Perspective

Memetics and neural models of conspiracy theories

Włodzisław Duch[1],*

[1]Department of Informatics, Faculty of Physics, Astronomy and Informatics, and Neurocognitive Laboratory, Center for Modern Interdisciplinary Technologies, Nicolaus Copernicus University, Toruń, Poland

*Correspondence: wduch@umk.pl

https://doi.org/10.1016/j.patter.2021.100353

THE BIGGER PICTURE Conspiracy theories are widespread. So far, research in this area has been focused on psychological, sociological, and political science perspectives. Brain processes facilitating formation of conspiracy theories are largely unknown. In neural systems, a meme may be represented by a quasi-stable associative memory network attractor state. Creation of memes with numerous fake associations distorts relations between stable memory states. Simulations of neural network models trained with competitive Hebbian learning (CHL) on stationary and non-stationary input data show the formation of distorted memory states. In non-stationary situations, rapid learning with high plasticity followed by stepwise decrease of plasticity leads to many states with overlapping attraction basins, distorting patterns in associative memory. Such system-level models may be used to understand conditions under which memplexes with distorted memory patterns arise, representing deeply settled conspiracy beliefs.

1 2 3 4 5 **Proof-of-Concept:** Data science output has been formulated, implemented, and tested for one domain/problem

SUMMARY

Memetics has so far been developing in social sciences, but to fully understand memetic processes it should be linked to neuroscience models of learning, encoding, and retrieval of memories in the brain. Attractor neural networks show how incoming information is encoded in memory patterns, how it may become distorted, and how chunks of information may form patterns that are activated by many cues, forming the foundation of conspiracy theories. The rapid freezing of high neuroplasticity (RFHN) model is offered as one plausible mechanism of such processes. Illustrations of distorted memory formation based on simulations of competitive learning neural networks are presented as an example. Linking memes to attractors of neurodynamics should help to give memetics solid foundations, show why some information is easily encoded and propagated, and draw attention to the need to analyze neural mechanisms of learning and memory that lead to conspiracies.

INTRODUCTION

Conspiracy theories are part of a much wider subject: formation of beliefs, creation of memes, distorted memories, twisted worldviews, or in general investigating ways in which learning fails to represent the data faithfully. In recent article by Seitz and Angel "Belief formation – A driving force for brain evolution,"[1] the authors write: "The topic of belief has been neglected in the natural sciences for a long period of time". They divide beliefs into empirical, relational, and conceptual, discussing large brain areas involved in the formation of beliefs. Bayesian models of belief propagation are used to model details of perceptual processes and relate them to connectomes.[2] The artificial neural network community has focused on faithful learning methods, but there is another, neglected side of learning and memory formation. When the training data are not learned perfectly, what types of errors may one expect, and how will they influence the performance of an artificial system? Can analysis of artificial systems help to understand how biological brains learn incoming information, transforming it into memes that are likely to be transmitted in a distorted form to other brains? The world view that we use to guide our behavior is based on a network of associative memory states. Consolidation of new memory states in the neocortex may occur quite quickly if they are well connected to other memory states.[3] Several lines of research lead to this conclusion: animal studies, association of places with items in mnemotechnics, behavioral studies on the use of schemas for rapid learning, and building of cognitive maps. Neural models of schemas and sequences of associations may be based on attractor states in neural networks.[4] Each episodic or semantic memory state is based on activations of synchronized, distributed networks of brain regions. It is encoded in relation to the existing activation patterns and may be modified when new patterns are learned.

Using functional magnetic resonance imaging (fMRI) evoked by natural movies, Huth et al.[5,6] have created a "semantic atlas," showing patterns of brain activations for categories of hundreds of objects and actions. These patterns are evoked by stimuli that provide sufficient cues to recall specific objects, such as body parts, animals, furniture, or types of actions. This process may be described using the language of dynamical systems for networks of elements representing neurons. The Hopfield network[7] was the simplest associative memory model encoding information in activation patterns of network nodes. In such recurrent networks, internal feedback changes activity patterns with time; this process is referred to as neurodynamics. All kinds of memory states (semantic, episodic, procedural, and working) are called attractors of neurodynamics[4] because initial patterns of neural network activations are attracted by the network dynamics toward one of the quasi-stable memory patterns. Usually only a small subset of neurons are highly active in each pattern, synchronizing their activity sending signals through strong mutual connections. In biologically motivated attractor networks, memory states are not stable, and neural noise, fatigue, and other processes lead to desynchronization, decrease activity of some neurons, and recruit others, forming different neural patterns. Transitions between neural patterns define trajectory of brain state changes in the space of neural activations. In artificial systems we can visualize it to observe neurodynamics of model networks[8] and transform it to dimensions that are meaningful at the mental level.[9] fMRI scans provide snapshots of the whole brain activation with temporal resolution of about 1 s and spatial resolution of about 1 mm, while measurement of electric potentials using electroencephalographic or magnetoencephalographic techniques provides millisecond temporal resolution but spatial resolution that is less than 1 cm.

Seitz et al.[10] presented a general theoretical model of formations of empirically grounded and metaphysical beliefs. In their view, the process of attraction is described by the verb "believing," and the endpoint, the final activation quasi-stable state, is called a "belief" and is interpreted as a mental construct. Beliefs are based on sensory perception and attribution of a personal value in an emotionally loaded process. High-level formula relates beliefs to incoming signals, ambient noise, current and previous valuation, learning, and prediction errors. Changes of neural activation in real brains depend on current knowledge schemas, history of previous activations (priming), general emotional state, specific context cues that invoke memories, and many other factors. Transitions that happen frequently increase probability of association between different activation patterns[11] and may not only create strong associations but distort or even completely blend different memories, creating false memories.[12,13] Understanding abnormal belief formation in neuropsychological disorders is an important challenge,[14] but neuropsychiatry needs precise hypotheses and models at the level of neural networks.

In some cases, memories may become easily activated in various contexts, leading to false associations and schemas that develop into conspiracy theories. While there is a large body of literature on conspiracy theories written by historians, philosophers, psychologists, sociologists, or political scientists (Routledge has a whole series of books on conspiracy theories; see also the review by Douglas et al.[15]), our understanding of the brain mechanisms is completely lacking. The best explanations that we have relate beliefs in conspiracy theories to personality traits, mental disorders, or the need to find a simple satisfactory explanation.

Memetics, introduced in the 1976 book *The Selfish Gene* by Richard Dawkins,[16] tried to explain cultural information transfer and persistence of certain ideas in societies. Memes may be understood as sequences or information structures that tend to replicate in a society. Despite great initial popularity of memetic ideas, and the desperate need of mathematical theories to underpin social science, theories connecting neuroscience and memetics have never been developed. *The Journal of Memetics* was discontinued in 2005 after 8 years of electronic publishing. Memetic ideas were relegated into a set of vague philosophical and psychological concepts of little interest to neuroscience. In evolutionary computing, memetic ideas have inspired many new developments, combining global search with focus on interesting local regions.[17] The *Memetic Computing* journal was established in 1989, a whole series of books on *Advances in Memetic Algorithms. Studies in Fuzziness and Soft Computing* is published by Springer. Research on memetic computing is focused on optimization problems, while here we are interested in the process of formation of memories.

The lack of efforts to understand distortions of information transmission and memory storage in biological learning systems is certainly related to the lack of theoretical models, and to the experimental difficulties in searching for memes in brain activity. McNamara[18] has argued that neuroimaging technology may be used to trace memes in the brain and to measure how they change over time. Following Heylighen and Chielens'[19] *memotype* and *mediotype* distinction, they propose to distinguish *i-memes*, internal activation of the central nervous system, from the external transmission/storage of information structures, the *e-memes* existing in the world (for example, created by marketing, or various media advertisements). One should distinguish clearly abstract information structure of memes, and their implementation in the brain or in artificial cognitive system. Internal representations of i-memes are created by forming memory states that link neural responses resulting from e-meme perception to behavioral (motor) responses that are necessary for replication of memes, linking sensory, memory, and motor subsystems in the brain. Sets of memes forming *memeplexes* determine world views, including culture, values, and religions, predisposing people to accept and propagate selected memes. Brain research has made a great progress in understanding schemas in the last decade.[20] Perhaps the time is ripe to make some progress along these lines to link the concept of memes with memory mechanisms that facilitate their spread. This could have very important social and educational implications.[21]

In the fascinating book *Why People Believe Weird Things*, Michel Shermer writes about 25 fallacies that lead people to believe in conspiracy theories and other bizarre things.[22] Brains are predisposed to perceive various observed patterns as meaningful information (pareidolia), search for explanations and form theories, referring to the long-term episodic and semantic memory. The conceptual framework that is needed to interpret new observations, including memes, is activated by various cues that invoke memory associations. Memes that are strongly encoded certainly influence most mental processes. Observations that agree with established individual beliefs will lead to strong activations of brain networks, thanks to the mutual co-activations of memeplex patterns, creating additional memes that make the whole memeplex even stronger. Contradicting arguments, facts, or observations will arouse only transient weak activations of brain networks and will be ignored. Worse than that, mentioning or presenting anything that may retrieve memes will only increase their influence, contributing to stronger encoding and easier arousal of false associations. The ''levels of processing'' paradigm in memory research has now found support in neuroimaging of deep and shallow episodic memory encoding, modulated by a number of neurotransmitters and linked to emotional arousal.[23] Research on forgetting shows that retrieval of competing memory traces may lead to interference and suppression of weaker patterns.[24] If conspiracy memes are already deeply encoded, they will distort formation of memory for contradicting facts. Although these facts may briefly activate brain patterns, the presence of strong memes will redirect these activations away toward conspiracy memes, preventing their understanding in a broader context.

