


FIGURE 7 | How part of the multiple alignment shown in Figure 6 may be realized in SP-neural—showing two of the attributes from column 0 in the multiple alignment and with "animal" and "mammal" pattern assemblies corresponding to patterns from columns 4 and 3—with an associated pattern assembly for "reptile." The conventions are the same as in **Figure 3**.

receiving excitatory signals from the pattern for "animal," one would expect excitatory signals to flow to pattern assemblies for the other main groups of animals, including reptiles. Ultimately, "reptile" would fail to feature in any winning NAMA because of evidence from the neural symbols "furry," "purrs," and "white-bib."

7. REPETITION AND RECURSION

Like any good database or dictionary, the repository of Old patterns in SP-abstract should only contain one copy of any given SP pattern. But in something like *Jack Sprat could eat no fat, His wife could eat no lean*, the words *could*, *eat*, and *no* each occur twice. With an example like this, it seems reasonable to suppose that there is only one stored pattern for each of the repeated words, and likewise for the many other examples of entities that are repeated within something larger, witness the many legs of a centipede.

In SP-abstract, this apparent difficulty has been overcome by saying that each SP pattern in a multiple alignment is an *appearance* of the pattern, not the pattern itself—which allows us to have multiple instances of a pattern in a multiple alignment without breaking the rule that the repository of Old patterns should contain only one copy of each pattern. But in SP-neural, it is not obvious how to create an "appearance" of a pattern assembly that is not also a physical structure of neurons and their interconnections—but the speed with which we can understand natural language seems to rule out what appears to be the relatively slow growth of new neurons and their interconnections.

How we can create new mental structures quickly arises again in other connections, as discussed in Section 11. If we duck these questions for the time being and return to parsing, it may be argued that with something like *Jack Sprat could eat no fat, His wife could eat no lean*, the first instance of *could* is represented only for the duration of the word by the stored pattern for *could*, so that the same pattern can be used again to represent the second instance of *could*—and likewise for *eat* and *no*. But it appears that this line of reasoning does not work with a recursive structure like *the very very very fast car*.

Native speakers of English know that with a phrase like *the very very very fast car*, the word *very* may in principle be repeated any number of times. This observation, coupled with the observation that recursive structures are widespread in English and other natural languages, suggests strongly that the most appropriate parsing of the phrase is something like the multiple alignment shown in **Figure 8**. Here, the repetition of *very* is represented via three appearances of the pattern "`ri ri1 ri #ri i #i #ri`," a pattern which is self-referential because the inner pair of symbols "`ri #ri`" can be matched with the same two symbols, one at the beginning of the pattern and one at the end. Because the recursion depends on at least two instances of "`ri ri1 ri #ri i #i #ri`" being "live" at the same time, it seems necessary for SP-neural to be able to model multiple appearances of any pattern.

That conclusion, coupled with the above-mentioned arguments from the speed at which we can speak, and the speed

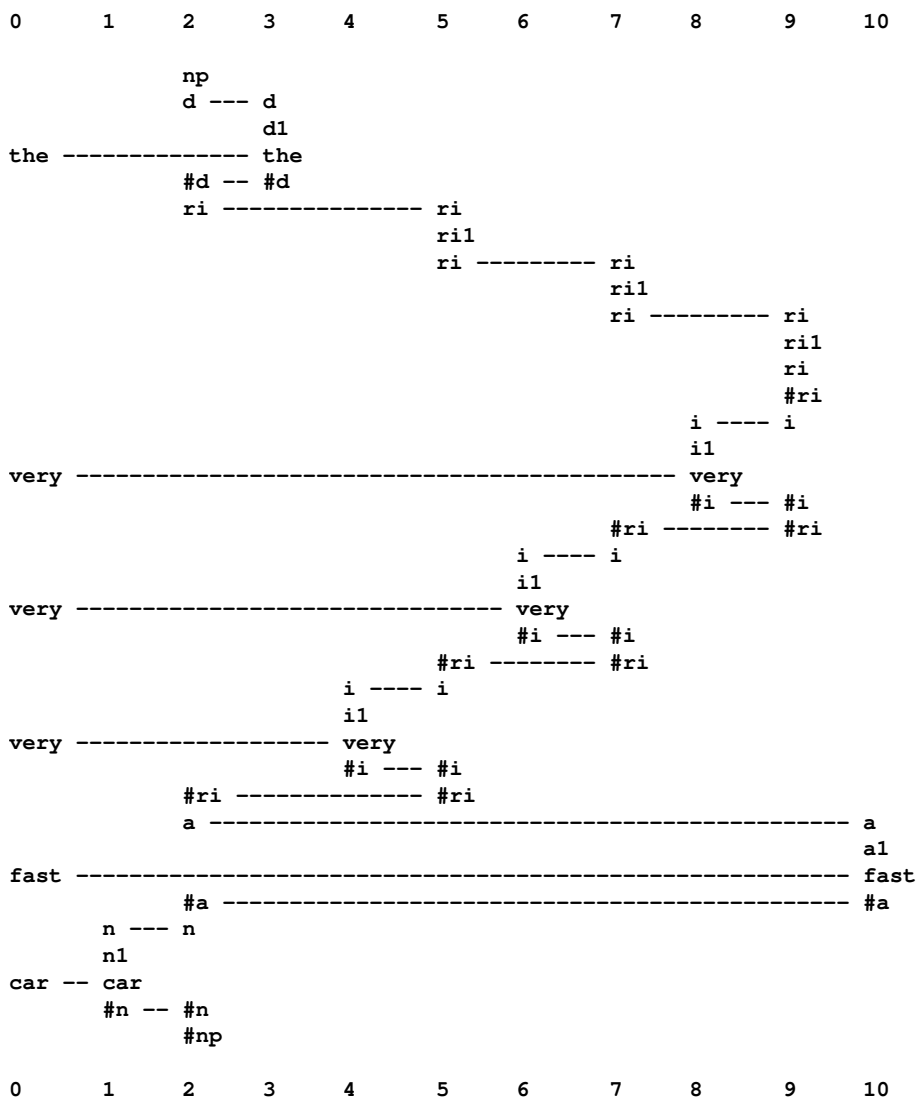


FIGURE 8 | A multiple alignment produced by the SP computer model showing how recursion may be modeled in the SP system. Adapted from Figure 3 in Wolff (2016), with permission.

with which we can imagine new things, argues strongly that SP-neural—and any other neural theory of cognition—must have a means of creating new mental structures quickly. It seems unlikely that these things could be done via the growth of new neurons and their interconnections.

The tentative answer suggested here is that, in processes like parsing or pattern recognition, including examples with recursion like that shown in **Figure 8**, virtual copies of pattern assemblies may be created and destroyed very quickly via the switching on and switching off of synapses (Section 11). Clearly, more detail is needed for a fully satisfactory answer.

Pending that better answer, **Figure 9** shows tentatively how recursion may be modeled in SP-neural, with neural symbols and pattern assemblies corresponding to selected symbols and patterns in **Figure 8**. On the left of that figure, we can see how the neural symbol "very" connects with a matching neural symbol in the pattern assembly "i i1 very #i." Further right, we can see how the first and last neural symbols in "i i1 very #i" connect with matching neural symbols in the pattern assembly "ri ri1 ri #ri i #i #ri."

In the figure, the self-referential nature of the pattern assembly "ri ri1 ri #ri i #i #ri" can be seen in the neural connection between "ri" at the beginning of that pattern assembly and the matching neural symbol in the body of the same pattern assembly, and likewise for "#ri" at the end of the pattern assembly. Although it is unclear how this recursion may achieve the effect of repeated appearances of the pattern assembly at the speed with which we understand or produce speech, the analysis appears to be more reliable than what is described

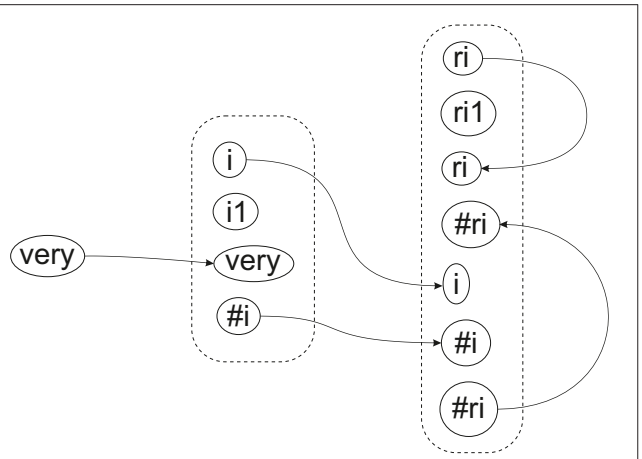


FIGURE 9 | A schematic example showing how recursive structures may be modeled in SP-neural.

in Wolff (2006, Section 11.4.2), especially Figure 11.10 in that section.

8. SP-NEURAL: AN OUTPUT PERSPECTIVE

An inspection of **Figure 3**—showing how, in SP-neural, a small portion of natural language may be analyzed by pattern assemblies and their interconnections—may suggest that if we wish to reverse the process—to create language instead of analysing it—then the innervation would need to be reversed: we may guess that two-way neural connections would be needed to support the production of speech or writing as well as their interpretation.

But a neat feature of SP-abstract is that one set of Old patterns, together with the processes for building multiple alignments, will support both the analysis and the production of language. So it is reasonable to suppose that if SP-neural works at all, a similar duality will apply to pattern assemblies and their interconnections, without the need for two-way connections amongst pattern assemblies and neural symbols (but see Section 8.3).

Of course, speaking or writing would need peripheral motor processes that are different from the peripheral sensory processes required for listening or reading, but, more centrally, the processes for analysing language or producing it may use the same mechanisms[17].

The reason that SP-abstract, as expressed in the SP computer model, can work in "reverse" so to speak, is that, from a multiple alignment like the one shown in **Figure 2**, a code pattern like "`S 0 2 4 3 7 6 1 8 5 #S`" may be derived, as outlined in Section 3.6. Then, if that code pattern is presented to the SP system as a New pattern, the system can recreate the original sentence, "`f o r t u n e f a v o u r s t h e b r a v e`," as shown in **Figure 10**.

[17]Of course, things are a little more complicated with output processes because sensory feedback is normally an important part of speaking or writing.

8.1. An Answer to the Apparent Paradox of "Decompression by Compression"

That the SP system should be able to reconstruct a sentence that was originally compressed by means of the same system (Section 8) may seem paradoxical. How is it that a system that is dedicated to information compression should be able, so to speak, to drive compression in reverse?

A resolution of this apparent paradox is described in Wolff (2006, Section 3.8). In brief, the key to the conjuring trick is to ensure that, after the sentence has been compressed, there is enough residual redundancy in the code pattern to allow further compression, and to ensure that this further compression will achieve the effect of reconstructing the sentence.

8.2. Meanings in the Analysis and Production of Language

Of course, parsing a sentence (as shown in Section 3.5) or constructing a sentence from a code pattern (as shown in Section 8) are very artificial applications with natural language. Normally, when we read some text or listen to someone speaking, we aim to derive meaning from the writing or the speech. And when we write or speak, it seems, intuitively, that the patterns of words that we are creating are derived from some kind of underlying meaning that we are trying to express.

It is envisaged that, in future development of SP-abstract and the SP computer model, the ID-symbols in code patterns will provide some kind of bridge between syntactic forms and representations of meanings, thus facilitating the processes of understanding the meanings of written or spoken sentences and of creating sentences to express particular meanings.

As noted at the end of Section 3.8.2, there are preliminary examples of how, with the SP computer model, a sentence may be analyzed for its meaning (Wolff, 2006, Section 5.7, Figure 5.18), and how the same sentence may be derived from a representation of its meaning (Wolff, 2006, Figure 5.19).

8.3. But There Are Projections from the Sensory Cortex to Subcortical Nuclei

Although as we have seen earlier in Section 8, SP-neural, via principles established in SP-abstract, provides for the creation of language, and other kinds of knowledge, without the need for efferent connections from the cortex back along the path of afferent nerves, there is evidence that such connections do exist:

> "Neurons of the cerebral cortex send axons to subcortical regions .... Subcortical projections are to those nuclei in the thalamus and brainstem that provide ascending sensory information. By far the most prominent of these is to the thalamus: the neurons of a primary sensory cortex project back to the same thalamic nucleus that provides input to the cortex. This system of descending connections is truly impressive because the number of descending corticothalamic axons greatly exceeds the number of ascending thalamocortical axons. These connections permit a particular sensory cortex to control the activity of the very neurons that relay information to it." (Squire et al., 2013, p. 509).

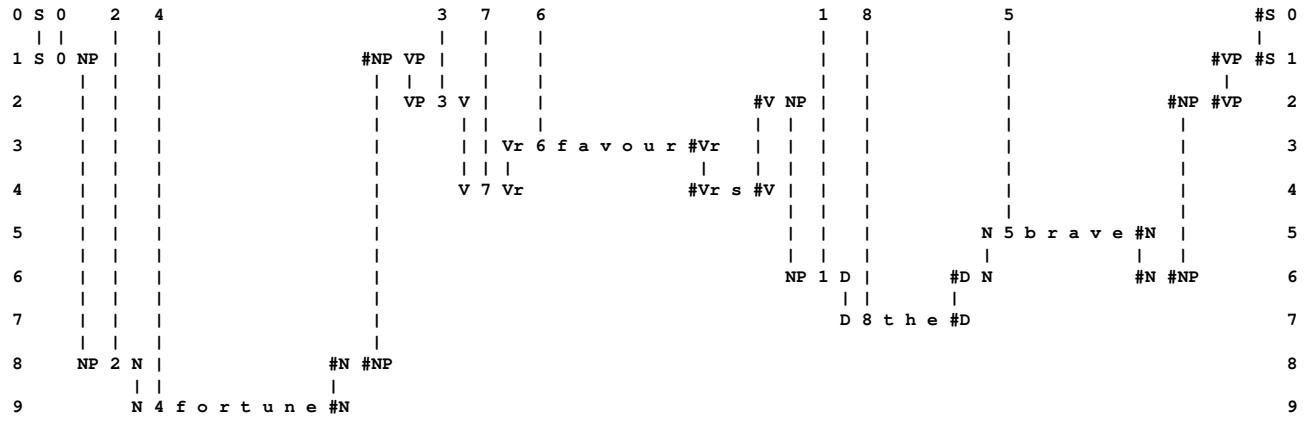


FIGURE 10 | The best multiple alignment produced by the SP computer model with the same Old patterns as for the multiple alignment shown in Figure 2 but with the New pattern comprising an appropriate sequence of ID-symbols, "S 0 2 4 3 7 6 1 8 5 #S," as described in the text.

But the descending nerves described in this quotation may have a function that is quite different from the creation of sentences or other patterns of activity. One possible role for such nerves may be "the focussing of activity so that relay neurons most activated by a sensory stimulus are more strongly driven and those in surrounding less well activated regions are further suppressed." (Squire et al., 2013, p. 509).

9. THE POSSIBLE ROLES FOR INHIBITION IN SP-NEURAL

A familiar observation is that, if something like a fan is switched on near us, we notice the noise for a while and then come to ignore it. And if, later, the fan is switched off, we notice the relative quiet for a while and then cease to be aware of it. In general, it seems that we are relatively sensitive to changes in our environment and relatively insensitive to things that remain constant.

It has been accepted for some time that the way we adapt to constant stimuli is due to inhibitory neural structures and processes in our brains and nervous systems, that inhibitory structures and processes are widespread in the animal kingdom, and that they have a role in reducing the amount of information that we need to process (von Békésy, 1967).

Regarding the last point, it is clearly inefficient for anyone to be constantly registering, second-by-second, the noise of a nearby fan: "noise, noise, noise, noise, noise, ..." and likewise for the state of relative quietness when the fan is switched off. In terms of information theory, there is *redundancy* in the second-by-second recurrence of the noise (or quietness), and we can eliminate most or all of the redundancy—and thus compress the information—by simply recording that the noise is "on" and that it is continuing (and likewise, *mutatis mutandis*, for quiet). This is the "run-length encoding" technique for compression of information,[18] it is essentially what adaptation does, and, in neural tissue, it appears to be mediated largely by "lateral" inhibition.

With lateral inhibition in sensory neurons, there are inhibitory connections between neighboring neurons so that, when they are both stimulated, they tend to inhibit each other, and thus reduce their rates of firing where there is strong uniform stimulation. But inhibition is reduced where strong stimulation gives way to weaker stimulation, leading to a local swing in the rate of firing (Ratliff et al., 1963; see also Wolff, 2006, Section 2.3.1; there is more about lateral inhibition in Squire et al., 2013, p. 505). There are similar effects in the time dimension. Again, Barlow (1982) says, in connection with neurons in the mammalian cortex that receive inputs from both eyes, "... it is now clear that input from one eye can, and frequently does, inhibit the effects of input from the other eye, ..." (p. 147).

Taking these observations together, we may abstract a general rule: *When, in neural processing, two or more signals are the same, they tend to inhibit each other, and when they are different, they don't.* The overall effect should be to detect redundancy in information and to reduce it, whilst retaining non-redundant information, in accordance with the central principle in the SP theory—that much of computing and cognition may, to a large extent, be understood as information compression.

In a similar vein: "Lateral inhibition represents the classic example of a general principle: most neurons in sensory systems are best adapted for detecting changes in the external environment. ... As a rule, it is change which has the most significance for an animal ... This principle can also be explained in terms of information processing. Given a world that is filled with constants—with uniform objects, with objects that move only rarely—it is most efficient to respond only to changes." (Squire et al., 2013, p. 578).

In view of the widespread occurrence of inhibitory mechanisms in the brain[19], and in view of their apparent

[18] See "Run-length encoding," *Wikipedia*, bit.ly/21JlB1T, retrieved 2016-03-04.

[19] "These [aspiny or sparsely spiny nonpyramida] interneurons constitute approximately 15–30% of the total population of cortical neurons, and they appear to be mostly GABAergic, representing the main components of inhibitory cortical circuits ...." (Squire et al., 2013, p. 45); "Synaptic inhibition in the mammalian brain

importance for the compression of information, and thus for selective advantage (Wolff, 2014d, Section 4), it is pertinent to ask what role or roles they may play in SP-neural. Here are some possibilities:

- *Low-level sensory features*. At relatively "low" levels in sensory processing, it appears that, as noted above, lateral inhibition has a role in identifying such things as boundaries between uniform areas, meaning lines. It may also have a role in identifying other kinds of low-level sensory features mentioned in Section 4.1.
- *Information compression via the matching and unification of patterns (ICMUP)*. As noted in Section 3.2, SP-abstract, and the SP computer model, is founded on the principle that information compression may be achieved by the matching and unification of patterns (ICMUP). Here, there appear to be these possible roles for inhibition:
  - As we have seen, lateral inhibition can have the effect of inhibiting signals from neighboring sensory neurons when they are receiving stimulation which is the same of nearly so. This may be seen as an example of ICMUP.
  - In accordance with the abstract general rule suggested above, inhibitory processes may serve as a means of detecting redundancy between a New pattern assembly like "t a b l e" and an Old pattern assembly like "N 9 t a b l e #N":
    - We may suppose that there are inhibitory links between neighboring neural symbols in the Old pattern assembly so that, if all of the neural symbols in the body of that Old pattern assembly (i.e., "t a b l e") are stimulated, or most of them, then mutual inhibition amongst those neural symbols will suppress their response. And, as with lateral inhibition in sensory neural tissue, inhibition in one area can mean enhanced responses at the boundaries with neighboring areas, which, in this case, would be the ID-symbols "N" and "9" on the left, and "#N" on the right. Then, excitatory signals from "N" and "#N" may go on to higher-level patterns that contain nouns, as suggested by the broken-line links from those two neural symbols in **Figure 3**. Since there is no link to export excitatory signals from "9," no such signals would be sent.
    - Alternatively, we may suppose that a stored pattern assembly like "N 9 t a b l e #N" has a background rate of firing and that, when matching stimulation is received for the neural symbols "t a b l e," the background rate of firing in the corresponding neural symbols in "N 9 t a b l e #N" is reduced, with an associated upswing in the rates of firing of the neural symbols "N" and "9" and "#N," as before.
- *Sharpening choices amongst alternatives*. As mentioned in Section 4.5, the process of forming neural analogs of multiple alignments (NAMAs) means identifying one or two of the most excited pattern assemblies, with structures below them that feed excitation to them. Here, inhibition may play a part by enhancing the status of the most excited pattern assemblies and suppressing the rest. How inhibition may achieve that kind of effect is discussed quite fully by von Békésy (1967, Chapters II and V), and also in Shamma (1985).

More information and discussion about the possible roles of inhibition in the cerebral cortex may be found in Isaacson and Scanziani (2011).

10. UNSUPERVISED LEARNING IN SP-NEURAL

This section considers how the learning processes in SP-abstract, which are outlined in Sections 3.4, 3.7, may be realized in SP-neural.

It seems likely that neural structures for the detection of "low level" features like lines and corners in vision, or formant ratios and transitions in hearing, are largely inborn[20], although "It is a curious paradox that, while [Hubel and Wiesel] have consistently argued for a high degree of ontogenetic determination of structure and function in the visual system, they are also the authors of the best example of plasticity in response to changed visual experience." (Barlow, 1982, p. 150), and "It has ... been shown convincingly that the orientation preference of cells can be modified, ..." (Barlow, 1982). Also, "In the somatosensory system, if input from a restricted area of the body surface is removed by severing a nerve or by amputation of a digit, the portion of the cortex that was previously responsive to that region of the body surface becomes responsive to neighboring regions ...." (Squire et al., 2013, p. 508).

But it is clear that most of what we learn in life is at a "higher" level which, in SP-neural, will be acquired via the the creation and destruction of pattern assemblies, as discussed in the following subsections.