Science systematically tries to falsify hypotheses by performing experiments, but, from the evolutionary perspective, falsification is simply too dangerous. In slowly changing environments, stability of beliefs is more important, even at the price of wide acceptance of meaningless taboos and superstitions. Even today, educational systems in most countries do not encourage skeptical thinking. Religious leaders and conservative politicians strongly oppose instating skepticism into the educational system, in fear of destabilization of established world views. There is little or no penalty for accepting false beliefs by individuals. Mutual support within groups of believers gives a boost to distorted views of reality, leading to bizarre conspiracy theories.

The discussion presented above shows that fake news and conspiracy theories tap into basic brain mechanisms of memory and learning. The complexity of the belief formation processes has discouraged scientists from approaching this important problem. Obviously, no simple computational model is going to explain all facts related to formation and preservation of human beliefs, and in particular of conspiracy theories. This should not discourage us from forming testable hypotheses based on neurodynamics. After all, simple neural network models introduced by Hopfield[7] and Kohonen,[25] despite being only loosely inspired by neurobiology, have found a number of applications in computational psychology and psychiatry. The central role of large-scale neural dynamics as a basis for understanding brain processes is now well recognized.[26,27] The two main goals of this paper are thus to show that memetics may be based on solid theoretical foundations grounded in neurodynamical models, and that learning using simple memory models may help to understand the process of formation of conspiracy theories. Although only simple competitive learning models are used in this paper, it should open the road toward application of more complex neural models that link memetics with neuroscience. Of course, psychological and social factors prepare the neural system for specific action, but, as Sapolsky stressed in his book, ''you can't begin to understand things like aggression, competition, cooperation, and empathy without biology.''[28]

The next section introduces memetics and discusses representation of information in the brain. It includes an attempt to define memes in a similar way to how genes are defined. It is followed by a section on competitive learning models of memory formation. These models are used to illustrate some mechanisms of memory distortions. Final conclusions and remarks about implications of network simulations for the theory of memetics are presented in the final section.

MEMETICS AND INFORMATION IN THE BRAIN

Subjective information

Ultimately all thoughts and beliefs result from neurodynamics. The flow of neural activation through neural systems is determined by many biological factors, including brain connectivity, concentration of neurotransmitters, emotional arousal, priming effects, and brain stem activity. Information is acquired and internalized in the brain through direct observation of patterns in the world, including communication with people and animals, and indirectly through various media, texts, and physical symbols of all sorts. Brains provide material support for mental processes, understanding and remembering symbols, ideas, and stories. Memes are units of information that spread in cultural environments, information granules that prompt activation of patterns in brains molded by particular subculture. Therefore the same information may become a meme in some brains, and may be ignored by other brains.

Understanding is a process that requires association of new information with what has already been learned. New things are learned on the basis of what is already known by the system. This is a general principle behind brain activity: information gain should be measured as a change induced in cognitive systems.[29] Patterns are encoded in memory depending on the context, sequence of events, attention devoted to these patterns, association with known facts, properties of already encoded information, and general mental state during the encoding process. The definition of Shannon information as entropy does not capture the intuitive meaning of the value of information for the cognitive system. The amount of optimal restructuring of the internal model of the environment (optimal in the minimum length description sense[30]) resulting from new observation (i.e., a new meme added to the memeplex) is a good subjective measure of the quantity of meaningful information carried out by this observation. Pragmatic information that captures the subjective meaning of information is based on the difference between algorithmic information before and after observation is made.[29] Itti and Baldi used a similar idea to define the amount of surprise, measured as the relative entropy or Kullback-Leibler (KL) divergence, between the posterior and prior distributions of beliefs in Bayesian models.[31]

Memes as patterns of brain activity

Organisms replicate preserving most of their properties thanks to genes that are copied with great precision during cell divisions. However, the way from genes to phenotypes is long and indirect. Many factors may influence the final development. At some stage, the replication process is facilitated by, or may even require, ''extended phenotype,''[32] specific environment, or constructions such as nests, burrows, or hospitals. In *The*

Selfish Gene,[16] Richard Dawkins introduced the idea of memes using analogy with genes: depending on the cultural environment, some ideas, news, melodies, videos, or behaviors are imitated and replicated. This process has some properties that are analogous with biological evolution: inheritance, mutation, variation, cooperation, and competition. Memetics emerged as a field of study in the 1990s to explore such analogies. Although categorization of various phenomena into discrete units may be criticized, phenotypes of some organisms may also be very diverse, with large numbers of species belonging to one family, such as 10 million arthropods or thousands of nudibranchs that have unique forms. Cultural phenomena may look quite different, but the mechanism of their proliferation may be based on memes that have certain structures. Dawkins sees the discovery by Lorenz of imprinting, a behavior pattern, as akin to an anatomical organ.[32] He quotes the suggestion of his colleague N.K. Humphrey that memes should be ‘‘physically residing in the brain’’, and are not only metaphors. However, attempts to create a scientific theory of memes as neural processes have not been successful.

Memes are hard to objectively characterize or measure. In 1981, C.J. Lumsden and E.O. Wilson wrote a book[33] on co-evolution of genes, minds, and culture, explaining how genes and epigenetic rules determine perceptions and influence cultural evolution, and how they lead to development of specific cognitive functions, including various types of memory, explaining social behavior. They have used the term ‘‘culturgene’’ (in later writings, Wilson himself used ‘‘meme’’) to describe the process of gene-culture translation using a mathematical model similar to models in population genetics. One of the goals was to establish ‘‘causal connections between semiotics and biology.’’ Memes were linked to the nodes of semantic memory. In his next book, *Consilience* (1998), Wilson wrote,[34] ‘‘If the connections can be established empirically, then future discoveries concerning the nodes of semantic memory will correspondingly sharpen the definition of memes. Such an advance will enrich, not replace, semiotics.’’

The concept of a gene has significantly changed in recent years. Genes, once defined as sequences of DNA base pairs that code proteins, are now understood as distributed DNA and RNA templates, with exons on different chromosomes, ‘‘encoding a coherent set of potentially overlapping functional products.’’[35] Precise definition of a gene is difficult because they are structures of partially mutable, highly organized molecular matter living in specific network of complex processes. They exist because a highly specialized environment facilitates their replication. Strong coupling of all elements in this environment makes the concept of a gene rather fuzzy: it is not a simple DNA sequence but a complex pattern in the whole network of processes, active only in certain situations controlled by epigenetic factors. The whole system is responsible for replication of information.

In memetics, information structures that reflect part of mental content based on a network of memes are called memeplexes. They evolve in response to enculturation and exposure to observed patterns. Specific cultural behaviors, learned concepts, word meanings, collocations, or phrases describing ideas may be treated as memes. Some are very rare and difficult to acquire, while others spread quickly with ease. Mental content can be much wider than just the network of memes. Memetics should position itself in respect to the theory of communication, language acquisition, and neural theories of learning.

Consider now a representation of a meme in the brain of an individual. It is a memory pattern recalled frequently, a state of the whole brain network that arises in many contexts. Wilson thought that it is a node in the semantic memory, but it does not have properties that define semantic information that arises from filtering of episodic memory. Semantic information is learned slowly and is remembered for a long time, while memes are quickly acquired and may be soon forgotten. Semantic memories are based on well-established pathways of brain activations, allowing us to understand meaning of words and concepts.[36,37] They provide conceptual framework for general understanding of the world. Conceptual spaces have been introduced by Gärdenfors[38] as a geometric framework for representing information at the conceptual level, bridging symbolic and neural representations. Concepts are characterized in terms of perceptual and abstract qualities that are treated as separate dimensions. Features are defined as subsets along one or more dimensions, and concepts that have many features form convex shapes in conceptual spaces. This approach has been quite successful in cognitive science, and similar ideas have been developed in cognitive linguistics. Fauconnier wrote a book on mental spaces[39] and another book with Turner on conceptual blending.[40] However, such conceptual models ignore the neurobiological basis of memory.