10.1. Creating Old Pattern Assemblies

Let us suppose that a young child hears the speech equivalent of "t h e b i g h o u s e" in accordance with the example in Section 3.4. As we have seen, when the repository of Old patterns is empty or nearly so, New patterns are stored directly as Old patterns, somewhat like a recording machine, but with the addition of ID-symbols at their beginnings and ends.

It seems unlikely that a young child would grow new neurons to store a newly-created Old pattern assembly like "A 1 t h e b i g h o u s e #A," as discussed in Section 3.4. It seems much more likely that a pattern assembly like that would be created by some kind of modification of pre-existing neural tissue comprising sequences or areas of unassigned neural symbols with lateral connections between them as suggested in Section 4.2. Pattern assemblies would be created by the switching on and off of synapses at appropriate points, in

is mediated principally by GABA receptors." (Squire et al., 2013, p. 169); "One of the great mysteries of synaptic integration is why there are so many different types of inhibitory interneurons. ... more than 20 different types of inhibitory interneuron have been described in the CA1 region of the hippocampus alone." (Squire et al., 2013, p. 249).

[20]"For all systems except the olfactory, the receptor neurons you were born with are the ones you will live with." (Squire et al., 2013, p. 503).

a manner that is more like a tailor cutting up pre-woven cloth than someone knitting or crocheting each item from scratch.

In accordance with the labeled line principle (Section 5.3), the meaning of each symbol in a newly-created pattern assembly would be determined by what it is connected to, as described in Section 10.2.

Similar principles would apply when Old patterns are created from partial matches between patterns, as described in Section 3.4.

10.2. Creating Connections between Pattern Assemblies

As with the laying down of newly-created Old patterns (Section 10.1), it seems unlikely that connections between pattern assemblies, like those shown in **Figure 3**, would be created by growing new axons or dendrites. It seems much more likely that such connections would be established by switching on synapses between each of the two neurons to be connected and pre-existing axons or dendrites, somewhat like the making of connections in a telephone exchange (see Section 11).

This idea, together with the suggestions in Section 10.1 about how Old pattern assemblies may be created, is somewhat like the way in which an "uncommitted logic array" (ULA)[21] may, via small modifications, be made to function like any one of a wide variety of "application-specific integrated circuits" (ASICs)[22], or how a "field-programmable gate array" (FPGA)[23] may be programmed to function like any one of a wide variety of integrated circuits.

10.3. Destruction of Pattern Assemblies and Their Interconnections

In the SP theory, patterns and pattern assemblies are never modified—they are either created or destroyed. The latter process occurs mainly in the process of searching for "good" grammars to describe a given set of New patterns, as outlined in Section 3.7. At each stage, when a few "good" grammars are retained in the system, the rest are discarded. This means that any pattern assembly in one or more of the "bad" grammars that is not also in one or more of the "good" grammars may be destroyed.

It seems likely that, in a process that may be seen as a reversal of the way in which pattern assemblies and their interconnections are created, the destruction of a pattern assembly does not mean the physical destruction of its neurons. It seems more likely that all neural connections from the pattern assembly are broken by switching off relevant synapses (Sections 10.3, 11) and that its constituent neurons are retained for later use in other pattern assemblies.

10.4. Searching for Good Grammars

It must be admitted that, apart from the remarks in forgoing subsections about the creation and destruction of pattern assemblies and their inter-connections, it is unclear how, in SP-neural, one may achieve anything equivalent to the process of searching the abstract space of possible grammars that has been implemented in the SP computer model.

One possibility is to simplify things as follows. Instead of evaluating whole grammars, as in the SP computer model, it may be possible to achieve roughly the same effect by evaluating pattern assemblies in terms of their effectiveness or otherwise for the economical encoding of New information and, periodically, to discard those pattern assemblies that do badly.

10.5. What about Hebbian Learning?

Readers familiar with issues in AI or neuroscience may wonder what place, if any, there may be in SP-neural for the concept of "Hebbian" learning. This idea, proposed by Hebb (1949), is that:

> "When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased." (p. 62).

Variants of this idea are widely used in versions of "deep learning" in artificial neural networks (Schmidhuber, 2015) and have contributed to success with such systems[24].

But in Wolff (2016, Section V-D) I have argued that:

- The gradual strengthening of neural connections which is a central feature of Hebbian learning (and deep learning) does not account for the way that people can, very effectively, learn from a single occurrence or experience (sometimes called "one-trial" learning)[25].
- Hebb was aware that his theory of learning with cell assemblies would not account for one-trial learning and he proposed a "reverberatory" theory for that kind of learning (Hebb, 1949, p. 62). But, as noted in Wolff (2016, Section V-D), Milner has pointed out (Milner, 1996) that it is difficult to understand how this kind of mechanism could explain our ability to assimilate a previously-unseen telephone number: for each digit in the number, its pre-established cell assembly may reverberate; but this does not explain memory for the *sequence* of digits in the number. And it is unclear how the proposed mechanism would encode a phone number in which one or more of the digits is repeated.
- One-trial learning is consistent with the SP theory because the direct intake and storage of sensory information is bedrock in how the system learns (Section 3.4).
- The SP theory can also account for the relatively slow learning of complex skills such as how to talk or how to play tennis at a

[21] See "Gate array," *Wikipedia*, bit.ly/1UdB46j, retrieved 2016-03-20.

[22] See "Application-specific integrated circuit," *Wikipedia*, bit.ly/1pUs2y8, retrieved 2016-03-20.

[23] See "Field-programmable gate array," *Wikipedia*, bit.ly/1Hgi9iH, retrieved 2016-03-20.

[24] See, for example, "Don't despair if Google's AI beats the world's best Go player," *MIT Technology Review*, bit.ly/1p7Wzb7, 2016-03-08; and "Google unveils neural network with "superhuman" ability to determine the location of almost any image", *MIT Technology Review*, bit.ly/1p5qmSe, 2016-02-24.

[25] It may be argued that Hebbian learning may apply in such cases because a single experience may be mentally rehearsed. But that begs the question of how the one experience is remembered between when it first occurred and the first rehearsal—and likewise later on. And, while rehearsal may be helpful in some cases, it seems that there are many things we do remember after a single experience, without rehearsal.

high standard—because of the complexity of the abstract space of possible solutions that needs to be searched.

Does this mean that Hebbian learning is dead? Probably not:

- In some forms, the phenomena of "long-term potentiation" (LPT) in neural functioning seem to be linked to Hebbian types of learning (Squire et al., 2013, pp. 1022–1023).
- Gradual strengthening of neural connections may have a role to play in SP-neural because some such mechanism is needed to record, at least approximately, the frequency of occurrence of neural symbols and pattern assemblies (Sections 3.4, 4.5).

11. THE PROBLEMS OF SPEED AND EXPRESSIVENESS IN THE CREATION AND DESTRUCTION OF NEURAL STRUCTURES

A general issue for any neural theory of the representation and processing of knowledge, is how to account for the speed with which we can create neural structures, and, probably, destroy them, bearing in mind that such structures must be sufficiently versatile to accommodate the representation and processing of a wide range of different kinds of knowledge. This issue arises mainly in the following connections:

- *One-trial learning*. In keeping with the remarks above about one-trial learning (Section 10.5), it is a familiar feature of everyday life that we can see and hear something happening—a football match, a play, a conversation, and so on—and then, immediately or some time later, give a description of the event. This implies that we can lay down relevant memories at speed.
- *The learning of complex knowledge and skills*. If we accept the view of unsupervised learning which is outlined in Sections 3.4, 3.7, and 10, then it seems necessary to suppose that pattern assemblies are created and destroyed during the search for one or two grammars that provide a "good" description of the knowledge or skills that is being learned—and it seems likely that the creation and destruction of pattern assemblies would be fast.
- *The interpretation of sensory data*. In processes like the parsing of natural language or, more generally, understanding natural language, and in processes like pattern recognition, reasoning, and more, it seems necessary to create intermediate structures like those shown in **Figure 2**, and for those structures to be created at speed.
- *Speech and action*. In a similar way, it seems necessary for us to create mental structures fast in any kind of activity that requires thought, such as speaking in a way that is meaningful and comprehensible, most kinds of sport, most kinds of games, and so on.
- *Imagination*. Most people have little difficulty in imagining things they are unlikely ever to have seen—such as a cat with a coat made of grass instead of fur, or a cow with two tails. We can create such ideas fast and, if we like them well enough, we may remember them for years.

One possible solution, which is radically different from SP-neural, is to suppose that our knowledge is stored in some chemical form such as DNA, and that the kinds of mental processes mentioned above might be mediated via the creation and modification of such chemicals. Another possibility is that learning is mediated by epigenetic mechanisms, as outlined in Baars and Gage (2010, Section 7.4). Without wishing to prejudge what the primary mechanism of learning may be, or whether perhaps there are several such mechanisms, this paper focusses on SP-neural and how it may combine speed with expressiveness, as seems to be required for the kinds of functions outlined above.

At first sight, the problem of speed in the creation of neural structures is solved via the long-established idea that we can remember things for a few seconds via a "short-term memory[26]" that is distinct from "long-term memory[27]," and "working memory[28]." But there is some uncertainty about the extent to which these three kinds of memory may be distinguished, one from another, and there is considerable uncertainty about how they might work, and how information may be transferred from one kind of memory to another.

As a proffered contribution to discussions in this area, the suggestion here is that, in any or all of short-term memory, working memory, and long-term memory, SP-neural may achieve the necessary speed in the creation of new structures, combined with versatility in the representation and processing of diverse kinds of knowledge, by the switching on and off of synapses in pre-established neural structures and their interconnections, as outlined in Sections 10.1, 10.2.

With regard to possible mechanisms for the switching on and off of synapses:

- It appears that, in the entorhinal cortex between the hippocampus and the neocortex, there are neurons that can be switched "on" and "off" in an all-or-nothing manner (Tahvildari et al., 2007), and we may suppose that synapses have a role to play in this behavior.
- "The efficacy of a synapse can be potentiated through at least six mechanisms" (Squire et al., 2013, Caption to Figure 47.10) and it is possible that at least one them has the necessary speed, especially since "[Long-term potentiation] is defined as a persistent increase in synaptic strength ... that can be induced *rapidly* by a brief burst of spike activity in the presynaptic afferents." (emphasis added) (Squire et al., 2013, p. 1016).
- "[Long-term depression] is believed by many to be ... a process whereby [Long-term potentiation] could be reversed in the hippocampus and neocortex ...." (Squire et al., 2013, p. 1023).
- "... it is now evident that [Long-term potentiation], at least in the dentate gyrus, can either be ... stable, lasting months or longer." (Abraham, 2003, Abstract), although there appears to be little or no evidence with a bearing on whether or not there might be an upper limit to the duration of long-term potentiation.
- There is evidence that the protein kinase Mζ (PKMζ) may provide a means of turning synapses on and off, and thus perhaps storing long-term memories (Ogasawara and Kawato, 2010).

[26] "Short-term memory," *Wikipedia*, bit.ly/1RzAVHN, retrieved 2016-04-04.
[27] "Long-term memory," *Wikipedia*, bit.ly/1M9uPhh, retrieved 2016-04-04.
[28] "Working memory," *Wikipedia*, bit.ly/1PQq0UA, retrieved 2016-04-04.

With all these possible mechanisms, key questions are: do they act fast enough to account for the speed of the phenomena described above; and can they provide the basis for memories that can last for 50 years or more.

12. ERRORS OF OMISSION, COMMISSION, AND SUBSTITUTION

A prominent feature of human perception is that we have a robust ability to recognize things despite disturbances of various kinds. We can, for example, recognize a car when it is partially obscured by the leaves and branches of a tree, or by falling snow or rain.

One of the strengths of SP-abstract and its realization in the SP computer model is that, in a similar way, recognition of a New pattern or patterns is not unduly disturbed by errors of omission, commission, and substitution in those data (Wolff, 2006, Chapter 6, Wolff, 2013, Section 4.2.2). This is because of the way the SP computer model searches for a global optimum in the building of multiple alignments, so that it does not depend on the presence or absence of any particular feature or combination of features in the New information that is being analyzed.

In its overall structure, SP-neural seems to lend itself to that kind of robustness in recognition in the face of errors in data. But the devil is in the detail. In further development of the theory, and in the development of a computer model of SP-neural, it will be necessary to clarify the details of how that kind of robustness may be achieved. In shaping this aspect of SP-neural, the principles that have been developed in SP-abstract are likely to prove useful and, with empirical evidence from brains and nervous systems, they may serve as a touchstone of success.

13. CONCLUSION

As was mentioned in the Introduction, SP-neural is a tentative and partial theory. That said, the close relationship between SP-neural and SP-abstract, the incorporation into SP-abstract of many insights from research on human perception and cognition, strengths of SP-abstract in terms of simplicity and power (Section 3.8.1), and advantages of SP-abstract compared with other AI-related systems (Section 3.8.3)—lend support to SP-neural as it is now as a conceptual model of the representation and processing of knowledge in the brain, and a promising basis for further research.

Naturally, we may have more confidence in some parts of the theory than others. Arguably, the parts that inspire most confidence are these:

- *Neural symbols and pattern assemblies*. *All* knowledge is represented in the cerebral cortex with *pattern assemblies*, the neural equivalent of patterns in SP-abstract. Each such pattern assembly is an array of *neural symbols*, each of which is a single neuron or a small cluster of neurons—the neural equivalent of a symbol in SP-abstract. Topologically, each array has one or two dimensions, perhaps parallel to the surface of the cortex.
- *Information compression via the matching and unification of patterns*. As in SP-abstract, SP-neural is governed by the overarching principle that many aspects of perception and cognition may be understood in terms of information compression via the matching and unification of patterns.
- *Information compression via multiple alignment*. More specifically, SP-neural is governed by the overarching principle that many aspects of perception and cognition may be understood via a neural equivalent of the powerful concept of *multiple alignment*.
- *Unsupervised learning*. As in SP-abstract, unsupervised learning in SP-neural is the foundation for other kinds of learning—supervised learning, reinforcement learning, learning by imitation, learning by being told, and so on. And as in SP-abstract, unsupervised learning in SP-neural is achieved via a search through alternative grammars to find one or two that score best in terms of the compression of sensory information. As noted in Section 10.5, this is quite different from the kinds of "Hebbian" learning that are popular in artificial neural networks.
- *Problems of speed and expressiveness in the creation of pattern assemblies and their interconnections*. To account for the speed with which we can assimilate new information, and the speed of other mental processes (Section 11), it seems necessary to suppose that pattern assemblies and their interconnections may be created from pre-existing neural structures by the making and breaking of synaptic connections, somewhat like the making and breaking of connections in a telephone exchange, or the creation of a bespoke electronic system from an "uncommitted logic array" (ULA) or a "field-programmable gate array" (FPGA).

As with SP-abstract, areas of uncertainty in SP-neural may be clarified by casting the theory in the form of a computer model and testing it to see whether or not it works as anticipated. It is envisaged that this would be part of a proposed facility for the development of the SP machine (Wolff and Palade, 2016), a means for researchers everywhere to explore what can be done with the SP machine and to create new versions of it.

At all stages in its development, the theory may suggest possible investigations of the workings of brains and nervous systems. And any neurophysiological evidence may have a bearing on the perceived validity of the theory and whether or how it may need to be modified.

AUTHOR CONTRIBUTIONS

This is part of a long term research programme by JGW, developing the SP theory of intelligence. SP-neural, which is the subject of this paper, was first outlined in Chapter 11 of "Unifying Computing and Cognition" and, in this paper, has been considerably refined and developed.

FUNDING

The research is funded by CognitionResearch.org.

ACKNOWLEDGMENTS

I'm grateful to referees for constructive comments on earlier drafts of this paper.

REFERENCES

Abraham, W. C. (2003). How long will long-term potentiation last? *Philos. Trans. R. Soc. B* 358, 735–744. doi: 10.1098/rstb.2002.1222

Attneave, F. (1954). Some informational aspects of visual perception. *Psychol. Rev.* 61, 183–193. doi: 10.1037/h0054663

Baars, B. J., and Gage, N. M. (2010). *Cognition, Brain, and Consciousness: Introduction to Cognitive Neuroscience, 2nd Edn*. Amserdam: Elsevier.

Barghout-Stein, L. (1999). *On Differences between Peripheral and Foveal Pattern Masking*. Technical report, University of California, Berkeley, Master's thesis. Available online at: bit.ly/1SCoUO4

Barlow, H. B. (1959). "Sensory mechanisms, the reduction of redundancy, and intelligence," in *The Mechanisation of Thought Processes* (London: Her Majesty's Stationery Office), 535–559.

Barlow, H. B. (1969). "Trigger features, adaptation and economy of impulses," in *Information Processes in the Nervous System*, ed K. N. Leibovic (New York, NY: Springer), 209–230.

Barlow, H. B. (1982). David Hubel and Torsten Wiesel: their contribution towards understanding the primary visual cortex. *Trends Neurosci.* 5, 145–152. doi: 10.1016/0166-2236(82)90087-X

Barrow, J. D. (1992). *Pi in the Sky*. Harmondsworth: Penguin Books.

de Penning, H. L. H., d'Avila Garcez, A. S., Lamb, L. C., and Meyer, J.-J. C. (2011). "A neural-symbolic cognitive agent for online learning and reasoning," in *Proceedings of the International Joint Conferences on Artificial Intelligence* (Palo Alto, CA), 1653–1658.

Drew, T., Võ M. L.-H., and Wolfe, J. M. (2013). The invisible gorilla strikes again: sustained inattentional blindness in expert observers. *Psychol. Sci.* 24, 1848–1853. doi: 10.1177/0956797613479386

d'Avila Garcez, A., Besold, T. R., de Raedt, L., Földiák, P., Hitzler, P., Icard, T., et al. (2015). "Neural-symbolic learning and reasoning: contributions and challenges," in *Proceedings of the AAAI Spring Symposium on Knowledge Representation and Reasoning, 2015*, ed T. Walsh (Stanford, CA), 18–21.

d'Avila Garcez, A. S. (2007). "Advances in neural-symbolic learning systems: modal and temporal reasoning," in *Perspectives of Neural-Symbolic Integration*, eds B. Hammer and P. Hitzler (Heidelberg), 265–282.

d'Avila Garcez, A. S., Lamb, L. C., and Gabbay, D. M. (2009). *Neural-Symbolic Cognitive Reasoning*. Heidelberg: Springer.

Gold, M. (1967). Language identification in the limit. *Inform. Control* 10, 447–474. doi: 10.1016/S0019-9958(67)91165-5

Gross, C. G. (2002). Genealogy of the "Grandmother Cell". *Neuroscientist* 8, 512–518. doi: 10.1177/107385802237175

Hebb, D. O. (1949). *The Organization of Behaviour*. New York, NY: John Wiley & Sons.

Herculano-Houzel, S. (2012). The remarkable, yet not extraordinary, human brain as a scaled-up primate brain and its associated cost. *Proc. Natl. Acad. Sci. U.S.A.* 109(Suppl. 1), 10661–10668. doi: 10.1073/pnas.1201895109

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., and Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* 532, 453–458. doi: 10.1038/nature17637

Isaacson, J. S., and Scanziani, M. (2011). How inhibition shapes cortical activity. *Neuron* 72, 231–243. doi: 10.1016/j.neuron.2011.09.027

Komendantskaya, E., Lane, M., and Seda, A. K. (2007). "Connectionist representation of multi-valued logic programs," in *Perspectives of Neural-Symbolic Integration*, eds B. Hammer and P. Hitzler (Heidelberg), 283–313. doi: 10.1007/978-3-540-73954-8_12

McCorduck, P. (2004). *Machines Who Think: A Personal Inquiry Into the History and Prospects of Artificial Intelligence, 2nd Edn*. Natick, MA: A. K. Peters Ltd.

Milner, P. M. (1996). Neural representations: some old problems revisited. *J. Cogn. Neurosci.* 8, 69–77. doi: 10.1162/jocn.1996.8.1.69

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

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

Ogasawara, H., and Kawato, M. (2010). The protein kinase mζ network as a bistable switch to store neuronal memory. *BMC Syst. Biol.* 4:181. doi: 10.1186/1752-0509-4-181

Ratliff, F., Hartline, H. K., and Miller, W. H. (1963). Spatial and temporal aspects of retinal inhibitory interaction. *J. Opt. Soc. Am.* 53, 110–120. doi: 10.1364/JOSA.53.000110

Roy, A. (2013). An extension of the localist representation theory: grandmother cells are also widely used in the brain. *Front. Psychol.* 4:300. doi: 10.3389/fpsyg.2013.00300

Schmidhuber, J. (2015). Deep learning in neural networks: an overview. *Neural Netw.* 61, 85–117. doi: 10.1016/j.neunet.2014.09.003

Shamma, S. A. (1985). Speech processing in the auditory system II: lateral inhibition and the central processing of speech evoked activity in the auditory nerve. *J. Acoust. Soc. Am.* 76, 1622–1632. doi: 10.1121/1.392800

Simons, D. J., and Ambinder, M. S. (2005). Change blindness: theory and consequences. *Curr. Direct. Psychol. Sci.* 14, 44–48. doi: 10.1111/j.0963-7214.2005.00332.x

Squire, L. R., Berg, D., Bloom, F. E., du Lac, S., Ghosh, A., and Spitzer, N. C. (eds.). (2013). *Fundamental Neuroscience, 4th Edn*. Amsterdam: Elsevier.

Stratton, G. M. (1897). Upright vision and the retinal image. *Psychol. Rev.* 4, 182–187. doi: 10.1037/h0064110

Tahvildari, B., Fransén, E., Alonso, A. A., and Hasselmo, M. E. (2007). Switching between "on" and "off" states of persistent activity in lateral entorhinal layer III neurons. *Hippocampus* 17, 257–263. doi: 10.1002/hipo.20270

von Békésy, G. (1967). *Sensory Inhibition*. Princeton, NJ: Princeton University Press.

Wolff, J. G. (1988). "Learning syntax and meanings through optimization and distributional analysis," in *Categories and Processes in Language Acquisition*, eds Y. Levy, I. M. Schlesinger, and M. D. S. Braine (Hillsdale, NJ: Lawrence Erlbaum), 179–215. Available online at: bit.ly/ZIGjyc

Wolff, J. G. (2006). Medical diagnosis as pattern recognition in a framework of information compression by multiple alignment, unification and search. *Decis. Support Syst.* 42, 608–625. doi: 10.1016/j.dss.2005.02.005

Wolff, J. G. (2006). *Unifying Computing and Cognition: the SP Theory and Its Applications*. Menai Bridge: CognitionResearch.org. ISBNs: 0-9550726-0-3 (ebook edition), 0-9550726-1-1 (print edition). Distributors, including Amazon.com. Available online at: bit.ly/WmB1rs

Wolff, J. G. (2007). Towards an intelligent database system founded on the SP theory of computing and cognition. *Data Knowl. Eng.* 60, 596–624. doi: 10.1016/j.datak.2006.04.003

Wolff, J. G. (2013). The SP theory of intelligence: an overview. *Information* 4, 283–341. doi: 10.3390/info4030283

Wolff, J. G. (2014a). Application of the SP theory of intelligence to the understanding of natural vision and the development of computer vision. *SpringerPlus* 3, 552–570. doi: 10.1186/2193-1801-3-552

Wolff, J. G. (2014b). Autonomous robots and the SP theory of intelligence. *IEEE Access* 2, 1629–1651. doi: 10.1109/ACCESS.2014.2382753

Wolff, J. G. (2014c). Big data and the SP theory of intelligence. *IEEE Access* 2, 301–315. doi: 10.1109/ACCESS.2014.2315297

Wolff, J. G. (2014d). *Information Compression, Intelligence, Computing, and Mathematics*. Technical report, CognitionResearch.org. Available online at: bit.ly/1jEoECH

Wolff, J. G. (2014e). The SP theory of intelligence: benefits and applications. *Information* 5, 1–27. Available online at: bit.ly/1lcquWF

Wolff, J. G. (2016). The SP theory of intelligence: its distinctive features and advantages. *IEEE Access* 4, 216–246. doi: 10.1109/ACCESS.2015.2513822

Wolff, J. G., and Palade, V. (2016). *Short Proposal for the Development of the SP Machine*. Technical report, CognitionResearch.org. Available online at: bit.ly/1SKAjhZ

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

APPENDIX

Cell Assemblies and Pattern Assemblies

The main differences between Hebb's (1949) concept of a "cell assembly" and the SP-neural concept of a "pattern assembly" are:

- The concept of a pattern assembly has had the benefit of computer modeling of SP-abstract—reducing vagueness in the theory and testing whether or not proposed mechanisms actually work as anticipated. These things would have been difficult or impossible for Hebb to do in 1949.
- Cell assemblies were seen largely as a vehicle for recognition, whereas, as neural realizations of SP "patterns," pattern assemblies should be able to mediate several aspects of intelligence, including recognition.
- Anatomically, pattern assemblies are seen as largely flat groupings of neurons in the cerebral cortex (Section 4.2), whereas cell assemblies are seen as structures in three dimensions.
- As described below, a fourth difference between cell assemblies and pattern assemblies is in how structures may be shared.

With regard to the last point, possible models for sharing of structures are illustrated in **Figure A1**.

In literal sharing, structures B and C in the figure both contain structure A. In sharing by copying, structures B and C each contains a copy of structure A. While in sharing by reference, structures B and C each contains a reference to structure A, in much the same way that a paper like this one contains references to other publications.

From Hebb's (1949) descriptions of the cell assembly concept, it is difficult to tell which of these three possibilities are intended.

By contrast with the concept of a pattern assembly in SP-neural, sharing is almost always achieved by means of neural "references" between structures. For example, a noun like "table" is likely to have neural connections to the many grammatical contexts in which it may occur, as suggested by the two broken-line connections from each of "N" and "#N" in the pattern assembly for "table" shown in **Figure 3**. Notice that, in this example, the putative direction of travel of nerve impulses is not relevant—it is the neural connection that counts.

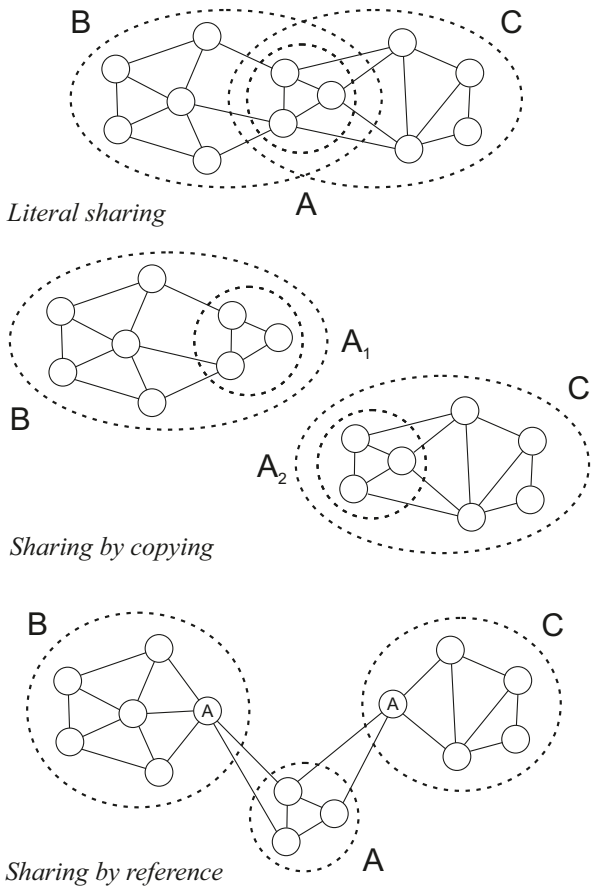


FIGURE A1 | Three possible ways in which cell assemblies may be shared, as described in the text. Adapted with permission from Wolff (2006, Figure 11.3).

In the SP system, it is intended that literal sharing should be impossible and that sharing by copying may only occur on the relatively rare occasions when the system has failed to detect the corresponding redundancy, and not always then.

frontiers in Psychology

HYPOTHESIS AND THEORY
published: 10 August 2017
doi: 10.3389/fpsyg.2017.01297

Linking Neural and Symbolic Representation and Processing of Conceptual Structures

Frank van der Velde[1,2]*, *Jamie Forth*[3], *Deniece S. Nazareth*[1] and *Geraint A. Wiggins*[4]

[1] Department of Cognitive Psychology and Ergonomics, Faculty of Behavioural, Management and Social sciences, University of Twente, Enschede, Netherlands, [2] Institute of Psychology (IOP), Leiden, Netherlands, [3] Department of Computing, Goldsmiths University of London, London, United Kingdom, [4] Computational Creativity Lab, Cognitive Science Group, School of Electronic Engineering and Computer Science, Queen Mary University of London, London, United Kingdom

We compare and discuss representations in two cognitive architectures aimed at representing and processing complex conceptual (sentence-like) structures. First is the Neural Blackboard Architecture (NBA), which aims to account for representation and processing of complex and combinatorial conceptual structures in the brain. Second is IDyOT (Information Dynamics of Thinking), which derives sentence-like structures by learning statistical sequential regularities over a suitable corpus. Although IDyOT is designed at a level more abstract than the neural, so it is a model of cognitive function, rather than neural processing, there are strong similarities between the composite structures developed in IDyOT and the NBA. We hypothesize that these similarities form the basis of a combined architecture in which the individual strengths of each architecture are integrated. We outline and discuss the characteristics of this combined architecture, emphasizing the representation and processing of conceptual structures.

Keywords: cognitive architecture, memory representation, hebbian learning, compositional learning, incremental learning, *In situ* representations

OPEN ACCESS

Edited by:
Leonid Perlovsky,
Harvard University and Air Force Research Laboratory, United States

Reviewed by:
Antonio Chella,
University of Palermo, Italy
Dorina Rajanen,
University of Oulu, Finland

***Correspondence:**
Frank van der Velde
f.vandervelde@utwente.nl

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *31 July 2016*
Accepted: *17 July 2017*
Published: *10 August 2017*

Citation:
van der Velde F, Forth J, Nazareth DS and Wiggins GA (2017) Linking Neural and Symbolic Representation and Processing of Conceptual Structures. Front. Psychol. 8:1297. doi: 10.3389/fpsyg.2017.01297

1. INTRODUCTION

The ability to represent and process conceptual structures, as found in language processing, reasoning, and in generating conceptual representations from visual and auditory perception, are key elements of human cognition. They can be studied with the aim to understand human cognition and its relation to the brain. But they can also be targets for the development of artificial cognitive systems. These aims can be combined to various degrees, because a cognitive architecture that provides an understanding of a (neural) cognitive process can also be used in artificial systems, and, conversely, the way in which an artificial system processes complex information can reveal aspects of human processing as well.

Here, we discuss and relate the different representations used in two cognitive architectures, one neural and one symbolic, in which complex conceptual structures can be represented and processed. That is, we discuss and illustrate the different ways in which complex conceptual structures are represented or learned in the two architectures and how these representations could be related.

In particular, we aim to outline how combined representations could be developed, for use in a combined architecture in which aspects of our neural and symbolic architectures are integrated. We hypothesize that such a combined architecture could serve as a model of human conceptual processing and its relation to the brain. When implemented, it could also serve as a new artificial

architecture in which forms of neural (parallel) hardware and neural and symbolic forms of learning and processing could be integrated. We are as yet at the beginning of the integration of our architectures, which is also a reason why we focus on issues of representation here.

The neural representation in our integration is that used in the Neural Blackboard Architecture (NBA), which is aimed to represent and process conceptual structures in language (e.g., van der Velde and de Kamps, 2006, 2010), reasoning and other cognitive domains (van der Velde, 2016a). The NBA assumes that conceptual representations in the brain consist of dedicated network structures, or neural assemblies, that develop over time and that can be distributed over wide areas in the brain and cortex. A fundamental characteristic of these network-like conceptual representations is that they are always content addressable, whether they are activated in isolation or whether they are parts of more complex (and even hierarchical) conceptual structures, such as sentences in language.

The NBA provides "neural blackboards" that afford the representation and processing of complex conceptual structures based on neural assembly conceptual representations in specific cognitive domains. Examples are neural blackboards for sentence structures, phonological structures, sequences, and relations as used in reasoning. In each domain, a dedicated neural blackboard will provide a range of specialized structural elements that can bind in a neural manner to the neural assemblies (e.g., representing "words" in language). The neural bindings, implemented with neural circuits, allow the creation and processing of more complex cognitive structures (e.g., "sentences") in a combinatorial manner.

The symbolic representation in our integration is that used in Information Dynamics of Thinking (IDyOT). IDyOT derives (e.g., sentence-like) structures by learning statistical sequential regularities over a linguistic corpus (Wiggins, 2012b; Wiggins and Forth, 2015; Forth et al., 2016). IDyOT is unusual as a machine learning formalism in that it is symbolic in nature, but it generates and gives explicit semantics to its own symbols, in a bottom-up learning process, which is optimized by a general, data-independent principle of information efficiency, conceptualized as predictive accuracy. These symbols correspond with concepts in the semantics of the system. Another unusual aspect of IDyOT's operation is that both representations and sequential models are optimized simultaneously with respect to the prediction accuracy of the models, causing a trade-off between overfitting and accuracy that we propose as a model of the corresponding trade-off in human cognition. The explanation of this process is a novel contribution of the current paper.

The representational links between IDyOT and NBA concern the nature of the dedicated structural elements that allow processing and representation of complex conceptual structures, the way these elements could be activated during processing, and the underlying semantics of the architectures in the form of conceptual spaces that possess a geometrical structure (Gärdenfors, 2000, 2014).

In the NBA, the dedicated structural elements form the neural blackboards. The kinds of elements used and the way they are activated derive from analyses of the cognitive domains at hand, as in the sentence NBA (e.g., van der Velde and de Kamps, 2006, 2010). However, the combination of NBA with IDyOT provides the possibility to derive these structural elements by learning from real corpora. Conversely, the NBA could provide a neural implementation of the more higher-level formal account as provided by IDyOT. Thus, IDyOT potentially supplies a higher-level formal account and learning abilities to the operations of the NBA. Conversely, the NBA provides a route toward a neural implementation of IDyOT, which could also form the basis of in parallel operating hardware.

2. THEORETICAL POSITION AND NOVELTY

Our theoretical position here is that the representations used in NBA and IDyOT are in fact two different representations of the same thing, at different levels of abstraction, but with focus on similar representational affordances. In the following sections, we describe the representations, and the relations between them—but, as always, to understand the representations it is necessary also to understand the processes that work over them.

The novelty in the current paper lies in several places, primarily in the thorough-going comparison between the representations and corresponding processes in the two architectures. The entire description of IDyOT memory construction is also novel, and we present a novel simulation of neural activity based on the NBA, which allows for a detailed comparison with brain activity observed in human (sentence) processing. To the best of our knowledge, such a detailed potential comparison between human brain activity and simulated model activity is not available in the case of high-level cognitive processing, such as sentence comprehension. This also strongly motivates the integration of our architectures, because that would endow the NBA with the learning capabilities of IDyOT, based on real corpora (as outlined below). In turn, the dynamics and structure of the NBA would then allow a comparison between the representations and underlying processing as learned by IDyOT and human brain activity.

The structure of the paper is as follows. In the next two sections we briefly describe the representations used in NBA and IDyOT in turn, also giving detail of processing where appropriate. In the sections that follow, we discuss a number of specific links between NBA and IDyOT and the potential benefits of their integration.

3. NEURAL BLACKBOARD ARCHITECTURE

In our outline of Neural Blackboard Architecture, or NBA for short, we focus on the representation of concepts (e.g., underlying words) in the architecture and the representational structures that are used to integrate concepts in more complex cognitive structures, such as relations and sentences.

The basis of concept representation in the NBA are "neural assemblies," as proposed by Donald Hebb (1949). In the view of Hebb, these neural assemblies develop over time by interconnecting the neurons in the brain that are involved

in processing (sensory) information and generating actions related to the concept they represent. However, unlike Hebbian assemblies, conceptual representations in the NBA are not only associative. Instead, they can (and mostly will) contain relational structures as well.

Figure 1 illustrates a neural assembly representation of *cat*. It would be distributed over different areas in the cortex and brain, depending on the kind of information involved, including networks processing perceptual information about cats and networks that can produce specific actions (e.g., pronouncing the word "cat"). But also networks representing emotional content or associations related to cats belong to the assembly, and networks that instantiate relations, such as *cat is pet*.

The combination of perception and action in the assembly structure of a concept entails that both the patterns (and activation) of "incoming" and "outgoing" connections determine the meaning of a concept. For example, neurons observed in the medial temporal cortex responded to a person whether the person or his or her name was presented (Quian Quiroga, 2012). These "perception" networks do in fact belong to the assembly structure of a concept, because without them the concept could not be activated (or was not learned). But this would capture only part of the role and (hence) meaning of the conceptual representation involved. Equally important would be the effect of these neurons on downstream processing (van der Velde, 2015).

The notion that conceptual representations interconnect sensory information processing and action generation underscores their role in producing behavior. The ability to produce behavior is a crucial aspect of cognition (and hence of every neuro-cognitive model) because the evolution of cognition depended on the ability to produce behavior. Without advocating a behavioristic view of cognition (e.g., as the basis of modeling cognition) we do argue that the prime role of cognition is to intervene in the reflex (cf. Shanahan, 2010). In this view, the need for a connection structure that transfers sensory activity to motor activity should always be at the background of a neural-cognitive model.

So, in **Figure 1**, it is not just the gray oval that represents the concept *cat*, but instead the entire network structure to which it is connected. The gray oval could play the role of a higher-level representation of the concept in the sense that it interconnects the concept to other networks. But it would be wrong to see this as the "genuine" encoding of the concept. Without the networks to which it is connected, the gray oval does not encode anything.

An important feature of conceptual representations given by neural assemblies is that they are "*in situ*." This entails that they cannot be copied and transported to create more complex structural representations with them (e.g., as found in language or reasoning). Instead, the same assembly (or a part thereof) is always activated when the concept it represents is tokened. One consequence of this kind of representation is that an assembly can develop and grow over time, as originally discussed by Hebb.

Another direct consequence of the *in situ* nature of a neural assembly is that the concept it represents is content addressable. This entails that the same assembly (or part thereof) will be activated when sufficient information about the concept it represents is available (e.g., perceived), even when the concept is part of a (complex) sentence structure.

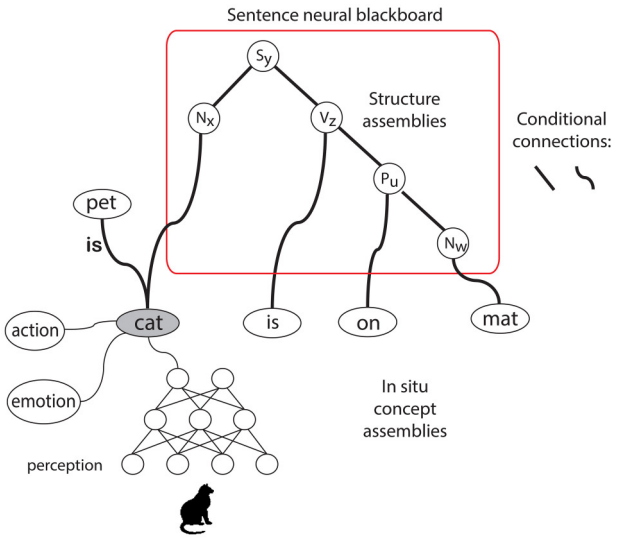


FIGURE 1 | Left: Neural assembly representation of *cat*. **Right**: Sentence structure in the Neural Blackboard Architecture: a sentence neural blackboard (temporarily) interconnects the "*in situ*" concept representations (given by neural assemblies) of *cat*, *is*, *on*, and *mat* to form the sentence structure *cat is on mat*. The thick line connections represent "conditional connections." They can be opened by gating circuits that are either activated by sustained activity in working memory neural populations (representing binding) or by neural control circuits (e.g., performing parsing operations). N, noun; P, preposition; S, sentence; V, verb.

Huth et al. (2016) give an indication of the *in situ* nature of conceptual representation in the brain. They measured brain activity related to words when people were listening to stories (in an fMRI scanner). The parts of the cortex that responded to the words (after statistical analysis) were much larger compared to previous studies in which only individual words were presented. The analysis divided the left hemisphere (LH) into 192 distinct functional areas, 77 of which were semantically selective. The right hemisphere was divided into 128 functional areas, 63 of which were semantically selective (even though the RH is usually regarded as not being involved in language). Remarkably, the organization of these areas was quite similar over the different (7) subjects involved in the study. Furthermore, next to these semantic areas, other areas also responded to other aspects of words (e.g., Broca's area).

Because the study was focused on semantic representation, the words observed in the study were categorized into 12 semantic domains. These domains tiled the cortex in terms of the 77 areas in the LH and the 63 areas in the RH referred to above. Inspection of the data reveals that semantic domains are generally represented in different tiles, distributed over the LH and/or RH cortex.

The semantic representation as observed by Huth et al. (2016) seems to be in line with the Hebbian assembly hypothesis, in that these representations would have arisen over time, and would (partly) be determined by the context in which the concepts were

processed. This could explain why, e.g., the same visual concept (e.g., *colored*) activates areas near the visual cortex but also in the prefrontal cortex. This pattern of activation could reflect different parts of the assembly of the concept, and their selective activation would then be determined by the context (visual processing vs. motor behavior) in which the concept is used and learned. The fact that a concept generates activation in different cortical areas is in line with the assembly representation as illustrated in **Figure 1**.

3.1. Neural Blackboards as Connection Paths

If concepts are represented and distributed as *in situ* assemblies, the question arises of how they could be combined to represent more complex cognitive structures, such as relations or sentences.

The key notion of the NBA is that more complex cognitive structures are formed by providing (temporal) *connection paths* between the assemblies (concepts) they contain, in relation with the structure they express. These (temporal) connection paths are formed and controlled in "neural blackboards."

For example, in the case of language, the NBA provides a connection structure (or connection path) that allows arbitrary words in a given language to be (temporarily) interconnected in accordance with the structure of the sentence. The words in this case are the network structures (neural assemblies) as described by Huth et al. (2016). The neural blackboards in the NBA provide a "small world" network structure that would allow the *in situ* and distributed concept assemblies ("words") to be interconnected using a limited set of intermediary "hubs and sub-hubs," given by the structure assemblies and their potential bindings in the blackboards. Small world networks are found in a wide variety of natural and man-made structures because they allow arbitrary interconnectivity with minimal means. They also play an important role in the brain (Shanahan, 2010).

Figure 1 illustrates how in the NBA a sentence can be formed with *in situ* concepts encoded by neural assemblies. The *in situ* assemblies for *cat*, *is*, *on*, and *mat* are bound to a "neural blackboard" to form the sentence *cat is on mat*.

Figure 1 illustrates the very basic aspects of the neural blackboards that the NBA uses to encode relations between *in situ* concept assemblies. In the case of language there are (at least) two neural blackboards involved. One is a phonological blackboard, which is not illustrated here. The other is the sentence blackboard which encodes sentence structures, as illustrated here with the sentence *cat is on mat*.