Concept learning in real brains is a result of complex neurodynamics and changes of neuronal pathways due to neuroplasticity. Conceptual or mental spaces, although highly influential, are not the best simplification of this process. An alternative has been offered by clusterization of neural activation patterns that may be represented by fuzzy prototypes rather than combination of features.[41] Mental events are not restricted to concepts, they are shadows of neurodynamics, metaphorically speaking.[9] Episodic memories are learned quickly, invoke associations that induce chains of mutations, and depend on cultural environment and social interactions. Semantic memory is largely restricted to concepts, states that are deeply entrenched in the brain networks, and have associated phonological representation. Episodic memory recalls brain state at the time of actual experience, and involves imagery, emotions, and behavior, all aspects of experience. Events are memorized without the need for repetition, especially in the case of emotional arousal that increases neuroplasticity. Episodic memories are mental events linked together in experiential spaces. Recently, a new model of spatial and non-spatial memory spaces based on topological schemas of representations of events derived from neuronal spiking activity has been formulated.[42] Such models may bridge the conceptual and neural levels of brain processes.

Memes may be considered at several levels: as abstract units of cultural information, that exist physically in electronic or printed media, called e-memes by McNamara.[18] Information becomes a meme only if some brains are ready to store it as i-memes, and transmit it further. While almost all work in memetics has been focused on e-memes, this paper is an attempt to define and understand formation of i-memes.

Using the language of neurodynamics a meme is defined as a quasi-stable associative memory attractor state, with robust attractor basin. Brain activation $A(w)$ prompted by stimulus w

(a word, set of words, seeing a symbol) may rapidly evoke activation corresponding to meme $A(w) \rightarrow A(M(w))$. The same attractor state may be activated by many different stimuli, including purely internal activations. For simple visual percepts, such as shapes of objects, similarity between brain activations $A(M)$ in the inferotemporal cortical area have been directly compared, using fMRI neuroimaging, with the similarity of the shapes of these objects.[43] Significant similarity has also been found in the fMRI patterns of whole-brain activity when people perceive and think about specific objects,[5,6,44] showing how meaning of concepts is encoded in distributed activity of the brain. Such encoding may be used for brain-based vector representation of the semantic meaning in natural language processing (NLP) algorithms.[45] Similarity between memes corresponding to perceived objects $M_i \Leftrightarrow O_i$, may be roughly compared with some measures of similarity between object properties. Therefore, similarity between brain activities $A(M_1)$ and $A(M_2)$ that represent two memes M_1 and M_2 evoked by objects O_1,O_2 (percepts, cues, words) should be directly related to some measures of object similarity:

$$\mathrm{S}_a(A(M_1),A(M_2)) \sim \mathrm{S}_\mathrm{o}(O_1,O_2). \qquad \text{(Equation 1)}$$

McNamara[18] hopes to detect the signature patterns of new memes by analyzing the neurodynamics of learning novel name-action associations for abstract category names, looking at the changes of the brain connectivity profiles. This may be a useful strategy for abstract categories, or for simple percepts, but general search for signatures of memes using neuroimaging techniques will be very difficult. Activation patterns may significantly differ for individual people, depending on their memeplexes. For the same person, distribution of fMRI activations may change at different times of the day. Transcranial magnetic stimulation (TMS) disrupting the function of the left inferior frontal gyrus has already been used to alter belief formation in favor of remembering more bad news.[46] Such brain stimulation may be used to change acceptance of memes that would normally be ignored.

Memes are difficult to extract from the whole network of brain activities. They exist as transient patterns in neurodynamics. Memory patterns arise due to the functional connectivity of neurons. In this dynamic process, brain regions that may be physically connected in a direct or an indirect way exchange information forming synchronized global states. Connectomics is still a new field, developing methods to describe details of structural and functional connectivity, and network neuroscience is using this knowledge to create dynamical models of cognitive and affective processes.[47] Structural brain connectivity is formed by genetics and developmental processes, and, thanks to neuroplasticity, shaped by life experiences, learning processes, social interactions, and culture.

Understanding how brain connectivity and other factors encode beliefs, filter incoming information, distort it, and transmits it further is certainly a grand challenge. Complex information processing in the brain has not yet been understood in sufficient detail to allow for development of comprehensive theories of such processes. Techniques based on fMRI do not offer sufficient temporal resolution, while electroencephalography and related techniques do not offer spatial resolution to follow precisely dynamical changes during mental processes. However, some insights based on simple memory models may be gained. New information added to the memeplex (existing pool of interacting memes, or attractor states) becomes distorted, changes the memeplex, and is replicated further. Once a set of distorted memory states is entrenched, it becomes a powerful force, attracting and distorting all information that has some association with these states, creating even broader basins of attractors. Encoding of information in this way enhances the memeplex and is one of the reasons why conspiracy theories are so persistent.

Concepts in brains and in computers

In the NLP field, word meaning is approximated using correlations between co-occurrence with several adjacent words. Vectors storing these correlation coefficients $\mathrm{C}(w)$ represent words w by averaging over many contexts restricted to a specific meaning of a given word (this requires annotation of large text corpora). From the human point of view, faithful representation of word meaning should require similar ordering of distances $\mathrm{D}(\mathrm{C}(w_1),\mathrm{C}(w_2))$ between vectors $\mathrm{C}(w_1)$,$\mathrm{C}(w_2)$ representing words w_1, w_2, as shown by dissimilarities $\mathrm{DS_a}(A(w_1), A(w_2))$ between brain activations $A(w)$ when concepts associated with these words are invoked:

$$\mathrm{DS}_a(A(w_1),A(w_2)) \sim \mathrm{D}(\mathrm{C}(w_1),\mathrm{C}(w_2)) \qquad \text{(Equation 2)}$$

Each vector $\mathrm{C}(w)$ attempts to approximate the meaning of the word that is encoded in the distribution of brain activity.[44,45] Without priming effects[48] and association of words with existing memory patterns, only a very coarse representation is possible. Brain activations strongly depend on context, and therefore the distance function $\mathrm{D}(\mathrm{C}(w_1),\mathrm{C}(w_2);\ \mathrm{cont})$ should be context dependent. The whole process is dynamic, with spreading of neural activations responsible for priming related concepts and providing feedback that becomes part of the new pattern encoding. Meaning is thus connected to the activation of many subnetworks in the brain, memory of sensory qualities, and motor affordances. A dynamical approach to the NLP vector model has not yet been fully developed, although some steps in this direction have been made.[36] Despite our efforts (Duch, unpublished) to describe dog breeds in terms of skin, head, and body features derived from databases and semi-structured texts describing dogs, it was not possible to categorize accurately dog breeds only by their features. Using images (or just silhouettes) of dogs leads to more accurate and faster identification of dog breeds. Brain activity evoked by hearing or reading words evokes internal imagery at a high level of invariant, multimodal object recognition. Similarity functions between objects $\mathrm{S_o}(O_1,O_2)$ based only on correlations between verbal descriptors cannot do justice to estimations of similarity of brain activations. Finer discrimination may require recall of lower-level sensory qualities, referring to particular shapes, colors, movements, voice timbre, or tastes. Vector representation based on word correlations does not reflect essential properties of the perception-action-naming activity of the brain,[49] and it does not even contain structural description in terms of object features or phonology. More details on word representation in the brain and its relation to the vector model may be found in Binder et al.[45] Words have phonological representations that serve as labels pointing to internalized knowledge about their meaning. Representation of percepts arising from sensory imagery is a minimal requirement for NLP systems capable of semantic interpretation of concepts.

Competitive learning models are introduced next and then used to illustrate the process of learning that leads to memes based on distorted relations.

COMPETITIVE LEARNING AND WEIRD BELIEFS

Conspiracy theories have serious consequences for politics, especially environmental policies, with the anti-vaccine movement becoming a threat to global health. They facilitate growth of political extremists and dangerous religious sects.[22] Conspiracy theories are investigated mainly by sociologists and psychologists, focusing on hidden networks controlling political and economic factors that are poorly understood. Instead of analyzing why and how brains form weird, distorted views of reality, they invent vague concepts and construct theories that are impossible to connect with brain research. While there are many psychological reasons for formation of such beliefs, so far there have been no attempts to create a cognitive theory supported by computational models, capable of generating testable hypotheses. In the past, secret societies were rather rare, but now media try to stir controversy discussing genetically modified organisms, vaccines, AIDS, miracle cures, unidentified flying objects, prophecies, assassinations, airplane crushes, and other such issues, despite plausible explanations based on scientific arguments or on common-sense consensus.