The need for both a phonological and a sentence blackboard derives from the productivity of natural language. Language has (at least) a two tier productive structure (Jackendoff, 2002) in which first phonemes form words and then words (or word-phoneme combinations) form sentences. The combination of (familiar) phonemes allows the generation of a very large set of words, which can grow continuously in life. These words (including novel but phonetically regular words) can then be combined to give a practically unlimited set of sentences. Yet, it is important to realize that this two tier productivity is restricted to the languages we are familiar with. In the NBA, that means languages for which we have developed neural blackboards (van der Velde and de Kamps, 2015a).

van der Velde and de Kamps (2006, 2010) explain the structure and operations of the neural blackboards in detail. Here, we address a number of main issues, focusing on representational structures in the sentence blackboard. The composite structural elements of the sentence blackboard are "structure assemblies," as illustrated in **Figure 1**. They can bind to concept assemblies (or to "word assemblies" in the phonological blackboard) and they can bind to each other to generate the structure of the sentence (e.g., *cat is on mat*).

The thick-line connections in the blackboards play a crucial role in the process of generating and representing a sentence structure. These connections are "conditional connections," consisting of gating circuits. To operate as a connection, the gates in the connections have to be opened or activated. This ensures that activation does not flow without control in the neural blackboards, that is, the connections in the blackboards are not associative. The gates can be activated by working memory (WM) activation, representing a binding, and by control circuits, which represent (e.g., syntactic) operations in the architecture. We will discuss these operations in more detail later on.

So, the *in situ* assembly *cat* is bound (via the phonological blackboard) to a "Noun" structure assembly *Nx* in the sentence blackboard. Binding is achieved by working memory activation that opens the gates between the assemblies involved. To this end, the sentence blackboard has a number of Noun assembles which can all potentially bind to each of the Word assemblies in the phonological blackboard (via a matrix or tensor-like connection structure, see below). All bindings in all neural blackboards are of this kind. A specific binding in the "connection matrix" between assemblies is achieved by activating a specific working memory, which consists of sustained activation in a population of neurons. Once activated (by the mutual activation of the assemblies it binds), the population remains active on its own for a while due to "reverberating" activity (e.g., Amit, 1989). So, in this way, *cat* will bind to *Nx*. Similarly, *is* will bind to the Verb structure assembly *Vz*, *on* to the Preposition structure assembly *Pu* and *mat* to *Nw* (again, via the phonological blackboard).

Thus, to represent sentences based on *in situ* words (concepts), the NBA builds a connection path (structure) in the sentence (and phonological) blackboard, in accordance with the syntactic structure of the sentence. These sentences can be novel sentences based on familiar words (or even novel words based on familiar phonemes), and they can include hierarchical structures like (e.g., center) embedding (van der Velde and de Kamps, 2006, 2010). Once a connection structure is built it can be used to produce behavior, because it constitutes a connection path between the *in situ* concept assemblies it interconnects. In turn, this entails that it forms a (temporal) connection path between all perception and action structures embedded in these concept assemblies, thus forming a path between perception and action as the basis for behavior.

4. IDYOT: THE INFORMATION DYNAMICS OF THINKING

4.1. Overview

IDyOT (Information Dynamics of Thinking: Wiggins, 2012b; Wiggins and Forth, 2015; Forth et al., 2016) implements Baars' Global Workspace Theory (GWT; Baars, 1988), affording a computational model of a hypothetical cognitive architecture. At the functional level[1], a number of *generators* sample from a complex statistical model of sequences (explained below), performing Markovian prediction from context (Wiggins and Forth, 2015; Forth et al., 2016). Each generator indexes a string of symbols, forming a *chunk*, a final substring of the overall memory model, expressed as symbols, whose origin is explained below. Each indexed string serves as a context for prediction of the next (as yet unsensed) symbol; predictions are expressed as distributions over the alphabet used to express the input. A chunk is integrated into the memory and Global Workspace (which may be thought of as an AI blackboard: Corkill, 1991) when it meets a criterion based on information content. The upshot of this design is that IDyOT's primary cognitive operation is perceptual chunking. **Figure 2** gives a functional overview.

IDyOT maintains a cognitive cycle that continually predicts what is expected next, from a statistical model, expressed in terms of self-generated symbols that are given semantics by perceptual experience; it is thus focused on sequence. Perceptual input is matched against generators' predictions, and where a match leads to a larger increase in uncertainty than other current matches, the corresponding generator's chunk is flushed into the Global Workspace, and stored in memory, linked in sequence with the previous chunk. Chunks that fail to win are forgotten after a fixed period, the duration of which is question of the research. The model entails that, for perception to work, at least some generators must be working in all perceptual modalities at all times; otherwise no generator would be predicting for input in a newly active modality to match against. This activity may account for otherwise unexplained electrical brain activity that is not directly concomitant with perceived events, and it may be responsible for spontaneous creativity (Wiggins and Bhattacharya, 2014).

4.2. Representation, Memory, and Prediction in IDyOT

Each chunk, having been recorded, is associated with a symbol in a higher-level model, which adds to the overall predictive model. Each symbol corresponds with a point in a conceptual space (Gärdenfors, 2000, 2014) associated with its own layer, and each such point corresponds with a region or subspace of the conceptual space of the layer below, defined by the lower-level symbols in the chunk. Thus, there are two parallel representations: one symbolic and explicitly sequential; and one continuous and non-sequential, but encoding sequential information. The former provides evidence from which the latter is derived, while the latter provides semantics for the former.

[1]The formal implementation of this functional behavior is somewhat different in actuality. However, the description given here is easier to understand in isolation.

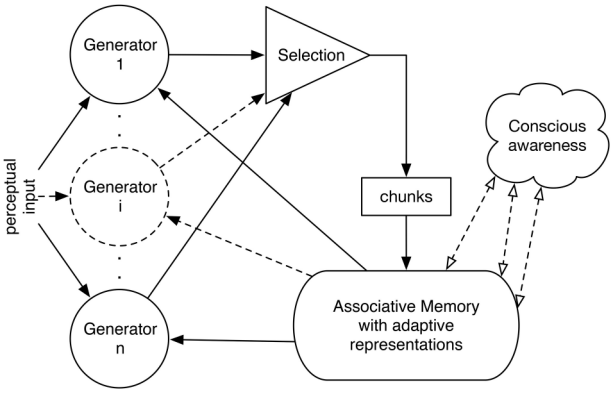


FIGURE 2 | Overview of the IDyOT (Information Dynamics of Thinking) architecture. Generators synchronized to perceptual input sample, given previous perceptual input (if any), from a first-order, multidimensional Markov model to predict the next symbol in sequence, which is matched with the input. Predicted symbols that match are grouped in sequence until a chunk is detected on grounds of its information profile. The generator then stores the chunk, as described in §7.1.3 and resets its chunk, which is the sum of the structured hierarchical memory and a detector that searches for salient information, shown as "conscious awareness" here. This allows the resulting chunk of sequence to be stored in the memory, to become part of the statistical model and thence to be used subsequently.

For grounding (or, more precisely, *tethering*: Sloman and Chappell, 2005), the lowest-level conceptual spaces are *a priori* defined by the nature of their sensory input (inspired by human biology: for example, auditory input models the output of the Organ of Corti); higher-level ones are inferred from the lower levels using the information in the sequential model. Structures may be grouped together in categories, according to similarity in their conceptual space, giving them semantics in terms of mutual interrelation. Using this, a consolidation phase allows membership of categories to be optimized, by local adjustment, in terms of the predictive accuracy of the overall model. Theoretically, the layering of models and its associated abstraction into categories can proceed arbitrary far up the constructed hierarchy (Wiggins, 2012b; Wiggins and Forth, 2015). Forth et al. (2016) provide an account of the representation of timing in IDyOT; these aspects, however, are beyond the scope of the current paper.

In general, the stimuli to which IDyOT will respond are sequences of atomic percepts. All the dimensions of music, pitch, timbre, amplitude and time, which also feature in speech, are used for prediction, as has been demonstrated in IDyOM (Pearce, 2005; Pearce et al., 2012), as can any other transduced signal. This demands a more powerful Markov model than is common in cognitive science language modeling. Conklin and Witten (1995) proposed a *viewpoint*-based approach that allows a set of interacting features, associated by means of sequences of multi-dimensional symbols, to perform multi-dimensional prediction. This is the system used in IDyOM and adapted for multidimensional language models by Wiggins (2012a). A key contribution of the viewpoint idea is the ability to superpose

distributions from different features with weights determined by their entropy (Pearce et al., 2005).

Given Conklin's notion of viewpoint (Conklin and Witten, 1995) and the associated mathematics, it becomes possible also to represent propositional meaning within the statistical framework: to do this, one incorporates representations of the meaning (perhaps drawn from another sensory modality, e.g., describing in language a scene representation derived from visual input) in the statistical model (Eshghi et al., 2013). Here, we presuppose a rich, multisensory input which allows associations to be constructed between different sensory modalities, on the basis of co-occurrence.

A key scientific advantage of this representations is that its symbols are (directly or indirectly) explicable in terms of IDyOT's perceptual input, and a record of that perception is maintained. Thus, its status as a cognitive model is more easily tested than in equally powerful, but less semantically transparent, learning systems, such as deep neural networks.

4.3. Summary: the Principles of IDyOT

In summary, the IDyOT model is based on 6 principles. Notations used in the current description of IDyOT are presented in the **Table 1**.

1. The fundamental function of cognition is to efficiently process sensory information so as to predict what is to happen next in the world.
2. Predictions are made by classifying events (§§6.1.3,7.1.3), counting likelihoods of short sequences, and building a literal model of the experience of the organism in these terms (§6.1.1). Predictions are expressed as distributions over alphabets of events.
3. Events are identified by chunking sensory input (§7.1.3).
4. The cognitive system always strives to maintain the optimal representation of its memory. Optimality is expressed in terms of the mean number of bits required to represent each symbol in the memory: smaller is better.
5. Meaning is constructed internally to the cognitive system, and incrementally, and consists in associations between symbols in the IDyOT memory (§§6.1.3,7.1.4).
6. Because the model maps directly to experience, it is learned incrementally (§7.1). This has the following consequences:
   a. Meanings attributed to symbols depend on the order of events that the model learns (§7.1).
   b. It is necessary from time to time to re-optimize the model, after an extended phase of incremental learning. This is termed *memory consolidation*. One consequence is that meanings can change retrospectively as the system learns.

TABLE 1 | Notation used in the current description of IDyOT.

$\aleph(v)$	The alphabet associated with viewpoint v.
$D_{t,A}$	The distribution that constitutes IDyOT's prediction at time point t over alphabet A.
$H(D)$	The estimated entropy of distribution D, over alphabet A: $\mathrm{H(D)} = -\sum_{s \in A} p(s) \log_2 p(s)$.
$h(D, s)$	The estimated information content of symbol s drawn from distribution D over alphabet A: $\mathrm{h(D,s)} = -\log_2 \mathrm{p(s)}$.
S_A	The conceptual space (Gärdenfors, 2000) associated with alphabet A.
$R_{A,s}$	The region of S_A that corresponds with the symbol $s \in A$.

5. NBA AND IDYOT AS COMPLEMENTARY APPROACHES TO REPRESENTATION

Although the NBA is a neural architecture whereas IDyOT is primarily a symbolic one, they are functionally and structurally related.

In particular, chunking plays a key role in this relation between the two architectures. Perceptual chunking is the key operation of IDyOT, but it is also the underlying principle of structure formation in the NBA. The neural blackboards in the NBA not only interconnect information or provide a workspace in which information can interact and compete, they also form larger chunks of the information presented to them. These chunks arise during information processing and competition and are represented with the structure assemblies that characterize a given blackboard.

In this way, the two approaches are strongly mutually complementary: IDyOT can provide the structural elements that would be needed in a neural blackboard representation, instead of deriving them from a laborious and perhaps faulty analyses. The way in which IDyOT derives these structural elements is much more direct and secure than the engineering approach in NBA, because the elements derived by IDyOT are based on learning mechanisms using real corpora. These learning methods could also be used to develop the structural elements of a phonological neural blackboard and for neural blackboards of other languages than English.

In turn, the NBA provides a direct neural implementation of the structures as learned by IDyOT. This offers the possibilities for fast hardware implementations combined with processing abilities based on dynamic competitions in the neural blackboards. The dynamics in neural blackboards also strengthen functional processing in the architecture. For example, they can play a role in sentence processing, in the generation of behavior (e.g., answering questions) or in ambiguity resolution. They also reduce the constraints that need to be learned to perform these tasks.

In the next sections we address a number of relations between the representations used in the NBA and IDyOT in more detail.

6. STRUCTURAL ELEMENTS IN NEURAL BLACKBOARDS OR WORKSPACES

The first relation between NBA and IDyOT concerns the role of neural blackboards or a workspace. In both architectures, special operators (or neural circuits) process and generate special forms of information. But to account for the productivity of human cognition there has to be a way in which the information processed or generated by special processors is interrelated

and combined. A neural blackboard or workspace allows these interactions to occur, with the special processors feeding into and competing within them. The role of neural blackboards or workspace in both architectures is also related to the small-world network structures that would allow different brain processors (areas) to interconnect with each other in a flexible way.

Blackboards play a role in classical computation (Corkill, 1991), in which they allow the representation of generic forms of information that can be stored and retrieved at will (in line with the characteristics of symbolic information processing). In contrast, the neural blackboards in the NBA are not generic in this sense. They do not represent arbitrary information which can be stored and retrieved at will. Instead, the information that can be stored in a given neural blackboard is determined by the nature of its composite structural elements, which depends on the kind of process the neural blackboard is involved in. For example, the structural elements of the neural sentence blackboard are different from those in the phonological neural blackboard: the sentence neural blackboard has main assemblies and sub assemblies for specific syntactic structural elements (e,g., "clause" or "preposition"), which are not found in the phonological neural blackboard. As a consequence, the neural sentence blackboard cannot (by itself) represent phonological structures. This is why the blackboards in the NBA are referred to as *neural* blackboards, to emphasize their internal and selective neural structure.

The workspace in IDyOT is symbolic. But the composite structural elements in the workspace, learned by IDyOT, are related to the composite structural elements in the neural blackboards of the NBA.

In the NBA, however, the composite structural elements (or 'structure assemblies') are engineered, derived from an analysis of the domain (e.g., language) for which the neural blackboards are used. In contrast, the structural elements in IDyOT that provide a representation of phonological and sentential structures are learned from a real corpus.

It would be a huge advantage for an architecture as the NBA if the structures in neural blackboards could be learned from real corpora instead of being designed. In return, the NBA could then offer a neural (parallel and dynamic) implementation of the structures as learned by IDyOT. The following subsections illustrate, for the first time, how learning proceeds in IDyOT and how structural elements as learned in IDyOT can be implemented in a neural and dynamic manner.

6.1. IDyOT Memory: Encoding Sequential Structure and Conceptual Meaning

6.1.1. Overview

Because IDyOT's learning process is incremental, as opposed to the one-shot learning of most statistical learning systems, there is diachronic development of meaning in its memory. As a result, it is difficult to see how the system works from a static, descriptive perspective. Therefore, we begin with a static description of the representation and how it is used, so that the reader has a clear idea of where the incremental process is heading. Related, but different descriptions are given by Wiggins and Forth (2015), with respect to the dynamics of lexical disambiguation, and by Forth et al. (2016) with respect to timing in music and language. First, then, the reader is asked to focus on the data structure presented, and to postpone the question of how it is constructed to §7.1. The "viewpoint" terminology used in the following description was originated by Conklin (1990) and Conklin and Witten (1995).

IDyOT's conceptual representation consists of two components, both of which are learned. The primary component is a sequence of events with separable features (*viewpoints*), annotated with chunk extents, which themselves form a sequence, and to which chunking is then applied recursively, up to a limit which is a parameter of the system (see §7.1.3); we say that a symbol at level i *subtends* a sequence at level $i-1$; **Figure 3** illustrates this. The shortest possible event is a multidimensional object that describes a simultaneous moment as sensed by IDyOT, at a sampling rate which is a parameter of the system, but of which 40 Hz is a preferred value, in terms of all the sensory modalities available to it. The examples given here are taken from auditory processing; however, there is no implication that this should be the only modality available.

6.1.2. Sequence

The sequence memory consists of symbols, beginning at the lowest representational level, and recorded sequentially in perceived time, as abstractly illustrated in **Figure 3**. Higher layers in the hierarchy constitute abstractions of the sequences that their symbols subtend, in lower layers. Thus, once the memory is constructed, there are in general three directions of possible prediction from any given context: up, with increased abstraction, down, with decreased abstraction, and forwards in perceived sequence. The theory does not currently consider the complicating possibility of reasoning backwards, nor of subsequent conscious reinterpretation; reinterpretation should be layered on top of this. The structure so produced, combined with the contextualized distributions afforded by the transition matrices, is similar in nature to a Dynamic Bayesian Network (Pearl, 1999).

For a concrete example, consider speech input. The lowest-level representation of this would be spectral and highly granular, and therefore prohibitively expensive in memory. Since the basic symbols would, in a full example, be sensory inputs, for a human-like IDyOT, retention of the very lowest levels of memory should be fleeting, modeling echoic memory, and therefore our example is more realistic, beginning, like Wiggins and Forth (2015), at the somewhat artificial level of phonemes, pitch and amplitude: these constitute our *basic viewpoints*.

Consider the extremely simple example sentence, "John loves Mary" in **Figure 4**, which illustrates the idea in multiple dimensions. For example only, we use emoji to denote the semantics of the sentence: these are presented at the same time as the example sentence is being spoken. This could be represented by viewpoint "emoji" in **Figure 4A**, which should be thought of as alongside the other viewpoints, together constituting level 0 in **Figure 4B**, simulating more complex world experience. We can consider not just single viewpoints, but also their cross products (known in Conklin's system as a *linked* viewpoint), whose alphabet consists of pairs constructed from the two source

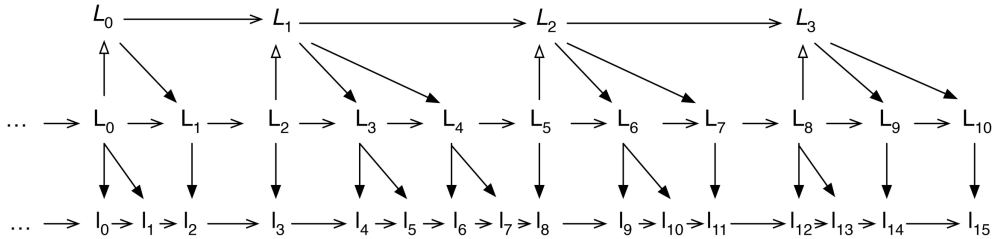


FIGURE 3 | Simplified illustration of the abstract structure of IDyOT memory. The three generators are working at the level denoted by labels with upper case L; these have been derived from the lower case l labels, below, and the generators are engaged in working on the next level up, denoted by labels with upper case italic L. Arrows with empty heads denote abstraction; arrows with solid heads denote concretisation; and arrows with open heads denote temporal sequence, though note that the diagram shows only sequence, and does not represent time. Recall that each generator's chunk subtends the sequence from its pointer to the end of the memory. Finally, note that each arrow denotes a range of possible next labels, with an associated distribution, and that generators can work at any and all of these levels. The diagram is simplified by showing only one of the alternative labels that exist at each level; thus, each of the abstraction and concretisation arrows should be thought of as a range of choices, governed by a distribution derived from observed likelihood.

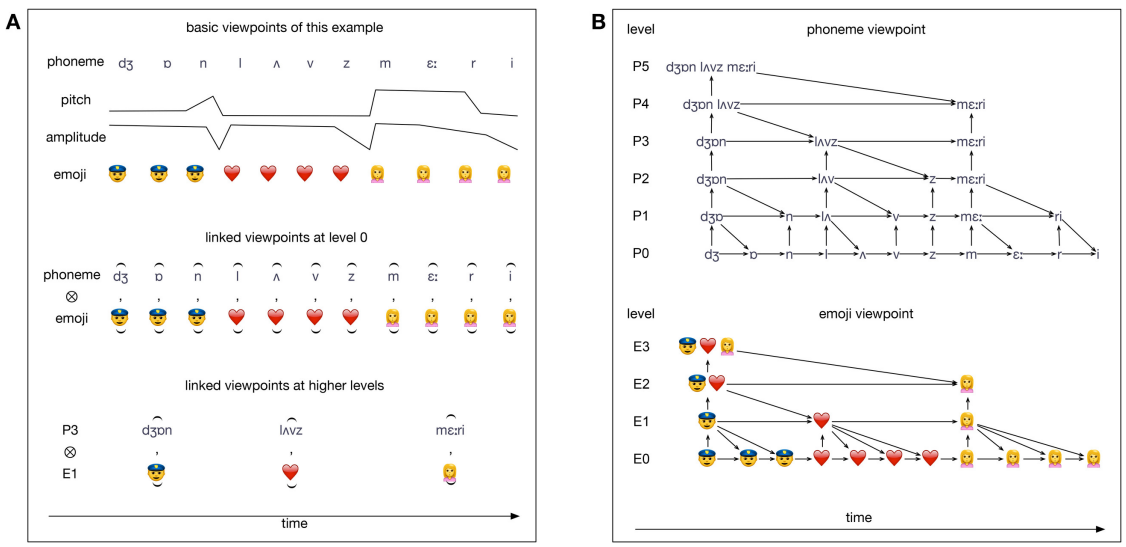


FIGURE 4 | Two perspectives on IDyOT memory. **(A)** An illustration of the parallel basic viewpoints for the sentence "John loves Mary," expressed in phonemes with associated voice pitch and amplitude signals, and semantics represented (for the purposes of example only) by emoji sequences. The top group are the *basic* viewpoints, as directly transduced (again, for the purposes of example); the middle example shows a *linked* viewpoint at the basic level; and the bottom example shows a linked viewpoint that encodes the discovered association between the semantic representations and the spoken words; such links can only take be generated when the two source viewpoints are aligned in time. **(B)** The hierarchical memory structures resulting from sensing of the sentence, "John loves Mary," in terms of the individual phoneme and emoji viewpoints by a fully trained IDyOT. Note that this does not correspond with the standard syntactic parse, and nor is it the same as a MERGE style parse of the words. Associated with each layer of the tree, L, is a continuous, time-variant conceptual space, S_L, (Gärdenfors, 2000) of timbre; this is a complex Hilbert space, whose points are time-slices in a spectral representation, such as a Fourier transform. Each stimulus at level 0 corresponds with a temporal trajectory (of variable length) in that space, while the corresponding structures at level 1 are points in a different, abstract space. S_{i+1} is related to S_i by spectral (e.g., Fourier) transformation, following Chella et al. (2007). Then, the sound /dZ/ is represented in full spectral detail at level 0, but in summary form, as a point, at level 1, as are /o/ and /n/. At level 1, further trajectories connect the more abstract representations, and thus the temporal detail of the individual sounds is abstracted, allowing (for example) the same word to be recognized regardless of how long the vowel takes. Expectations as to timing are generated from the various examples of each sound in each context in the memory (Forth et al., 2016).

alphabets. This, of course, generates a combinatorial explosion of viewpoints.