The language of memetics is descriptive and does not help to explain deeper reasons why some information become memes and others are forgotten.[16,50] Conspiracy theory may be treated as a memeplex that is easily activated by various pieces of information, giving it meaning consistent with the memeplex responses. From a neurobiological perspective, learning requires adaptation, changing functional connectivity, and adjusting the physical structure of the brain. Learning is thus energy consuming, and requires effort that should be carried out only when there are potential benefits. Simple explanations of complex phenomena thus have a great advantage even when they are quite naive, as long as they do not lead to behaviors that are obviously harmful or significantly decrease chances for reproduction. Evolutionary Darwinian adaptations are established only after several generations and have noticeable influence on human beliefs only if they affect large subpopulations. Evolutionary factors explain slow changes in approaches to human freedom, caste and racial divisions, abandonment of slavery, attitudes toward children (selling children into slavery continued until the nineteenth century), etc. Why do some people easily fall for conspiracy theories and other stay skeptical? The field of neural networks, aiming at achieving perfection in learning, paid little attention to distortions of learning and its effects on memory states.

There are many scenarios that may lead to formation of distorted views of observations, and it is not possible to create a neural model that takes into account all factors identified in the literature on this topic. Slow and steady environmental pressures lead to changes of attitude and may redefine the whole world view. Here I will focus on a rather common situation that arises as the result of rapid decrease in neuroplasticity. Emotional arousal coming from the uncertainty of important information (e.g., rumors that something potentially life threatening has happened) leads to confusion and strong anxiety (the rumors may not be true; it is not clear what has really happened). High emotions and stress are linked to release of large amounts of neurotransmitters and neuromodulators from the brain stem nuclei, through the ascending pathways, activating serotonin, norepinephrine, acetylcholine, and dopamine systems. Strong arousal increases brain plasticity, facilitating rapid learning of all potentially relevant cues.[51] Emotionally salient stimuli evoke selective attention, adding more brain states that are closely linked to those that have initially been created. Such states arise when input signals partially overlap, and they share some properties, either related to perception or associations recalled from memory. In attractor networks, similar states share a subset of active neurons. In the visualizations below, each brain state is represented by a small circle, and similarity of brain patterns is represented by distance between the circles that represent them.

Information that arouses emotions and strong neuroplasticity leads to rapid learning. Priming effects[48] direct attention to search for more information on the same topic, sharing some features (activating similar brain regions) with the initial information. Dynamical systems perspective on behavioral priming in attractor networks has been presented in Krpan.[52] After some time, emotions subside, arousal will lessen, sources of neurotransmitters will be depleted, and neuroplasticity will decrease. Thus, the recipe to create a memeplex based on distorted beliefs is, first, priming by uncertain information and strong emotional arousal, followed by selective memorization of information that matches initial impression, and decrease of neuroplasticity that may result from information overload. A short period of acute stress may potentiate learning, but, when it lasts longer, neuroplasticity decreases. The brain network is left with a memeplex based on selected memories frozen in its associative memory. All future information related to the initial event will be associated and interpreted in view of what has been memorized at that period, setting foundations for conspiracy theory.

This scenario may be reproduced in many unsupervised competitive learning neural models,[25] including adaptive resonance theory (ART) models that regulate neuroplasticity using the vigilance parameter.[53,54] Many other competitive learning models based on Hebbian learning have been presented.[55] The DemoGNG 2.2 Java package, written by Bernd Fritzke and Hartmut S. Loos,[56] implements winner-take-all learning in self-organizing map (SOM), competitive Hebbian and hard competitive learning, neural gas, growing neural gas, growing grid, and other algorithms.[57] In all these algorithms, activity of units representing neurons is compared with the input, and those units with the best match adapt their parameters, increasing their activation. Neurons in the neighborhood of a winner are also allowed to adapt, depending on their distance from the winner. If there is no clear match, constructive algorithms add new neurons, allowing the network to grow.

The rapid freezing of high neuroplasticity (RFHN) model described here is based on the following assumptions:

- Emotions and uncertain stressful situations at the beginning of learning lead to high neuroplasticity.
- High neuroplasticity is imitated in the model by large learning rates (due to the primary neurotransmitters), and by a broad neighborhood of the winner neuron for each input pattern (due to the diffuse neuromodulation and volume learning).
- The network tries to reflect associations between input vectors, adapting neuron parameters (usually codebook vectors) to approximate distribution of information contained in the presented input vectors.
- Sudden decrease of the uncertainty and emotional arousal is mirrored by the decrease of learning rates and neighborhood sizes, leading to distortions of complex relations between input items.
- Slow forgetting that follows rapid freezing is based on memory reactivations, and contributes to the retention of memory states represented by the highest number of neurons only, forming clusters of nodes with large and strong basins of attraction that link many states.
- Clusters of neurons that are frequently activated and thus easily replicated represent memes.
- Conspiracy theories are characterized by memplexes, numerous strong memes, with many neurons encoding information that has never been presented, forming distorted associations between facts.

As a result, these networks do not reflect real observations. The role of emotions in susceptibility to fake news has been verified in a recent experiment.[58] The RFHN model may be simulated using several competitive learning models. In fact, all such models show similar behavior; therefore, only the results of SOMs[25] and the neural gas model with competitive Hebbian learning (NG-CHL)[56] are shown below for illustration.

The basic idea of competitive learning is to approximate the activity of neural cell assemblies by neurons (units) that serve as codebook vectors **W**(*t*). They represent receptive fields, adapting to the probability density of the incoming signals. Each neuron receives input signals and competes with other neurons using the winner-takes-most (or takes all) principle, leaving only a small subset of active units that are updated. The winning neural assembly is represented by a vector $\mathbf{W}^{(c)}$ (*t*) and a small group of vectors in its direct neighborhood *O*(*c*). SOM starts with a fixed two-dimensional grid of neurons. Learning proceeds by identifying the most similar codebook vector to the current observation **X**(*t*), and updating the codebook vector and vectors in its immediate physical neighborhood according to the formula:

$$\begin{gathered}\text{For } \forall i \in (0) \\ \mathbf{W}^{(i)}(t+1) = \mathbf{W}^{(i)}(t) + h(r_i, r_c, t)\left[\mathbf{X}(t) - \mathbf{W}^{(i)}(t)\right]\end{gathered} \qquad \text{(Equation 3)}$$

where the neighborhood is usually assumed to be Gaussian:

$$h(r, r_c, t, \varepsilon, \sigma) = \varepsilon(t)\exp\left(-\|r - r_c\|^2 / \sigma^2(t)\right) \qquad \text{(Equation 4)}$$

The size of this neighborhood is decreased from the initial value of dispersion σ_i to the final value σ_f according to the formula:

$$\sigma(t) = \sigma_i\left(\frac{\sigma_f}{\sigma_i}\right)^{t/t_{\max}} \qquad \text{(Equation 5)}$$

The maximal age t_{max} determines the annealing schedule. The learning rate is similarly decreased by:

$$\varepsilon(t) = \varepsilon_i\left(\frac{\varepsilon_f}{\varepsilon_i}\right)^{t/t_{\max}} \qquad \text{(Equation 6)}$$

The SOM model has been used with success in many applications; for example, it works quite well, in comparison with other neural models, for explanation of details of orientation and ocular dominance columns in the visual cortex.[59]

The NG-CHL algorithm does not have such fixed initial grid topology as does SOM, and new neurons are recruited for encoding input patterns. At each adaptation step, a connection between the winner and the second-nearest unit is created, if it does not already exist. The newly created or existing selected edges are refreshed receiving age = 0, while the ages of other edges emanating from the winner neurons are increased by 1. The reference age is gradually changed from T_i to T_f according to:

$$T(t) = T_i\left(\frac{T_f}{T_i}\right)^{t/t_{\max}} \qquad \text{(Equation 7)}$$

Edges that are not refreshed for more than *T*(*t*) steps are removed. This simulates the forgetting mechanism.

The following computational experiments have been done to illustrate the RFHN model:

- Training SOM and NG-CHL on stationary data concentrated in two distinct areas, with initial high plasticity and rapidly decreasing learning rates.
- Training SOM and NG-CHL on non-stationary data from observations that move and suddenly change, with initial high plasticity and rapidly decreasing learning rates.
- Retraining the model after malformation of relations has already occurred, using temporally increased plasticity.