At each layer, there is a first-order Markov model, which allows prediction of the next item in sequence; Wiggins (2012b) explains the importance of this prediction with respect to creativity. Predictions, expressed as distributions over the alphabet of the relevant layer, may be generated for any point at the leading edge of the hierarchical memory structure as it is generated: thus, higher-level, abstract predictions are current at the same time as surface-level ones, and this is how long-term dependency in language, music, and narrative is managed.

6.1.3. Meaning

IDyOT is unusual as a symbolic learning system because it does not use symbols with predefined meanings. Rather, symbols are grounded in perception, and their meaning is determined either

in terms of synchronic relations between sensory modalities, or in terms of the diachronic sequence chunks that they subtend. In either case, meaning is placed in context of the conceptual spaces (Gärdenfors, 2000, 2014) associated with the viewpoints and the alphabets built above them. To summarize very briefly, conceptual spaces are low-dimensional geometrical spaces that afford judgments of similarity or betweenness. An example is the familiar color spindle, which has regions corresponding with colors of the spectrum, in which Euclidean distance models similarity (Gärdenfors, 2000, 2014). Different perceptual phenomena exhibit different geometries (for example, musical pitch is a spiral, Shepard, 1964), and methods for deriving these properties are a rich area of future research; Tenenbaum et al. (2011) propose various candidate statistical structures. In the higher layers of IDyOT memory, because a symbol subtends a sequence of symbols below it, it must be possible to map a trajectory of points or regions in a lower space to a single point in a higher one; this suggests that spectral representations are a promising route; Chella et al. (2008) and Chella (2015) suggest methods.

The conceptual spaces in IDyOT are important, because they afford the similarity measures that categorize chunks together in the incremental chunking and representation process, which we describe in §7.1.

6.2. NBA: binding sequential structures and concepts

The abstract structure of IDyOT memory, as illustrated in **Figure 3**, consist of learned components, organized in hierarchical layers. They form the link between the learning mechanisms of IDyOT and the neural blackboard structures of the NBA.

Figure 5 illustrates these neural blackboard structures in more detail, with the structure the sentence *cat sees cat*, to compare the encoding of sequential structures in IDyOT and the NBA.

The red and black thick lines in the figure illustrate the (crucial) conditional connections in the NBA, which consist of gating circuits. In the NBA, each concept assembly (e.g., of a noun) is connected to a set of structure assemblies of the same kind (all *Ni* assemblies in the case of a noun) with gating circuits. (In fact, the words need to be represented in a phonological blackboard first, to enhance the productivity of the architecture, being able to represent novel but phonologically regular words, and to reduce the number of conditional connections in the architecture.) In turn, each structure assembly consists of a "main assembly," such as *N*1, and (a set of) sub assemblies, such as *n* or *t*. The connection between a main assembly and a sub assembly consists of a gating circuit as well.

Structure assemblies of different kinds, such as *V*1 and *N*2, are connected by their sub assemblies of the same kind. Here, by their *t* (theme) sub assemblies, which represents the fact that a verb can have a theme (object). This connection (red line) also consists of a gating circuit, which can be activated by a WM neural population. This results in the binding of the two connected sub assemblies and hence their main assemblies, which last as long as this WM population is active. When two sub assemblies are bound in this way, activation can flow from one of the main assemblies to the other, by opening the gates between these main assemblies and their sub assemblies.

The gating circuits operate by disinhibition (*di*), as illustrated in **Figure 5**. When *N*1 is active, it activates a neuron (or neuron population) *X* and an inhibitory neuron (or population) *i*. The latter inhibits *X*, which blocks the flow of activation. But when *i* itself is inhibited (by neuron or population *di*), activation can flow from *N*1 (via *X*) to *n*.

Gating circuits can be disinhibited (or "activated") in one of two different ways. In the case of gating circuits between main assemblies and sub assemblies (the black connections in **Figure 5**), the activation results from an external control circuit that activates the *di* population. This is how syntactical operations affect binding in the blackboard. A control circuit could have recognized that *sees cat* represent a verb and a theme. It then activates all *di* populations in the gating circuits between all *Vi* and *Nj* assemblies and their *t* assemblies. As a result, the active *Vi* and *Nj* will activate their *t* sub assembly.

Gating circuits between sub assemblies and between word and main assemblies (the red connections in **Figure 5**) are activated by (specific) "working memory" (WM) populations. A WM population remains active for a while, after initial activation, by reverberating activation in the population (e.g., Amit, 1989). An active WM population binds the assemblies to which it is connected. **Figure 6** illustrates how this is achieved in the NBA. **Figures 6A–C** illustrate the same binding process with increasing detail. In **Figure 6A**, the binding between the *t* sub assemblies of *V*1 (or $V1-t$) and *N*2 ($N2-t$) in **Figure 5** is repeated. **Figure 6B** illustrates that this binding is based on a "connection matrix," which consists of columns and rows of "connection nodes," which are illustrated in **Figure 6C**.

Each specific $Vi - t$ and $Nj - t$ pair of sub assemblies is interconnected in a specific connection node, located in a connection matrix dedicated to binding $Vi - t$ and $Nj - t$ sub assemblies. In general, when two assemblies *Xi* and *Yj* (e.g., $Vi-t$ and $Nj-t$) are concurrently active in the processing of a sentence, they activate a WM population in their connection node by means of a gating circuit, as illustrated in **Figure 6C**. In turn, the active WM population disinhibits a gating circuit by which activation can flow from *Xi* to *Yj*, and another such circuit, not show in (c), by which activation can flow from *Yj* to *Xi*. As long as their WM population is active, *Xi* and *Yj* are "bound" because activation will flow from one to the other whenever one of them is (initially) activated.

The NBA allows any noun to bind to any verb in any thematic role using dedicated connection matrices. Also, the NBA has structure assemblies that can bind to other structure assemblies, such as *S*1 in **Figure 5** or clause structure assemblies. In this way, hierarchical sentence structures can be represented, such as relative or complement clauses.

6.2.1. Sentence Structure as Connection Path

To form a sentence structure, the structure assemblies have to bind to each other. This process is regulated by control circuits that build a sentence structure in line with the (syntactical) relations in the sentence (van der Velde and de Kamps, 2010).

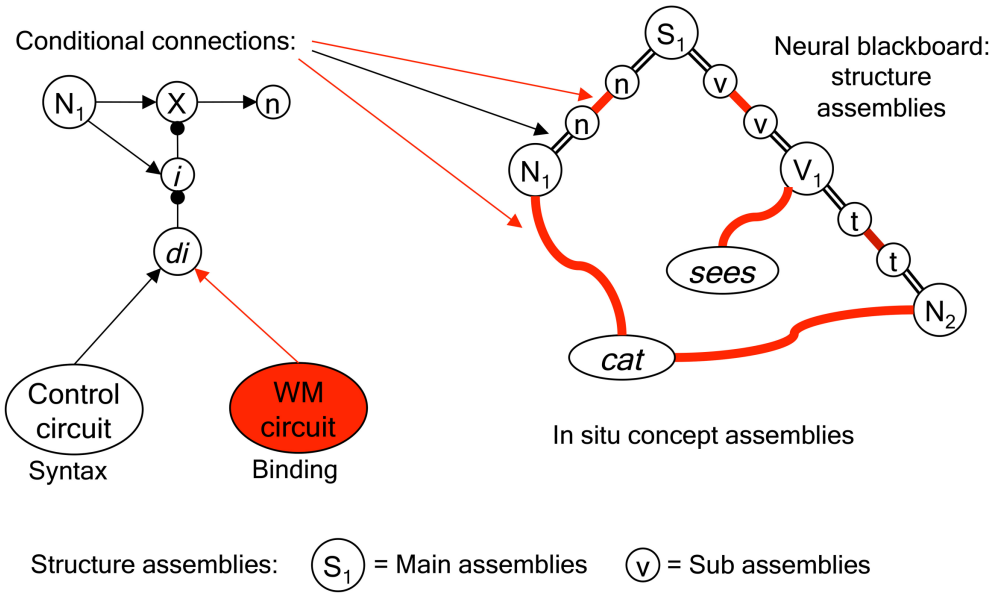


FIGURE 5 | Left: conditional connections. N, n, noun; *i*, inhibition; *di*, dis-inhibition; WM, working memory. **Right**: Representation of *cat sees cat* in the sentence neural blackboard. S, sentence; V, v, verb; t, theme (object).

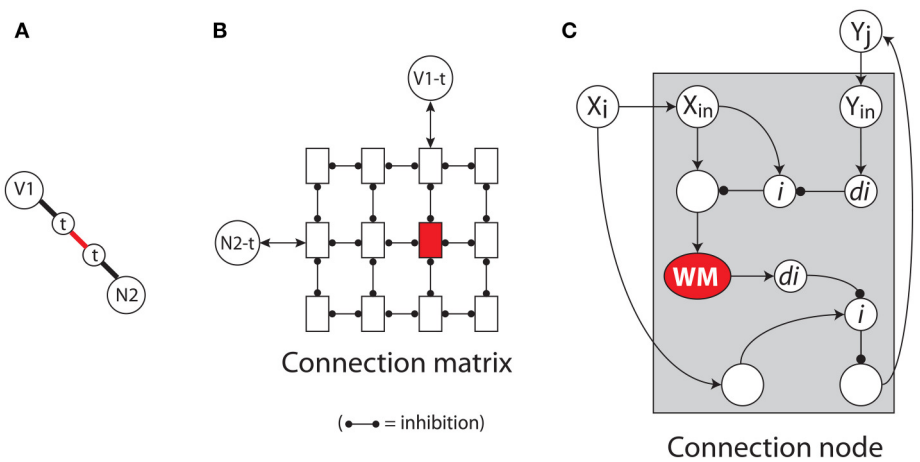


FIGURE 6 | (A) Conditional connections. N, noun; V, verb, t, theme (object) . **(B)** Connection matrix. **(C)** Connection node. *i*, inhibition; *di*, dis-inhibition; WM, working memory.

So, with *cat sees cat* in **Figure 5**, the control circuits will recognize *cat* as the subject of the sentence, expressed by the binding of *N*1 to the 'Sentence' structure assembly *S*1, and *sees* as the verb of the main clause, expressed by binding *V*1 with *S*1.

But then, the control circuits will recognize the second occurrence of *cat* as the object of the sentence. This seems to pose a problem, because that would seem to require a copy (different token) of *cat* to bind as the object to the verb. Indeed, symbol manipulation represents the sentence *cat sees cat* with two tokens of *cat*. But in the NBA, a given concept assembly can bind to different structure assemblies at the same time, allowing the creation of sentence structures in which words are repeated, as illustrated in **Figure 5**. However, the concept assemblies remain *in situ* in this way, so words in sentence structures are always content addressable and grounded. This example illustrates how the NBA solves the "problem of two" posed by Jackendoff (2002).

The sentence structures in the NBA (as illustrated in **Figures 1**, **5**) and IDyOT (e.g., *John loves Mary* in **Figure 4**) are structurally similar. The sentence in IDyOT is derived from its learning principles, as outlined above, and it can be represented in the NBA in the manner illustrated in **Figure 5**.

As we argued, the representational similarities between IDyOT and NBA would offer a basis for combining the learning mechanisms of IDyOT, based on real corpora, with the parallel

and dynamic implementation of the NBA. The dynamics in the neural blackboard can in fact be used to solve forms of (e.g., sentence) ambiguity (van der Velde and de Kamps, 2015b), which in turn offers the possibility of further reduction of the constraints that would have to be learned to represent and process complex cognitive structures.

7. PROCESSING OF SEQUENTIAL STRUCTURES

A second link between the NBA and IDyOT concerns the processing of sequential information. Based on its learning mechanism, IDyOT derives probabilistic choices between structural interpretations of the processed information, in the form of transition matrices. Based on learning, predictions can be made that influence further processing of the input sequence.

The NBA uses similar kinds of information to train control circuits that selectively activate the neural blackboards, as illustrated in **Figure 5**. Control circuits have been implemented with feedforward networks (van der Velde and de Kamps, 2010) and, more recently, with reservoirs (Jaeger and Haas, 2004) consisting of "sequence nodes" (van der Velde, 2016a).

Similar to the connection nodes in **Figure 6**, each sequence node has a column structure with gating circuits that control the activation of the node. This activation depends on three sources: previously activated sequence nodes (hence forming a chain of nodes in the reservoir, representing sequential order), external activation generated by the (ongoing) input sequence, and activation already generated in the neural blackboard. The latter includes the predictions generated in the neural blackboard in the course of processing an input sequence, as in the resolution of ambiguity (van der Velde and de Kamps, 2015b).

The reservoir can, for example, learn to answer the question *Where is cat?* with the sentence *Cat is on mat* in **Figure 1**. The reservoir can learn to do this by recognizing the sequence *Where - localizer - noun - Agent*. Here, the sequence *Where - localizer - noun* is based on transforming the question *Where is cat?* in a more general form (with *is* = *localizer* and *cat* = *noun*). The *Agent* in the sequence is derived from the activation of the neural blackboard representation of *cat is on mat*, because *cat* in the question *Where is cat?* activates its *in situ* neural assembly (**Figure 1**) and thus the part of the neural blackboard representation of *cat is on mat* to which the assembly *cat* is bound. In this way, the reservoir can learn to reactivate the sentence representation of *cat is on mat* in the neural blackboard, to generate the answer *mat* (van der Velde, 2016b).

But, for example, the transformation of the question *Where is cat?* into the more general form *Where - localizer - noun?*, learned by the reservoir in the NBA, is based on an analysis. In contrast, the learning mechanism of IDyOT can provide the information to train the reservoir in the NBA, based on real corpora. Conversely, the distinction between structured neural blackboards and the control reservoirs in the NBA can strongly reduce the number of contingencies that have to be learned over time, as illustrated with the ease with which the reservoir can learn to recognize *Where - localizer - noun - Agent* (van der Velde, 2016b).

The more elaborate learning mechanism of IDyOT would thus have to be integrated with the NBA, and eventually be implemented with neural reservoirs that interact with the neural blackboard in the NBA. The learning process in IDyOT is outlined in more detail below, again for the first time.

7.1. The IDyOT Incremental Learning Process

7.1.1. Initial State

Initially, IDyOT has no memory, no symbols, and only inputs. Input is in terms of percepts conceptualized as symbols representing continuous real-world phenomena at whatever level of abstract is chosen: here, phonemes, pitch, amplitude, and observed meaning (emoji).

7.1.2. Chunks and Labels

Given a low-level, predefined conceptual space, S_v (which initially has no geometry, but learns it as more data is received), for each low-level viewpoint, v, IDyOT labels the mutually discriminable points in S_v with symbols, building an alphabet, $\aleph(v)$, and, separately, builds a chain of these symbols as the input sequence proceeds; this may be thought of as the chain l_i in **Figure 3**. Simultaneously, IDyOT builds a first order transition matrix of the chain; this will allow the construction of successive distributions over $\aleph(v)$, $D_{t,\aleph(v)}$, as time, t, proceeds. Each symbol is considered in relation to the symbols already created, in terms of their corresponding points: a quasi-Euclidean distance (*norm*), in S_v, may be computed between them. At the same time, the space is progressively partitioned into regions whose points are nearest to each point in the sequence, as in a Voronoi Tessellation (Aurenhammer, 1991). This tessellation, possibly modified by a parameter which creates a gap between the regions (**Figure 7**), forms the basis of similarity comparison. Points in (non-zero) gap regions form new seeds. This process will, of course, produce initially inaccurate predictions and labelings, but as sufficient data is processed, these early errors fade into statistical obscurity, propelled by the memory consolidation process described below.

However, this simple mechanism would not account for the human propensity to perceive what is expected, because S_v, the conceptual space associated with v, is static. The distribution, $D_{0,\aleph(v)}$, describes IDyOT's expectation at this point; it is derived from the transition matrix for v. Each region, $R_{v,s}$, where $s \in \aleph(v)$ in the Voronoi tessellation of S_v is now expanded or contracted, by changing the position of each plane dividing the space, in proportion to the relative likelihood of the symbols corresponding with the points to whose connecting line the plane is perpendicular. A parameter, whose value is the subject of study, determines the degree of variation; an interesting possibility is that this value is related to entropy of the distribution, as was found empirically to be case in a related application of distributions in IDyOM (Pearce et al., 2005), where distributions containing more information influence the outcome more. Thus, the less expected a phoneme, s, is, the smaller its $R_{v,s}$ temporarily becomes, and so a phoneme that is both imprecisely articulated and unexpected may be misidentified as one near it, which

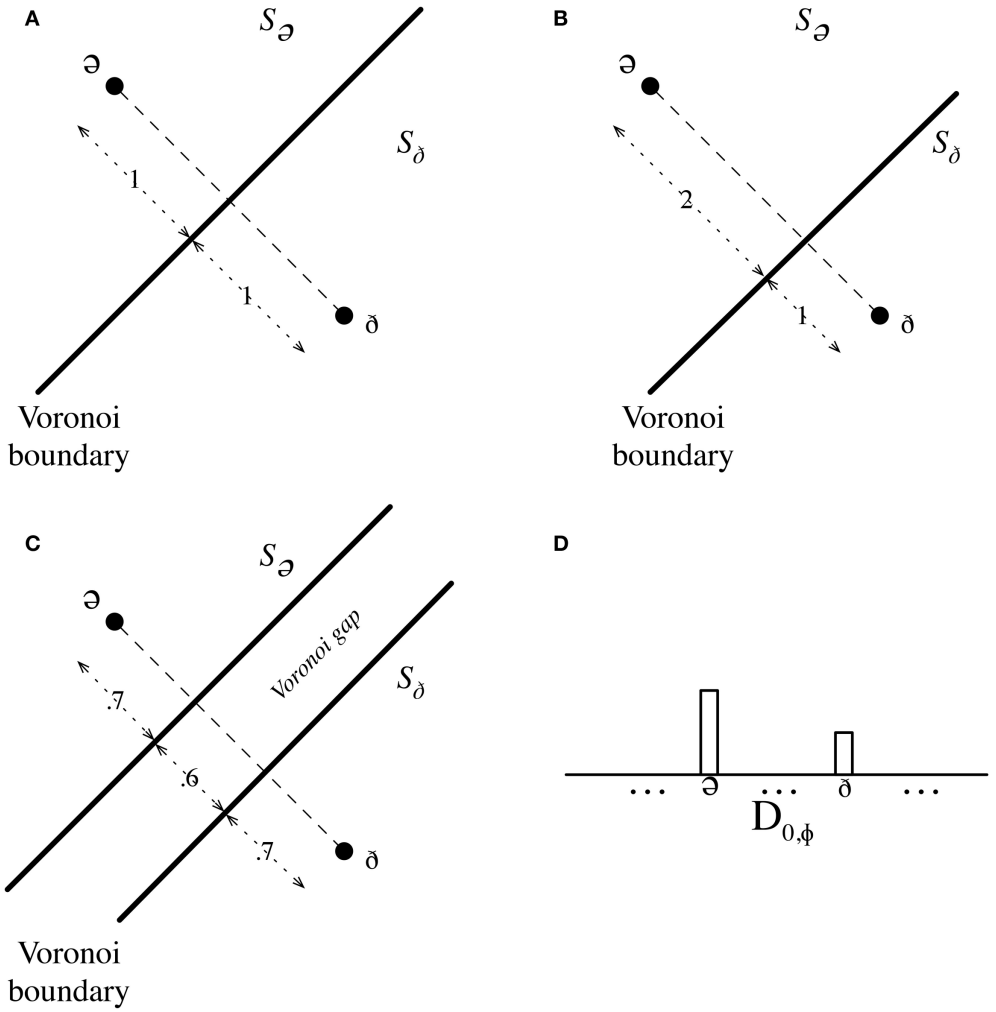


FIGURE 7 | Modulating conceptual regions according to expectedness. **(A)** The unmodified Voronoi tessellation of the conceptual space of phonemes, S_Φ, showing the boundary between $R_{S_\Phi,ə}$ and $R_{S_\Phi,ð}$. **(B)** The modified tessellation; note that the distances from the labeled points to the boundary have changed in proportion to the relative likelihoods in $D_{0,\Phi}$. **(C)** The tessellation with a non-zero gap. **(D)** Schematic partial representation of the distribution, $D_{0,\Phi}$ showing (imaginary) proportions for /ə/ and /ð/.

is more likely in the distribution (**Figure 7**). IDyOT behaves like a human in this context: it commits to memory incorrect perceptions, as if they were correct.