The number of neurons in the brain is extremely large, so it is instructive to check how the number of network nodes in simulations will affect distributions. For the stationary experiments, 10,000 nodes have been used, with initial parameters randomly distributed, and signals coming from two separated circular areas. This should represent two alternative situations that are monitored. For the non-stationary situation, all parameters were initially concentrated in the rectangular patch, simulating situations in which restricted domain has already been learned and is stable. Then the patch moves across the whole domain, providing new input patterns (observations) from the areas it covers. When the edge of the domain is reached, the patch jumps to the other side.

CONSPIRACIES AND MEMORY DISTORTIONS

The algorithms used here are stochastic, so results may differ after each run. This is actually desired, because exposing a

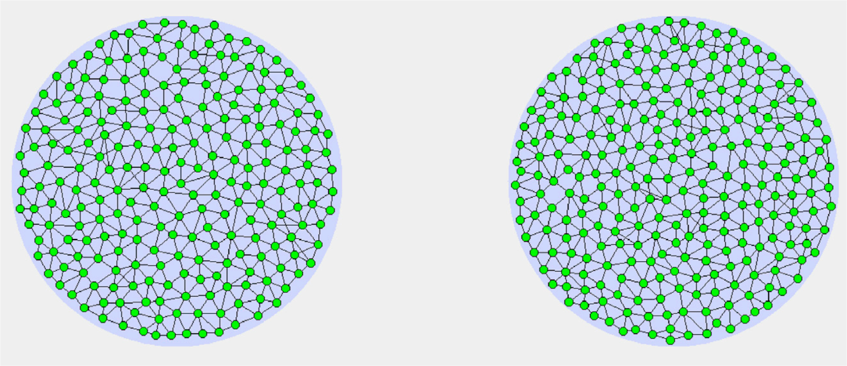

Figure 1. SOM network learning slowly stationary uniform samples drawn from double circles approximates these two circular distributions correctly

group of people with similar memplexes will also lead to different changes. At this level of modeling, only qualitative results may be expected. Each fragment of information (signal, or chunk) is represented by a dot in figures below. Associations between these fragments determine mutual distance in graphs, reflecting similarity of encoded information chunks in the neural model. Chunks of information that appear in the same context (or batch of signals provided as input) become strongly associated.

Stationary situation

Perfect representation of all signals should cover two distinct circular areas (Figure 1). A good solution that requires slow learning with 500,000 steps is shown below. The domain and relations (represented by edges) of input patterns are represented fairly well.

Training 100 × 100 SOM network, with initial $\sigma_i = 5$, $\sigma_f = 0.01$, $\varepsilon_i = 1$, $\varepsilon_f = 0.001$, for 10,000 steps, did not pull all parameters of neurons toward the data area. Despite high density of neurons, some gaps have been left and were not removed by further retraining. This effect comes from the dynamics of learning with shrinking neighborhoods. There is a greater chance for neurons near the edge to be pulled toward high-density areas by many neurons that are selected as winners than to be pulled toward the data in the gap area. Moreover, in the space where no samples ever appeared, many neurons are placed, and this will lead to false associations and confabulations (Figure 2). These effects are random due to the stochastic nature of learning. The resulting map has the same character, although details differ every time it is simulated.

The NG-CHL model with initial high plasticity and rapidly decreasing learning rates has also produced big gaps and high-density areas, as seen in Figure 3. Forgetting parameters have been set to $edge_i = 20$ and $edge_f = 200$. Further retraining with fast forgetting creates even bigger gaps. Many input patterns are therefore associated with high-density clusters acting as memes. Associations with other input patterns are based more on stereotypes (clusters) rather than faithful observations.

Non-stationary situations

Information that reaches us through media or social networks is fragmented. If it is interesting or emotionally exciting, more sources are searched for. Learning in non-stationary situations is much more difficult and therefore distortions in representation are much stronger. In the figures below, a dark rectangle moves randomly across the whole area and the training data that should be learned appear only inside its area.

Using the same parameters as for the stationary case, SOM started with high plasticity that was rapidly decreased in 10,000 steps. The map in Figure 4 shows very strong concentration of nodes that point to the initial patterns. The network did not learn much during the later part of the training. It has ignored most of the facts coming after the rapid learning period, creating one big sink for all associations. Such a network will interpret

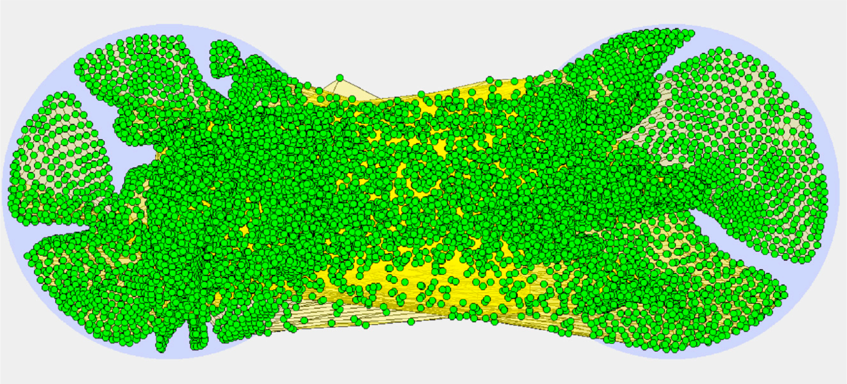

Figure 2. SOM network learning the same distribution as in Figure 1, with fast decrease of plasticity, covers areas where no samples appeared and leaves large gaps in the data space

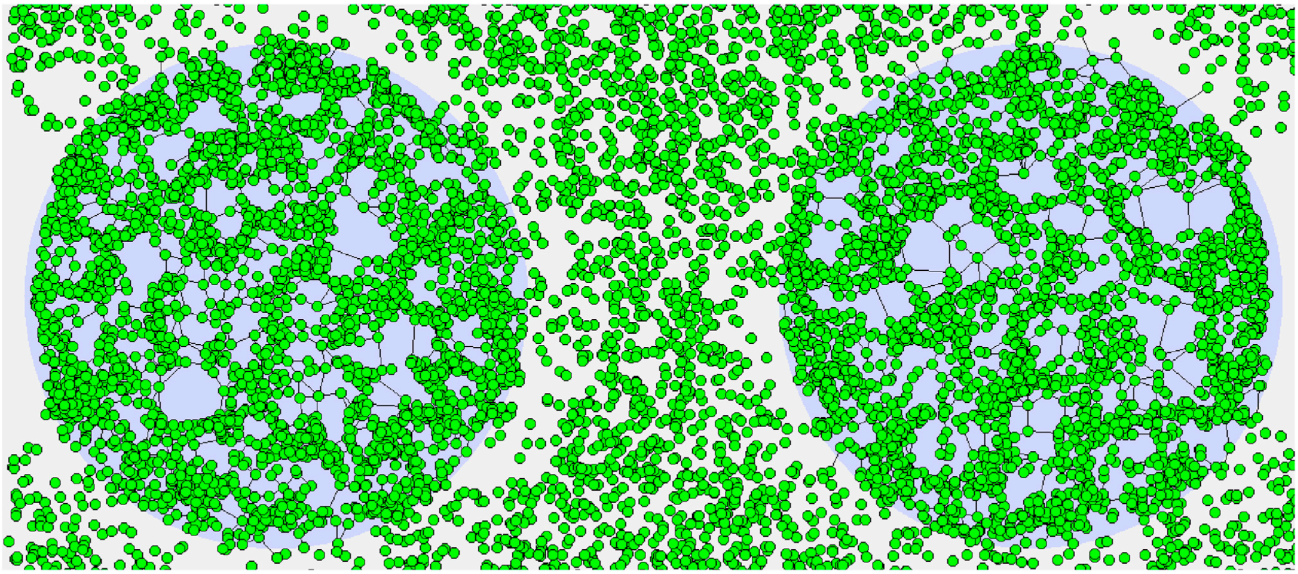

Figure 3. The neural gas model with fast decrease of plasticity creates even stronger distortion of original distribution than the SOM map in Figure 2, leaving many gaps and covering empty space densely

most input data as similar to what it has seen in the critical period of high plasticity.

Further training with increased plasticity may somehow repair the distorted view, although, even after a very long training (Figure 5), a strong meme that has been formed in the center is still present. A large basin of attraction for this meme will lead to its frequent activation even by irrelevant input patterns. After additional 100,000 steps with slow annealing of the central sink may disentangle to some degree, providing a distorted, but more diversified, map.

The NG-CHL algorithm may also create completely distorted representation. After 40,000 steps with rapid decrease of plasticity, it has created two separate memplexes, each with several strong memes that are used to interpret all incoming patterns (Figure 6).