7.1.3. Chunking: Competition and Boundary Entropy

Each new symbol, indexing a point in S_v, is available to all generators associated with this viewpoint (see **Figure 3**). As the transition matrix is populated, predictions can be made of likelihood, and as IDyOT's memory develops, progressively more informed predictions may be made using the probabilistic network afforded by the layered memory. Thus, the entire context will influence $D_{t,\aleph(v)}$ it any time point t. Again, initially, these predictions will not be particularly accurate; as more data is received they will improve. As each new label appears, therefore, a new distribution is generated, and its entropy, $H(D_{t+1,\aleph(v)})$ can be calculated and compared with $H(D_{t,\aleph(v)})$. On the basis of empirical evidence from computational linguistics and music cognition (e.g., Sproat et al., 1994; Brent, 1999; Pearce et al., 2010; Rohrmeier et al., 2015), at each time step, IDyOT's agents compete for global workspace access, the largest positive change being the winner. If no agent registers an increase in entropy, there is no winner, and no change in the memory; IDyOT proceeds to the next input stimulus.

Thus, IDyOT achieves hierarchical perceptual chunking.

7.1.4. Layer Formation and Abstraction

Following the identification of a chunk in memory, IDyOT must decide whether to generate a new label or to label this chunk with an existing symbol, on grounds of similarity. In the former case, a new label is generated, at level L_i in **Figure 3**, and it is added to memory, along with pointers to the lower level chunk that it subtends; also, the transition matrix for the upper layer is updated. A further transition matrix, of which one exists for each pair of contiguous levels, is also updated with the new

symbol and transition. In this connection, a higher-level symbol is deemed to connect down to any symbol in its chunk, while any lower level symbol is deemed to connect to any symbol in whose chunk it appears. It is implicit in this process that each symbol in an IDyOT memory chain may be subtended by more than one symbol at the immediately higher level, and it may subtend more than one symbol below. Transition matrices for these upward connections, too, must be maintained.

Returning to the example: the higher level sequence has a transition matrix, and so its entropy can be determined, symbol-wise, as above, and therefore the same boundary test as above can be applied. If a new chunk at this level is detected, then the same process applies, and so on up the layers of the network, using the same principle of similarity measurement as above. This first generates level L_i in **Figure 3**, and then on beyond the scope of that simple example.

This recursive process constructs a tree from the very lowest level of representation up to the highest possible abstraction, as shown for our concrete example, in **Figure 4**. Although this simple example has focused on only one aspect of the stimulus, it is important to recall that, in a fully implemented IDyOT, all modalities of perception would be active simultaneously, and synchronized (Forth et al., 2016) in such a way as to interrelate simultaneous stimuli. Thus, the association between, for example, the word "orange", the sound [ɒrɪnʒ], and appropriate representations of the corresponding color, fruit, pop star and politics, could be learned, as illustrated in **Figure 4**.

8. FURTHER RELATIONS BETWEEN NBA AND IDYOT

8.1. Conceptual spaces

The semantics underlying the IDyOT and NBA representations are derived from the conceptual spaces with which they interact. In turn, the conceptual spaces play a role in processing in both architectures. For IDyOT, the role of conceptual spaces is illustrated in **Figure 4**. In the NBA, representations of conceptual structures (relations, propositions, sentences) are based on content addressable concept representations, which directly and selectively activate conceptual structures in neural blackboards. Also, conceptual domains and relations are needed to influence sequential processing in the control reservoirs of the NBA (van der Velde, 2016a).

McGregor et al. (2015) outline a basis for a geometrical conceptual space, with interpretable spaces and dimensions derived from observed co-occurrence statistics in a large corpus. Conceptual relations and domains can be obtained by the techniques described by McGregor et al. (2015) and by the metric based on a semantic map as derived by van der Velde et al. (2015). This semantic map also consists of a co-occurrence matrix, derived from human categorizations. The metric provided a similar concept-cluster structure as derived from reduction techniques. But it also revealed the possibility of deriving bridges between conceptual domains based on metric violations.

The geometrical nature of such a conceptual space provides a natural representation for the content addressable concept representations underling the combined IDyOT-NBA architecture. Furthermore, the geometrical nature of this conceptual space and the neural blackboard mechanisms of the IDyOT-NBA architecture provide the possibilities of new forms of hardware implementations that can circumvent the limitations of the Von Neumann Architecture, on which symbolic computation is standardly based.

8.2. Brain and Computation

As referred to in our introduction, the processing of conceptual structures can be studied with the aim to understand human cognition and its relation to the brain. Or they can be targets for the development of artificial cognitive systems. We argue that a combined IDyOT-NBA architecture can address both aims.

Because learning in IDyOT is based on information found in real corpora, it derives structures and processes based on human information processing and generation. In this way, the NBA structures and processes derived from IDyOT will be based on human information processing as well. The neural implementation of the NBA then allows a comparison between the structures and processes of the combined architecture with those observed in brain research.

An example of how the combined architecture can be related to neuro-cognitive processing is presented in **Figure 8**. The figure illustrates a novel simulation of NBA activity, with the processing of the sentence *Bill-Gates has met two very tired dancers in Dallas*, with *Bill-Gates* as one noun (BG). Activation of "main assemblies" (MA), "sub assemblies" (SA) and binding in working memory (WM) are shown, because they determine the representation structure of the sentence in the sentence neural blackboard of the NBA (van der Velde and de Kamps, 2006). Also shown is the overall activation of all assemblies and circuits, consisting of more than 300 neural populations in all (marked "Total"; red line). The neural populations are simulated with Wilson and Cowan population dynamics (Wilson and Cowan, 1972).

Using intracranial measurements, Nelson et al. (2014) observed that binding of words and phrases produces an increase and then decrease of activity (e.g., because binding related activation will reduce after binding). The NBA activation simulates this effect, and also indicates why it occurs, i.e., which structures and processes are related to this effect. In particular, total neural activity first increases when a new word is presented (as illustrated by the increase of total activity at the location of the black vertical bars, that indicate the presentation times of the words). But then, total activity drops, due to the binding of the presented word to previously presented words and phrases in the developing sentence structure in the sentence neural blackboard of the NBA. Occasionally, activity does not decline, as with *Bill has* or *very tired*, which results from the fact that *Bill* is the first word, which cannot bind to other words yet, and *very* does not bind to the previous word *two*.

Hence, the simulation illustrates the close relation between neural dynamics and the representation structures underlying processed sentences in the NBA. The aim of the integration of NBA with IDyOT is to develop these representation structures by learning from real corpora. In this way, machine learning

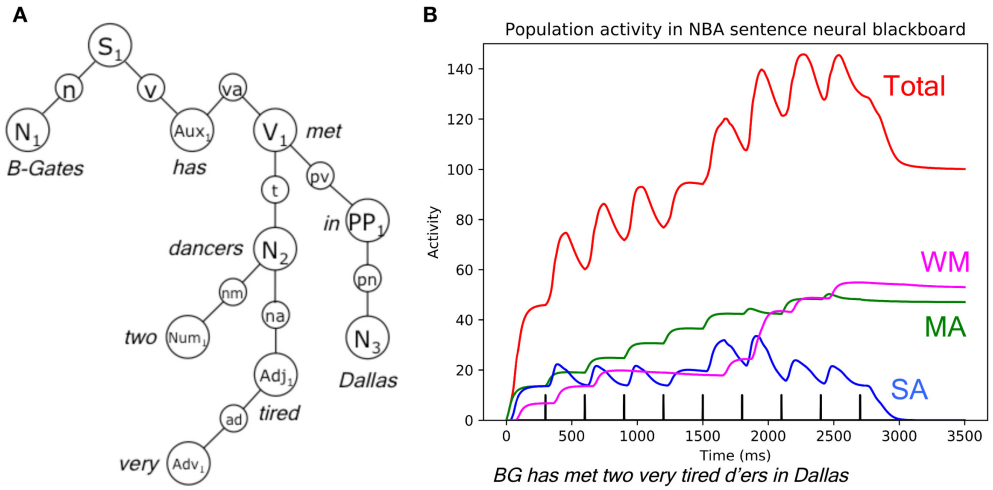


FIGURE 8 | (A): NBA structure of *Bill-Gates has met two very tired dancers in Dallas*, with *B(ill)-Gates* as one noun. Aux, va, auxiliary verb; Adj, na, adjective; Adv, ad, adverb; N, n, noun; Num, nm, numerator; PP, pv, pn, preposition; S, sentence; V, v, verb. **(B)**: Neural activity in the NBA when *Bill-Gates has met two very tired dancers in Dallas* is processed. BG, Bill Gates; d'ers, dancers; MA, main assemblies; SA, sub assemblies; WM, working memory. 'Total' (red activity) is the sum of the activation of all neural populations in the NBA structure of this sentence (over 300 populations), simulated with Wilson-Cowan population dynamics. The words of the sentence are presented at the times indicated with the vertical bars. The last bar signals the end of input activation.

could be related to brain activity observed in human cognitive processing.

Furthermore, the NBA predicts the existence of "connection" fields (or matrices) with special roles, such as "agent" and "theme" (object) in which bindings between (e.g.,) arbitrary verbs and nouns as (agent or theme) arguments can occur. Recent fMRI observations indicated the existence of (agent and theme) areas in the cortex that are selectively activated when nouns function, respectively, as agents or themes of verbs (Frankland and Greene, 2015). The activation patterns in these areas also concur with the activation patterns produced in the NBA. These areas could form a neural substrate for (parts of) a Global Workspace, in which competitions between neural structural representations could occur.

The combined IDyOT-NBA architecture also targets the development of artificial cognitive systems. Recently, Lake et al. (2016) argued that, despite recent successes, Deep Learning does not capture essential characteristics of human learning and processing. One of the difficulties for Deep Learning concerns compositional (combinatorial) processing, in which structured information is processed in terms of already familiar constituents and partial structures.

A crucial feature of compositional processing is the interaction between specialized processors and domains in which these processors, and the information they process, can interact, compete, and be combined. This is what the neural blackboards and the workspace in NBA and IDyOT are about. Because the combined architecture can develop and activate these structures based on learning from real corpora, it can address key features of human cognitive processing.

The combined architecture can also address new demands on computing power because the NBA can be implemented fully as a system operating in parallel, based on dynamic interactions. Of course, processing will be sequential when input is presented in a sequential manner. Also, the dynamic interactions will proceed in time as well. But each of the components (e.g., connection nodes in the connection matrices) will operate in parallel with all other components, and their interactions are based on direct dynamical activation and competition. When implemented in hardware, this allows the system to operate at minimal levels of power, with fast processing speeds.

9. CONCLUSION

We have presented two knowledge representations, used in two cognitive architectures, the NBA and IDyOT, that both aim to account for conceptual representation and processing in productive forms of cognition. Although the architectures differ in that the NBA is neural and IDyOT is symbolic, they are also similar in many ways. Both assume that conceptual representations consist of structures in which all aspects related to a concept are interconnected. Both assume that processing with representations occur in blackboards or a workspace, in which these representations can interact and can be (re)combined. And both rely on the principles of chunking to generate higher-level structural representations based on the more elementary ones.

Finally, the relations between both architectures combined with their different bases provide unique opportunities for a complementary integration. The NBA could provide a neural implementation of the processing and representation of higher level conceptual representations and IDyOT could provide the learning mechanisms by which the more elementary

representations needed for this implementation could be derived from human cognitive (corpus) material.

AUTHOR CONTRIBUTIONS

FvV and DN wrote the sections on NBA. GW wrote the sections on IDyOT, based on discussions with JF. The rest was a joint effort.

ACKNOWLEDGMENTS

This work funded by the ConCreTe (Concept Creation Technology) project, which acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET grant number 611733.

REFERENCES

Amit, D. (1989). *Modeling Brain Function*. Cambridge, MA: Cambridge University Press.

Aurenhammer, F. (1991). Voronoi diagrams – a survey of a fundamental geometric data structure. *ACM Comput. Surv.* 23, 345–405. doi: 10.1145/116873.116880

Baars, B. J. (1988). *A Cognitive Theory of Consciousness*. New York, NY: Cambridge University Press.

Brent, M. R. (1999). An efficient, probabilistically sound algorithm for segmentation and word discovery. *Mach. Learn.* 34, 71–105. doi: 10.1023/A:1007541817488

Chella, A. (2015). "A cognitive architecture for music perception exploiting conceptual spaces," in *Applications of Conceptual Spaces: The Case for Geometric Knowledge Representation*, number 359 in Synthese Library. Cham Heidelberg; New York, NY; Dordrecht; London: Springer.

Chella, A., Dindo, H., and Infantino, I. (2007). Imitation learning and anchoring through conceptual spaces. *Appl. Artif. Intell.* 21, 343–359. doi: 10.1080/08839510701252619

Chella, A., Frixione, M., and Gaglio, S. (2008). A cognitive architecture for robot self-consciousness. *Artif. Intell. Med.* 44, 147–154. doi: 10.1016/j.artmed.2008.07.003

Conklin, D. (1990). *Prediction and Entropy of Music*. Master's Thesis, Department of Computer Science, University of Calgary.

Conklin, D., and Witten, I. (1995). Multiple viewpoint systems for music prediction. *J. New Music Res.* 24, 51–73. doi: 10.1080/09298219508570672

Corkill, D. D. (1991). Blackboard systems. *AI Expert* 6, 40–47.

Eshghi, A., Purver, M., and Hough, J. (2013). "Probabilistic induction for an incremental semantic grammar," in *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, (Potsdam: Association for Computational Linguistics), 107–118.

Forth, J., Agres, K., Purver, M., and Wiggins, G. A. (2016). Entraining IDyOT: time in the information dynamics of thinking. *Front. Psychol.* 7:1575. doi: 10.3389/fpsyg.2016.01575

Frankland, S. M., and Greene, J. D. (2015). An architecture for encoding sentence meaning in left mid-superior temporal cortex. *Proc. Natl. Acad. Sci. U.S.A.* 112, 11732–11737. doi: 10.1073/pnas.1421236112

Gärdenfors, P. (2000). *Conceptual Spaces: The Geometry of Thought*. Cambridge, MA: MIT Press.

Gärdenfors, P. (2014). *Geometry of Meaning*. Cambridge, MA: MIT Press.

Hebb, D. O. (1949). *The Organisation of Behaviour*. New York, NY: Wiley.

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., and Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* 532, 453–458. doi: 10.1038/nature17637

Jackendoff, R. (2002). *Foundations of Language*. Oxford: Oxford University Press.

Jaeger, H., and Haas, H. (2004). Harnessing nonlinearity: predicting chaotic systems and saving energy in wireless communication. *Science* 304, 78–80. doi: 10.1126/science.1091277

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. (2016). Building machines that learn and think like people. *Behav. Brain Sci.* doi: 10.1017/S0140525X16001837. [Epub ahead of print].

McGregor, S., Agres, K., Purver, M., and Wiggins, G. A. (2015). From distributional semantics to conceptual spaces: a novel computational method for concept creation. *J. Artif. Gen. Intell.* 6, 55–86. doi: 10.1515/jagi-2015-0004

Nelson, M. J., El Karoui, I., Rangarajan, V., Pallier, C., Parvizi, J., Cohen, L., et al. (2014). "Constituent structure representations revealed with intracranial data," in *Society for Neuroscience Annual Meeting* (Washington, DC: Poster).

Pearce, M. T. (2005). *The Construction and Evaluation of Statistical Models of Melodic Structure in Music Perception and Composition*. Ph.D. Thesis, Department of Computing, City University, (London).

Pearce, M. T., Conklin, D., and Wiggins, G. A. (2005). "Methods for combining statistical models of music," in *Computer Music Modelling and Retrieval*, ed U. K. Wiil (Heidelberg: Springer Verlag), 295–312.

Pearce, M. T., Müllensiefen, D., and Wiggins, G. A. (2010). The role of expectation and probabilistic learning in auditory boundary perception: a model comparison. *Perception* 39, 1367–1391. doi: 10.1068/p6507

Pearce, M. T., and Wiggins, G. A. (2012). Auditory expectation: the information dynamics of music perception and cognition. *Top. Cogn. Sci.* 4, 625—652. doi: 10.1111/j.1756-8765.2012.01214.x

Pearl, J. (1999). "Bayesian networks," in *The MIT Encyclopedia of the Cognitive Sciences*, eds R. A. Wilson and F. C. Keil (Cambridge, MA: MIT Press), 72–74.

Quian Quiroga, R. (2012). Concept cells: the building blocks of declarative memory functions. *Nat. Rev. Neurosci.* 13, 587–597. doi: 10.1038/nrn3251

Rohrmeier, M., Zuidema, W., Wiggins, G. A., and Scharff, C. (2015). Principles of structure building in music, language and animal song. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 370:20140097. doi: 10.1098/rstb.2014.0097

Shanahan, M. (2010). *Embodiment and the Inner Life*. Oxford: Oxford University Press.

Shepard, R. N. (1964). Circulatory in judgments of relative pitch. *J. Acoust. Soc. Am.* 36, 2346–2353. doi: 10.1121/1.1919362

Sloman, A., and Chappell, J. (2005). "The altricial-precocial spectrum for robots," in *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, Edinburgh.

Sproat, R., Shih, C., Gale, W., and Chang, N. (1994). "A stochastic finite-state word-segmentation algorithm for chinese," in *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, Las Cruces, NM. 66–73.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D. (2011). How to grow a mind: statistics, structure, and abstraction. *Science* 331, 1279–1285. doi: 10.1126/science.1192788

van der Velde, F. (2015). Communication, concepts and grounding. *Neural Netw.* 62, 112–117. doi: 10.1016/j.neunet.2014.07.003

van der Velde, F. (2016a). Concepts and relations in neurally inspired *in situ* concept-based computing. *Front. Neurorobot.* 10:4. doi: 10.3389/fnbot.2016.00004

van der Velde, F. (2016b). "Learning sequential control in a neural blackboard architecture for *in situ* concept reasoning," in *Proceedings of NeSy 2016: Neural-Symbolic Learning and Reasoning*, eds T. R. Besold, W. Tabor, L. Serafini and L. Lamb (New York, NY), 1–11.

van der Velde, F., and de Kamps, M. (2006). Neural blackboard architectures of combinatorial structures in cognition. *Behav. Brain Sci.* 29, 37–70. doi: 10.1017/S0140525X06009022

van der Velde, F., and de Kamps, M. (2010). Learning of control in a neural architecture of grounded language processing. *Cogn. Syst. Res.* 11, 93–107. doi: 10.1016/j.cogsys.2008.08.007

van der Velde, F., and de Kamps, M. (2015a). The necessity of connection structures in neural models of variable binding. *Cogn. Neurodyn.* 9, 359–370. doi: 10.1007/s11571-015-9331-7

van der Velde, F., and de Kamps, M. (2015b). "Combinatorial structures and processing in neural blackboard architectures," in *Cognitive Computation:*

Integrating Neural and Symbolic Approaches, eds T. R. Besold, A. d'Avila Garcez, G. F. Marcus and R. Miikulainen (Montreal, OC).

van der Velde, F., Wolf, R. A., Schmettow, M., and Nazareth, D. S. (2015). "A semantic map for evaluating creativity," in*Proceedings of the Sixth International Conference on Computational Creativity (ICCC 2015)*, eds H. Toivonen, S. Colton, M. Cook, and D. Ventura (Park City, UT: Brigham Young University), 94–101.

Wiggins, G., and Bhattacharya, J. (2014). Mind the gap: an attempt to bridge computational and neuroscientific approaches to study creativity. *Front. Hum. Neurosci.* 8:540. doi: 10.3389/fnhum.2014.00540

Wiggins, G. A. (2012a). "I let the music speak: cross-domain application of a cognitive model of musical learning," in *Statistical Learning and Language Acquisition*, eds P. Rebuschat and J. Williams (Amsterdam, NL: Mouton De Gruyter). 463–495.

Wiggins, G. A. (2012b). The mind's chorus: creativity before consciousness. *Cogn. Comput.* 4, 306–319. doi: 10.1007/s12559-012-9151-6

Wiggins, G. A., and Forth, J. C. (2015). "IDyOT: a computational theory of creativity as everyday reasoning from learned information," in *Computational Creativity Research: Towards Creative Machines*, eds T. R. Besold, M. Schorlemmer, and A. Smaill (Atlantis/Springer: Atlantis Thinking Machines). 127–150.

Wilson, H. R., and Cowan, J. D. (1972). Excitatory and inhibitory interactions in localized populations of model neurons. *Biophys. J.* 12, 1–24. doi: 10.1016/S0006-3495(72)86068-5

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

frontiers in Psychology

PERSPECTIVE
published: 23 May 2017
doi: 10.3389/fpsyg.2017.00678

The Representation of Motor (Inter)action, States of Action, and Learning: Three Perspectives on Motor Learning by Way of Imagery and Execution

Cornelia Frank[1,2]* *and Thomas Schack*[1,2,3]

[1] Neurocognition and Action – Research Group, Faculty of Psychology and Sports Science, Bielefeld University, Bielefeld, Germany, [2] Cognitive Interaction Technology – Cluster of Excellence, Bielefeld University, Bielefeld, Germany, [3] Research Institute for Cognition and Robotics (CoR-Lab), Bielefeld University, Bielefeld, Germany

Learning in intelligent systems is a result of direct and indirect interaction with the environment. While humans can learn by way of different states of (inter)action such as the execution or the imagery of an action, their unique potential to induce brain- and mind-related changes in the motor action system is still being debated. The systematic repetition of different states of action (e.g., physical and/or mental practice) and their contribution to the learning of complex motor actions has traditionally been approached by way of performance improvements. More recently, approaches highlighting the role of action representation in the learning of complex motor actions have evolved and may provide additional insight into the learning process. In the present perspective paper, we build on brain-related findings and sketch recent research on learning by way of imagery and execution from a hierarchical, perceptual-cognitive approach to motor control and learning. These findings provide insights into the learning of intelligent systems from a perceptual-cognitive, representation-based perspective and as such add to our current understanding of action representation in memory and its changes with practice. Future research should build bridges between approaches in order to more thoroughly understand functional changes throughout the learning process and to facilitate motor learning, which may have particular importance for cognitive systems research in robotics, rehabilitation, and sports.