Maps created with rapid decrease of high plasticity are quite unstable. In Figure 7, another solution is shown with four larger memeplexes that completely distort the view of the input patterns. It is quite difficult to create faithful representations of input patterns for non-stationary signals. Very long training times with several hundred-thousand iterations are needed to achieve this. Although central beliefs may be similar within a group of conspiracy believers, a number of subgroups may emerge.

In rapidly changing situations, it is much more likely that a distorted view will be learned instead of a faithful representation of reality. Gaining experience in changing environments obviously takes more time, as can be observed in many domains such as medicine, where initial background knowledge is slowly structured into high competence by the working environment.

CONCLUSIONS

Belief formation may be investigated at the biological and psychological levels. Predispositions for accepting distorted views of reality may come as a side effect of education and life experiences and therefore are rather hard to investigate. Accepting simple explanations is rewarding and creates pleasant feelings of understanding. Complex explanations require a lot of effort and a long time to understand them fully. A simple (although inadequate) explanation is always better than to have no explanation at all, saving energy required for learning and creating a (false) impression of reducing uncertainty. Many papers have been written on this subject from a psychological perspective.[15] The European Union supports European Cooperation in Science and Technology (COST) networking action on Comparative Analysis of Conspiracy Theories (COMPACT), which gathers researchers in history, sociology, psychology, and political sciences interested in conspiracy theories.[60]

From a biological perspective, beliefs have been defined as “the neural product of perception of objects and events in the external world”[14] or “the neuropsychic product of fundamental brain processes that attribute affective meaning to concrete objects and events and of an affirmative internal affective state reflecting personal meaning,”[10] but what are these fundamental brain processes, and why do people believe in conspiracy theories? Because mechanisms of memory formation in the brain work the way they do. Neurodynamics helps to understand the conditions under which large basins of attraction, called memes, are created in memory networks, and how and why they form memplexes that lead to the distorted associations. This is an important step toward linking memetics with theoretical and experimental brain science. Perhaps patterns of brain signals corresponding to memes can be measured,[18] and computer simulations should help to define most suitable experimental conditions. With the advent of highly detailed brain simulations and neuroimaging techniques, we should be able to understand precisely the mechanism behind false memory formation.

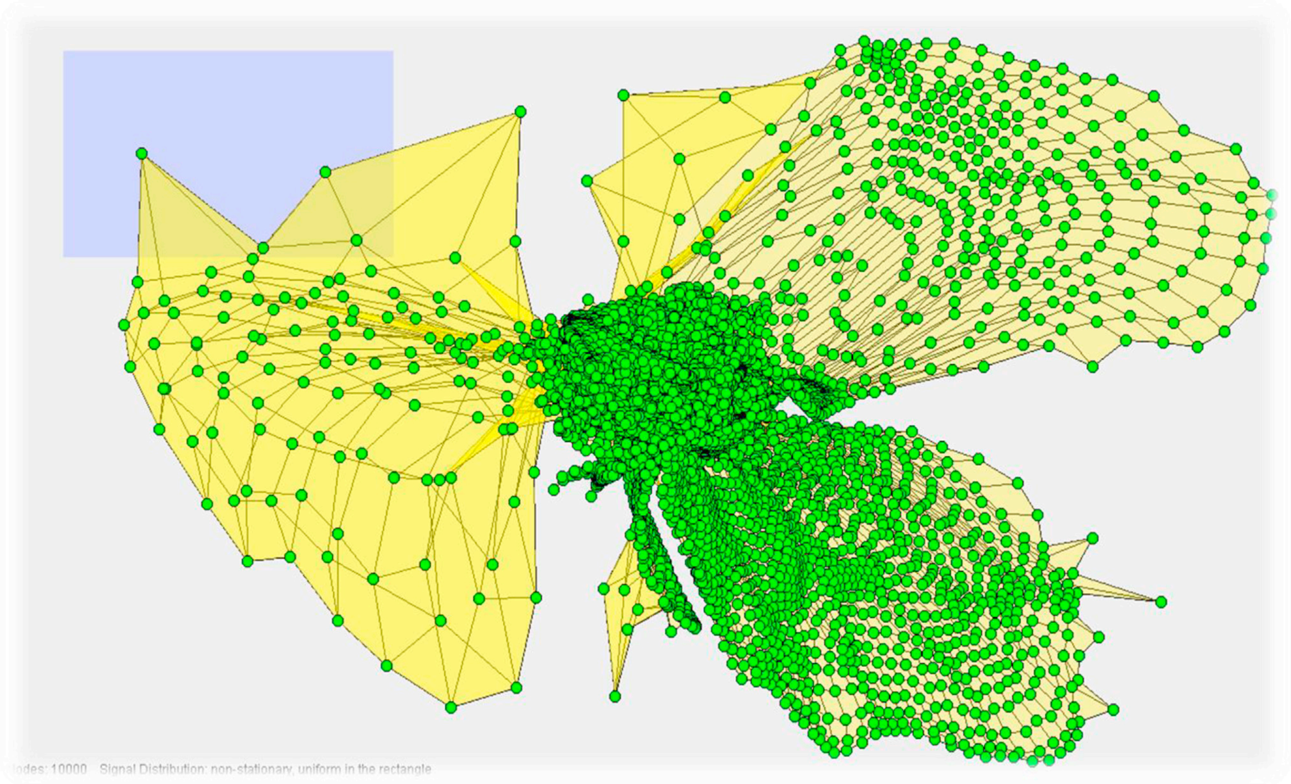
lodes: 10000 Signal Distribution: non-stationary, uniform in the rectangle

Figure 4. SOM with rapidly decreasing plasticity for non-stationary distribution
Samples come here from a moving square (seen in the left corner) and with very slow learning are uniformly distributed in the whole rectangle, but fast learning leads to completely distorted associations.

However, it should be possible to repeat the experiments on artificial distributions with maps based on texts in some restricted domain. Each network node will represent than a word, and distances between words will be based on their similarity in a given context. Such models should allow for semi-realistic analysis of formation of distorted world views.

What lessons can we draw from computational experiments with competitive learning? The RFHN model presented here is very

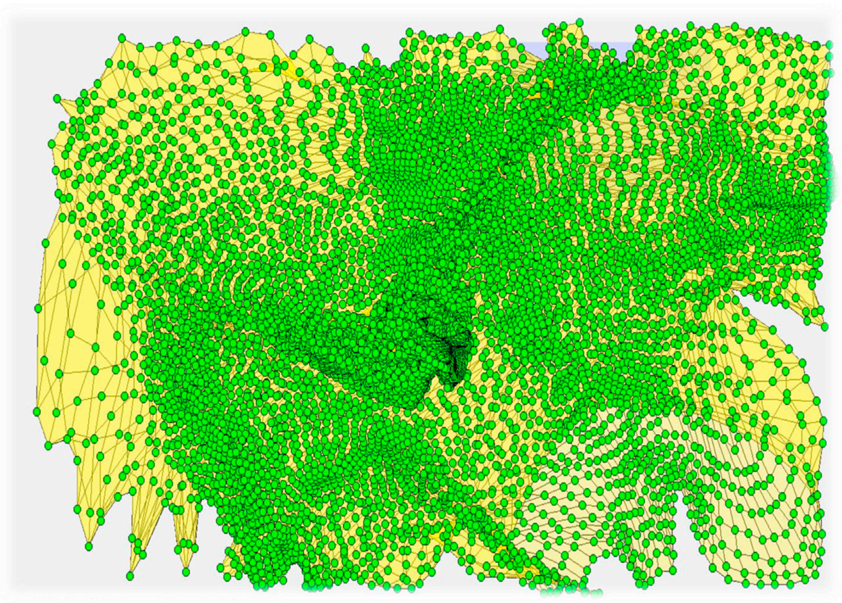

Figure 5. Non-stationary case, neural gas map as in Figure 4, followed by long, slow training (100,000 steps) only partially recovers uniform distribution, leaving large concentration of the codebook vectors in the middle