Keywords: motor imagery, motor memory, simulation, s-states, intelligent systems, functional equivalence

OPEN ACCESS

Edited by:
Tarek Richard Besold,
University of Bremen, Germany

Reviewed by:
Michael Ziessler,
Liverpool Hope University, United Kingdom
Franz Mechsner,
Northumbria University, Germany

***Correspondence:**
Cornelia Frank
cornelia.frank@uni-bielefeld.de

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *11 August 2016*
Accepted: *13 April 2017*
Published: *23 May 2017*

Citation:
Frank C and Schack T (2017) The Representation of Motor (Inter)action, States of Action, and Learning: Three Perspectives on Motor Learning by Way of Imagery and Execution. Front. Psychol. 8:678. doi: 10.3389/fpsyg.2017.00678

INTRODUCTION

Learning in intelligent systems is a result of direct and indirect interaction with the environment. To understand how intelligent systems learn to adequately act in a given environment with respect to a particular task, thereby adapting, is of particular relevance to cognitive science disciplines such as psychology, biology, and computer science (e.g., Pfeifer and Bongard, 2007; Wolpert et al., 2011; Abrahamsen and Bechtel, 2012; Pacherie, 2012; Engel et al., 2013, 2015). This capability of goal-directed motor (inter)action changes and develops with practice, transitioning from unskilled into skilled motor (inter)action, and resulting in refined planning and execution

of motor (inter)actions (e.g., Meinel and Schnabel, 2007; Schmidt and Wrisberg, 2008; Magill, 2011; Schmidt and Lee, 2011). Interestingly, advancing our understanding of intelligent systems' actions and their acquisition remains a significant endeavor to this day, especially in view of prospective applications in various settings such as robotics, psychology, sports, and rehabilitation. For instance, the development of intelligent interactive technical platforms which are to assist humans requires a thorough understanding of natural, intelligent forms of (inter)action and their acquisition, respectively (e.g., Pfeifer and Bongard, 2007; Schack and Ritter, 2009, 2013; Di Nuovo et al., 2013; De Kleijn et al., 2014). Understanding learning by way of different states of action (e.g., imagery or execution) and related functional changes within the motor action system, particularly with regards to action representation may help to advance in this direction. Here, we overview the literature on learning by imagery and execution from three perspectives, namely the performance, the brain, and the mind perspective.

STATES OF (INTER)ACTION AND LEARNING

An action reflects "a set of mechanisms that are aimed at producing activation of the motor system for reaching a goal" (Jeannerod, 2004, p. 376). Similarly, interaction may be considered as sets of mechanisms of several individuals acting together, which are aimed at producing activations of all motor systems involved for reaching a shared goal. (Inter)actions can overt as well as covert actions, that is executed, imagined or observed actions (Jeannerod, 2001, 2004). Given the principle of functional equivalence (Finke, 1979; Johnson, 1980; Jeannerod, 1994, 1995) and the simulation theory (Jeannerod, 2001, 2004, 2006), executed, imagined, and observed actions are all suggested to be actions, as each draws on the same action representation. While 'actual' actions involve both a covert (e.g., planning) and an overt (e.g., execution) stage of action, 'simulated' actions such as imagery imply a covert stage of action only (i.e., simulation state; s-state; Jeannerod, 2001). To this extent, each of the different types of s-states to some degree involves the activation of the motor action system. That is, any form of executed or simulated state of action is considered an action, regardless of whether it includes covert stages of action only or both covert and overt stages of action. Given the principle of functional equivalence, the repeated use of any of these states as means of practice should lead to functional changes within the motor action system and to learning. Accordingly, mental types of practice have been suggested to be effective means to induce learning (e.g., Jeannerod, 1994, 1995, 2001, 2004).

To date, it is widely accepted that humans can learn by way of different states of (inter-)action, but their unique potential to induce changes in the motor action system is still being debated (e.g., Driskell et al., 1994; Allami et al., 2014; Di Rienzo et al., 2016; Frank et al., 2016). Interestingly, while evidence on the functional equivalence of executed and imagined actions is vast (e.g., Finke, 1979; Johnson, 1980; Jeannerod, 1994, 1995, 2001; Decety, 1996, 2002; Jeannerod and Frak, 1999), only little is known about how learning by execution or imagery works. Furthermore, it is unclear what the similarities and differences of these ways of learning are, particularly with regards to changes in action representation. In other words, research has yet to systematically examine the differential effects of learning by way of different states of action.

In this perspective paper, we focus on learning by way of imagery and execution, and discuss it from a perceptual-cognitive point of view on action representation. For this purpose, we review learning by way of imagery and execution from three different levels of analyses. First, we examine the literature from the performance perspective (here: in terms of changes in motor behavior), followed by the brain perspective (here: in terms of changes in neurophysiological representations of motor action), and finally by the mind perspective (here: in terms of changes in perceptual-cognitive representations of motor action). In doing so, we highlight the role of action representation within a motor hierarchy, and exemplify how such models could advance our understanding of learning, enabling links between neurophysiological approaches and motor control and learning theories. Finally, we discuss potential future directions to advance research comparing learning by way of execution, imagery, and other states of action.

THE PERFORMANCE PERSPECTIVE ON IMAGERY AND EXECUTION: LEARNING AS CHANGES IN MOTOR PERFORMANCE

The systematic use of different states of action for practice and their contribution to the learning of complex motor actions has traditionally been approached by way of persisting performance improvements (e.g., Schmidt and Lee, 2011). Similarly, researchers investigating the influence of mental practice traditionally have focused on motor performance (e.g., Corbin, 1967a,b; for reviews and meta-analyses, see Richardson, 1967a,b; Feltz and Landers, 1983; Feltz et al., 1988; Hinshaw, 1991; Grouios, 1992; Driskell et al., 1994). From this, mental practice has shown to be more effective than no practice, but less effective than physical practice (e.g., Feltz and Landers, 1983; Feltz et al., 1988; Driskell et al., 1994). Driskell et al. (1994), for instance, conducted a meta-analysis on the effects of mental practice in comparison to irrelevant practice and physical practice, reporting an overall average effect size of $d = 0.53$[1] for mental practice, and an effect size of $d = 0.78$ for physical practice. Moreover, combined mental and physical practice has been suggested to be as effective as or superior to physical practice (e.g., Corbin, 1967b; McBride and Rothstein, 1979; Hall et al., 1992; Gomes et al., 2014). From this perspective, mental practice is considered a potentially effective means to promote learning.

[1]Effect sizes reported throughout this chapter refer to Cohen's d (Cohen, 1992).

THE BRAIN PERSPECTIVE ON IMAGERY AND EXECUTION: LEARNING AS CHANGES IN NEUROPHYSIOLOGICAL ACTION REPRESENTATION

In search of answers to the question why learning by way of different states of action works (e.g., Heuer, 1985; Murphy, 1990; Murphy et al., 2008), neurocognitive approaches have evolved, considering learning from within (e.g., Jeannerod, 2001, 2004). Neurocognitive approaches highlight the role of action representation in the learning of complex motor actions from a neurophysiological perspective. So far, the adaptation of the brain (i.e., neurophysiological and -anatomical changes) as a result of physical practice has received a great deal of attention (e.g., Wadden et al., 2012). From this, multifaceted insights into central changes within the motor action system have been provided regarding the neural aspects of learning a motor action, and the neural plasticity of the brain, respectively (for a recent meta-analysis, see Hardwick et al., 2013; for reviews, see also e.g., Doyon and Ungerleider, 2002; Ungerleider et al., 2002; Doyon and Benali, 2005; Kelly and Garavan, 2005; Halsband and Lange, 2006; Dayan and Cohen, 2011). In the context of the principle of functional equivalence and the simulation theory (Jeannerod, 2001, 2004, 2006), the study of action representation from a neurophysiological point of view has received tremendous research interest (for overviews, see e.g., Decety, 2002; Guillot et al., 2014). While considerable research attention has been directed to comparing the different states of action, such as the imagery and the execution of an action (e.g., Decety, 1996, 2002; Jeannerod and Frak, 1999), only few studies exist that compare learning by way of imagery and execution and respective changes in the brain (e.g., Pascual-Leone et al., 1995; Jackson et al., 2003; Nyberg et al., 2006; Zhang et al., 2012, 2014; Allami et al., 2014; Avanzino et al., 2015; for a review, see Di Rienzo et al., 2016).

For instance, Pascual-Leone et al. (1995) investigated plastic changes in the human motor action system resulting from physical and mental practice, using transcranial magnetic stimulation. Interestingly, while the authors found physical practice to be superior to mental practice in terms of performance improvement in a key pressing task, both physical and mental practice led to the same plastic changes, namely an equally increased size of the cortical representation for the finger muscle groups involved. From this, the authors concluded that mental practice modulates the neural circuits involved in learning, potentially by forming a cognitive model of the motor action. Jackson et al. (2003) investigated cerebral functional changes in the brain as induced by mentally practicing foot movements employing positron emission tomography and compared these changes to those induced by physically practicing foot movements (Lafleur et al., 2002). Similar to the findings reported by Lafleur et al. (2002) on physical practice effects, the authors found mental practice to be associated with functional cerebral reorganization in the right medial orbitofrontal cortex. From the lack of striatum activation after mental practice, however, the authors suggest that the re-organization rather relates to the planning and the anticipation of motor actions than to its motor execution. More recently, Zhang et al. (2014) examined changes in functional connectivity in resting state as a result of mental practice, using functional magnetic resonance imaging. The authors reported alterations in cognitive and sensory resting state networks in various brain systems after learning by way of motor imagery (i.e., mental practice), while no alterations in connectivity were found in the control condition (i.e., no practice). From this, the authors concluded that modulation of resting-state functional connectivity as induced by mental practice may be associated with attenuation in cognitive processing related to the formation of motor schemas. These neurophysiological studies on learning as induced by mental practice and/or physical practice show that both mental and physical practice lead to significant changes in action-related brain activation during skill acquisition. At the same time, however, they reveal distinct differences pointing to a hierarchy in learning by way of different states of action (for more details, see discussion section).

From a neurophysiological perspective, learning can be considered as neurophysiological reorganization, with the neurophysiological representation of motor action functionally developing over the course of the learning process. This seems to hold for both learning by execution and learning by imagery. Neurophysiological studies as the ones exemplified above provide valuable multifaceted insights into the functional changes in brain activation as a result of physical and mental practice. Findings elucidating neurophysiological changes associated with motor learning as induced by mental and physical practice, however, do not necessarily allow for specific conclusions regarding action representation and its relation to motor control. Therefore, it seems important to link these approaches to models and theories of motor control and learning, particularly those emphasizing the role of action representation, in order to be able to draw specific conclusions about changes of the motor action system during learning. To put it differently: Given the functional reorganization of neurophysiological features in the brain, is there a functional reorganization of perceptual-cognitive representations of motor (inter)action in the mind as part of a functional stratification on various levels within the motor action system?

THE MIND PERSPECTIVE ON IMAGERY AND EXECUTION: LEARNING AS CHANGES IN PERCEPTUAL-COGNITIVE ACTION REPRESENTATION

According to perceptual-cognitive approaches (e.g., theory of anticipative behavioral control: Hoffmann, 1993; theory of event coding: Hommel et al., 2001; simulation theory: Jeannerod, 2001) and the original idea of a bidirectional link between an action and its effects (i.e., ideomotor theory: James, 1890), actions are primarily guided by cognitively represented perceptual effects. Drawing on the seminal work of Bernstein (1967) and his idea of a model of the desired future, motor actions can be considered

as being stored in memory as well-integrated representational networks or taxonomies comprised of perceptual-cognitive units (i.e., basic action concepts; BACs) that guide action execution (cf. cognitive action architecture approach/ CAA-A: for an overview, see Schack, 2004; Schack and Ritter, 2009). Moreover, these networks of BACs are suggested to change throughout the process of motor learning by way of perceptual-cognitive scaffolding, resulting in a more elaborate perceptual-cognitive representation.

Based on research relating to CAA-A (e.g., Schack and Mechsner, 2006), experts, as compared to novices, have been shown to hold structured representations. A functionally structured representation is comprised of groupings of perceptual-cognitive units (i.e., groupings of BACs) that relate to the same (sub-)functions of the action, and thus reflect the functional phases of the motor action (cf. Göhner, 1992, 1999; Hossner et al., 2015). Schack and Mechsner (2006), for instance, examined representational networks of the tennis serve in experts and non-experts, using the structural dimensional analysis of mental representations (SDA-M). Results elicited that skilled individuals held functionally structured representations relating well to the biomechanical demands of the task (i.e., reflecting clearly the three movement phases pre-activation, strike, and final swing), whereas unskilled individuals' representations were unstructured. This has been shown to generalize to motor skills of different complexities (e.g., manual action: Braun et al., 2007; gait: Schega et al., 2014; Stöckel et al., 2015; dance: Bläsing, 2010).

With regards to learning, action representations have been shown to functionally adapt in the direction of an elaborate representation during motor learning (Frank et al., 2013). Findings revealed that, together with improvements in golf putting performance, representations changed with practice, developing toward more functional ones, with groupings of perceptual-cognitive units (i.e., groupings of BACs) relating more closely to the same (sub-)functions of the action itself (i.e., preparation, forward swing, and impact). Drawing on the finding that novices' perceptual-cognitive representations of complex action develop and adapt with practice, Frank et al. (2014) addressed the development of one's representation according to type of practice, comparing physical practice (i.e., repeated motor execution), mental practice (i.e., repeated motor imagery) and their combination. While motor performance reflected the well-known pattern of magnitude of improvement according to type of practice (i.e., combined practice > physical practice > mental practice > no practice), mental practice, either solely or in combination with physical practice, led to even more elaborate representations compared to physical practice only. Representation structures of the groups practicing mentally became more similar to a functional expert structure, whereas those of the physical practice group revealed less development. Building on these findings, Frank et al. (2016) further examined the perceptual-cognitive background of performance changes that occur within the motor action system as a result of mental and physical practice, employing a mobile eye-tracking system to investigate gaze behavior (i.e., the quiet eye; e.g., Vickers, 1992, 1996, 2009). Combined practice led both to more developed representation structures and to more elaborate gaze behavior prior to the execution of the putt, with final fixations prior to the onset of the putting movement (i.e., the quiet eye) being longest for this group and better developed representation structures relating to longer quiet eye durations after learning. Accordingly, the quiet eye might reflect a predictive mode of control that initiates a cognitively demanding process of motor planning based on the representation available (for details on a perceptual-cognitive perspective on the quiet eye, see Frank and Schack, 2016).

More recently, learning as it relates to interaction was investigated by examining representational frameworks of interaction and their development with mental practice (Frank et al., under review). The impact of a team action imagery intervention on futsal player's shared representations of team-specific tactics was investigated. Mental practice consisted of practicing four team-specific tactics (i.e., counter-attack, play making, pressing, transitioning) by imagining team actions in specific game situations for three times a week over the course of 4 weeks. Results revealed representational networks of team action becoming more similar to those of experts after mental practice. This study indicates that the imagery of team actions can have a significant impact on players' representational networks of interaction in long-term memory.

From this line of studies, the learning of a motor action can be considered as perceptual-cognitive reorganization, with the perceptual-cognitive representation of action functionally developing throughout the learning process. This research furthermore indicates that the perceptual-cognitive reorganization taking place during learning depends on the state of action used for practice. Learning by way of imagery differs from learning by way of execution, with practice through imagery promoting the functional development of a perceptual-cognitive action representation (perceptual-cognitive explanation of mental practice), while not necessarily transferring one-to-one to motor performance. This points to a differential influence of mental and physical practice with regards to different levels of action organization, with mental practice operating primarily on higher levels within the motor action system, particularly during early skill acquisition (for a more detailed discussion, see Frank, 2014). This approach, particularly together with neurophysiological approaches, may add to the picture of potential basic mechanisms that underlie each type of practice, an issue still being highly debated (e.g., Annett, 1995; Jackson et al., 2001; Munzert et al., 2008; Murphy et al., 2008; Cumming and Williams, 2012; Glover and Dixon, 2013). By complementing existing evidence from a performance and a brain perspective on learning by mental and physical practice (e.g., Driskell et al., 1994; Allami et al., 2014), these findings contribute to a better understanding of the adapting motor action system, by disentangling changes on various levels within the motor action system during learning.

DISCUSSION AND CONCLUSION

While there is ample evidence on the functional equivalence between different states of action (such as the imagery and the execution of an action; e.g., Decety, 1996, 2002; Jeannerod and

Frak, 1999), research addressing the similarity or difference with respect to the influence that each of the states of action has on the motor action system during learning has remained scarce to date. Meanwhile, more and more researchers have claimed to take into account potential differences between the states of action and their contribution to motor control and learning, as these might be as well (or in particular) meaningful to fully understand the motor action system (e.g., Munzert et al., 2009; Wakefield et al., 2013; O'Shea and Moran, 2017). Given that each state of action differs to some degree, the repeated use of imagery or execution is likely to differ in their influence on the motor action system. In other words, while the repeated use of imagery and execution of an action is suggested to result in learning, learning is likely to differ as a function of the state of action used for practice.

Here, we outlined learning by way of imagery and execution from three perspectives. While there is ample evidence from the performance perspective (for a review, see e.g., Driskell et al., 1994), the research from a brain perspective (for a review, see e.g., Di Rienzo et al., 2016), and from a mind perspective (e.g., Frank et al., 2016) on action representation as it relates to learning by imagery and execution has just started to gain momentum. Despite these initial steps, the potential of imagery and execution to induce changes within the motor action hierarchy during learning, however, remains to be explored more thoroughly. Interestingly, although sometimes not explicitly introduced as the theoretical background of their studies, (indirect or direct) conclusions about the formation of action representations are drawn from the brain changes observed, linking neurophysiological findings to hierarchical motor control and learning theories: for instance, Pascual-Leone et al. (1995, p. 1043) discussed that repeated imagery may help establish a cognitive model of the motor action; Zhang et al. (2014, p. 4) state that motor schemas have developed; Jackson et al. (2003, p. 1178) discuss from the lack of striatum activation after mental practice, that the re-organization relates to the planning and the anticipation of motor actions rather than to its motor execution. By doing so, each of the studies implicitly refers to a theoretical background of motor control and learning, and alludes to some form of representational format in memory. However, the results of these studies have not yet been discussed in the light of hierarchical models of action organization, focusing on higher and lower levels of action representation, as the one delineated in the present perspective paper. By suggesting that mental practice helps promote a 'cognitive model,' 'attenuated cognitive processing,' and the 'planning and the anticipation of actions,' these findings are in line with the perceptual-cognitive explanation of mental practice and the idea that the repeated use of imagery particularly helps establish perceptual-cognitive representations of action (Frank et al., 2014, 2016).

Future studies may place more emphasis on the role of action representation and compare learning by way of imagery and learning by execution with regards to brain- and mind-related changes on different levels within motor action system. For instance, related research disentangling neurophysiological representations of actions within a motor hierarchy (e.g., Grafton and Hamilton, 2007), research on the degree of abstractness of neurophysiological representation of actions (e.g., Tucciarelli et al., 2015; Wurm and Lingnau, 2015; Turella et al., 2016), or research on neurophysiological representations' structural geometry across states of action (Zabicki et al., 2016) in conjunction with perceptual-cognitive approaches to motor learning might be promising avenues to better understand learning across states of action. In a recent study, for instance, Zabicki et al. (2016) investigated imagined and executed actions using a multivariate approach and a representational similarity analysis to neurophysiological representations of action, highlighting a similar structural geometry as well as distinct differences in action representation between the two states of action. Using such approaches together with hierarchical, perceptual-cognitive ones in the realm of motor cognition might help to further approach the phenomenon of action representation in motor control and learning and the unique potential of imagery and execution to induce changes on different levels within the motor action system during learning.

In sum, research directly comparing the two modes of learning has remained scarce to date, with many studies focusing on one mode only (e.g., imagery: Zhang et al., 2014; execution: Lafleur et al., 2002). Furthermore, most of the studies conducted so far focus on the potential similarities that learning by way of motor imagery may share with learning by way of motor execution, thereby disregarding potential differences across learning types, such as a differential influence on various levels within the motor action system. And finally, the brain and the mind perspective have been considered merely isolated, investigating neurophysiological representations or perceptual-cognitive representations. Accordingly, three main challenges may have to be addressed by future studies in order to advance research comparing learning by way of execution, imagery, and observation, and thus to more thoroughly understanding intelligent systems and learning by different states of action. First, research comparing learning by different states of action should be conducted in a systematic manner, employing research designs that allow for examining states of action both in isolation as well as in combination (cf. four group design in mental practice research, e.g., Corbin, 1967b; Hall et al., 1992). Second, research questions and hypotheses should be directed toward the differences between learning by different states of action, and thus going beyond the traditional focus on the functional equivalence between the states of action, and the potential similarities across learning types, toward a hierarchical view of the motor action system. Third, learning by different states of action should be approached in future research by integrating findings and methods from different disciplines (e.g., Moran et al., 2012) such as the ones exemplified above in order to approach the problem from distinct, but complementary perspectives.