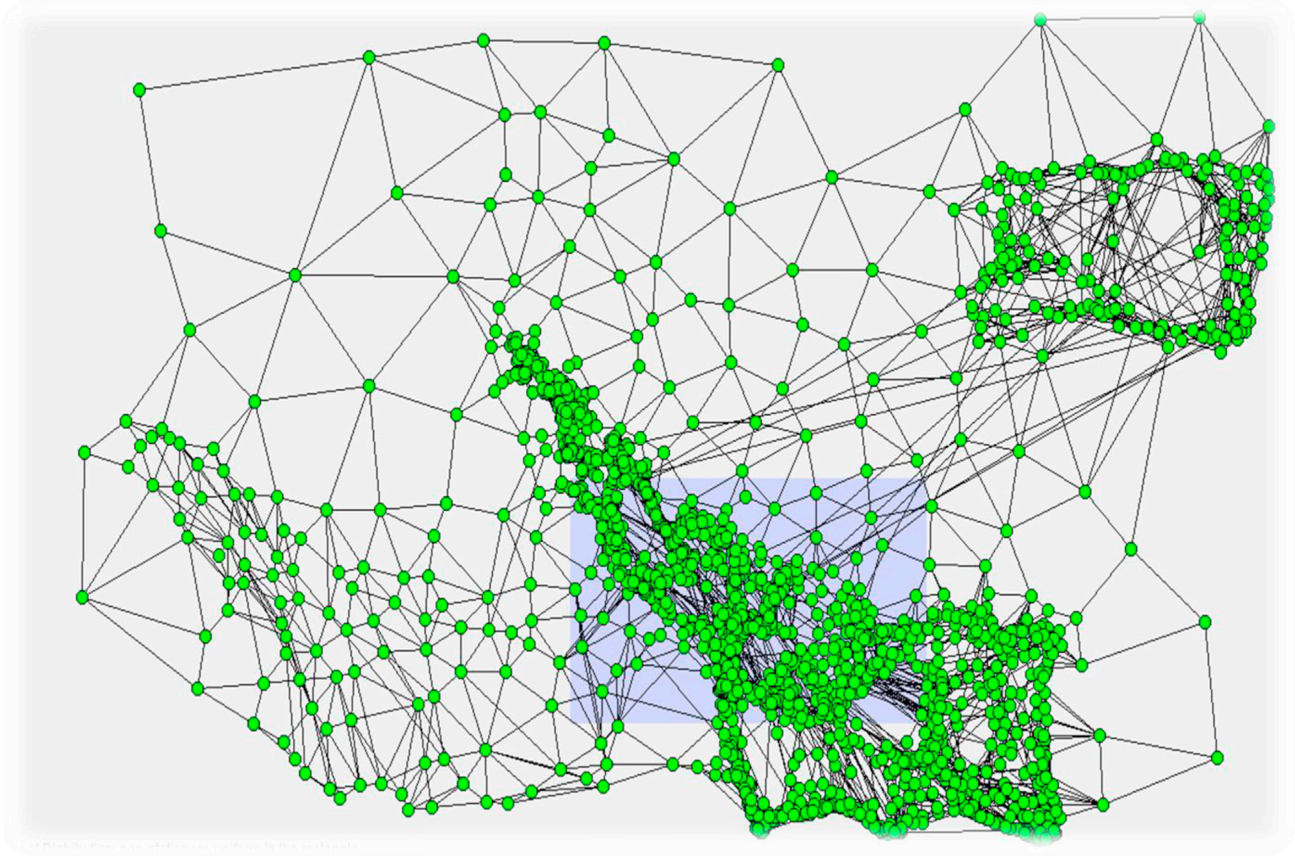

Figure 6. In the non-stationary case, neural gas created two densely connected structures and did not encode signals from many areas

simple, but it seems that all types of competitive learning models show similar behavior. More complex models with high-dimensional input patterns almost certainly will have even bigger problems with faithful representation of input patterns using the rapid freezing of neuroplasticity scenario, and will lead to large attractor basins that can be interpreted as memes. Slow learning leads to faithful representations, but, if the information is false (for example, frequently repeated in media), it may also end in conspiracy theory.

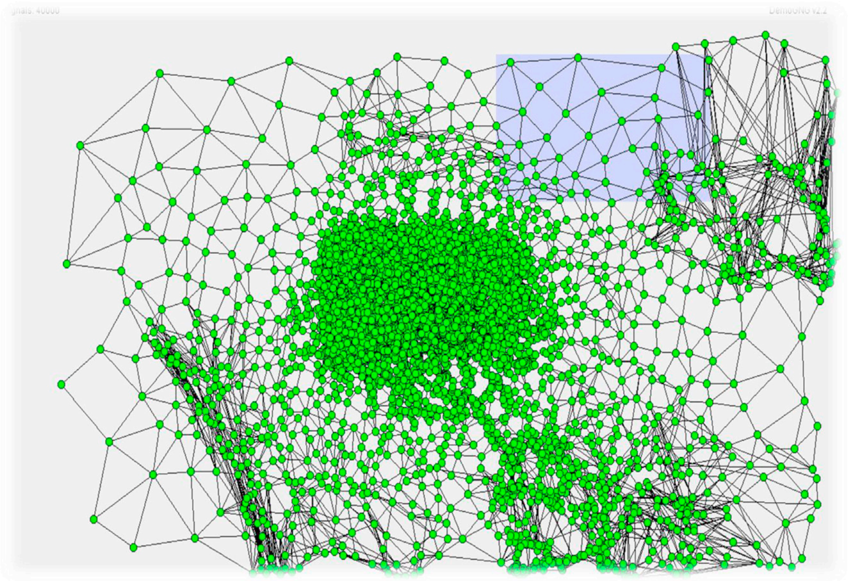

Figure 7. Another neural gas map for non-stationary case, showing how unstable such learning may be

A lie repeated 10,000 times becomes truth, as in the famous Big Lie propaganda technique. On the other hand, after formation of memplexes, slow long learning may lead to some improvement of the veracity of information represented, especially if neuroplasticity is enhanced by emotional arousal.

Although factors that contribute to the individual mental state and influence formation of memories are very diverse, people that subscribe to specific subcultures share many common beliefs and contribute to replication of specific memes. In such subcultures, memes, units of cultural transmission, may become viral because they complement already existing episodic memories, extending memplexes that are common in such populations, adding new, easily excitable elements strongly associated with already memorized memes. Creation of such realistic models is a big challenge.

The contributions of this paper are 2-fold. First, memetics theory has been developed in social sciences but a link to neuroscience has been missing. Linking memes to attractors of neurodynamics should help to give memetics solid foundations. Second, analysis of formation of weird beliefs is very important, but so far there have been no models of brain processes that could explain the creation of such beliefs. Simulations presented here should draw attention to the need for analysis of the type of distortions that are common in neural networks. Of course, more complex neural models will be needed to allow for predictions that could be compared with the results of neuroimaging and behavioral experiments, but even such coarse models based on competitive learning networks may serve as an illustration of putative processes responsible for formation of various conspiracy theories. Our next step is to perform such simulations on real data from the newspapers. Other computational models, such as ART[53] and associative self-organizing network (ASON), that have been used to explain emergence of false memories[61] can be used to model memes and formation of conspiracy theories. A lot of information about memory distortions from cognitive, psychiatric, neuropsychological, neurobiological, and sociocultural perspectives is in the book *Memory Distortion*, edited by D. Schacter.[62]

ACKNOWLEDGMENTS

This work started when I worked as Nanyang Visiting Professor in the School of Computer Engineering, Nanyang Technological University, Singapore, and has now been supported by the Polish National Science Center grant UMO-2016/20/W/NZ4/00354.

REFERENCES

1. Seitz, R.J., and Angel, H.-F. (2020). Belief formation – a driving force for brain evolution. Brain Cogn. *140*, 105548.
2. Friston, K.J., Parr, T., and de Vries, B. (2017). The graphical brain: belief propagation and active inference. Netw. Neurosci. *1*, 381–414.
3. Tse, D., Langston, R.F., Kakeyama, M., Bethus, I., Spooner, P.A., Wood, E.R., Witter, M.P., and Morris, R.G.M. (2007). Schemas and memory consolidation. Science *316*, 76–82.
4. Amit, D.J. (1992). Modeling Brain Function: The World of Attractor Neural Networks (Cambridge University Press).
5. 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.
6. Huth, A.G., Nishimoto, S., Vu, A.T., and Gallant, J.L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron *76*, 1210–1224.
7. Hopfield, J.J. (2007). Hopfield network. Scholarpedia *2*, 1977.
8. Dobosz, K., and Duch, W. (2010). Understanding neurodynamical systems via fuzzy symbolic dynamics. Neural Netw. *23*, 487–496.
9. Duch, W. (2019). Mind as a shadow of neurodynamics. Phys. Life Rev. *31*, 28–31.
10. Seitz, R.J., Paloutzian, R.F., and Angel, H.-F. (2018). From believing to belief: a general theoretical model. J. Cogn. Neurosci. *30*, 1254–1264.
11. St Jacques, P.L., Olm, C., and Schacter, D.L. (2013). Neural mechanisms of reactivation induced updating that enhance and distort memory. PNAS *110*, 19671–19678.
12. Loftus, E.F. (2017). Eavesdropping on memory. Annu. Rev. Psychol. *68*, 1–18.
13. Roediger, H.L., and Mcdermott, K.B. (1995). Creating false memories—remembering words not presented in lists. J. Exp. Psychol. Learn *21*, 803–814.
14. Seitz, R.J. (2021). Beliefs: a challenge in neuropsychological disorders. J. Neuropsychol. https://doi.org/10.1111/jnp.12249.
15. Douglas, K.M., Uscinski, J.E., Sutton, R.M., Cichocka, A., Nefes, T., Ang, C.S., and Deravi, F. (2019). Understanding conspiracy theories. Polit. Psychol. *40*, 3–35.
16. Dawkins, R. (1976). The Selfish Gene (Oxford Uni. Press).
17. Gupta, A., and Ong, Y.-S. (2019). Memetic Computation: The Mainspring of Knowledge Transfer in a Data-Driven Optimization Era (Springer International Publishing).
18. McNamara, A. (2011). Can we measure memes? Front. Evol. Neurosci. *3*. https://doi.org/10.3389/fnevo.2011.00001.
19. Heylighen, F., and Chielens, K. (2009). Evolution of culture, memetics. In Encyclopedia of Complexity and Systems Science, B. Meyer, ed. (Springer), pp. 3205–3220.
20. Gilboa, A., and Marlatte, H. (2017). Neurobiology of schemas and schema-mediated memory. Trends Cogn. Sci. *21*, 618–631.
21. van Kesteren, M.T.R., and Meeter, M. (2020). How to optimize knowledge construction in the brain. Npj Sci. Learn. *5*, 1–7.
22. Shermer, M. (2007). Why People Believe Weird Things: Pseudoscience, Superstition, and Other Confusions of Our Time (Souvenir Press).
23. Galli, G. (2014). What makes deeply encoded items memorable? Insights into the levels of processing framework from neuroimaging and neuromodulation. Front. Psychiatry *5*, 61.
24. Wimber, M., Alink, A., Charest, I., Kriegeskorte, N., and Anderson, M.C. (2015). Retrieval induces adaptive forgetting of competing memories via cortical pattern suppression. Nat. Neurosci. *18*, 582–589.
25. Kohonen, T. (2001). Self-Organizing Maps, 3rd, ext (Springer).
26. Menon, V. (2011). Large-scale brain networks and psychopathology: a unifying triple network model. Trends Cogn. Sci. *15*, 483–506.
27. Ruppin, E. (1995). Neural modeling of psychiatric disorders. Network *6*, 635–656.
28. Sapolsky, R.M. (2017). Behave: The Biology of Humans at Our Best and Worst (Penguin Press).
29. Duch, W. (2007). Towards comprehensive foundations of computational intelligence. In Challenges for Computational Intelligence, *63*, W. Duch and J. Mandziuk, eds. (Springer Studies in Computational Intelligence), pp. 261–316.
30. Rissanen, J. (1978). Modeling by shortest data description. Automatica *14*, 465–658.

31. Itti, L., and Baldi, P.F. (2006). Bayesian surprise attracts human attention. . Advances in Neural Information Processing Systems, *vol. 19* (MIT Press), pp. 547–554.

32. Dawkins, R. (1999). The Extended Phenotype: The Long Reach of the Gene (Oxford University Press).

33. Lumsden, C.J., and Wilson, E.O. (1981). Genes, Mind, and Culture: The Coevolutionary Process (Harvard University Press).

34. Wilson, E.O. (1999). Consilience: The Unity of Knowledge (Reprint Edition) (Vintage).

35. Gerstein, M.B., Bruce, C., Rozowsky, J.S., Zheng, D., Du, J., Korbel, J.O., Emanuelsson, O., Zhang, Z.D., Weissman, S., and Snyder, M. (2007). What is a gene, post-ENCODE? History and updated definition. Genome Res. *17*, 669–681.

36. Duch, W., Matykiewicz, P., and Pestian, J. (2008). Neurolinguistic approach to natural language processing with applications to medical text analysis. Neural Netw. *21*, 1500–1510.

37. Lamb, S. (1999). Pathways of the Brain: The Neurocognitive Basis of Language (J. Benjamins Pub. Co.).

38. Gärdenfors, P. (2004). Conceptual Spaces: The Geometry of Thought (MIT Press).

39. Fauconnier, G. (1994). Mental Spaces: Aspects of Meaning Construction in Natural Language (Cambridge University Press).

40. Fauconnier, G., and Turner, M. (2002). The Way We Think: Conceptual Blending and the Mind's Hidden Complexities (Basic Books).

41. Duch, W. (1997). Platonic model of mind as an approximation to neurodynamics. . Brain-like Computing and Intelligent Information Systems, *Chap. 20* (Springer), pp. 491–512.

42. Babichev, A., and Dabaghian, Y.A. (2018). Topological schemas of memory spaces. Front. Comput. Neurosci. *12*, 27.

43. Op de Beeck, H.P., and Baker, C.I. (2010). The neural basis of visual object learning. Trends Cogn. Sci. *14*, 22.

44. Mitchell, T.M., Shinkareva, S.V., Carlson, A., Chang, K.-M., Malave, V.L., Mason, R.A., and Just, M.A. (2008). Predicting human brain activity associated with the meanings of nouns. Science *320*, 1191–1195.

45. Binder, J.R., Conant, L.L., Humphries, C.J., Fernandino, L., Simons, S.B., Aguilar, M., and Desai, R.H. (2016). Toward a brain-based componential semantic representation. Cogn. Neuropsychol. *33*, 130–174.

46. Sharot, T., Kanai, R., Marston, D., Korn, C.W., Rees, G., and Dolan, R.J. (2012). Selectively altering belief formation in the human brain. PNAS *109*, 17058–17062.

47. Sporns, O., and Betzel, R.F. (2016). Modular brain networks. Annu. Rev. Psychol. *67*, 613–640.

48. McNamara, T.P. (2005). Semantic Priming. Perspectives from Memory and Word Recognition (Psychology Press).

49. Pulvermuller, F. (2003). The Neuroscience of Language. On Brain Circuits of Words and Serial Order (Cambridge University Press).

50. Distin, K. (2005). The Selfish Meme (Cambridge University Press).

51. Tyng, C.M., Amin, H.U., Saad, M.N.M., and Malik, A.S. (2017). The influences of emotion on learning and memory. Front. Psychol. *8*, 1454.

52. Krpan, D. (2017). Behavioral priming 2.0: enter a dynamical systems perspective. Front. Psychol. https://doi.org/10.3389/fpsyg.2017.01204.

53. Grossberg, S. (2012). Adaptive resonance theory: how a brain learns to consciously attend, learn, and recognize a changing world. Neural Netw. *37*, 1–47.

54. Grossberg, S. (2021). Conscious Mind, Resonant Brain (Oxford Universisty Press).

55. Xu, R., and Wunsch, D., II (2005). Survey of clustering algorithms. IEEE Trans. Neural Netw. *16*, 645–678.

56. Fritzke, B., and Loos, H.S. (2017). DemoGNG 2.2. http://www.demogng.de/.

57. Martinetz, T.M., and Schulten, K.J. (1994). Topology representing networks. Neural Netw. *7*, 507–522.

58. Martel, C., Pennycook, G., and Rand, D.G. (2020). Reliance on emotion promotes belief in fake news. Cogn. Res. Princ. Implic. *5*, 47.

59. Erwin, E., Obermayer, K., and Schulten, K. (1995). Models of orientation and ocular dominance columns in the visual cortex: a critical comparison. Neural Comput. *7*, 425–468.

60. (1995). COST action: comparative analysis of conspiracy theories (COMPACT). https://www.cost.eu/actions/CA15101/.

61. van Dantzig, S., and Postma, E.O. (2004). A connectionist model of false memories. Proc. 26th Annual Conference of the Cognitive Science Society, 1375–1380.

62. Schacter, D.L., Fischbach, G.D., and Coyle, J.T. (1997). Memory Distortion: How Minds, Brains, and Societies Reconstruct the Past (Harvard University Press).

About the Authors

Wlodzislaw Duch is the head of the Neuroinformatics and Artificial Intelligence group at Nicolaus Copernicus University, Toruń, Poland. He obtained an MSc (1977) in theoretical physics; a PhD in quantum chemistry (1980); a postdoc at USC, Los Angeles (1980–1982); and a DSc in applied math (1987). He worked at the Max-Planck-Institute, Munich, Germany; Nanyang Technological University, Singapore; several places in Japan; and other countries. He served as the President of the European Neural Networks Society, is an International Neural Network Society Fellow, and is a member of the high-level expert group of European Institute of Innovation & Technology (EIT). He has published over 360 peer-reviewed papers, co-authored six, and co-edited 21 books. Search his name for details.