To systematically examine learning by different states of action from various perspectives focusing on both the similarities and the differences across higher and lower levels of action organization may contribute to a better understanding of the

motor action system. Complementing both the performance and the brain perspective by a mind perspective may lead to advancing our understanding of intelligent systems in general, and the learning of (inter)action across states of action in particular, in order to better be able to design training tools that facilitate motor (re)learning. Future research should therefore build bridges between the perspectives in order to more thoroughly understand functional changes throughout learning across states of action, and to subsequently address specific levels within the motor action hierarchy as part of individualized coaching in robotics, rehabilitation, or sports settings (e.g., Hülsmann et al., 2016).

AUTHOR CONTRIBUTIONS

Conception and draft of perspective paper (CF and TS).

FUNDING

This research/work was supported by the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC277) at Bielefeld University, which is funded by the German Research Foundation (DFG). Furthermore, we acknowledge support for the Article Processing Charge by the German Research Foundation and the Open Access Publication Fund of Bielefeld University.

REFERENCES

Abrahamsen, A., and Bechtel, W. (2012). "History and core themes," in *Cambridge Handbook of Cognitive Science*, eds K. Frankish and W. Ramsey (Cambridge: Cambridge University Press), 9–28. doi: 10.1017/CBO9781139033916.003

Allami, N., Brovelli, A., Hamzaoui, E. M., Regragui, F., Paulignan, Y., and Boussaoud, D. (2014). Neurophysiological correlates of visuo-motor learning through mental and physical practice. *Neuropsychologia* 55, 6–14. doi: 10.1016/j.neuropsychologia.2013.12.017

Annett, J. (1995). On knowing how to do things: a theory of motor imagery. *Cogn. Brain Res.* 3, 65–69. doi: 10.1016/0926-6410(95)00030-5

Avanzino, L., Gueugneau, N., Bisio, A., Ruggeri, P., Papaxanthis, C., and Bove, M. (2015). Motor cortical plasticity induced by motor learning through mental practice. *Front. Behav. Neurosci.* 9:105. doi: 10.3389/fnbeh.2015.00105

Bernstein, N. A. (1967). *The Co-ordination and Regulation of Movements*. Oxford: Pergamon Press.

Bläsing, B. (2010). "The dancer's memory: Expertise and cognitive structures in dance," in *The Neurocognition of Dance*, eds B. Bläsing, M. Puttke, and T. Schack (London: Psychology Press), 75–98.

Braun, S. M., Beurskens, A. J. H. M., Schack, T., Marcellis, R. G., Oti, K. C., Schols, J. M., et al. (2007). Is it possible to use the SDA-M to investigate representations of motor actions in stroke patients? *Clin. Rehabil.* 21, 822–832.

Cohen, J. (1992). A power primer. *Psychol. Bull.* 112, 155–159.

Corbin, C. (1967a). Effects of mental practice on skill development after controlled practice. *Res. Q.* 38, 534–538.

Corbin, C. (1967b). The effects of covert rehearsal on the development of a complex motor skill. *J. Gen. Psychol.* 76, 143–150. doi: 10.1080/00221309.1967.9710383

Cumming, J., and Williams, S. E. (2012). "The role of imagery in performance," in *The Oxford Handbook of Sport and Performance Psychology*, ed. S. M. Murphy (New York, NY: Oxford University Press), 213–232.

Dayan, E., and Cohen, L. G. (2011). Neuroplasticity subserving motor skill learning. *Neuron* 72, 443–454. doi: 10.1016/j.neuron.2011.10.008

De Kleijn, R., Kachergis, G., and Hommel, B. (2014). Everyday robotic action: lessons from human action control. *Front. Hum. Neurosci.* 8:13. doi: 10.3389/fnbot.2014.00013

Decety, J. (1996). Do imagined and executed actions share the same neural substrate? *Cogn. Brain Res.* 3, 87–93.

Decety, J. (2002). "Is there such a thing as functional equivalence between imagined, observed and executed action?," in *The Imitative Mind: Development, Evolution, and Brain Bases*, eds A. N. Meltzoff and W. Prinz (Cambridge: Cambridge University Press), 291–310.

Di Nuovo, A. G., Marocco, D., Di Nuovo, S., and Cangelosi, A. (2013). Autonomous learning in humanoid robotics through mental imagery. *Neural Networks* 41, 147–155. doi: 10.1016/j.neunet.2012.09.019

Di Rienzo, F., Debarnot, U., Daligault, S., Saruco, E., Delpuech, C., Doyon, J., et al. (2016). Online and offline performance gains following motor imagery practice: a comprehensive review of behavioral and neuroimaging studies. *Front. Hum. Neurosci.* 10:315. doi: 10.3389/fnhum.2016.00315

Doyon, J., and Benali, H. (2005). Reorganization and plasticity in the adult brain during learning of motor skills. *Curr. Opin. Neurobiol.* 15, 161–167. doi: 10.1016/j.conb.2005.03.004

Doyon, J., and Ungerleider, L. G. (2002). "Functional anatomy of motor skill learning," in *Neuropsychology of Memory*, eds L. R. Squire and D. L. Schacter (New York, NY: Guilford), 225–238.

Driskell, J., Copper, C., and Moran, A. (1994). Does mental practice enhance performance? *J. Appl. Psychol.* 79, 481–492. doi: 10.1037/0021-9010.79.4.481

Engel, A. K., Friston, K., and Kragic, D. (2015). *Where's the Action? Toward Action-Oriented Views in Cognitive Science.* Strüngmann Forum Reports. Cambridge, MA: MIT Press.

Engel, A. K., Maye, A., Kurthen, M., and König, P. (2013). Where's the action? The pragmatic turn in cognitive science. *Trends Cogn. Sci.* 17, 203–209. doi: 10.1016/j.tics.2013.03.006

Feltz, D., and Landers, D. (1983). The effects of mental practice on motor skill learning and performance: a meta-analysis. *J. Sport Psychol.* 5, 25–57. doi: 10.1123/jsp.5.1.25

Feltz, D. L., Landers, D. M., and Becker, B. J. (1988). "A revised meta-analysis of the mental practice literature on motor skill learning," in *Enhancing Human Performance, Part III: Improving Motor Performance*, ed. National Research Council (Washington, DC: National Academy Press).

Finke, R. A. (1979). The functional equivalence of mental images and errors of movement. *Cognit. Psychol.* 11, 235–264. doi: 10.1016/0010-0285(79)90011-2

Frank, C. (2014). Mental Representation and Learning in Complex Action: A Perceptual-Cognitive View on Mental and Physical Practice. Ph.D. thesis, Bielefeld University, Bielefeld.

Frank, C., Land, W. M., Popp, C., and Schack, T. (2014). Mental representation and mental practice: experimental investigation on the functional links between motor memory and motor imagery. *PLoS ONE* 9:e95175. doi: 10.1371/journal.pone.0095175

Frank, C., Land, W. M., and Schack, T. (2013). Mental representation and learning: the influence of practice on the development of mental representation structure in complex action. *Psychol. Sport Exerc.* 14, 353–361. doi: 10.1371/journal.pone.0095175

Frank, C., Land, W. M., and Schack, T. (2016). Perceptual-cognitive changes during motor learning: the influence of mental and physical practice on mental representation, gaze behavior, and performance of a complex action. *Front. Psychol.* 6:1981. doi: 10.3389/fpsyg.2015.01981

Frank, C., and Schack, T. (2016). In my mind's (quiet) eye: a perceptual-cognitive approach to the quiet eye – comment on vickers. *Curr. Issues Sport Sci.* 1:107. doi: 10.15203/CISS_2016.107

Glover, S., and Dixon, P. (2013). Context and vision effects on real and imagined actions: support from the common representation hypothesis of motor imagery. *J. Exp. Psychol.* 39, 1352–1364. doi: 10.1037/a0031276

Göhner, U. (1992). *Einführung in die Bewegungslehre des Sports, Teil 1: Die sportlichen Bewegungen*. Schorndorf: Hofmann.

Göhner, U. (1999). *Einführung in die Bewegungslehre des Sports, Teil 2: Bewegungslehre des Sports*. Schorndorf: Hofmann.

Gomes, T. V. B., Ugrinowitsch, H., Marinho, N., Shea, J. B., Raisbeck, L. D., and Benda, R. N. (2014). Effects of mental practice in novice learners in a serial positioning skill acquisition. *Percept. Mot. Skills* 119, 397–414. doi: 10.2466/23.PMS.119c20z4

Grafton, S. T., and Hamilton, A. F. (2007). Evidence for a distributed hierarchy of action representation in the brain. *Hum. Mov. Sci.* 26, 590–616. doi: 10.1016/j.humov.2007.05.009

Grouios, G. (1992). Mental practice: a review. *J. Sport Behav.* 15, 42–59.

Guillot, A., Di Rienzo, F., and Collet, C. (2014). "The neurofunctional architecture of motor imagery," in *Advanced Brain Neuroimaging Topics in Health and Disease: Methods and Applications*, eds T. D. Papageorgiou, G. I. Christopoulos, and S. M. Smirnakis (Rijeka: InTech), 421–443.

Hall, C., Buckolz, E., and Fishburne, G. (1992). Imagery and the acquisition of motor skills. *Can. J. Sport Sci.* 17, 19–27.

Halsband, U., and Lange, R. K. (2006). Motor learning in man: a review of functional and clinical studies. *J. Physiol. Paris* 99, 414–424. doi: 10.1016/j.jphysparis.2006.03.007

Hardwick, R. M., Rottschy, C., Miall, R. C., and Eickhoff, S. B. (2013). A quantitative meta- analysis and review of motor learning in the human brain. *NeuroImage* 67, 283–297. doi: 10.1016/j.neuroimage.2012.11.020

Heuer, H. (1985). Wie wirkt mentale Übung? [How does mental practice work?]. *Psychol. Rundsch.* 36, 191–200.

Hinshaw, K. (1991). The effects of mental practice on motor skill performance: critical evaluation and meta-analysis. *Imagin. Cogn. Pers.* 11, 3–35. doi: 10.2190/X9BA-KJ68-07AN-QMJ8

Hoffmann, J. (1993). *Vorhersage und Erkenntnis*. Göttingen: Hogrefe.

Hommel, B., Müsseler, J., Aschersleben, G., and Prinz, W. (2001). The theory of event coding (TEC): a framework for perception and action planning. *Behav. Brain Sci.* 24, 849–937. doi: 10.1017/S0140525X01000103

Hossner, E.-J., Schiebl, F., and Göhner, U. (2015). A functional approach to movement analysis and error identification in sports and physical education. *Front. Psychol.* 6:1339. doi: 10.3389/fpsyg.2015.01339

Hülsmann, F., Frank, C., Schack, T., Kopp, S., and Botsch, M. (2016). "Multi-level analysis of motor actions as a basis for effective coaching in virtual reality," in *Proceedings of the International Symposium on Computer Science in Sport: Advances in Intelligent Systems and Computing*, Vol. 392, eds P. Chung, A. Soltoggio, C. Dawson, Q. Meng, and M. Pain (Heidelberg: Springer), 211–214. doi: 10.1007/978-3-319-24560-7_27

Jackson, P. L., Lafleur, M. F., Malouin, F., Richards, C. L., and Doyon, J. (2001). Potential role of mental practice using motor imagery in neurologic rehabilitation. *Arch. Phys. Med. Rehabil.* 82, 1133–1141. doi: 10.1053/apmr.2001.24286

Jackson, P. L., Lafleur, M. F., Malouin, F., Richards, C. L., and Doyon, J. (2003). Functional cerebral reorganization following motor sequence learning through mental practice with motor imagery. *Neuroimage* 20, 1171–1180. doi: 10.1016/S1053-8119(03)00369-0

James, W. (1890). *The Principles of Psychology*. New York, NY: Holt. doi: 10.1037/11059-000

Jeannerod, M. (1994). The representing brain: neural correlates of motor intention and imagery. *Behav. Brain Sci.* 17, 187–245. doi: 10.1017/S0140525X00034026

Jeannerod, M. (1995). Mental imagery in the motor context. *Neuropsychologia* 33, 1419–1432. doi: 10.1016/0028-3932(95)00073-C

Jeannerod, M. (2001). Neural simulation of action: a unifying mechanism for motor cognition. *Neuroimage* 14, S103–S109. doi: 10.1006/nimg.2001.0832

Jeannerod, M. (2004). Actions from within. *Int. J. Sport Exerc. Psychol.* 2, 376–402. doi: 10.1080/1612197X.2004.9671752

Jeannerod, M. (2006). *Motor Cognition: What Actions Tell the Self*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780198569657.001.0001

Jeannerod, M., and Frak, V. (1999). Mental imaging of motor activity in humans. *Curr. Opin. Neurobiol.* 9, 735–739. doi: 10.1016/S0959-4388(99)00038-0

Johnson, P. (1980). The functional equivalence of imagery and movement. *Q. J. Exp. Psychol. A Hum. Exp. Psychol.* 34, 349–365. doi: 10.1080/14640748208400848

Kelly, A. M., and Garavan, H. (2005). Human functional neuroimaging of brain changes associated with practice. *Cereb. Cortex* 15, 1089–1102. doi: 10.1093/cercor/bhi005

Lafleur, M. F., Jackson, P. L., Malouin, F., Richards, C. L., Evans, A. C., and Doyon, J. (2002). Motor learning produces parallel dynamic functional changes during the execution and imagination of sequential foot movements. *Neuroimage* 16, 142–157. doi: 10.1006/nimg.2001.1048

Magill, R. A. (2011). *Motor Learning and Control: Concepts and Applications*. New York, NY: McGraw-Hill.

McBride, E. R., and Rothstein, A. L. (1979). Mental and physical practice and the learning and retention of open and closed skills. *Percept. Mot. Skills* 49, 359–365. doi: 10.2466/pms.1979.49.2.359

Meinel, K., and Schnabel, G. (2007). *Bewegungslehre, Sportmotorik*. Aachen: Meyer & Meyer.

Moran, A., Guillot, A., MacIntyre, T., and Collet, C. (2012). Re-imagining motor imagery: Building bridges between cognitive neuroscience and sport psychology. *Br. J. Psychol.* 103, 224–247. doi: 10.1111/j.2044-8295.2011.02068.x

Munzert, J., Lorey, B., and Zentgraf, K. (2009). Cognitive motor processes: the role of motor imagery in the study of motor representations. *Brain Res. Rev.* 60, 306-326. doi: 10.1016/j.brainresrev.2008.12.024

Munzert, J., Zentgraf, K., Stark, R., and Vaitl, D. (2008). Neural activation in cognitive motor processes: comparing motor imagery and observation of gymnastic movements. *Exp. Brain Res.* 188, 437–444. doi: 10.1007/s00221-008-1376-y

Murphy, S., Nordin, S., and Cumming, J. (2008). "Imagery in sport, exercise, and dance," in *Advances in Sport Psychology*, ed. T. S. Horn (Champaign, IL: Human Kinetics), 297–324.

Murphy, S. M. (1990). Models of imagery in sport psychology: a review. *J. Mental Imagery* 14, 153–172.

Nyberg, L., Eriksson, J., Larsson, A., and Marklund, P. (2006). Learning by doing versus learning by thinking: an fMRI study of motor and mental training. *Neuropsychologia* 44, 711-717. doi: 10.1016/j.neuropsychologia.2005.08.006

O'Shea, H., and Moran, A. (2017). Does motor simulation theory explain the cognitive mechanisms underlying motor imagery? A critical review. *Front. Human Neurosci.* 11:72. doi: 10.3389/fnhum.2017.00072

Pacherie, E. (2012). "Action," in *Cambridge Handbook of Cognitive Science*, eds K. Frankish and W. Ramsey (Cambridge: Cambridge University Press), 92–111. doi: 10.1017/CBO9781139033916.008

Pascual-Leone, A., Dang, N., Cohen, L. G., Brasil-Neto, J. P., Cammarota, A., and Hallett, M. (1995). Modulation of muscle responses evoked by transcranial magnetic stimulation during the acquisition of new fine motor skills. *J. Neurophysiol.* 74, 1037–1045.

Pfeifer, R., and Bongard, J. (2007). *How the Body Shapes the Way We Think: A New View of Intelligence*. Cambridge: MIT Press.

Richardson, A. (1967a). Mental practice: a review and discussion, part I. *Res. Q.* 38, 95–107.

Richardson, A. (1967b). Mental practice: a review and discussion, part II. *Res. Q.* 38, 263–273.

Schack, T. (2004). The cognitive architecture of complex movement. *Int. J. Sport Exerc. Psychol.* 2, 403–438. doi: 10.1080/1612197X.2004.9671753

Schack, T., and Mechsner, F. (2006). Representation of motor skills in human long-term memory. *Neurosci. Lett.* 391, 77–81. doi: 10.1016/j.neulet.2005.10.009

Schack, T., and Ritter, H. (2009). "The cognitive nature of action: functional links between cognitive psychology, movement science and robotics," in *Progress in Brain Research. Mind and Motion: The Bidirectional Link between Thought and Action*, eds M. Raab, J. Johnson, and H. Heekeren (Amsterdam: Elsevier), 231–252. doi: 10.1016/S0079-6123(09)01319-3

Schack, T., and Ritter, H. (2013). Representation and learning in motor action: bridges between experimental research and cognitive robotics. *New Ideas in Psychol.* 31, 258–269. doi: 10.1016/j.newideapsych.2013.04.003

Schega, L., Bertram, D., Foelsch, C., Hamacher, C., and Hamacher, D. (2014). The influence of visual feedback on the mental representation of gait in patients with THR: a new approach for an experimental rehabilitation strategy. *Appl. Psychophysiol. Biofeedback* 39, 37–43. doi: 10.1007/s10484-014-9239-8

Schmidt, R., and Wrisberg, C. (2008). *Motor Learning and Performance. A Situation-Based Learning Approach*. Champaign, IL: Human Kinetics.

Schmidt, R. A., and Lee, T. D. (2011). *Motor Control and Learning: A Behavioral Emphasis*. Champaign, IL: Human Kinetics.

Stöckel, T., Jacksteit, R., Behrens, M., Skripitz, R., Bader, R., and Mau-Moeller, A. (2015). The mental representation of the human gait in young and older adults. *Front. Psychol.* 6:943. doi: 10.3389/fpsyg.2015.00943

Tucciarelli, R., Turella, L., Oosterhof, N. N., Weisz, N., and Lingnau, A. (2015). MEG multivariate analysis reveals early abstract action representations in the lateral occipitotemporal cortex. *J. Neurosci.* 35, 16034-16045. doi: 10.1523/JNEUROSCI.1422-15.2015

Turella, L., Rumiati, R., and Lingnau, A. (2016). Hierarchical organization of action encoding within the human brain. *J. Vis*. 16:24. doi: 10.1167/16.12.24

Ungerleider, L. G., Doyon, J., and Karni, A. (2002). Imaging brain plasticity during motor skill learning. *Neurobiol. Learn. Mem.* 78, 553–564. doi: 10.1006/nlme.2002.4091

Vickers, J. N. (1992). Gaze control in putting. *Perception* 21, 117–132. doi: 10.1068/p210117

Vickers, J. N. (1996). Visual control when aiming at a far target. *J. Exp. Psychol.* 22, 342–354. doi: 10.1037/0096-1523.22.2.342

Vickers, J. N. (2009). "Advances in coupling perception and action: The quiet eye as a bidirectional link between gaze, attention, and action," in *Progress in Brain Research. Mind and Motion: The Bidirectional Link between Thought and Action,* eds M. Raab, J. G. Johnson, and H. R. Heekeren (Amsterdam: Elsevier), 279–288. doi: 10.1016/S0079-6123(09)01322-3

Wadden, K. P., Borich, M. R., and Boyd, L. A. (2012). "Motor skill learning and its neurophysiology," in *Skill Acquisition in Sport*, eds N. J. Hodges and A. M. Williams (London: Routledge), 247–265.

Wakefield, C., Smith, D., Moran, A. P., and Holmes, P. (2013). Functional equivalence or behavioural matching: a critical reflection on 15 years of research using the PETTLEP model of motor imagery. *Int. Rev. Sport Exerc. Psychol.* 6, 105–121. doi: 10.1080/1750984X.2012.724437

Wolpert, D. M., Diedrichsen, J., and Flanagan, J. R. (2011). Principles of sensorimotor learning. *Nat. Rev. Neurosci.* 12, 739–751. doi: 10.1038/nrn3112

Wurm, M. F., and Lingnau, A. (2015). Decoding actions at different levels of abstraction. *J. Neurosci.* 35, 7727–7735. doi: 10.1523/JNEUROSCI.0188-15.2015

Zabicki, A., de Haas, B., Zentgraf, K., Stark, R., Munzert, J., and Krüger, B. (2016). Imagined and executed actions in the human motor system: testing neural similarity between execution and imagery of actions with a multivariate approach. *Cereb. Cortex*. doi: 10.1093/cercor/bhw257 [Epub ahead of print].

Zhang, H., Long, Z., Ge, R., Xu, L., Jin, Z., Yao, L., et al. (2014). Motor imagery learning modulates functional connectivity of multiple brain systems in resting state. *PLoS ONE* 9:e85489. doi: 10.1371/journal.pone.0085489

Zhang, H., Xu, L., Zhang, R., Hui, M., Long, Z., Zhao, X., et al. (2012). Parallel alterations of functional connectivity during execution and imagination after motor imagery learning. *PLoS ONE* 7:e36052. doi: 10.1371/journal.pone.0036052

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

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: info@frontiersin.org | +41 21 510 17 00

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership