Neural Networks 122 (2020) 338–363

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

A review on neural network models of schizophrenia and autism spectrum disorder

Pablo Lanillos [a,1,*], Daniel Oliva [b,1], Anja Philippsen [c,1], Yuichi Yamashita [d], Yukie Nagai [c], Gordon Cheng [b]

[a] Donders Institute for Brain, Cognition and Behavior, Radboud University, Nijmengen, The Netherlands
[b] Institute for Cognitive Systems, Technical University of Munich, Arcisstraße 21, Munich, Germany
[c] International Research Center for Neurointelligence, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan
[d] Department of Functional Brain Research, National Center of Neurology and Psychiatry, 4-1-1 Ogawa-Higashi, Kodaira, Tokyo, Japan

ARTICLE INFO

Article history:
Received 5 April 2019
Received in revised form 18 September 2019
Accepted 23 October 2019
Available online 13 November 2019

Keywords:
Neural networks
Schizophrenia
Autism spectrum disorder
Computational psychiatry
Predictive coding

ABSTRACT

This survey presents the most relevant neural network models of autism spectrum disorder and schizophrenia, from the first connectionist models to recent deep neural network architectures. We analyzed and compared the most representative symptoms with its neural model counterpart, detailing the alteration introduced in the network that generates each of the symptoms, and identifying their strengths and weaknesses. We additionally cross-compared Bayesian and free-energy approaches, as they are widely applied to model psychiatric disorders and share basic mechanisms with neural networks. Models of schizophrenia mainly focused on hallucinations and delusional thoughts using neural dysconnections or inhibitory imbalance as the predominating alteration. Models of autism rather focused on perceptual difficulties, mainly excessive attention to environment details, implemented as excessive inhibitory connections or increased sensory precision. We found an excessively tight view of the psychopathologies around one specific and simplified effect, usually constrained to the technical idiosyncrasy of the used network architecture. Recent theories and evidence on sensorimotor integration and body perception combined with modern neural network architectures could offer a broader and novel spectrum to approach these psychopathologies. This review emphasizes the power of artificial neural networks for modeling some symptoms of neurological disorders but also calls for further developing of these techniques in the field of computational psychiatry.

© 2019 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

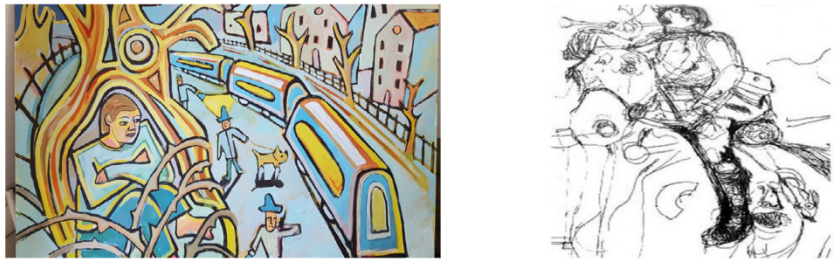
In the world, there is a prevalence of schizophrenia (SZ) that ranges between four and seven per 1000 individuals (between thirty and fifty million people) (Saha, Chant, Welham, & McGrath, 2005) and a prevalence of Autism Spectrum Disorder (ASD) that ranges between six and 16 per 1000 children (between 1 of 150 and 1 of 59 children) (Baio et al., 2018). SZ and ASD have in common that they both cause deficits in social interaction and are characterized by perceptual peculiarities. While ASD has its onset in early childhood, SZ is typically diagnosed in adults, although in very rare cases, appears during development (Rapoport, Chavez, Greenstein, Addington, & Gogtay, 2009). Similar neural bases have been observed for both disorders (Pinkham, Hopfinger, Pelphrey, Piven, & Penn, 2008), which has even led to the suggestion that some SZ cases might be part of the autism spectrum (King & Lord, 2011). In fact, there are similarities such that both pathologies show atypical sensorimotor integration and perceptual interpretation. However, there are also striking differences between these disorders. A common symptom of SZ is the occurrence of hallucinations or delusions, in contrast to ASD which is characterized by atypical non-verbal communication and emotional reciprocity. Furthermore, a few savant syndrome cases were reported in ASD individuals with extraordinary skills like painting (Treffert, 2009). Fig. 1 depicts, in an artistic way, the reality perceived by two individuals in the spectrum of these disorders.

For both disorders, neurological, genetic and environmental factors have been suggested, but to date the actual causes and underlying cognitive processes remain unclear. A major challenge in diagnosis is their heterogeneity and non-specificity. Heterogeneity means that symptoms, prognosis and treatment responses vary significantly between different subjects. Non-specificity expresses that a single biological basis can be underlying different phenotypes (multifinality) and different biological bases can

* Corresponding author.
E-mail addresses: p.lanillos@donders.ru.nl (P. Lanillos), daniel.oliva@tum.de (D. Oliva), anja@ircn.jp (A. Philippsen), yamay@ncnp.go.jp (Y. Yamashita), nagai.yukie@mail.u-tokyo.ac.jp (Y. Nagai), gordon@tum.de (G. Cheng).

[1] Authors contributed equally.

https://doi.org/10.1016/j.neunet.2019.10.014
0893-6080/© 2019 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).



(a) (b)

Fig. 1. Artistic pieces representing different perceptions of the world. (a) Hunted, ©2019 Henry Cockburn, a SZ diagnosed artist. (b) Drawing by Nadia Chomyn at the age of 5, a gifted ASD diagnosed child, reprinted from Selfe (2012), ©2012 Lorna Selfe.

result in a single phenotype (equifinality). Non-specificity, as a biological abnormality related to a psychiatric disorder, can be found in many other neurological disorders (C.-D. G. of the Psychiatric Genomics Consortium et al., 2013; Redish & Gordon, 2016).

Computational modeling of psychopathologies or *Computational Psychiatry* is one of the potential key players (Montague, Dolan, Friston, & Dayan, 2012; Redish & Gordon, 2016; Wang & Krystal, 2014) to tackle heterogeneity and non-specificity, and to better understand the cognitive processes underlying these disorders. Eventually, computational models might help to obtain a deeper understanding of theoretical models, generate new hypothesis or even suggest new treatments. There are different levels of descriptions or units of analysis to study these disorders, which encompass from genes to molecules, to cells, to circuits, to physiology, and then to behavior. *"Computational Psychiatry provides some of the tools to link these levels"* (Adams, Huys, & Roiser, 2016).

In particular, neural network models serve, due to their analogy to biological neurons, as a tool to test and generate hypotheses on possible neurological causes (Huys, Moutoussis, & Williams, 2011). Artificial neural networks cannot only be useful from the data-driven point of view (e.g., fitting a model to fMRI[2] data), but can also be used as a simplified model of the human brain to replicate and predict human behavior and to investigate which modifications in the connectionist models cause a specific alteration in the behavior.

1.1. Artificial neural network modeling of psychopathologies

Artificial Neural Networks (ANNs or NNs) were first introduced in the 1950s as an attempt to provide a computational model of the inner processes of the human brain (Rosenblatt, 1958). Nevertheless, their potential was not fully unraveled until the last decades because of limited computational power and data shortage (Schmidhuber, 2015). Due to the inspiration from biological processes of our brain and their connectionist nature, these technologies have also opened a door to new research fields that combine disciplines, such as neuroscience and psychology with artificial intelligence and robotics. Within the field of cognitive neuroscience, neural networks are already used as a tool for getting insights into the complex structures of our brain and gaining a better understanding of how learning, memory or visual perception might work on a neural level (Crick, Mitchison, et al., 1983; Spitzer, 1995).

In the late 80s and early 90s, neural networks were used for the first time related to psychiatry, trying to imitate psychological disorders (Cohen & Servan-Schreiber, 1992; Hoffman, 1987). Early efforts in compiling ANN models for cognitive disorders can be found in Reggia, Ruppin, and Berndt (1996) and in Gustafsson and Paplinski (2004), in particular, for autism. Due to immense advances in computational power, 20 years later, computational modeling using ANNs and deep learning is becoming a powerful asset to aid the investigation of this type of disorders. The challenge is to translate findings from behavioral or neurological studies at different levels of description in a coherent way into a mathematical connectionist model.

ANN models can process a vast amount of information, cope with non-linearities in the data, and the structure of ANNs makes it possible to systematically test which parameter modifications cause effects similar to the symptoms of psychiatric disorders. Furthermore, these ANN models and their alterations may be directly implemented in artificial agents (e.g., robots) filling the last level: comparing the behavior of such agents with behaviors observed in patients (Cheng et al., 2007; Pfeifer & Bongard, 2006). In this way, existing hypotheses from neuroscience and psychology could be tested, and new hypotheses on potential causes could be formulated.

1.2. Purpose and content overview

This historical review aims at serving as a reference for computational neuroscience, robotics, psychology and psychiatry researchers interested in modeling psychopathologies with neural networks. This work extends general computational modeling reviews (Anticevic, Murray, & Barch, 2015; Gustafsson & Paplinski, 2004; Moustafa, Misiak, & Frydecka, 2017; Reggia et al., 1996; Valton, Romaniuk, Steele, Lawrie, & Seriès, 2017) by focusing on neural network models for SZ and ASD with detailed explanation of the alterations on a neural level and their associated symptoms, including their technical architectures as well as their mathematical formulation. For completeness, we also included Bayesian and predictive processing models due to their similarities to ANNs and their relevance inside the neuroscience community. Actually, conceptually, ANN and Bayesian models often take similar approaches to model psychiatric disorders (see Section 4.3 and Section 5.5).

We start in Section 2 with an introduction to the mentioned disorders, listing their main characteristics and symptoms based on the latest Diagnostic and Statistical Manual of Mental Disorders (DSM-5) descriptions.

For readability and due to the heterogeneity of the reviewed methods, in Section 3, we first summarize and discuss the main modeling approaches and hypotheses which are referenced in the literature. Afterwards, Section 4 and Section 5 present a comprehensive review of models of SZ and ASD, respectively, organized by the type of modeling approach. To help the reader, we summarized the content of Section 4 and Section 5 into two

[2] fMRI: functional magnetic resonance imaging.

tables: Table 1 for SZ and Table 2 for ASD. Finally, in Section 6 we discuss the reviewed works and compile recommendations for future research on ANNs for computational psychiatry, in particular for ASD and SZ.

2. Pathologies and their symptoms

SZ and ASD are disorders that change the way we perceive and act in the world. Atypicalities in perception and in cognitive process cause difficulties in connecting with the world, in particular for social interaction. Since the first reports of autistic symptoms (Kanner et al., 1943), both conditions have been closely related. Before ASD was recognized as a separate disorder, subjects with ASD were often diagnosed as schizophrenic instead (Kanner et al., 1943). Also nowadays, these two pathologies remain strongly connected as both are associated with atypicalities in sensory processing and information processing, and due to their strong heritability (Aukes et al., 2008; Daniels et al., 2008; Sandin et al., 2017).

2.1. Schizophrenia

SZ is a serious psychiatric disorder that affects a person's feelings, social behavior and perception of reality. Its biological causes are still unknown, but genetic and environmental factors, i.e., prenatal stress, traumatic experiences or drug use, can be key factors for the development of this disorder. Its symptoms are usually divided into positive symptoms and negative symptoms (Sims, 1988). Positive symptoms correspond to the presence of abnormal functions, for instance, hallucinations and delusions. Negative symptoms, corresponding to decreased function, are a lack of the normal function such as diminished emotional expression. Positive symptoms are more apparent and generally respond better to medication. Negative symptoms are more subtle and less responsive to pharmacological treatment. Below some of the most characteristic symptoms of SZ taken from the DSM-5 (Association et al., 2013) are listed.
Positive symptoms:

1. *Delusions*: have convinced beliefs that are not real, and cannot be changed despite clear evidence.
2. *Hallucinations*: perceive things that do not exist as real, without an external stimulus.
3. *Disorganized thinking*: difficulty to keep track of thoughts, drift between unrelated ideas during speech.
4. *Disorganized or abnormal movements*: difficulties to perform goal-directed tasks, catatonic (stopping movement in unconventional posture) or stereotyped (repetitive) movements.

Negative symptoms:

1. *Diminished emotional expression*: reduced expression of emotions through speech, facial expressions or movements.
2. *Avolition*: lack of interests, inaction.
3. *Alogia*: diminished speech output.
4. *Anhedonia*: diminished ability to experience pleasure.
5. *Asociality*: lack of interest in social interaction.

Multiple reports have also associated *self-other disturbances* to SZ. This means that schizophrenic patients can perceive own and external actions or feelings, but may have problems differentiating them. This could be part of the explanation for auditory hallucinations and struggles during social interaction. Van der Weiden and colleagues published an extensive review (van der Weiden, Prikken, & van Haren, 2015) on possible causes for this disorder. Finally, in more severe cases, motor disorders have been reported (Morrens, Hulstijn, Lewi, De Hert, & Sabbe, 2006), such as stereotypical and catatonic behavior.

SZ is investigated by many researchers because of its prevalence and its devastating effects on patients, which can have life-changing consequences on the patient's relationships and social situation. Moreover, its close relation with the inner workings of self-perception and self-other distinction, raises the interest of researchers from multiple areas such as psychology, neuroscience, cognitive science and even developmental robotics.

2.2. Autism spectrum disorder

ASD is a prevalent developmental disorder that has a behavior-based diagnosis due to its still unclear biological causes. It was first introduced in the 1940s by Kanner et al. (1943), who presented the cases of eleven children "whose condition [differed] so markedly and uniquely from anything reported so far", some of them being previously diagnosed as schizophrenic. Actually, the term *autistic* was originally used for describing symptoms in schizophrenic patients. This kind of disorder mainly affects individual's social interaction, communication, interests and motor abilities. It is often referred to as a heterogeneous group (spectrum) of disorders, as individuals typically show distinct combinations of symptoms with varying severity. Nevertheless, there are some characteristic attributes that are commonly associated with ASD, which we have listed from the DSM-5 (Association et al., 2013).

Deficits in social communication and interaction:

1. *Impairment in socio-emotional reciprocity*: struggle to share common interests and emotions, reduced response or interest in social interaction,
2. *Deficits in non-verbal communication*: problems integrating verbal and nonverbal communication, and using and understanding gestures or facial expressions,
3. *Problems to maintain relationships*: problems or absence of interest in understanding relationships and adjusting behavior.

Abnormal behavior patterns, interests or activities:

1. *Stereotyped movements or behavior*: repetitive motor movements or speech,
2. *Attention to sameness*: adherence to routines, distress because of small changes,
3. *Fixated and restricted interests*: strong attachment to certain objects, activities or topics,
4. *Hyper- or hyporeactivity to sensory input*: indifference to pain, repulsive response to certain sounds or textures, visual fascination.

Deficits in social interaction are often the most obvious symptoms of ASD. Hence, for a long time, ASD was mainly considered as a disorder of *theory of mind*, suggesting that individuals with ASD are characterized by absence or weakening of their ability to reason about the beliefs and mental states of others in social contexts (Baron-Cohen, 1997). Actually, early identification of individuals with ASD has focused on non-verbal communication interaction, mainly observing attention and gaze behaviors using standardized tests, such as the Autism Diagnostic Observation Schedule (ADOS) (Lord et al., 2012). Whereas this explanation could account for a vast amount of symptoms that become obvious in development and socialization of children with ASD, it was mainly criticized due to its failure to explain similarly prominent non-social symptoms such as restricted interests, desire for sameness or excellent performance in specific areas.

An alternative was suggested in the 90s with the *weak central coherence* theory (Frith & Happé, 1994; Happé & Frith, 2006). It sees the underlying causes of ASD in the perceptual domain, namely in difficulties to integrate low-level information with higher-level constructs. This "inability to integrate pieces of information into coherent wholes (central coherence)", stated in Frith (2003), could offer explanations for the aforementioned deficits and also be extended to an explanation of social deficits. An even broader view is provided by the Bayesian brain hypothesis which suggests general deficits in the processing of predictions and sensory information, and can be applied to non-visual perception as well as motor abilities.

ASD is thought to be caused by genetic disorders and environmental factors and evidence points at high heritability (Sandin et al., 2017). Furthermore, recent studies, using a computer model of the human fetus, have also highlighted the importance of intrauterine embodied interaction on the development of the human brain and in particular cortical representation of body parts (Yamada et al., 2016). Some authors have suggested that preterm infants might have a higher risk of enduring such developmental disorders.

3. Modeling approaches and hypotheses

ASD and SZ are among the psychiatric disorders which are most commonly investigated using computational modeling. A reason might be the unclear underlying cognitive mechanisms of these disorders which computational models might help to unravel. The studies we discuss in this review often take similar approaches for modeling ASD and SZ. In fact, these two disorders share certain symptoms, such as deficits in social communication and motor impairments manifesting as decreased response or repetitive and stereotyped movements. Although, perceptual atypicalities in both disorders are usually differentiated in that SZ involves perceptual experiences that occur without an external stimulus (e.g., hallucinations) whereas ASD is more typically characterized by hypersensitivity to certain stimuli from the environment, there is some overlap. For instance, hypersensitivity can be also found in SZ patients (Robbins, 1993). Furthermore, both disorders present less sensitivity to some visual illusions (Happé, 1996; Notredame, Pins, Deneve, & Jardri, 2014). Despite of all these similarities, it is still under debate how these two disorders relate to each other (Wood, 2017).

In computational modeling, similarities between modeling approaches are not primarily motivated by the similarities in symptoms. In fact, studies modeling SZ focused mainly on delusions and hallucinations which are not predominant in ASD. Similarities, instead, can be found in the suggested biological causes and in the type of altered neural network parameters.

There are three main biological causes that are commonly employed in computational models: neural dysconnections,[3] imbalance of excitation and inhibition, and alterations of the precision of predictions or sensory information.

3.1. Dysconnection hypotheses

Especially for SZ, one of the most discussed theories is the idea of functional disconnections (Friston, 1998; Lynall et al., 2010). The main motivation is that SZ cannot be explained by an impairment of a single brain region, but only by a (decreased) interaction between multiple brain regions (Friston, 1998). Disconnections or underconnectivity are also discussed as a potential cause of ASD (Anderson et al., 2010; Frith, 2004; Just, Cherkassky, Keller, & Minshew, 2004), but more recent evidence also points at increased connectivity (Keown et al., 2013; Supekar et al., 2013) or a distortion of patterns of functional connectivity (Hahamy, Behrmann, & Malach, 2015).

In the discussed studies for SZ, dysconnection is primarily implemented by an increased pruning of synapses (Hoffman & Dobscha, 1989; Hoffman et al., 2011; Hoffman & McGlashan, 1997). Such a pruning is a normal developmental process between adolescence and early adulthood (Huttenlocher et al., 1979). Computational models using Hopfield networks (Hoffman & Dobscha, 1989) or feed-forward networks (Hoffman et al., 2011; Hoffman & McGlashan, 1997) demonstrate that too strong pruning can cause fragmented recall or the recall of new patterns, which can be related to the symptom of hallucinations in SZ.

Notably, the SZ symptoms replicated with connection pruning focus solely on hallucinations or delusions and might not be appropriate for modeling ASD. In fact, in a biological context, it might be more appropriate to disturb connections between neurons instead of simply cutting them. This idea was followed by Yamashita and Tani (2012) who induced noise between different hierarchies of neurons (suggested by Friston & Frith, 1995). They demonstrated in a robotic experiment that this leads to the emergence of inflexible, repetitive motor behavior similar to catatonic symptoms in SZ. This motor behavior could also be present in ASD.

Just a single study focused on dysconnection in ASD. Park and colleagues (Ichinose et al., 2017; Park et al., 2019) showed, using a spiking neural network, that local over-connectivity, especially locally in the prefrontal cortex (Courchesne & Pierce, 2005), can account for the emergence of aberrant frequency patterns of neural connections in patients with ASD.

3.2. Excitation/inhibition imbalance

An excitation/inhibition (E/I) imbalance is among the most commonly referenced biological evidence for SZ as well as for ASD (Canitano & Pallagrosi, 2017; Rubenstein & Merzenich, 2003; Snijders, Milivojevic, & Kemner, 2013; Sun et al., 2012). E/I imbalance was found in many neurobiological studies on SZ and ASD. Although it is not clear how exactly E/I imbalance translates to changes in cognition and behavior (Canitano & Pallagrosi, 2017), it seems to be linked to core symptoms of both disorders such as hallucinations (Jardri et al., 2016) and social interaction deficits (Yizhar et al., 2011).

An unanswered question is also of which quality this imbalance is. A recent review of studies regarding ASD found evidence for increased inhibition as well as for increased excitation (Dickinson, Jones, & Milne, 2016). Conflicting results in various brain regions might arise by differences in measurements and their reliability. The most commonly used mechanisms are magnetic resonance spectroscopy which allows to measure the cortical levels of glutamate or GABA, measurements of gamma-band activity (which is hypothesized to be connected to inhibition) or the analysis of the number of glutamate or GABA receptors in post-mortem studies (Dickinson et al., 2016). Another possible interpretation of these conflicting results is that both, increases and decreases, in inhibition and excitation are present in ASD. This hypothesis was put forward by Nagai, Moriwaki, and Asada (2015), suggesting that both impairments share a common underlying mechanism. Their model could show that increased inhibition and increased excitation can simulate the local or global processing bias of ASD, respectively.

Furthermore, Gustafsson (1997) also connected E/I imbalance to the local processing style of ASD. He implemented increased

[3] Note that *dis*connection usually refers to a lack of connection whereas *dys*connection describes atypical connectivity which might include decreased as well as increased connectivity.

inhibition in a self-organizing map, in particular, stronger inhibition in the surrounding of receptive fields which led to over-discrimination.

For SZ, although E/I imbalance is commonly associated to SZ in the literature, only the approach from Jardri and Deneve (2013) explored E/I imbalance as a modeling mechanism. In their model, a stronger excitation or insufficient inhibition caused circular belief propagation: bottom-up and top-down information are confused with each other which might cause hallucinations and delusions. This model was recently supported by some experimental evidence (Jardri, Duverne, Litvinova, & Denève, 2017).

3.3. Hypo-prior theory and aberrant precision account

The increasing popularity of the Bayesian view on the brain in recent years resulted in a trend of explaining psychiatric disorders as a cause of the failure of correctly integrating perceived low-level sensory information (bottom-up information) with high-level prior expectations (top-down information). These approaches are inspired by diminished susceptibility of subjects with psychiatric disorders to visual illusions (Notredame et al., 2014) and the well-known symptom of hypersensitivity to certain stimuli (e.g., Lucker, 2013).

Problems in the integration of top-down and bottom-up information can be explained by an inadequate estimation of the precision of these signals. A decreased precision of the prior causes a weaker reliance on predictions and, hence, a relatively stronger reliance on sensory input. This so-called hypo-prior theory was first suggested by Pellicano and Burr for ASD in 2012 (Pellicano & Burr, 2012). Similarly, an increased precision of the bottom-up signal can account for the same consequences (Lawson, Rees, & Friston, 2014). Despite some initial evidence in favor of an overrating of sensory information (Karvelis, Seitz, Lawrie, & Seriès, 2018), it cannot be decided to date which of these theories is more compelling than the other. Possibly, both contribute to the observed phenomena.

For both, ASD and SZ, typically a weaker influence of predictions and a higher influence of sensory information is suggested (Karvelis et al., 2018; Lawson et al., 2014; Pellicano & Burr, 2012). Lawson and colleagues substantiated aberrant precision for ASD by basing it on hierarchical predictive coding. They argued that both hypo-priors and increased sensory noise might influence the perception on different levels of the cortical hierarchy, leaving open both hypotheses. In an endeavor to clarify how such theories differ for ASD and SZ, Karvelis et al. (2018) recently investigated how healthy individuals, scored for traits of ASD and SZ, use prior information in a visual motion perception task. ASD traits were associated with increased sensory precision, whereas SZ traits did not correlate.

However, it might be intuitively plausible that also an overrating of top-down information can account for the occurrence of hallucinations (Powers III, Kelley, & Corlett, 2016). In a recent review, Sterzer et al. (2018) noticed that too strong as well as too weak priors explain psychosis. They suggested that the way that priors are processed might differ depending on the sensory modality or the hierarchical level of processing, yielding inconsistent theories and findings.

In line with this idea, computational models for ASD often suggest that an impairment might be present in both extremes (Idei et al., 2017; Philippsen & Nagai, 2018). In Idei et al. (2017), repetitive movement could be replicated by an aberrant estimation of sensory precision, leading to inflexible behavior, either due to sameness of intentional states (increased sensory variance) or due to high error signals and misrecognition (decreased sensory variance). Similarly, Philippsen and Nagai (2018) suggest that too strong as well as too weak reliance on the sensory signal may impair the internal representation of recurrent neural networks. Thus, for SZ as well as for ASD, too strong as well as too weak reliance on priors or sensory information seem to be valid modeling approaches.

3.4. Alternative modeling approaches

There are alternative theories used in the discussed computational models. Synaptic gain, for instance, has been evaluated for SZ (Cohen & Servan-Schreiber, 1992) as well as for ASD (Dovgopoly & Mercado, 2013). In fact, a reduction of synaptic gain might be related to reduced precision of prior beliefs as discussed in Adams (2018).

Less biologically inspired approaches can also be found in the literature and focus more on replicating behavioral data using known engineering techniques in ANN. For instance, deficits in generalization capabilities are modeled in neural networks by modifying the number of neurons (Cohen, 1994), changing the training time (Dovgopoly & Mercado, 2013) or introducing regularization factors (Ahmadi & Tani, 2017; Dovgopoly & Mercado, 2013).

4. ANN models of schizophrenia

In the following section, we present a comprehensive description of the most important ANN models of SZ. The majority of approaches focuses on positive symptoms of SZ, such as hallucinations and delusional behavior, e.g., Hoffman and McGlashan (1997) and Horn and Ruppin (1995). Nevertheless, there have been also approaches targeting other symptoms, for instance attention characteristics (Cohen & Servan-Schreiber, 1992) and movement disorders (Yamashita & Tani, 2012). An overview of the most important models is presented in Table 1.

4.1. Hopfield networks: memory

4.1.1. Memory overload

In 1987, Ralph E. Hoffman, professor of psychiatry from Yale, presented the earliest neural network model of SZ (Hoffman, 1987), inspired by the suggestions of Crick et al. (1983), who explored the function of dreams using a neural network model. Hoffman tried to explain the causes of schizophrenic and maniac disorders with simulations using a Hopfield Network, an associative memory ANN that is usually employed to simulate the inner functioning of human memory (Hopfield, 1982) and to store binary memory patterns. It is a recurrent neural network that converges to fixed-point attractors. As a learning mechanism, the famous Hebbian rule, "cells that fire together wire together", is applied. In other words, connections between neurons that get activated with temporal causality are increased (Hebb, 1949). In order to model SZ, the author inspected the behavior of the network attractors after storing an increasing number of binary memories.

Results showed that by *increasing the number of binary memory patterns stored*, the network reaches "parasitic" states that do not correspond to previously stored memories. With higher numbers of memories or decreased storage capacity, the network's internal energy minima, that correspond to the stored memories, might influence each other and create additional deep minima (attractors) that do not correspond to any previously learned pattern. These minima might influence either only the information processing course (mind being controlled by outside force) or lead to convergence to "parasitic states", which are compared to hallucinations and delusional thoughts. This study did not use biological evidence to support its main thesis that SZ might be caused by memory overload and only compared behavioral observations. However, this model served as a stepping stone for a successor model (see Section 4.1.2).

Table 1
Overview of neural network models of schizophrenia.

Model type	Paper	Disorder characteristic	Biological evidence	Approach
Hopfield networks	Hoffman (1987)	Delusions, sense of mind being controlled by outside force	–	Storing of an excessive number of memories (memory overload)
	Hoffman and Dobscha (1989)	Hallucinations, delusions, sense of mind being controlled by outside force	Reduced connectivity in prefrontal cortex and other regions	Excessive connection pruning
	Horn and Ruppin (1995)	Delusions and hallucinations	Reactive synaptic regeneration in frontal cortex	Weakening of external input projections, increase of internal projections and noise levels, additional Hebbian component
Feed-forward NNs	Cohen and Servan-Schreiber (1992)	Disturbances of attention, representation of context	Abnormal dopamine activity in prefrontal cortex	Reduction of activation function gain in context-neurons
	Hoffman and McGlashan (1997)	Auditory hallucinations	Reduced connectivity in prefrontal cortex and other regions	Excessive connection pruning
	Hoffman et al. (2011)	Delusionary story reconstruction	Abnormal dopamine activity, cortical disconnections	Increased BP learning rates, excessive connection pruning in working memory
Predictive processing	Adams, Stephan, Brown, Frith, and Friston (2013)	Delusions and hallucinations, abnormal smooth pursuit eye movement	Abnormal neuromodulation of superficial pyramidal cells in high hierarchical levels	Abnormal precision computation in the free energy minimization scheme
Circular inference	Jardri and Deneve (2013)	Hallucinations and delusions	Disruption in the neural excitatory to inhibitory balance	Increased excitation/reduced inhibition in belief propagation
Recurrent NNs	Yamashita and Tani (2012)	Disturbance of self, feeling of being controlled by outside force, disorganized movements	Disconnectivities in hierarchical networks of prefrontal and posterior brain regions	Noise between context neuron hierarchies in MTRNN

4.1.2. Memory model with disconnections

Observations that show diminished metabolism in the prefrontal cortex (hypofrontality) of individuals with SZ led to the theory that excessive synaptic pruning might be the reason for the appearance of SZ between adolescence and early adulthood (Feinberg, 1982; Keshavan, Anderson, & Pettergrew, 1994). A decline in synaptic density is a normal developmental process (Huttenlocher, de Courten, Garey, & Van der Loos, 1982; Huttenlocher et al., 1979) which might have gone too far in the case of SZ. In 1989, Hoffman and Dobscha used a Hopfield network, arranged as a 2D grid, as a content-addressable memory to retrieve previously stored memories giving a similar input (Hoffman & Dobscha, 1989). A "neural Darwinism" principle was applied, which is a *pruning rule that erases connections depending on their weights and length (proximity of neurons in the grid)*. The concrete pruning rule is shown in Eq. (1), with $|T_{xy}|$ being the weight of the connection between neurons in coordinates (x, y) and (i, j), and $\hat{p}$ the pruning coefficient. The pruning coefficient determines the number of connections which are discarded. Fig. 2 illustrates a possible scenario for this pruning process.

$$|T_{xy}| = \hat{\boldsymbol{p}} \cdot [(i-x)^2 + (j-y)^2]^{0.5} \tag{1}$$

For a moderate level of pruning, the network is still able to perform the memory-retrieval task, but for connection reductions of 80% the network shows fragmented retrieval. This fragmentation was compared to thought disorders observed in SZ, which lead to incoherence, attention deficits or the feeling that one's mind is being controlled by an outside force. Furthermore, sometimes over-pruned areas converged to patterns not included in any of the stored memories. These were denominated as "parasitic foci". The authors compared these to hallucinations in SZ because they contained decodable information that does not belong to any stored memory. Occasionally, these parasitic regions extended on a larger area and persisted independently of the input, which was compared to delusional thoughts observed in patients.

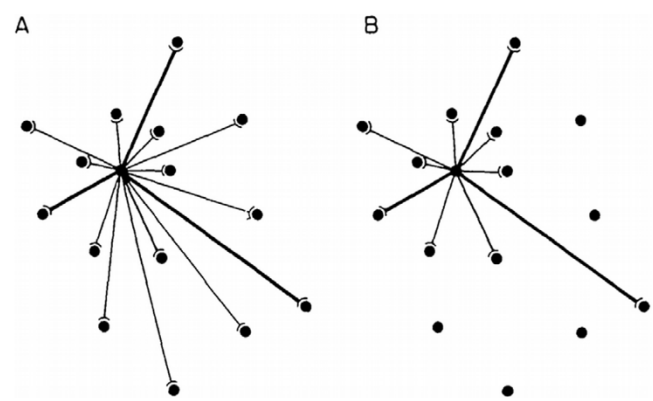


Fig. 2. Pruning rule used for the Hopfield Network in Hoffman and Dobscha (1989). The connections are pruned depending on the connection weight and the distance between the connected neurons. A: Connections before pruning. B: Connections after pruning.
Source: Reprinted from Hoffman and Dobscha (1989).

4.1.3. Memory model hippocampal region

In 1995, Horn and Ruppin (1995) and Ruppin, Reggia, and Horn (1996) also introduced a Hopfield-based network to replicate the positive symptoms of SZ. This model was based on the hypothesis by Stevens (1992) that schizophrenic symptoms might be caused by "reactive anomalous sprouting and synaptic reorganization taking place at the frontal lobes, subsequent to the degeneration of temporal neurons projecting at these areas". The hypothesis takes into account observations that showed atrophic changes in the temporal lobe, and at the same time increased dendritic branching in the frontal lobe of a significant number of schizophrenic patients. Essentially, the idea is that degenerations in temporal lobe regions that are connected to the frontal lobe regions might produce a compensatory reaction in

that area, namely increased receptor bindings (frontal lobe connections) and anomalous dendritic sprouting (increased influence from other cortical areas).

The work by Hoffman explained in the previous section suggested that hallucinations should always appear in combination with memory problems in patients because pruning clearly affects the network's memory retrieval performance. However, this is not always the case in patients. Following the hypothesis from Stevens, the model described in Horn and Ruppin (1995) would make hallucinations and intact memory capabilities compatible.

The model used in this paper was a Hopfield network taken from Tsodyks (1988) and Tsodyks and Feigel'man (1988), which is more appropriate for the storage of correlated patterns. This network is used for a pattern retrieval and recovery task, which means that in its original functionality, it receives an external input pattern and outputs the previously learned pattern that corresponds to it, given that a similar one was learned before.

Defining the connection strength (weight) between neuron i and j as W_{ij}, the learning rule is:

$$W_{ij\,new} = \boldsymbol{c}\; W_{ij\,old}\;,\; (c > 1) \tag{2}$$

$$W_{ij} = \frac{c_0}{N}\sum_{\mu=1}^{M}(\xi_i^{\mu} - p)(\xi_j^{\mu} - p) \tag{3}$$

where $\boldsymbol{c}$ is the internal projection parameter with value always > 1. Eq. (3) describes the initial configuration of the network weights, with $c_0 = 1$, p being the probability that a memory pattern is chosen to be 1, and ξ_i^{μ} one of the $M = \alpha N$ memory patterns.

The input of each neuron i at time step t is expressed as:

$$h_i(t) = \sum_j W_{ij}S_j(t-1) + \boldsymbol{e}\cdot\xi_i^1 \tag{4}$$

where e is the network input parameter with value 1 in normal conditions, which weights the incoming memory pattern, and S_j is the neuron output defined by a sigmoid function with noise level T and a fixed uniform threshold of all N neurons θ:

$$S_i(t) = \begin{cases} 1, & \text{with probability } \frac{1}{1+\exp(-(h_i(t)-\theta)/\boldsymbol{T})} \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

In order to simulate degenerated temporal lobe projections to the frontal lobe, the input is scaled down by decreasing parameter $e < 1$ in Eq. (4). In order to model increased receptor bindings and dendritic sprouting the parameter $\boldsymbol{c}$ in Eq. (3) and noise level $\boldsymbol{T}$ in Eq. (5) are increased. The parameter c scales the internal weights of the network and T influences the neuron activation. After performing these modifications, the network is still able to retrieve previously stored memories, but spontaneously converges to certain memories without a specific input stimulus.

An additional Hebbian learning rule during pattern retrieval on a lower time scale is used to account for increased dopamine levels observed in patients with SZ:

$$W_{ij}(t) = W_{ij}(t-1) + \frac{\gamma}{N}(\bar{S}_i - p)(\bar{S}_j - p) \tag{6}$$

where $\bar{S}_i$ is a variable that only becomes 1 if the neuron in question has been active during the last τ iterations. There are studies that have observed that dopamine activity increases may enhance Hebbian-like activity-dependent synaptic changes in the brain, and a high synaptic modification rate γ is used to replicate this effect, as this parameter influences how much the network's weights are changed during learning. This modification is used to imitate high dopamine levels observed in schizophrenia.

In total, four network modifications were tested on the presented architecture (Fig. 3): (1) *weakening of the network input parameter e*, (2) *increase of internal projections c*, (3) *increase of noise levels T*, and (4) *additional Hebbian learning rule* (Eq. (6)).

Combining the reactive modifications to a decrease of e (internal connections and external noise) with the described Hebbian rule (even with a small γ of 0.0025), the spontaneous retrievals are enhanced and get continuously triggered without a concrete retrieval input. This behavior is compared to long-term hallucinations or delusional beliefs characteristic of schizophrenic patients. This result would also fit with the effect of dopaminergic blocking agents (equivalent to reducing the effect of the Hebbian learning rule), which are used to reduce hallucinations in patients.

4.2. Feed-forward networks: context and language

4.2.1. Attention and context representation

In 1992 the first model based on feed-forward neural networks was introduced. The psychology professor Jonathan D. Cohen and neuroscientist David Servan-Schreiber (Cohen & Servan-Schreiber, 1992) presented an extensive analysis of a possible explanation for negative symptoms in SZ. More concretely, they focused on disturbances of attention and contextualization problems in schizophrenics, which were for instance reported in Garmezy (1977) and Lang and Buss (1965). Their main hypothesis was that schizophrenics fail to make an internal representation of context and that an abnormal amount of dopamine in the prefrontal cortex is the main cause (cf. Section 4.1.3 as a comparison). The authors refer to previous studies suggesting that the prefrontal cortex is the brain region responsible for maintaining an internal representations of context, and that patients with SZ show dysfunctions and abnormal dopamine levels in this area. In order to test the dopamine-theory of SZ, three experimental tasks were compared to three neural network models, obtaining similar results to empirical observations. They simulated reduced dopamine activity by decreasing the gain of the activation function (the activation function's slope), described by Eq. (7), in the neurons responsible for context representations. In this equation, we used the same nomenclature as in the original paper, where *net* is the added activation of all incoming connections, *bias* the neuron bias and *gain* the parameter that is modified. The mentioned idea of modifying the activation function's gain was based on studies that suggest that high dopamine levels potentiate the neurons' activation (inhibitory and excitatory) in the prefrontal cortex. The modification of the gain has a similar effect because higher gain values increase the activation function's slope, which means that even small neuron input values produce either very low neuron activations (equivalent to inhibitory signals) or high activations (equivalent to excitatory signals).

$$f(net) = \frac{1}{1+\exp(\boldsymbol{gain}\cdot net + bias)} \tag{7}$$

The first experiment, depicted in Fig. 4, was the *Stroop* task (Stroop, 1935), which consists of color words printed in different color inks that are presented to the participants. These words have either congruent stimuli (color and word are the same), conflicting stimuli (color and word contradict each other) or control stimuli (color words printed in black ink or the letters "*XXX*" printed in a certain color). The subjects must then either always name the letter's ink color or the written word. This exercise is used to test the participant's attention capacities, and schizophrenic subjects show overall slower reaction times and perform even worse when conflicting stimuli are shown (Henik & Salo, 2004). In order to feed the information in the network, the printed word's ink color and meaning were numerically coded. By reducing the gain on the *color naming* and *word reading* units from 1.0 (normal gain) to 0.6 they observed a delay in the

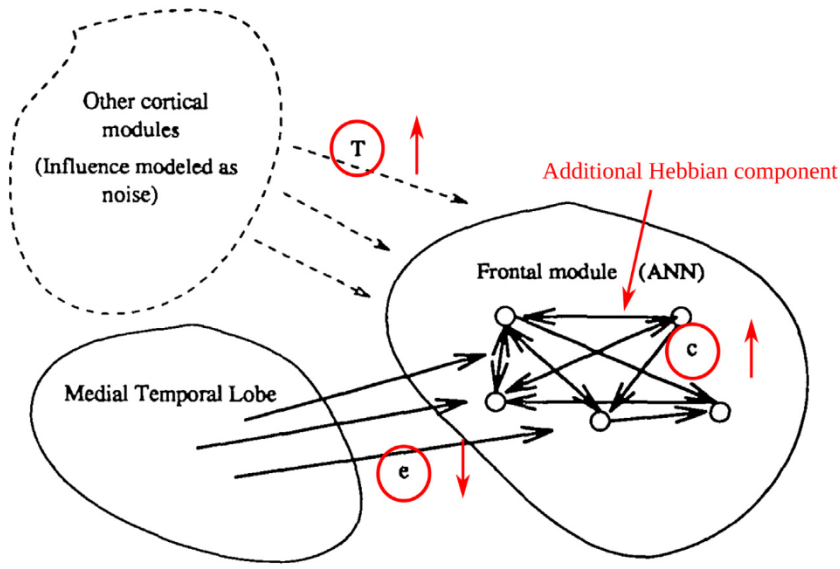


Fig. 3. Schematic illustration of the proposed model: An ANN models the frontal module, receiving input from internal connections c, external connections from the medial temporal lobe e and connections T from distant cortical modules modeled as external noise. Highlighted in red are the modifications made on the Hopfield network to imitate schizophrenic behavior: Decrease of external input projections, and increase of internal projections and external noise.
Source: Adapted from Horn and Ruppin (1995).

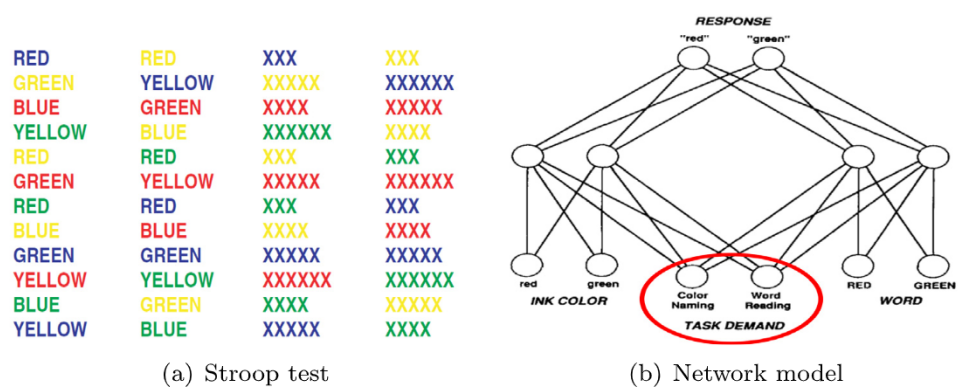


Fig. 4. Attention and context. (a) Stroop card test used for SZ, reprinted from Henik and Salo (2004) (b) Neural network model used for the *Stroop* task in Cohen and Servan-Schreiber (1992). Highlighted in red are the neurons with modified gain.

response time of the network to properly produce a correct answer, similar to what it was observed in schizophrenic diagnosed individuals.

The second experiment, shown in Fig. 5, implemented the *Continuous Performance Test (CPT)* (Rosvold, Mirsky, Sarason, Bransome Jr, & Beck, 1956) identical pair version (Cornblatt, Lenzenweger, & Erlenmeyer-Kimling, 1989). It measures participant's ability to detect repeated pattern of symbols in a longer sequence. Symbols are presented sequentially and the volunteers must detect when the pattern appears consecutively, words or numbers, e.g., "9903". In this experiment, schizophrenics usually struggle with the detection of longer patterns where previous symbols need to be taken into account. *Prior stimulus module* neurons were used to save the information about previous sequence symbols. To simulate schizophrenic behavior, the authors *reduced the gain of the activation-function* of the task context yielding to a higher miss-rate in concordance with schizophrenic empirical observations.

Finally, a lexical disambiguation task depending on context was modeled based on the original work from Chapman, Chapman, and Miller (1964) (see Fig. 6). Participants had to solve homonym conflicts (words with more than one meaning), taking into account the context of the sentence. In this case, schizophrenics show worse performances when the needed context to resolve ambiguity comes before the word in question. A similar approach than in the CPT experiment was taken: context neurons gain was manually reduced to 0.6 like in the previous experiments. It resulted in low performance for the schizophrenic model when the sentence context that was needed to interpret the ambiguous word was located at the beginning of the sentence.

4.2.2. Auditory processing

During a person's life, the number of neurons in the brain peaks during childhood and then decreases by a 30% to 40% in adolescence, which is also the period of time where SZ appears most frequently (adolescence/early adulthood) (Huttenlocher et al., 1979). Based on this observation and post-mortem findings which suggest neural deficits in the schizophrenic's cerebral cortex (Keshavan et al., 1994; Margolis, Chuang, & Post, 1994), Hoffman and McGlashan designed a feed-forward neural network capable of translating phonetic inputs into words (Hoffman & McGlashan, 1997). This model was inspired by Elman's (1990) model (Elman, 1990). As illustrated in Fig. 7(a) it consists of one hidden layer and a *temporal storage layer* that saves a copy of the hidden layer from the previous processing step.

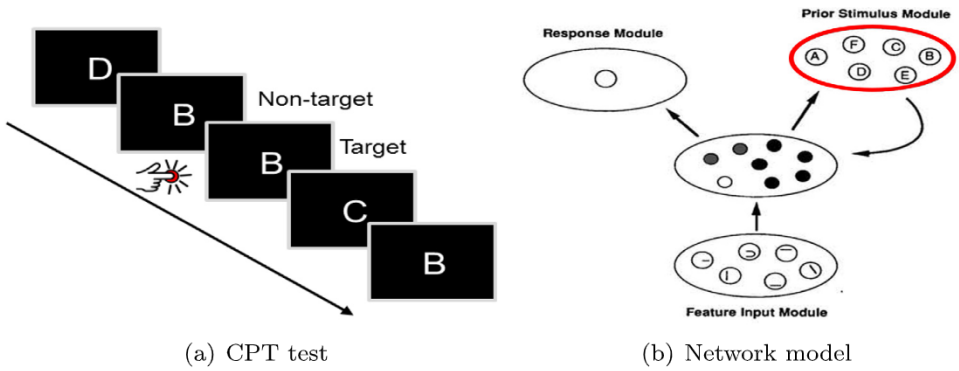


Fig. 5. Continuous performance test. (a) Simplified CPT Identical Pair test used (b) Neural network model for the *CPT* adapted from Cohen and Servan-Schreiber (1992). Highlighted in red are the neurons whose gain was decreased to model disturbed processing in the prior stimulus module.

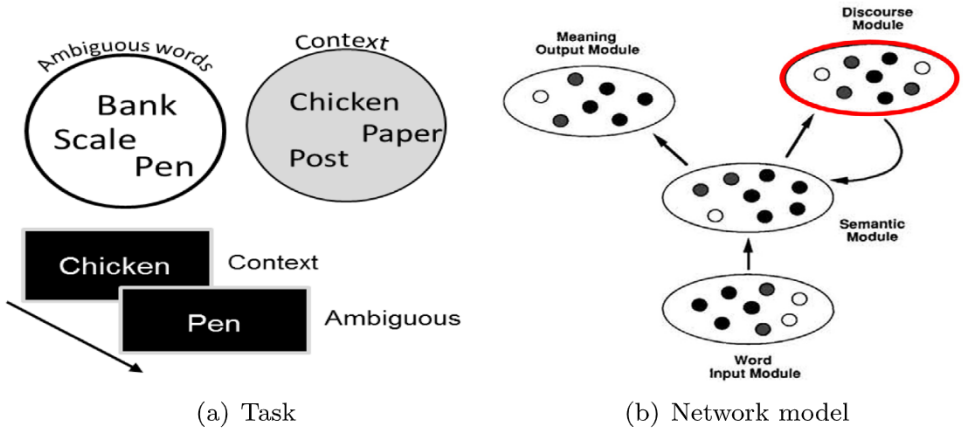


Fig. 6. Lexical disambiguation. (a) Task with context dependent meaning word. (b) Neural network model reprinted from Cohen and Servan-Schreiber (1992). Highlighted in red are the (context) neurons whose gain was reduced to 0.6. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

A *pruning rule* was used to set the value of *all connections below a certain threshold to zero*. After pruning approximately 30% of the connections, the word detection capabilities of the used network improved.[4] However, with excessive pruning the network starts to struggle with detection tasks and shows spontaneous responses during periods without input (shown in Fig. 7(b)). This last observation was associated to auditory hallucinations reported in patients with severe SZ. Furthermore, it supported the common theory that auditory hallucinations might be caused by false identification of own inner speech as externally generated.

In posterior tests with healthy patients, schizophrenics with auditory hallucinations showed reduced word detection capabilities compared to schizophrenics without such hallucinations, which fits with the previous simulations. Furthermore, a later review of this paper (Hoffman & McGlashan, 2001) highlighted that by applying active repetitive transcranial magnetic simulation (active rTMS) on the left temporoparietal cortex, a brain region usually associated to speech perception, hallucinations seem to be reduced. This further supports the hypothesis of a possible correlation between speech-processing disorders and auditory hallucinations.

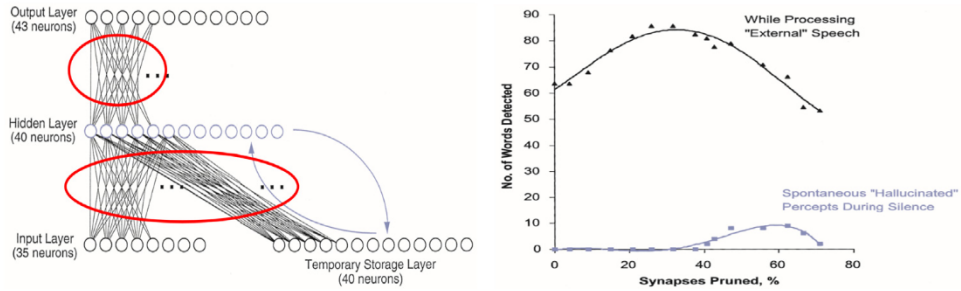
[4] Pruning is a bioinspired standard technique for improving generalization of the network. However, nowadays, dropout approaches have gained popularity over pruning.

4.2.3. Language processing

Another feed-forward model of SZ introduced by R. E. Hoffman and collaborators (Hoffman et al., 2011) uses a network called DISCERN (Grasemann, Miikulainen, & Hoffman, 2007; Miikkulainen, 1993; Miikkulainen & Dyer, 1991) that is able to learn narrative language and reproduce learned content, e.g., learn a story and reproduce it after feeding it with a fraction of the story.

Based on previous studies about SZ, eight different network modifications were tested: (1) *Working Memory (WM) disconnections* by pruning of connections with a weight below a certain threshold, (2) *Noise addition in working memory* by adding of Gaussian noise to WM neuron outputs, (3) *WM network gain reduction* by reducing the activation function's gain, (4) *WM neuron bias shifts* by increasing neuron bias and inducing an increased overall activation, (5) *Semantic network distortions* by adding noise to word representations in semantic memory, (6) *Excessive activation semantic networks* by increasing neuron outputs in semantic network, (7) *Increased semantic priming* by blurring semantic network outputs, (8) *Exaggerated prediction-error signaling (hyperlearning)* by increasing back-propagation learning rates.

The resulting network behaviors were compared to empirical results using a goodness-of-fit measure (GOF), which compared factors such as story recall success (successfully retelling story), agent confusions (switching of certain story characters), lexical errors and derailed clauses (false interpretation of certain sentences). The authors concluded that *(1) WM disconnections* with pruning and *(8) hyperlearning* best explain real-world data.



(a) Network model (b) Word detection results

Fig. 7. Auditory hallucinations (a) Neural network model used in Hoffman and McGlashan (2001). Input of the network are simulated phonetic codes, output are semantic features of the input word. Highlighted in red are the connections where the pruning rule was applied on to imitate schizophrenic symptoms. (b) Word detection results depending on connection pruning. Spontaneous detections are observed for excessive pruning.
Source: Reprinted from Hoffman and McGlashan (2001) with permission.

(a) Arcimboldo's painting (b) Tacher's illusion (c) Ocampo's painting (d) Dallenbach's illusion

Fig. 8. Visual illusions where the brain infers different interpretations depending on the prior information or context. (a) Ortaggi in una ciotola o l'Ortolano. G. ©Arcimboldo 1590. (b) Tacher illusion (Thompson, 1980). (c) Forever Always, ©Octavio Ocampo 1976. (d) Dallenbach's illusion 1952 (Dallenbach, 1951).

These results for WM disconnections further reinforce the previously presented theory by Hoffman and McGlashan in Hoffman and McGlashan (1997) that excessive connection pruning during human's adolescence might be one of the causes for this disorder. Moreover, the authors also suggested that over-learning in schizophrenic brains might cause modifications in previously stored memories, which might lead to delusional or erroneous convictions.

4.3. Bayesian approaches

Several important models of psychiatric disorders are based on the idea that the brain uses Bayesian inference as a basic principle. The Bayesian brain hypothesis describes the human brain as a generative model of the world that makes predictions about its environment and adapts its internal model depending on the observation provided by the senses. For SZ as well as for ASD it is suggested that patients might differ in the way they combine sensory inputs with prior information. The idea was highly influenced by Hermann Helmholtz's work in experimental psychology (Von Helmholtz, 1867) that dealt with the brain's capacity to process ambiguous sensory information. In his words: *"Visual perception is mediated by unconscious inferences"*.

Fig. 8 shows puzzle images that stress that perception depends on prior knowledge as well as sensory input. For instance, if we rotate Arcimboldo's painting by 180 degree instead of vegetables we will see a human face with a hat. Tacher's illusion can be broken by also rotating the upside down images, and we will see that both faces are different. In particular, mouth and eyes are inverted. In Ocampo's painting, we can see two old people from a larger distance but two mariachis when viewing the picture from close range. Finally, Dallenbach's illusion shows that even if you know that there is an animal looking at you in the picture, it is impossible to see it until the shape of the cow is highlighted. Afterwards you cannot stop seeing it. In essence, what we perceive not only depends on the raw sensory information, but also on our prior knowledge and predictions we have about the world.

The classical concept of Bayesian inference presents perception as computing the posterior belief from the sensory input (likelihood) and from the model prediction (prior belief) depending on their relevance. For instance, in the case of a very imprecise (highly variable) prior, the perception would shift more strongly to the direction of the sensory input. Fig. 9 illustrates these concepts assuming that the world is one-dimensional and can be described via Gaussian distributions.

4.3.1. Free-energy model of schizophrenia

Friston's free-energy model (Friston, 2010) describes the brain functionality as a dynamical inference network. It combined the Helmholtz machine ideas (Dayan, Hinton, Neal, & Zemel, 1995) with the hierarchical prediction error message passing (Rao & Ballard, 1999) and the Bayesian mathematical framework. Despite not being implemented as an ANN model, we included it in this review because it is considered one of the most relevant models in the computational neuroscience community. Furthermore, it serves for comparative purposes with predictive coding neural network implementations of psychiatric disorders (Philippsen & Nagai, 2018; Yamashita & Tani, 2008, 2012).

Under the free-energy principle, the brain is seen as a prediction machine that progressively constructs an internal model of the world which is constantly improved, based on the received sensory feedback and the resulting prediction error. Perception (posterior belief) then results from combining the brain's predictions (prior) with the sensory evidence (likelihood) as shown in Fig. 9. If the prior's precision is relatively higher than the precision of sensory evidence, the posterior will be more similar to the prior. In the opposite case, the posterior will be more close to sensory input. Therefore, precision weights the influence of prior and sensory evidence on the posterior belief.

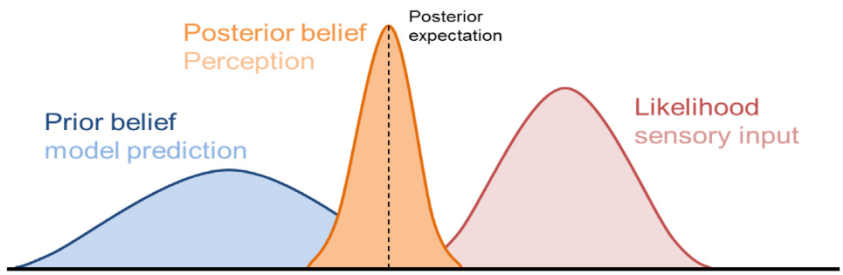


Fig. 9. Illustration of Bayesian inference: The posterior belief is generated by inference of prior belief and sensory evidence. Depending on the variance (precision) of prior and sensory evidence, the posterior belief will be influenced more by one of the previous.
Source: Adapted from Adams et al. (2013).

Mathematically, the internal model is updated by minimizing the negative *free energy* F a lower bound on the KL-divergence that quantifies the difference between the internal belief about the world and reality.

Assuming that $\vec{\mu}$ are the dynamical internal states of the brain, perception is then described as the adaption of $\vec{\mu}$ given the sensory observations by minimizing the free energy using the gradient descent method described in Eq. (8):

$$\dot{\vec{\mu}}(t) = D\vec{\mu}(t) - \frac{\partial F(\vec{s}, \vec{\mu})}{\partial \vec{\mu}} = D\vec{\mu}(t) - \frac{\partial \epsilon}{\partial \vec{\mu}} \Pi \epsilon \tag{8}$$

where D is a differential matrix operator that computes the currently expected hidden state, ϵ is the error between the predicted (sensory) input from the higher layer and the real input (observation) and Π is the inverse variance (precision) of the information. For instance, in humans, visual information would typically have higher precision than proprioceptive sensing for body localization (Hinz, Lanillos, Mueller, & Cheng, 2018).

Based on these concepts, Adams et al. (2013) built a computational model of SZ and analyzed in three different experiments: auditory pattern recognition (using the example of a bird recognizing its own song), an object eye-tracking task and a simulation of force-matching illusion. One of the core ideas was that a reduction of the precision at higher levels of the cortical hierarchy (i.e., reduced precision of prior beliefs) influenced the responses of the model. More concretely, decreases in prior precision (or, for the force-matching illusion, failure to reduce sensory precision) led to struggles in auditory pattern recognition, problems with eye-tracking with occlusion and attribution of agency. Furthermore, with an additional compensatory decrease of sensory precision (for the force-matching illusion, increase of prior precision), the model showed hallucination-like behavior during the auditory pattern recognition task and difficulties to distinguish self-touch and touch from others in the force-matching illusion.

Fig. 10 shows the experiment of auditory pattern recognition of a birdsong, showing how the precision in different cortical levels changes the response to surprising events. The first row describes a normal behavior to surprising events (the belief precision is high). In this case, when a chirp of the bird is omitted, the posterior perception contains an illusory (weakened) response at the point in the signal where sensory input is missing (white arrow at left plot). This effect might correspond to omission-related responses found in electrophysiological recordings of the brain (Nordby, Hammerborg, Roth, & Hugdahl, 1994). The middle and bottom rows correspond to abnormal behaviors in line with SZ findings, such as attenuation of omission-related responses and auditory hallucinations respectively.

4.3.2. Circular inference in Bayesian graphical models

In Jardri and Deneve (2013), Jardri and Denéve investigated how excitatory to inhibitory imbalance may relate to psychotic symptoms in schizophrenia, using belief propagation in a hierarchical Bayesian graphical model. In particular, it is shown that *a dominance of excitation causes circular belief propagation*: bottom-up sensory information and top-down predictions are reverberated, and therefore, may be confused with each other or taken into account multiple times. The model can account for the occurrence of erroneous percepts (hallucinations) and fixed false beliefs (delusions) in SZ.

In the graphical model, low hierarchical levels correspond to sensory experience and high levels to top-down predictions. Messages are passed between nodes in different hierarchical levels from lower to higher levels (bottom-up processing) and from higher to lower levels (top-down processing). The fact that connections exist in both directions raises an important challenge: to differentiate between *real* sensory information and sensory information which were simply *inferred* from top-down expectations. The authors suggest that such circular belief propagation in the Bayesian network is avoided if a careful balance between excitation and inhibition is maintained. A disruption of this balance can account for the appearance of schizophrenic symptoms.

Concretely, information between higher and lower levels are exchanged in the form of messages. For belief propagation, messages are passed recursively until convergence:

$$M_{ji}^{n+1} = \begin{cases} W_{ij}(B_i^n - \alpha_d M_{ij}^n) & \text{if } i \text{ is above } j \\ W_{ij}(B_i^n - \alpha_c M_{ij}^n) & \text{if } j \text{ is above } i, \end{cases} \tag{9}$$

where the term *above* means that the node i is in a higher hierarchical level than j. M_{ij}^n is the message sent from i to j at step n. W_{ij} is the connection strength, and α_d and α_c are the parameters that scale the inhibitory loops in upward and downward directions, respectively. B_i^n is the computed belief expressed as a log-odd ratio[5] and updated as:

$$B_i^{n+1} = \sum_i M_{ji}^{n+1}. \tag{10}$$

The authors experimented with the two α parameters in this framework, adjusting them between 1 (normal level of inhibition) and 0 (no inhibition). Simulated results show that equally impaired loops (same α below 1) are still able to arrive at a proper inference. Conversely, with unbalanced impaired upward loops ($\alpha_u < 1$) "over-estimation of the strength of sensory evidence and an underweighting of the prior" is produced. This is compatible with over-interpretation of sensory evidence and the reduced influence to illusions observed in schizophrenic patients.

The authors recently demonstrated in Jardri et al. (2017) that the circular inference model nicely fits decisions of SZ diagnosed

[5] Log-odd ratio: computed as the log of the ratio between the probability that a cause is present and that the cause is absent, thus, values around 0 describe uncertain states, positive values correspond to belief in presence, negative values to belief in absence.

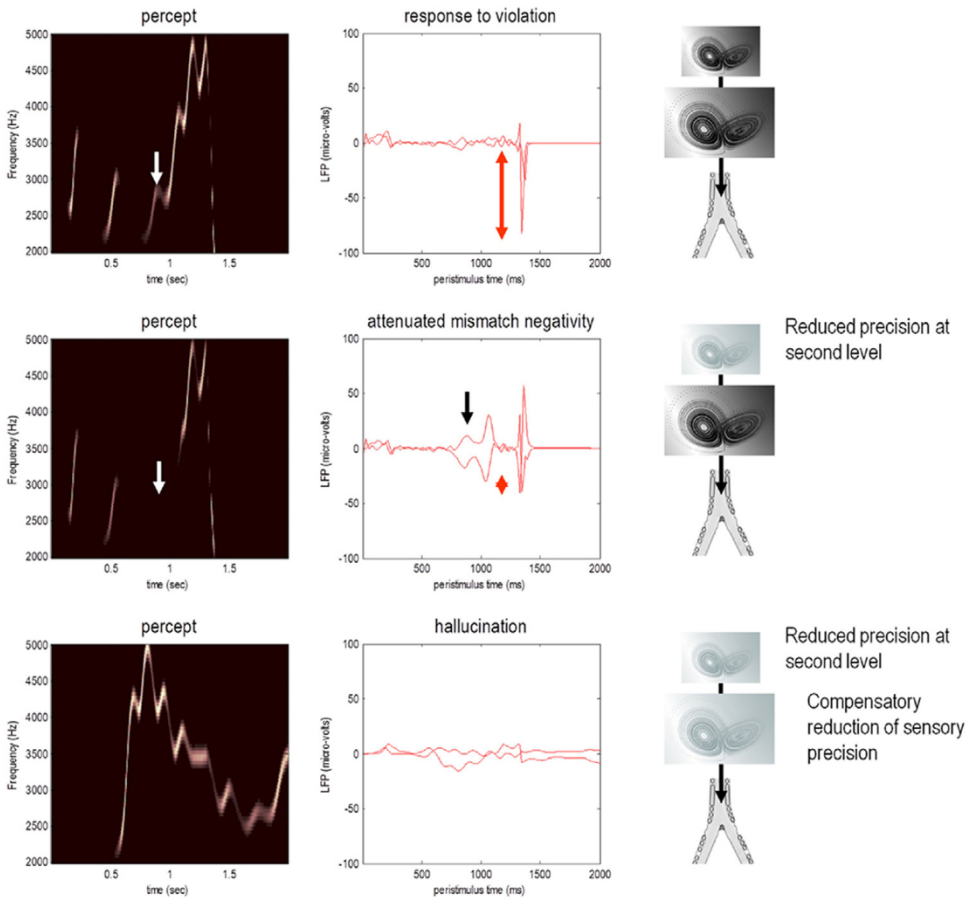


Fig. 10. Prediction sonograms of the auditory signal of a birdsong (left), prediction error with respect to stimulus (middle) and used model (right), when last three chirps are omitted. Top row: Unmodified model generated prediction error increases with the first missing chirp, which corresponds to normal behavior. Middle row: With reduced precision at second level the model is unable to predict the third chirp, and the prediction error for missing chirps is reduced. Bottom row: With compensatory sensory precision reduction in first level, there is a complete failure of perceptual inference. Despite the wrong predictions, almost no prediction error is generated due to missing precise sensory information. This behavior is compared to auditory hallucinations.
Source: Reprinted from Adams et al. (2013) with kind permission.

patients using the Fisher task as the experimental paradigm. The Fisher task permits the manipulation of the prior and the likelihood allowing comparisons with the Bayesian model predictions. Participants have to decide whether the fish captured comes from the left or the right lake. First, two boxes (left, right) with fish and different sizes are presented (prior): bigger box expresses higher probability. Secondly, the two lakes (left, right) are presented with fishes inside with two colors (red and black). The proportion of red fishes represent the likelihood of the observation. Finally, participants have to decide if the red fish comes from the left or the right. According to the participant's data and their proposed model, descending and ascending loops correlated with negative and positive SZ symptoms respectively.

4.4. Recurrent neural networks

In 2012, Yamashita and Tani presented a model of SZ using a recurrent neural network (RNN) (Yamashita & Tani, 2008) such as they are commonly used for the recognition and generation of time series. Specifically, in this study, the RNN is applied to the task of sensorimotor sequence learning in a humanoid robot: the robot learns to predict visual information and own motor movements in a scenario where it moves a cube on a surface.

The type of RNN they used is the Multiple Timescale Recurrent Neural Network (MTRNN), a special type of RNN that mimics the hierarchical structure of biological motor control systems. Human and animal motor movements are commonly suggested to be segmented into so-called "primitives" (Schaal, Kotosaka, & Sternad, 2000). These primitives can then be reused and combined to more complex motor sequences. The MTRNN contains neurons working at different timescales: fast context neurons (corresponding to the lower level of the hierarchy) learn the motion primitives and slow context units (corresponding to higher, more abstract levels) control the sequence of the primitives (see Fig. 11). This network is trained to perform prediction error minimization, i.e., to build an internal model of the world following the Bayesian brain idea. Training the network using the Backpropagation Through Time algorithm (BPTT), the robot learns multiple motions (grasping and moving an object) adapting to different object positions. It is also able to combine these actions into new action sequences by only training the slow context units. The trained network works as a predictor where the sensory input modulates the changes on the slow context units (goals) depending on the error.[6]

Eq. (11) describes the dynamics of each neuron at each layer:

$$\tau \dot{u}_{i,t} = -u_{i,j} + \sum_j w_{i,j} \cdot x_{j,t}. \tag{11}$$

[6] There is a strong parallelism between Multiple Timescale RNNs and the hierarchical model proposed by Friston.

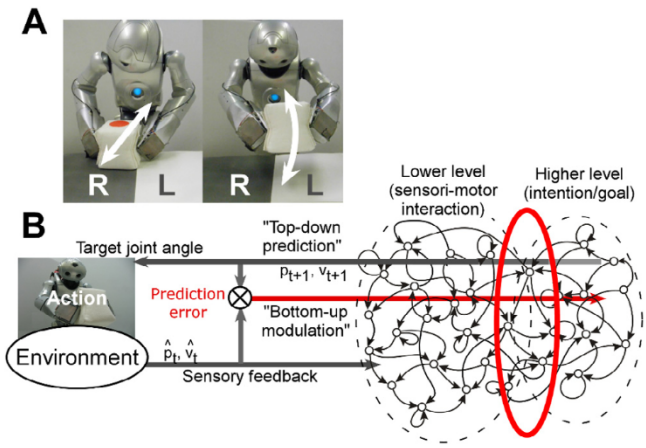


Fig. 11. (A) Tasks to be performed by the robot: when the object is on the Right move the object backward and forward, when the object is on the Left move the object up and down. (B) MTRNN network architecture. Highlighted with a red ellipse are the connections between fast and slow context units that are degraded with noise to imitate schizophrenic behavior.
Source: Adapted from Yamashita and Tani (2012).

In this formula, the membrane potential $u_{i,t}$ of neuron i at time step t is updated with the neural state $x_{j,t}$ of neuron j scaled with the (learnable) connection weights $w_{i,j}$. The time constant τ determines the update frequency of the neuron. A small time constant is used for fast context units, and a large time constant for slow context units.

Schizophrenics can have trouble to distinguish self-generated actions from others' actions and, in severe cases of SZ, patients can even have problems performing movements, and show repetitive or stereotypical behavior (van der Weiden et al., 2015). Based on observations that suggest that SZ may be caused by disconnections in hierarchical brain regions, mainly between prefrontal and posterior regions (Bányai, Diwadkar, & Érdi, 2011; Friston & Frith, 1995), *uniformly distributed random noise was added in the connections between fast and slow context units* highlighted with the red circle in Fig. 11. For the evaluation of the model a humanoid robot was used. It had the task of locating an object on a table in front of it and performed different actions depending on the object's position: if the object was located to the right, the robot was supposed to grab the object and move it back and forth three times. Otherwise, if the object was located to the left, the robot had to grab the object and move it up and down three times.

They showed that for a small degree of disconnection (small noise addition) the robot had no problems to perform the mentioned task. Nevertheless, increases of spontaneous prediction error were observed and abnormal state switching appeared in the intention-network (slow units). The authors compared these prediction errors to patient's problems in attribution of agency (when own movements are perceived as being executed by someone else). Schizophrenics might want to perform an action and have an internal prediction of the upcoming proprioceptive and external states. The increases of prediction error could be seen as incongruences between the intended actions and the results, which can give a person the feeling of not being able to control the consequences of its own actions or it may have problems to perceive these actions as self-generated. For more severe disconnections, the humanoid robot clearly struggled to perform the given task and showed disorganized sequences of movements. These observations were compared to more severe cases of SZ, where cataleptic (stopping) and stereotypical (repetitive) behaviors have been observed.

5. ANN models of autistic spectrum disorder

This section describes the most important ANN models of ASD. They focused on the atypical processing style suggested by the weak central coherence theory which could be summarized as excessive attention to detail. They replicated deficits in perception (Cohen, 1994; Dovgopoly & Mercado, 2013; Gustafsson, 1997; Nagai et al., 2015). Some also addressed atypicalities in memory structure and internal representations (McClelland, 2000; Philippsen & Nagai, 2018) and inflexibility in motor behavior (Idei et al., 2017). Although most studies suggested connections to social deficits in an indirect way, only one of the models made a direct connection to theory of mind, by modeling weak central coherence on the level of logical reasoning (OLaughlin & Thagard, 2000). An overview of the reviewed approaches is given in Table 2.

5.1. Feed-forward and simple recurrent neural networks

First, we describe approaches using simple connectionist models, typically feed-forward networks for classification tasks. Recurrent connections might be included at a structural level, but networks are not supposed to learn temporal sequences, which is why we refer to them as simple recurrent NN. These approaches mainly explored parameters of the network such as number of neurons or learning rate.

5.1.1. Generalization deficits through overfitting

The first neural network model of ASD to our knowledge was proposed by Ira L. Cohen in 1994 (Cohen, 1994). It was a feed-forward neural network trained with back-propagation and investigated basic properties of neural networks. Based on studies that suggested that individuals with autism have either too few or too many neurons and neuronal connections (e.g., Bauman, 1991), the influence of increased or reduced number of hidden neurons was analyzed. The evaluated task was to classify children with ASD and children with mental retardation into two groups, using features obtained via a diagnostic interview (Cohen, Sudhalter, Landon-Jimenez, & Keogh, 1993). Note that although the considered task was related to ASD, the chosen task is just taken as an example and is not crucial for the findings of this paper.

A training and a test set were used to analyze the network's accuracy and generalization abilities. The results were compared for an increasing number of hidden units and through different number of trials. The results showed that a small number of hidden neurons translates into low accuracy (high training error) and bad generalization (high testing error) and an increased number of hidden neurons improved the network's learning accuracy and generalization. When the *number of hidden neurons was largely increased*, its generalization ability decreased: the network learned too much details of the input data and was not able to adapt to new input data. An *increased number of training trials (longer training duration)* had a similar effect. For the training set, the network accuracy increased with longer training duration. However, with the test set, the network again showed signs of overfitting, as the accuracy decreased significantly.

Cohen compared these results qualitatively to the learning and behavioral characteristics of children with ASD. In particular, many individuals with ASD show great discrimination capabilities and have no problems with already learned routines, but have problems when trying to abstract information or when confronted with new situations.

Cohen extended this approach in 1998 (Cohen, 1998) to the generalization capability in the presence of extraneous inputs to the network (set to random values). In the task of classifying

Table 2
Overview of neural network models of ASD.

Model type	Paper	Disorder characteristic	Biological evidence	Approach
Feed-forward and simple recurrent NNs	Cohen (1994, 1998)	Generalization deficits due to excessive attention to detail	Abnormal neural density in various brain regions	Excessive or reduced number of neurons, increased training duration
	McClelland (2000)	Hyperspecificity of memory concepts	–	Excessive conjunctive coding
	Dovgopoly and Mercado (2013)	Deficits in visual categorization and generalization	Abnormalities in synaptic plasticity	Reduced learning rate, negative weight decay (anti-regularization)
Self-organizing maps	Gustafsson (1997)	Excessive attention to detail	Lateral inhibition enhances sensory perception	Excessive inhibitory lateral feedback
	Gustafsson and Papliński (2004)	Avoidance of novelty	–	Familiarity preference, higher weighting of close data points
	Noriega (2007)	Domain-based hypersensitivity	Early brain overgrowth in children with ASD	Variable (increasing) number of neurons, stronger/weaker attention to stimuli
	Noriega (2008)	Domain-based hypersensitivity	Early brain overgrowth in children with ASD	Propagation delays in neural weight updates
Convolutional NN	Nagai et al. (2015)	Local/global processing bias	Excitation/inhibition imbalance	Excitation/inhibition imbalance in visual processing
Spiking NNs	Park et al. (2019)	Atypical neural activity: High power in higher frequency bands and decreased signal complexity	Increased short-range connectivity in frontal cortex and atypicalities in resting-state EEG	Local over-connectivity
Predictive coding	Pellicano and Burr (2012)	Excessive attention to detail	–	Hypo-prior: lower precision of prior, stronger focus on sensory input
	Lawson et al. (2014)	Excessive attention to detail	Stronger activation in visual cortex than in prefrontal cortex in ASD	Hypo-prior or hyper sensory input: Precision imbalance that leads to excessive reliance on input
Recurrent NNs	Idei et al. (2017)	Stereotypical behaviors	–	Modification of variance estimation (sensory precision)
	Philippsen and Nagai (2018)	Reduced generalization capability, heterogeneity among subjects	–	Modification of reliance on external signal and of variance estimation (sensory precision)
	Ahmadi and Tani (2017)	Generalization deficits	–	Regularization
Other approaches	OLaughlin and Thagard (2000)	Weak coherence, Theory of Mind impairment	–	Impairment of coherence optimization in logical reasoning due to strong inhibition

happy and sad expressions of a simplified cartoon face, generalization was strongly impaired in the presence of extraneous inputs. This might suggest that networks trained for too long tend to attend more to non-relevant input information, instead of focusing on the more informative input neurons.

Note that although increased number of hidden neurons may replicate autistic traits as shown in Cohen (1994), this parameter did not cause generalization deficits neither in Cohen's follow-up work (Cohen, 1998) nor in a similar modeling study (Dovgopoly & Mercado, 2013).

5.1.2. Precision of memory representations

In McClelland (2000), James L. McClelland addressed the tendency of children with ASD to represent concepts in a too specific way, which results in difficulties to recognize two different instances of an object as the same category.

He suggested that in neural networks, this could be explained with the concept of excessive conjunctive coding. Typically, similar inputs to a neural network lead to similar neuron activation patterns. Such pattern overlaps can be useful for sharing existing knowledge and establishing associations. However, too strong associations can also cause interference. Conjunctive coding describes the reduction of such overlap by recoding the input patterns with neurons which only become active for particular combinations of elements. Assuming that what characterizes healthy human learning is a balance between generalization and discrimination, the representation of concepts in subjects with ASD could be characterized by *excessive conjunctive coding*. This would make a neural network loose the ability to generalize, as activation pattern overlaps cannot be exploited.

This idea was not tested experimentally, but the author used the neural network shown in Fig. 12 to explain his reasoning. McClelland presented the example of a semantic network used in McClelland, McNaughton, and O'reilly (1995), as a model of organization of knowledge in memory (see Fig. 12). This model was used to associate words with their meaning, e.g., "robin" and "can" trigger the outputs "grow", "move" and "fly" because these are the actions a "robin" can perform. The internal layer of the network (highlighted in red in Fig. 12) progressively learns to code the meaning of input words during learning. This means that "robin" and "canary" should cause a very similar activation pattern because a robin has much more in common with a canary than, for instance, a tree. The author suggests that hyperspecificity in perception and memory representations of ASD children might be caused by an abnormality during this process. Namely, excessive conjunctive coding in the internal layer is proposed as a mechanism: an excessive reduction of overlap between representations of similar concept might cause the reported hyperspecificity which would result in generalization deficits. No concrete network parameters are proposed, but it can be imagined that such an effect might be achieved by increasing the number of neurons in the internal layer. In this regard, the

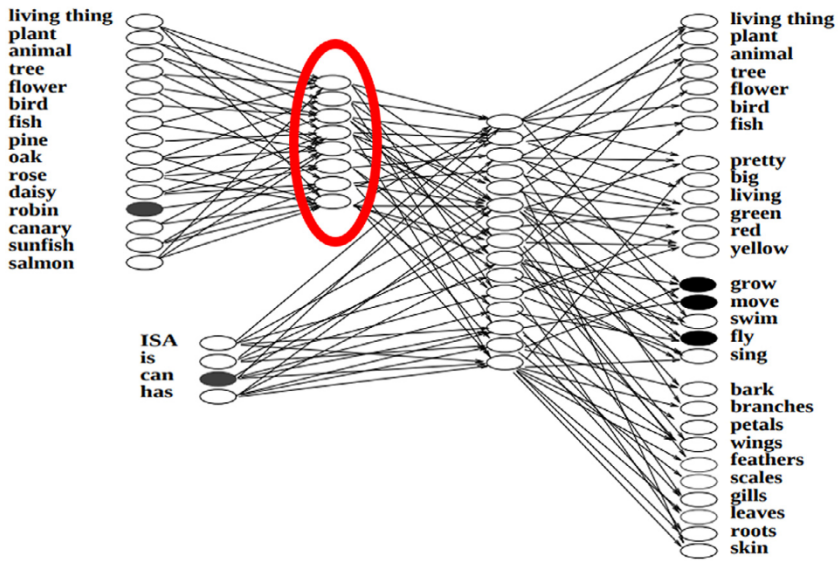


Fig. 12. Semantic network used to explain the conjunctive coding hypothesis. In the hidden layers, the feed-forward neural network generates internal representations of the inputs (highlighted in red). Words describing similar concepts should produce similar internal representations that overlap with each other. The author suggests that excessive conjunctive coding to avoid these overlaps could produce excessive discrimination, such as in autistic perception.
Source: Adapted from McClelland (2000).

approach is similar to Cohen's suggestion (Cohen, 1994), but extended to learning of representations.

5.1.3. Generalization and categorization abilities in visual perception

Dovgopoly and Mercado (2013) used an existing model of visual object perception (Henderson & McClelland, 2011) to replicate deficits in classification and generalization in ASD. The neural network was a feed-forward network, which modeled visual input processing via two pathways: the ventral cortical pathway (for object identification, including recurrent connections), and the dorsal cortical pathway (for processing of location-relevant information).

The authors replicated behavioral data from Church et al. (2010) and Vladusich, Olu-Lafe, Kim, Tager-Flusberg, and Grossberg (2010), separately on both visual pathways, which show deficits in generalization and prototype formation in children with high-functioning ASD. The experiment was the classification of random dot patterns as category or non-category stimuli (Church et al., 2010), or as category A or category B stimuli (Vladusich et al., 2010). After adjusting the parameters for replicating typical behavior, four different parameter modifications were tested individually to replicate the data from ASD children. Following evidence for abnormalities in synaptic plasticity in individuals with ASD (e.g., Auerbach, Osterweil, & Bear, 2011; Bourgeron, 2009), the first two parameters modified how weights in the network were updated.

First, the *learning rate was decreased*, which corresponds to reduced synaptic plasticity in biological neurons. As a result, network training takes longer and is more prone to lead to exhibit overfitting. Second, *generalization of the network was impaired by suppressing regularization* using negative weight decay. Weight decay is a method for regularizing neural networks and improving their generalization abilities by keeping the connection weights small (Krogh & Hertz, 1992). Typically, weight decay punishes large weights by adding a term $\lambda\vec{w}'\vec{w}$ to the error function. With a negative weight decay factor λ instead, anti-regularization is performed, encouraging the increase of weight magnitudes, and thus, over-complex classification rules. Third, they tested the influence of *increasing and decreasing the number of hidden neurons* similar to Cohen (1994, 1998), based on neurological evidence of an increased number of cortical minicolumns in the brain of individuals with ASD (Casanova et al., 2006). Finally, the authors adjusted the *gain of the neuron's activation function, to model the increased level of noise* that is hypothesized to underlie the relative increase in cortical excitation observed in ASD subjects (Rubenstein & Merzenich, 2003; Yizhar et al., 2011).

The gain G of the activation function, as displayed in Eq. (12), manipulates the slope of the activation function. A smaller gain reduces the slope, and makes the network more prone to pass noise instead of signal information to the next processing layers:

$$s(x) = \frac{1}{1 + \exp(-(G \cdot x + b))} \tag{12}$$

where x represents the input to the activation function and b is a bias term.

Good replications of the behavioral data were achieved with a decrease of learning rate and a negative weight decay. A negative weight decay also caused a high variability of generalization abilities, depending on the initial network weights, providing a potential explanation for the heterogeneity of findings between different studies. The gain of the activation function could not fully account for the generalization deficit. Also an increased number of neurons did not replicate the generalization deficit in ASD children, which contradicts previous findings from Cohen (1994). In fact, an increased number of hidden units seems to lead to generalization problems only under certain training circumstances (Caruana, Lawrence, & Giles, 2001), indicating that it is not a good candidate for explaining generalization difficulties in general.

5.2. Self-organizing maps

Self-organizing maps (SOMs) are ANNs that are usually used for unsupervised learning and clustering tasks. They model the functionality of cortical feature maps, which are spatially organized neurons that respond to stimuli and self-organize according to the features in stimuli. They are able to learn the relation of different input data such as different sensory inputs. Approaches for modeling ASD with SOMs typically investigate the formation of higher-level representations from sensory input.

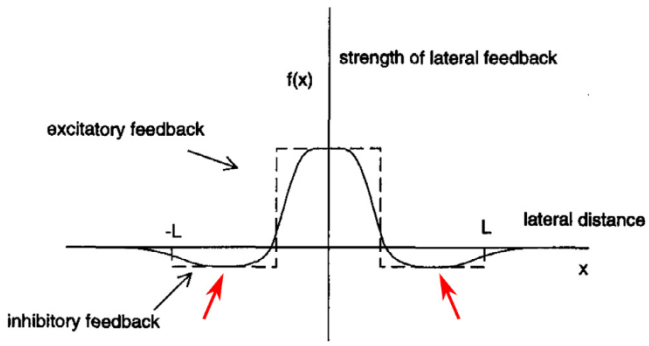


Fig. 13. Mexican-hat function of the SOM. It defines the strength of lateral connections depending on distance to current neuron. The red arrows point to the part that is modified to simulate autistic perception (excessive lateral feedback inhibition).
Source: Adapted from Gustafsson (1997).

5.2.1. Increased lateral feedback inhibition

Lennart Gustafsson presented two models of ASD using SOMs in Gustafsson (1997) and Gustafsson and Papliński (2004). Inspired by findings on weak central coherence in subjects with ASD and an enhanced ability to discriminate sensory stimuli (Frith & Happé, 1994), he suggested that alterations in the lateral feedback weights between the SOM neurons could result in atypicalities in perception (Mountcastle, 1957).

In a SOM, each neuron typically has excitatory connections to close neighbors and inhibitory connections to more distant neighbors. They tuned the Mexican-hat curve (Fig. 13) to induce stronger lateral feedback inhibition. Such activation patterns are similar to receptive fields in biological cortices and have been used to model center–surround operators in the visual cortex. Manipulating the lateral connections to achieve a stronger inhibition (such that the integral of the function in Fig. 13 becomes negative), the sensory discrimination ability of the network is increased. Neural columns focus on more narrow features during learning which slows down convergence and might lead to a fragmented feature map. However, excessive lateral inhibition will degrade discriminatory power and cause instabilities in information processing. This behavior is compared to autistic over-discrimination and may also explain fascination or fright of moving objects, due to the instability of its cortical feature maps.

5.2.2. Familiarity preference

In Gustafsson and Papliński (2004), Gustafsson and Papliński evaluated the effect of attention-shift impairment and avoidance of novelty on the formation of cortical feature maps. The used SOM received input stimuli from two sources (compared to two "dialects of a language"), each of which produces 30 different stimuli ("speech sounds") grouped in three clusters ("phonemes").

The computational model was run in four different modes. In the first mode, attention was always shifted to the source producing novel input (considered as normal learning). In the second mode, an attention-shift impairment was modeled by shifting attention to novel sources with a very low probability. The third mode implements familiarity preference: attention is shifted to novel sources only if the map is familiar with that source (measured as mean distance of the current stimulus to the map nodes). This map develops a preference over learning to the more familiar source. Finally, a model with both familiarity preference and attention-shift impairment was applied.

The simulation results showed that *familiarity preference* leads to precise learning of the stimuli from one of the sources (the source with lower variability) in expense of the other source. This might remind of ASD individuals' characteristic of learning in great detail a narrow field, which leads to increased discrimination and poor generalization. The authors also showed that this impairment can be counteracted by modifying the probabilities of stimuli presentation in response to the system, similar to early intervention in children's learning process. Maps learned with attention-shift were not impaired, whereas a combination of both mechanisms only sometimes led to an impairment. The authors concluded that, in contrast to speculations in previous work (Courchesne et al., 1994), familiarity preference, rather than attention-shift is a more likely cause for ASD.

5.2.3. Unfolding of feature maps and stimuli coverage

In 2007, Noriega (2007) modeled abnormalities in the feature coverage and the unfolding of feature maps in SOMs. Neurological evidence suggests abnormal brain development in children with ASD (Bauman & Kemper, 2005), typically reporting larger growth in young children, which gets reduced later in life (Aylward, Minshew, Field, Sparks, & Singh, 2002; Courchesne et al., 2001). These abnormalities were modeled by manipulating the number of network nodes during the training of the SOM where the structure emerges. Thus, the network dimension is temporarily increased.

Results showed that such disturbance in the physical structure of a SOM does not affect stimuli coverage, but impairs the unfolding of feature maps which might result in sub-optimal representations. Furthermore, the author models hyper- and hyposensitivity to stimuli in a similar way like (Gustafsson, 1997) using lateral interactions between neurons. *Hyper- or hyposensitivity* was modeled by adjusting the neuron weights toward the winner neuron, either with a positive factor (attraction, or hypersensitivity) or with a negative factor (repulsion, or hyposensitivity). This factor converges exponentially toward zero (normal sensitivity) during map formation. The authors showed that hypersensitivity to one of the input domains (stronger attention to this domain, i.e., restricted interests), improves the coverage of stimuli in this domain, but too strong hypersensitivity or a hyposensitivity to stimuli reduces coverage.[7]

One year later, Noriega extended his approach in Noriega (2008), investigating *propagation delays between neurons*. Unlike in normal SOMs where all neurons propagate the information instantaneously to all neighboring neurons, Noriega presented a biologically more realistic approach by introducing delays in the update. He shows that decreased propagation speed has a negative effect on stimuli coverage. As the delayed propagation causes the arrival of competing stimuli at the same time at a neuron, he also altered the way in which these competing stimuli are handled. In his experiments, a high *dilution factor*, meaning that incoming stimuli are averaged instead of being handled separately, decreased the stimuli coverage and also impaired the topological structure of the map.

5.3. Convolutional neural networks and inhibition imbalance

In 2015, Y. Nagai and colleagues presented an ANN network based on Fukushima's neocognitron (Fukushima, 1988, 2003; Fukushima & Miyake, 1982), seen as the basis for convolutional neural networks, to model visual processing in ASD (Nagai et al., 2015). The hypothesis considered was that there is an excitation/inhibition imbalance in ASD (Snijders et al., 2013; Sun et al., 2012; Yizhar et al., 2011).

[7] Hypersensitivity in Gustafsson (1997) was implemented as increased inhibition in the neighborhood of neurons (higher specificity of perception), whereas this approach interprets hypersensitivity as a stronger attraction of neighboring signals to signals from a specific domain.

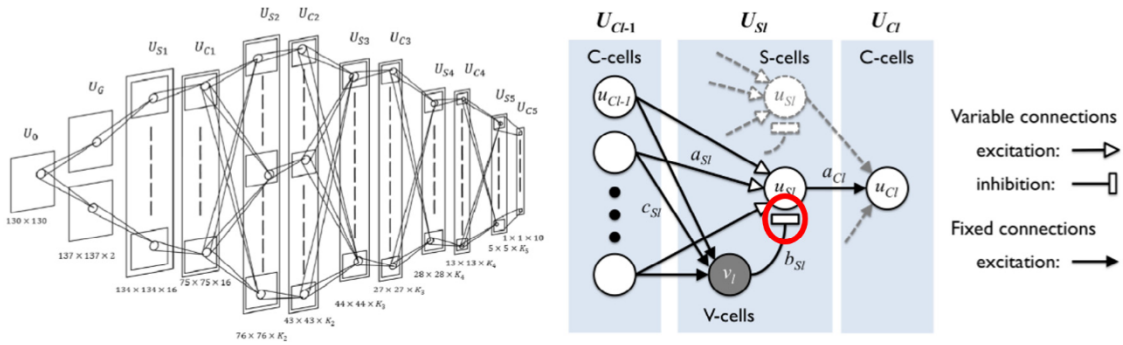


Fig. 14. Left: Overview of the neocognitron's structure. Right: Detailed view of the connections between C-cell layers U_C and S-cell layers U_S. Highlighted in red are the inhibitory connections that are modified to influence the ratio between inhibition and excitation. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)
Source: Adapted from Nagai et al. (2015).

The structure of the neocognitron for visual processing is illustrated in Fig. 14. The network is trained to recognize patterns by adjusting the weights between U_C and U_S layers. The S-cells in the U_S layers perform feature extraction. They receive excitatory input from the C-cells in the preceding layer, and inhibitory connections from the V-cells in the same layer. During training, the excitatory connections a_{Sl} are updated and the inhibitory connections b_{Sl} are calculated accordingly.

The network was trained for the recognition of numbers "0" to "9" in large or small size at different positions. After training, the model was tested with compound numbers (cf. Fig. 15 left) where a larger number is created from multiple smaller numbers. The trained network is able to detect both global (large number, here "2") and local (small numbers, here "3") patterns for $\alpha = 1$ and 0.9, but shows a preference for the global pattern, characteristics that correspond to observations with healthy individuals (Behrmann et al., 2006).

It is known that people with ASD perform differently in such a task, primarily focusing their attention on the details (i.e., the smaller number instead of the larger one). In order to simulate this local processing bias, an imbalance of excitatory and inhibitory connections was simulated by scaling the *inhibitory weight* b_{Sl} *with a factor* α.

The results show that a moderate increase of α, which corresponds to increasing inhibition, causes the network to rather detect local patterns, replicating the local processing bias in ASD. When reducing α (increasing excitation), the network does not show any processing bias, rather it looses its ability to differentiate patterns. These results fit with ASD symptoms of hyperesthesia (increased focus on detail) and hypoesthesia (no bias and general difficulty in pattern recognition) and suggest that excitation/inhibition imbalance could account for these symptoms.

5.4. Spiking neural networks and local over-connectivity

In Ichinose et al. (2017) and a follow-up study in Park et al. (2019), it was proposed to use spiking neural network as computational models to investigate the consequences of local over-connectivity, which was found in the prefrontal cortex of ASD brains (Courchesne & Pierce, 2005). The hypothesis considered was that local over-connectivity affects frequency patterns of neural activations.

A spiking neural network is more closely inspired by natural neural networks (Izhikevich, 2003). Whereas in standard artificial neural networks each neuron fires at every time step, neurons in a spiking network only fire if their potential (similar to the membrane potential of biological neurons) exceeds a certain threshold.

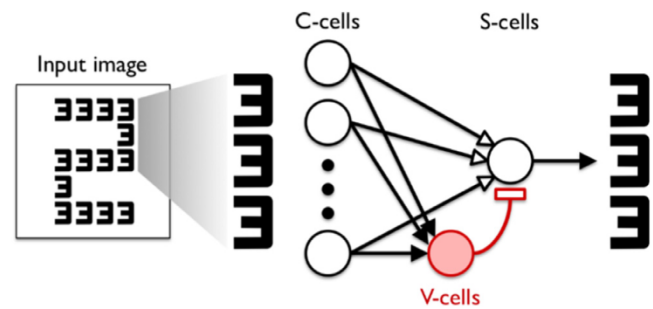


Fig. 15. The neocognitron is fed with a visual stimulus consisting of local patterns (here *3*) and global patterns (here *2*), which are incongruent. In normal conditions the network should be able to detect both local and global patterns.

Therefore, more complex firing patterns can occur ranging over various frequency bands, comparable to patterns visible in EEG.[8]

A number of studies found evidence that EEG signals of ASD brains tend to exhibit higher power in low-frequency and high-frequency bands of EEG (Wang et al., 2013) and that EEG resting-state activity has lower complexity (Bosl, Tierney, Tager-Flusberg, & Nelson, 2011). The authors suggest that these atypical EEG data might be explained by differences in how ASD brains, as opposed to TD brains, are connected. In particular, it has been found that the brains of people with ASD have an increased local connectivity, especially in the frontal cortex (Courchesne & Pierce, 2005).

The authors investigated this hypothesis with a spiking neural network by modifying the network's connection patterns and observing how the connectivity affected the emerged activation patterns. To manipulate the *degree of local over-connectivity in the network*, a parameter based on the small-world paradigm from Watts and Strogatz (1998) was used. By default, neurons are connected to six neighboring neurons in a ring lattice as displayed in Fig. 16 (left). A parameter p_{WS} expresses the probability for each of the connections to rewire to other neurons. Thus, p_{WS} determines the randomness of the network (Fig. 16), ranging from regular lattice structure ($p_{WS} = 0$) to random wiring ($p_{WS} = 1$). Medium values of p_{WS} around 0.2 describe "typically developed networks" with local clusters and some short-range connections between the clusters. Notably, the parameter from Watts and Strogatz (1998) keeps the overall number of connections in the network intact, such that differences emerge only due to differences in the network structure, not by the total number of neurons or neural connections.

8 EEG: Electroencephalography.

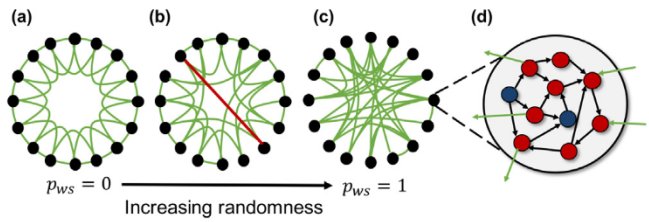


Fig. 16. Three different networks with different degrees of randomness. (a) is a locally over-connected network (corresponding to ASD individuals), (b) is a small-world network with many local clusters and a few longer connections (corresponding to typically developed individuals), (c) is a random network including many wide-range connections. (d) shows the structure of each single neuron group with excitatory (red) and inhibitory (blue) connections. Note that the number of nodes and edges in (a), (b) and (c) remains the same. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)
Source: Reprinted with permission from Park et al. (2019), originally based on Watts and Strogatz (1998).

Networks are formed by generating 100 groups of neurons, corresponding to the black nodes in Fig. 16. Each group contains 1000 spiking neurons: 800 excitatory and 200 inhibitory neurons, which have an increasing or decreasing effect on the firing probability of postsynaptic neurons, respectively. Neurons are mainly connected to neurons of the same neuron group (intra-group connections), and have connections to six neighboring groups according to Fig. 16 (inter-group connections). Different rewiring probabilities p_{WS} between 0 and 1 are used to determine the initial inter-group connectivity of the network.

After initialization, the network updates its connections according to the rules of spike-time-dependent plasticity (Izhikevich & Desai, 2003): the update of connection weights occurs depending on the timing of firing of the pre- and postsynaptic neurons. If the postsynaptic neuron fires within a certain time window *after the presynaptic neuron*, the weight of the connection is increased (corresponding to the biological process of *long term potentiation*). If the presynaptic neuron fires within a time window *after the postsynaptic neuron*, the connection weight is weakened (*long term depression*). During this learning period the connection weights self-organize. Tonic random input is presented to the network. After learning, the spontaneous activity of the neurons was recorded (in the absence of input), and compared to the graph-theoretical properties of the network.

The activation patterns were evaluated according to their frequency spectrum and the complexity of the time series, as measured by the multiscale entropy (Costa, Goldberger, & Peng, 2005). This measure rates the informative content of time series at different temporal scales. High complexity corresponds to the presence of long-range correlations on multiple scales in space and time, low complexity is computed for time-series with perfect regularity or randomness. The evaluation suggested that networks exhibiting local over-connectivity generate more oscillations in high-frequency bands and exhibit lower complexity in the signals than small-world networks. Findings of atypical resting-state EEG for people with ASD, thus, might be explained by local over-connectivity in their brains.

5.5. Bayesian approaches

There are promising models in the literature interpreting ASD on the basis of the Bayesian framework (for an introduction see Schizophrenia section). However, most of these approaches are only conceptual and still lack an implementation. Nevertheless, these approaches are able to explain a wide range of different symptoms which might be caused by an atypical integration of prediction and sensory information (Lawson et al., 2014; Pellicano & Burr, 2012).

The first approach utilizing the Bayesian brain hypothesis for explaining the non-social symptoms of ASD was proposed by Pellicano and Burr in 2012 (Pellicano & Burr, 2012). Their *hypo-prior hypothesis*[9] suggests that broader or less precise priors cause people with ASD to rely less on their predictions and stronger on sensory input which could explain the hypersensitivity of people with ASD. J. Brock broadened this idea (Brock, 2012) by proposing that hypersensitivity cannot only be caused by a reduced precision of the prior, but also by an increased precision of sensory input. Lawson et al. (2014) summarized these ideas, arguing that both modifications *reduced prior precision or increased sensory precision*, can cause the same functional consequences. They suggest that the cause could be aberrant precision in general: Expected precision of a signal is an important source of information that helps us to decide whether to rely on this signal or not. Aberrant precision of sensory input or prior predictions, thus, would alter the way in which we integrate these signals. The precision of the signals also can be considered as a weighting term of the prediction error: For a signal that is expected to be imprecise, a prediction error does not need to be corrected while a prediction error arising between signals that are expected to be very precise would need correction. People with ASD might have problems to accurately estimate this precision. Thus, they might, at the one extreme, try to minimize the prediction error too strongly, or, at the other extreme, fail to minimize the prediction error.

Finally, in Lawson, Mathys, and Rees (2017), Lawson and colleagues suggested that subjects with ASD overestimate the volatility of the environment. They conducted a behavioral experiment which demonstrated that ASD subjects are less surprised when encountering environmental changes. Using Hierarchical Gaussian Filters, they modeled the experimental findings computationally. The model parameter that best accounts for the differences found in ASD and neurotypical subjects was a meta-parameter which controlled learning about volatility of the environment. These results suggest that ASD subjects overestimate the probability of a change in the environmental conditions, and build less stable expectations. As a result, they might misinterpret an event with low probability which occurred by chance as an event that signifies a change in environmental conditions. Therefore, instead of being surprised in the case of an extraordinary event, they would be mildly surprised at all times.

5.6. Recurrent neural networks

The studies presented here follow the idea of predictive coding which can be seen as an implementation of the Bayesian brain idea: an RNN is used as an internal model of the world and its learning corresponds to the process of adapting network weights in order to perform prediction error minimization. The role of the network is to learn to predict sensory consequences, and integrates these predictions with the perceived sensory information.

5.6.1. Freezing and repetitive behavior in a robotics experiment

Idei and colleagues (Idei et al., 2017, 2018) used the stochastic continuous-time recurrent neural network (S-CTRNN) (Murata, Namikawa, Arie, Sugano, & Tani, 2013) model with parametric bias (PB) (Tani, 2003) to teach a robot to interact with a human

[9] In this article, we stick to the original definition of hypo-priors as a belief in low precision of priors and hyper-priors as a belief in high precision of priors. Note, however, that due to the hierarchical structure of the brain and the role of precision as a hyperparameter for the inference process it might be more appropriate to talk of hypo-priors as attenuated hyperpriors as argued in Friston, Lawson, and Frith (2013).

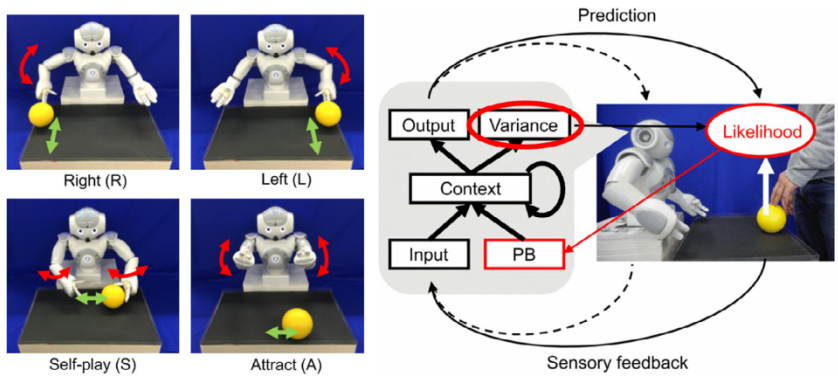


Fig. 17. Left: Overview of the interactive tasks the robot must perform. Right: Overview of the ANN model used for the experiments. Highlighted in red are the variance units where a constant K is added to increase or decrease the sensory precision in order to imitate autistic behavior. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)
Source: Adapted from Idei et al. (2017).

in a ball-playing game (similar to the schizophrenia model Yamashita & Tani, 2012). The S-CTRNN with PB learns to predict a time series of proprioceptive (joint angles) and vision features. From the current input, the network estimates the next time step (output) and its predicted precision (variance) as shown in Fig. 17. The state of the PB units reflects the intention of the network, i.e., the ball-playing pattern that the robot believes that they are currently engaged in.

The S-CTRNN was trained offline to perform certain tasks depending on a yellow ball's position, as depicted in Fig. 17 (left). Synaptic weights and biases of the network, as well as the internal states of the PB units are updated via the backpropagation through time (BPTT) algorithm in order to maximize the likelihood in Eq. (13). This equation describes that at time step t of training sequence s, the network output of the ith neuron (a normal distribution defined by the estimated mean (output) y and estimated variance v) properly reflects the desired input data $\hat{y}$.

$$L_{t,i}^{(s)} = -\frac{\ln(2\pi v_{t,i}^{(s)})}{2} - \frac{(\hat{y}_{t,i}^{(s)} - y_{t,i}^{(s)})^2}{2v_{t,i}^{(s)}} \tag{13}$$

After training, a recognition mechanism (via adaptation of the PB units, while keeping weights and biases fixed) enables the network to switch its behavior depending on the current situation.

To model ASD behavior, the *estimated variance (sensory precision) is modified* in the activation function of the variance units with the constant K in Eq. (14), where ϵ is the minimum value and $u_{t,i}^{(s)}$ is the output of the ith context unit time step t for movement sequence s.

$$v_{t,i}^{(s)} = \exp(u_{t,i}^{(s)} + K) + \epsilon \tag{14}$$

Experimental results with a humanoid NAO robot showed that for $K = 0$ the robot behaved normally. For increased variance (reduced precision), the robot seemed to ignore prediction error and performed stopping and stereotypic movements. For decreased variance (increased precision), the robot performed incorrect movement changes or concentrated on certain movements, which also led to sudden freezing and repetitive movements. These results fit with the disordered motor system reported in ASD (Gowen & Hamilton, 2013), but add the surprising insight that increased and decreased sensory precision may cause the same consequences.

5.6.2. Impairment in internal network representations

Another study using the S-CTRNN to model ASD characteristics is Philippsen and Nagai (2018). Using an S-CTRNN (Murata et al., 2013), the authors modify two parameters which control how the network makes predictions. In contrast to the other RNN model which concentrates on replicating behavioral patterns, this study investigates "invisible" features characterizing the network's learning process. More specifically, the authors evaluate *how attention to sensory input and deficits in the prediction of trajectory noise influence the internal representation* that a network acquires during learning. Internal representations are informative as they reflect the network's generalization capabilities (Boden, 2002; Yamashita & Tani, 2008): similar input pattern should cause an overlap in the corresponding context neuron activations (attractors in the RNN), whereas different patterns should be differentiated.

The network as displayed in Fig. 18 is trained to recognize and draw ellipses and "eight" shapes, located at four different (overlapping) positions of the input space (cf. Fig. 19(b)). Inputs and outputs are two-dimensional trajectories and the recurrent context layer comprises 70 neurons. Learning is modified in two ways: The parameter χ determines how much the network relies on external input, as opposed to its own prediction, i.e., χ gradually switches between open-loop ($\chi = 1$) and near-closed-loop ($\chi \approx 0$) control. The second parameter K is defined analogous to Idei et al. (2017) (see Eq. (14)) and manipulates the estimated variance such that networks with $K \neq 0$ over- or underestimate noisy variations in the signal. Unlike its usage in Idei et al. (2017), this manipulation is not performed after training, but already *during* the training process, to account for the developmental nature of ASD.

After training, the network's behavior is evaluated as the network's ability to reproduce the trained trajectories. The internal representations are evaluated by collecting the time course of activations of the context layer neurons while generating the trajectories.

A visualization of how the high-dimensional space (time steps $\times$ number of context neurons) is structured can be achieved by principal component analysis (PCA). The results indicated that networks tend to reuse internal representation structure for patterns located at the same position in the input space. Such an overlap is advantageous as similarities between patterns are coded. However, too strong overlap of the context activations indicates missing differentiation between the patterns which might lead to worse differentiation in a recognition task. Thus, the authors define "good" internal network representations as representations which strongly reflect the characteristics of the input data. Fig. 19(b) shows an example of how task performance (top) and internal representation quality (bottom) change depending on the external contribution parameter. The best internal

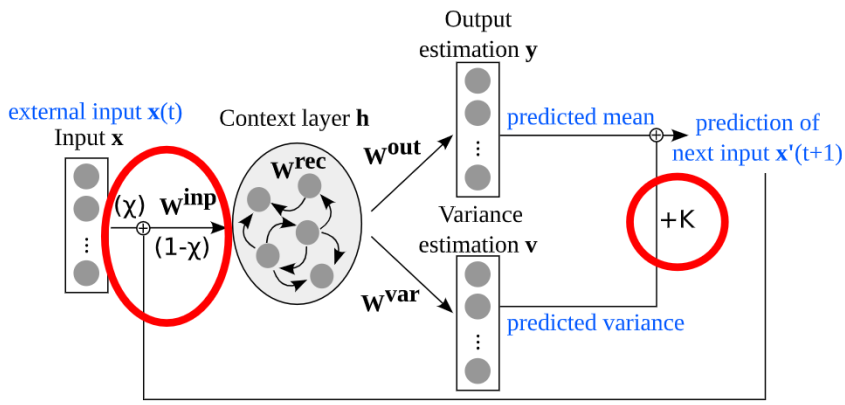


Fig. 18. The S-CTRNN used in Philippsen and Nagai (2018) with two parameter modifications.
Source: Adapted from Philippsen and Nagai (2018).

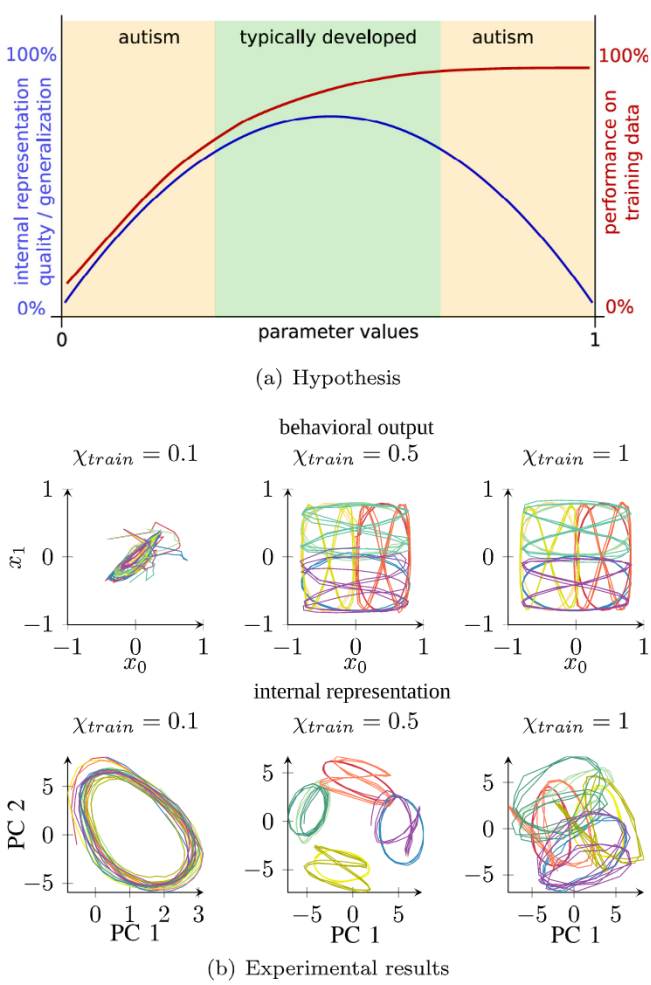


Fig. 19. Effect of changing the external contribution parameter of the S-CTRNN from Fig. 18 on behavioral output (top) and on internal representation quality, evaluated in the two-dimensional principal components (PC) space (bottom).
Source: Adapted from Philippsen and Nagai (2018).

representation quality is achieved with $\chi = 0.5$ (moderately integrating input and predictions), as the internal representation reuses activations but clearly differentiates trajectories at different input space positions. However, the performance in reproducing the trained behavior is comparable between $\chi = 0.5$ and $\chi = 1$ (relying stronger on input). These qualitative observations were also quantitatively verified in the high-dimensional space of neurons. How well the network is able to reproduce the learned patterns, thus, is not always reflected in the internal representation quality.

Interestingly, for the parameter χ, both extremes lead to an ASD-like impairment, as schematically depicted in Fig. 19(a). Typical development could correspond to the middle. Whereas the right-hand side would express high-functioning ASD where the performance in specific tasks might be intact, but representations might be too specific (overfitting). The left-hand side describes ASD with severe impairments also at a behavioral level. It can be, thus, imagined that heterogeneity in the ASD population, comprising opposite symptoms such as hyper- and hyposensitivity, does not necessarily be caused by different underlying mechanisms, but that a continuous modification of parameters could account for the variability.

5.6.3. Generalization ability in a variational Bayes recurrent neural network

In Ahmadi and Tani (2017), a novel recurrent network type is introduced, the variational Bayes predictive coding RNN (VBP-RNN). It differs from the S-CTRNN in that variance is not only coded on the output level, but also in the network's context neurons to enhance the network's ability to represent uncertainty in the data.

We do not discuss it in detail here, as this study is not focusing on modeling ASD, but on representing deterministic as well as probabilistic behavior in an RNN in a coherent way. The analogy to ASD is made in terms of the *meta-parameter W that performs a trade-off between reconstruction and regularization in the optimization (loss) function.* W switches between the typically minimized reconstruction error term ($W = 0$) and a regularization term that keeps the posterior distribution of the latent variables (i.e., the context units) similar to its prior. If the network is trained with $W = 0$, it develops deterministic dynamics and exhibits poor generalization capabilities. Values of $W > 0$ lead to more randomness in the network and improve generalization, but too high values result in a performance drop.

W could therefore model the spectrum of ASD: $W = 0$ is one extreme where the network solely relies on its top-down intentionality and fails to generalize, whereas too high values of W reflect performance impairment due to excessive randomness in the network. As this parameter controls how much regularization is performed, the approach is similar to Dovgopoly and Mercado (2013) where regularization was intentionally impaired.

5.7. Other approaches

In 2000, OLaughlin and Thagard (2000) used a connectionist model to simulate weak coherence, and to demonstrate how a failure of maximizing global coherence can cause deficits in theory of mind (Baron-Cohen, 1997). Their network model, a so-called constraint network, is hand-designed according to the task and does not strictly fit into an existing network category. The network performs logical reasoning and consists of a set of neurons, each of which corresponds to a logical element such as a belief (expressed as a sentence). Connections between them are set as excitatory and inhibitory, depending on whether two arguments support each other or are contradicting. Weights remain fixed, but the activations of neurons get updated depending on the connections to neighboring cells which can be excitatory (positive) on inhibitory (negative). A decaying factor lets the network's activation converge to a state after a certain amount of time. Positive activations are then interpreted as an acceptance of this belief, negative activations as a denial.

The authors showed that *a high level of inhibition, compared to excitation*, causes early activated association nodes in the network to suppress concurring hypotheses. The network, therefore, prefers more direct solutions, and makes wrong predictions. The overall *coherence* of the network, defined as the satisfaction of most constraints in the complete network, is not optimized, which can be considered as weak coherence.

6. Discussion and future directions

Artificial neural network models of SZ and ASD have been presented as a useful tool to fill the gap between theoretical models and biological evidence. Early works were biased by technical restrictions, but recent models are able to capture the same complexity as conceptual models, such as hierarchical Bayesian models. However, designing ANN architectures that are able to predict novel findings and through computational simulations contribute to clinical applications (e.g., diagnosis or therapy) remains a challenging task. In this sense, the model should (i) reproduce empirical behavioral findings, preferably in more than one domain, (ii) be supported by a process theory in which the abnormality used to reproduce empirical findings is realistic from the point of view of known neuropathology, and (iii) predict novel findings. Furthermore, addressing heterogeneity and nonspecificity is still one of the most important challenges of these two psychiatric disorders.

Due to the large overlap in SZ and ASD regarding biological evidence (e.g., E/I imbalance), similar hypotheses were discussed as a potential cause for both disorders. Computational models, however, still tend to focus on specific impairments of a specific disorder. To help the community, it is crucial that overarching neural network models are developed which connect ideas and results across different contexts (ASD, SZ or even other mental disorders).

In this section, we first discuss the quality of the discussed models in terms of how well they fit and predict empirical findings (Section 6.1). Secondly, we discuss the approaches from the point of view of multifinality and equifinality (Section 6.2). Thirdly, we emphasize the importance of testing the models in an embodied system (Section 6.3). Finally, we describe new promising directions to address with ANN models: developmental factors (Section 6.4.1); disorders of the self (Section 6.4.2); and state-of-the-art ANN architectures for future models of psychopathologies (Section 6.4.3).

6.1. Models quality: Empirical findings and predictability

Early SZ modeling works from Hoffman and Dobscha (1989) and Ruppin, Reggia, and Horn (1995) on Hopfield networks as well as the feed-forward approaches from Cohen and Servan-Schreiber (1992) and Hoffman and McGlashan (1997) lack the capabilities to generalize to a broader context: every experiment required a different ANN architecture. Hence, in terms of predictability of other symptoms, these approaches are not powerful enough. In particular, the work on auditory hallucinations (Hoffman & McGlashan, 1997) is far from replicating the brain mechanism and does not account for deficits in distinguishing self-produced sounds observed in SZ patients. However, the underlying discussion presented in those papers still provides valuable insights. They highlighted the connectivity factor between different cortical areas of the brain (either by gain reduction or pruning) specially in the context ones. Later works on RNN, such as Yamashita and Tani (2012), revisited this idea with hierarchical networks, with the same capability to generate parasitic states due to dynamic attractors. Pruning was substituted by noise injection. Interestingly, there are conceptual similarities between noise injection and precision reduction used in Bayesian approaches. Due to the more general architecture regarding sensorimotor integration, this RNN might be able to replicate other findings in earlier works such as hallucinations or performance in the Stroop task, however, this has not been experimentally demonstrated yet.

Bayesian approaches, such as predictive processing (Adams et al., 2013) and circular inference (Jardri & Deneve, 2013) have shown better quality in terms of predictability of new empirical findings. Their mathematical abstraction is more powerful and may be applicable to different types of experiments. For instance, within the free-energy optimization framework, eye-tracking deficits with occlusion and agency attribution disorders were investigated. The circular inference model with E/I imbalance predicted findings in decision-making tasks involving likelihoods (e.g., Fisher task). However, due to the conceptual design, their scalability is really poor for handling real sensory information. Here we find that ANNs, such as convolutional network approaches (Nagai et al., 2015) or Variational-Bayes RNN (Ahmadi & Tani, 2017) could better account for real sensory data input.

Just as Hopfield networks were applied for modeling SZ, some early models of ASD focused on SOM approaches. These models (Gustafsson, 1997; Gustafsson & Papliński, 2004; Noriega, 2007, 2008) could account for strong specificity in cortical representations or novelty avoidance. Despite that, they were highly linked to the specific network architecture, and thus, it is difficult to use these mechanisms to predict performance in other types of tasks. More general approaches were suggested using simple parameter modifications of feed-forward neural networks (Cohen, 1994, 1998; Dovgopoly & Mercado, 2013). These parameters rather utilize general engineering mechanisms of neural networks and, thus, are also applicable to different architectures (e.g., regularization was also used in a recent approach using RNNs Ahmadi & Tani, 2017). These studies mostly focused on replicating the specific symptom of generalization deficits, but may not be applicable to explaining a broader range of symptoms.

The reviewed models of SZ only addressed positive symptoms mainly hallucinations, delusions and abnormal movements. Self-other disturbances have been only discussed in the free-energy models and negative symptoms have been set aside. Within the ASD models only repetitive motor movements and hyper/hyporeactivity to sensory input were properly discussed. Furthermore, social communication and interaction deficits have been minimally addressed.

Interestingly, for ASD (Ahmadi & Tani, 2017; Idei et al., 2017; Pellicano & Burr, 2012; Philippsen & Nagai, 2018) as well as for SZ (Adams et al., 2013; Jardri & Deneve, 2013; Yamashita & Tani, 2012), the majority of recent approaches incorporate the idea of predictive coding (Rao & Ballard, 1999). In particular, Pellicano and Burr's paper (Pellicano & Burr, 2012) and novel hypotheses based on their theory (Lawson et al., 2017, 2014) significantly influence the recent developments. In terms of finding a general account for cognition, predictive coding and related approaches are the most promising candidates right now. Therefore, predictive coding based approaches can be considered a useful abstraction in developing a broader model that is able to integrate typical and atypical development in a coherent whole.

6.2. Multifinality, equifinality and heterogeneity

A challenge in modeling psychopathologies is the non-specificity of these disorders. Different biological bases may lead to the same symptom (equifinality). Therefore, many modeling mechanisms might be valid for modeling a single symptom. Accordingly, the studies reviewed here cover a wide range of approaches, using various pieces of biological evidence. This variety has its drawback: even if a model can explain some symptoms, we cannot judge whether this mechanism actually is comparable to what happens in the human brain or not.

The non-specificity of psychopathologies also means that a single biological basis can cause different symptoms (multifinality). Thus, instead of targeting single symptoms, it is important to develop models which explain several symptoms of a disorder. A good starting point is to first model typical behavior. One possible basis could be ANN models of sensorimotor integration. According to the majority of the computational models discussed in this manuscript, SZ and ASD are presented as disorders of sensory information fusion or interpretation. Thus, general ANN sensorimotor integration models that are able to fit human-like data (control and patient data) in different experimental paradigms such as body perceptual tests or decision making task could be extended to model psychopathologies.

Additionally, modeling mechanisms should not only cover various symptoms of a single disorder, but they may also be used for modeling similar symptoms in different disorders. For instance, hallucinations are present in several disorders but researchers used different ANN approaches to model them. Hallucinations produced by a loss of sensory input, like in the Charles Bonnet syndrome, were studied by modeling homeostasis in a Deep Boltzmann machine (DBM) for visual (Series, Reichert, & Storkey, 2010) and tactile inputs (Deistler, Yener, Bergner, Lanillos, & Cheng, 2019). However, homeostasis or DBMs were never studied for hallucinations in SZ, or discussed within circular inference or free-energy approaches (Adams et al., 2013).

Regarding heterogeneity, recent studies modeling ASD already acknowledge the nature of ASD as a spectrum. Instead of distinguishing between impaired and intact behavior as two categories, a continuous change in symptoms is suggested, leading to impairments of different severeness (Idei et al., 2017; Nagai, 2019) or even opposite types of impairments (Philippsen & Nagai, 2018). This offers a potentially more sophisticated view on heterogeneity in ASD.

6.3. Models validation on real robotic systems

We presented some works that employ robotics systems' validation as a useful servant for the behavior unit/level of analysis (Yamashita & Tani, 2012). The relevant aspect of these approaches is that the internal mechanism of the behavior is visible (Cheng et al., 2007). Furthermore, a connection can be made from rather perceptual or mechanistic impairments inside the system to difficulties in real interaction scenarios. For instance, Murata et al. (2013) replicated freezing and repetitive behaviors on a robot. Most of the discussed models, however, are solely data models. Closing the gap to real world embodied models could, therefore, help to validate how these models extend to other tasks.

ANN approaches can also focus on solving scalability to raw stimuli in other brain-inspired mathematical abstractions. For instance, Lanillos and Cheng (2018) and Oliver, Lanillos, and Cheng (2019) presented free-energy-based perception and action algorithms working on humanoid robots. They can be used to evaluate atypical behaviors related to body perception in SZ and ASD.

6.4. New directions

We identified the following three research directions that are still underrepresented in the discussed studies.

6.4.1. Developmental factors

Developmental factors are especially relevant for ASD as a developmental disorder, but also for SZ. Specially, to explain why many cases of SZ emerge during adolescence and early adulthood (Feinberg, 1982; Huttenlocher et al., 1979; Keshavan et al., 1994) and to investigate developmental factors which might contribute to the onset of SZ (Cannon, 2015). Current models only partially take the developmental process into account and focus more on modeling existing deficits in adult subjects with ASD. For instance, existing models assume an aberrant number of neurons (Cohen, 1994; Noriega, 2007) or differences in the neural connections (Ichinose et al., 2017; Park et al., 2019) during the development, or they change the way that learning proceeds by altering network regularization (Ahmadi & Tani, 2017; Dovgopoly & Mercado, 2013) or how information are integrated during learning (Philippsen & Nagai, 2018). However, these studies still cannot answer the question of which initial causes promote the appearance of ASD during the development. It might be beneficial to take even one step more back in development, back to the development of the human fetus. For instance, a recent study (Yamada et al., 2016) suggests that disordered intrauterine embodied interaction during fetal period is a possible factor for neuro-developmental disorders like ASD.

6.4.2. SZ and ASD as disorders of the self

One of the aspects not properly addressed in ANN computational modeling, neither for SZ nor for ASD, is how diagnosed individuals experience their body and self in comparison with control subjects. For instance, SZ patients have troubles differentiating self-produced actions. In fact, modeling the spectrum of differences in body experience could make several psychopathologies comparable. In addition to already described visual illusions, also body illusions can be investigated. Recently, Noel, Cascio, Wallace, and Park (2017) discussed how body perception differs between ASD and SZ individuals, suggesting a sharper boundary between self and other in ASD and a weaker boundary in SZ. This suggestion is based on experimental findings, for example, on peripersonal space in body illusions where "opposite" results were found: whereas individuals with SZ were more prone to have body illusions (Thakkar, Nichols, McIntosh, & Park, 2011), individuals with ASD showed a reduced illusionary effect (Cascio, Foss-Feig, Burnette, Heacock, & Cosby, 2012). Hence, the causes of these psychopathologies have a direct impact on the perception of our body and the self. In the case of patients diagnosed with SZ, this relation has been more intensively studied (Stanghellini, 2009) and some treatments include

embodiment therapies. Hence, models of the bodily or sensorimotor self (Hinz et al., 2018; Lanillos, Dean-Leon, & Cheng, 2017) that are able to explain body illusions would help to validate the hypothesis in a common framework. Behavioral measures like the proprioceptive drift or peripersonal space should be also predicted by the model. For instance, in Hinz et al. (2018), they used the perceptual drift as a measure to evaluate the validity of a predictive coding model (Lanillos & Cheng, 2018) for typical individuals.

6.4.3. ANN novel architectures for psychopathologies

In terms of neural network architectures, there is a further need of transferring the knowledge from state-of-the-art recurrent neural networks and deep learning to neurological disorders as it was performed, for instance, with the Neocognitron model of ASD (Nagai et al., 2015) or the MTRNN model of SZ (Yamashita & Tani, 2012). Theoretical ANN studies, computational psychiatry and neuroscience should be always be in contact to boost the feedback of those disciplines.

In opposition to Bayesian models that are implemented on a high abstraction level of the task, modern ANN approaches (Schmidhuber, 2015) are able to cope with real sensor data such as visual information. For instance, cross-modal learning architectures combined with hierarchical representation learning provide an interesting follow-up to early ANN studies on SZ and ASD. Furthermore, ANN models of Bayesian brain such as predictive coding (Yamashita & Tani, 2008) and circular inference are a basis for uniting both communities. In fact, recent advances in probabilistic NNs like Variational Autoencoders (Kingma & Welling, 2013) and Variational-RNN (Ahmadi & Tani, 2017; Fabius & van Amersfoort, 2014), provide the mathematical framework to deploy ANN versions of prominent plausible models of the brain such as the free-energy principle (Friston, 2010).

In this review, we showed the power of ANNs for modeling symptoms of neurological disorders. However, these techniques need to be further developed and refined in the future to play a key role in computational psychiatry and to contribute in clinical applications.

Acknowledgments

This work was supported by SELFCEPTION project (www.selfception.eu) European Union Horizon 2020 Programme (MSCA-IF-2016) under grant agreement no. 741941, Japan Science and Technology Agency CREST Cognitive Mirroring (grant no. JPMJCR16E2), and by JSPS, Japan KAKENHI Grant No. JP17H06039, JP18K07597 and JP18KT0021.

References

Adams, R. A. (2018). Bayesian Inference, predictive coding, and computational models of psychosis. In *Computational Psychiatry* (pp. 175–195). Elsevier.

Adams, R. A., Huys, Q. J., & Roiser, J. P. (2016). Computational psychiatry: towards a mathematically informed understanding of mental illness. *Journal of Neurology, Neurosurgery and Psychiatry*, *87*(1), 53–63.

Adams, R. A., Stephan, K. E., Brown, H. R., Frith, C. D., & Friston, K. J. (2013). The computational anatomy of psychosis. *Frontiers in Psychiatry*, *4*, 47.

Ahmadi, A., & Tani, J. (2017). Bridging the gap between probabilistic and deterministic models: a simulation study on a variational Bayes predictive coding recurrent neural network model. In *International conference on neural information processing* (pp. 760–769). Springer.

Anderson, J. S., Druzgal, T. J., Froehlich, A., DuBray, M. B., Lange, N., Alexander, A. L., et al. (2010). Decreased interhemispheric functional connectivity in autism. *Cerebral Cortex*, *21*(5), 1134–1146.

Anticevic, A., Murray, J. D., & Barch, D. M. (2015). Bridging levels of understanding in schizophrenia through computational modeling. *Clinical Psychological Science*, *3*(3), 433–459.

Association, A. P., et al. (2013). *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub.

Auerbach, B. D., Osterweil, E. K., & Bear, M. F. (2011). Mutations causing syndromic autism define an axis of synaptic pathophysiology. *Nature*, *480*(7375), 63.

Aukes, M. F., Alizadeh, B. Z., Sitskoorn, M. M., Selten, J.-P., Sinke, R. J., Kemner, C., et al. (2008). Finding suitable phenotypes for genetic studies of schizophrenia: heritability and segregation analysis. *Biological Psychiatry*, *64*(2), 128–136.

Aylward, E. H., Minshew, N. J., Field, K., Sparks, B., & Singh, N. (2002). Effects of age on brain volume and head circumference in autism. *Neurology*, *59*(2), 175–183.

Baio, J., Wiggins, L., Christensen, D. L., Maenner, M. J., Daniels, J., Warren, Z., et al. (2018). Prevalence of autism spectrum disorder among children aged 8 years – autism and developmental disabilities monitoring network, 11 sites, united states, 2014. *MMWR Surveillance Summaries*, *67*(6), 1.

Bányai, M., Diwadkar, V. A., & Érdi, P. (2011). Model-based dynamical analysis of functional disconnection in schizophrenia. *Neuroimage*, *58*(3), 870–877.

Baron-Cohen, S. (1997). *Mindblindness: An essay on autism and theory of mind*. MIT press.

Bauman, M. L. (1991). Microscopic neuroanatomic abnormalities in autism. *Pediatrics*, *87*(5), 791–796.

Bauman, M. L., & Kemper, T. L. (2005). Neuroanatomic observations of the brain in autism: a review and future directions. *International Journal of Developmental Neuroscience*, *23*(2–3), 183–187.

Behrmann, M., Avidan, G., Leonard, G. L., Kimchi, R., Luna, B., Humphreys, K., et al. (2006). Configural processing in autism and its relationship to face processing. *Neuropsychologia*, *44*(1), 110–129.

Boden, M. (2002). A guide to recurrent neural networks and backpropagation. *The Dallas Project*.

Bosl, W., Tierney, A., Tager-Flusberg, H., & Nelson, C. (2011). EEG Complexity as a biomarker for autism spectrum disorder risk. *BMC Medicine*, *9*(1), 18.

Bourgeron, T. (2009). A synaptic trek to autism. *Current Opinion in Neurobiology*, *19*(2), 231–234.

Brock, J. (2012). Alternative Bayesian accounts of autistic perception: comment on pellicano and burr. *Trends in Cognitive Sciences*, *16*(12), 573–574.

Canitano, R., & Pallagrosi, M. (2017). Autism spectrum disorders and schizophrenia spectrum disorders: excitation/inhibition imbalance and developmental trajectories. *Frontiers in Psychiatry*, *8*, 69.

Cannon, T. D. (2015). How schizophrenia develops: cognitive and brain mechanisms underlying onset of psychosis. *Trends in Cognitive Sciences*, *19*(12), 744–756.

Caruana, R., Lawrence, S., & Giles, C. (2001). Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. In *Advances in neural information processing systems* (pp. 402–408).

Casanova, M. F., van Kooten, I. A., Switala, A. E., van Engeland, H., Heinsen, H., Steinbusch, H. W., et al. (2006). Minicolumnar abnormalities in autism. *Acta Neuropathologica*, *112*(3), 287.

Cascio, C. J., Foss-Feig, J. H., Burnette, C. P., Heacock, J. L., & Cosby, A. A. (2012). The rubber hand illusion in children with autism spectrum disorders: delayed influence of combined tactile and visual input on proprioception. *Autism*, *16*(4), 406–419.

Chapman, L., Chapman, J. P., & Miller, G. A. (1964). A theory of verbal behavior in schizophrenia.. *Progress in Experimental Personality Research*, *72*, 49.

Cheng, G., Hyon, S.-H., Morimoto, J., Ude, A., Hale, J. G., Colvin, G., et al. (2007). CB: A humanoid research platform for exploring neuroscience. *Advanced Robotics*, *21*(10), 1097–1114.

Church, B. A., Krauss, M. S., Lopata, C., Toomey, J. A., Thomeer, M. L., Coutinho, M. V., et al. (2010). Atypical categorization in children with high-functioning autism spectrum disorder. *Psychonomic Bulletin & Review*, *17*(6), 862–868.

Cohen, I. L. (1994). An artificial neural network analogue of learning in autism. *Biological Psychiatry*, *36*(1), 5–20.

Cohen, I. (1998). Neural network analysis of learning in autism. *Neural Networks and Psychopathology*, 274–315.

Cohen, J. D., & Servan-Schreiber, D. (1992). Context, cortex, and dopamine: a connectionist approach to behavior and biology in schizophrenia.. *Psychological Review*, *99*(1), 45.

Cohen, I. L., Sudhalter, V., Landon-Jimenez, D., & Keogh, M. (1993). A neural network approach to the classification of autism. *Journal of Autism and Developmental Disorders*, *23*(3), 443–466.

Cornblatt, B. A., Lenzenweger, M. F., & Erlenmeyer-Kimling, L. (1989). The continuous performance test, identical pairs version: Ii. contrasting attentional profiles in schizophrenic and depressed patients. *Psychiatry Research*, *29*(1), 65–85.

Costa, M., Goldberger, A. L., & Peng, C.-K. (2005). Multiscale entropy analysis of biological signals. *Physical Review E*, *71*(2), 021906.

Courchesne, E., Karns, C., Davis, H., Ziccardi, R., Carper, R., Tigue, Z., et al. (2001). Unusual brain growth patterns in early life in patients with autistic disorder an MRI study. *Neurology*, *57*(2), 245–254.

Courchesne, E., & Pierce, K. (2005). Why the frontal cortex in autism might be talking only to itself: local over-connectivity but long-distance disconnection. *Current Opinion in Neurobiology*, *15*(2), 225–230.

Courchesne, E., Townsend, J., Akshoomoff, N. A., Saitoh, O., Yeung-Courchesne, R., Lincoln, A. J., et al. (1994). Impairment in shifting attention in autistic and cerebellar patients.. *Behavioral Neuroscience*, *108*(5), 848.

Crick, F., Mitchison, G., et al. (1983). The function of dream sleep. *Nature*, *304*(5922), 111–114.

Dallenbach, K. M. (1951). A puzzle-picture with a new principle of concealment. *The American Journal of Psychology*, 431–433.

Daniels, J. L., Forssen, U., Hultman, C. M., Cnattingius, S., Savitz, D. A., Feychting, M., et al. (2008). Parental psychiatric disorders associated with autism spectrum disorders in the offspring. *Pediatrics*, *121*(5), e1357–e1362.

Dayan, P., Hinton, G. E., Neal, R. M., & Zemel, R. S. (1995). The helmholtz machine. *Neural Computation*, *7*(5), 889–904.

Deistler, M., Yener, Y., Bergner, F., Lanillos, P., & Cheng, G. (2019). Tactile hallucinations on artificial skin induced by homeostasis in a deep boltzmann machine. arXiv preprint arXiv:1906.10592.

Dickinson, A., Jones, M., & Milne, E. (2016). Measuring neural excitation and inhibition in autism: different approaches, different findings and different interpretations. *Brain Research*, *1648*, 277–289.

Dovgopoly, A., & Mercado, E. (2013). A connectionist model of category learning by individuals with high-functioning autism spectrum disorder. *Cognitive, Affective, & Behavioral Neuroscience*, *13*(2), 371–389.

Elman, J. L. (1990). Finding structure in time. *Cogn. Science*, *14*(2), 179–211.

Fabius, O., & van Amersfoort, J. R. (2014). Variational recurrent auto-encoders. arXiv preprint arXiv:1412.6581.

Feinberg, I. (1982). Schizophrenia: caused by a fault in programmed synaptic elimination during adolescence?. *Journal of Psychiatric Research*, *17*(4), 319–334.

Friston, K. J. (1998). The disconnection hypothesis. *Schizophrenia Research*, *30*(2), 115–125.

Friston, K. (2010). The free-energy principle: a unified brain theory?. *Nature Reviews Neuroscience*, *11*(2), 127.

Friston, K. J., & Frith, C. D. (1995). Schizophrenia: a disconnection syndrome. *Clinical Neuroscience*, *3*(2), 89–97.

Friston, K. J., Lawson, R., & Frith, C. D. (2013). On hyperpriors and hypopriors: comment on pellicano and burr. *Trends in Cognitive Sciences*, *17*(1), 1.

Frith, U. (2003). *Autism: Explaining the enigma*. Blackwell Publishing.

Frith, C. (2004). Is autism a disconnection disorder?. *The Lancet Neurology*, *3*(10), 577.

Frith, U., & Happé, F. (1994). Autism: Beyond "theory of mind". *Cognition*, *50*(1–3), 115–132.

Fukushima, K. (1988). Neocognitron: A hierarchical neural network capable of visual pattern recognition.. *Neural Networks*, *1*(2), 119–130.

Fukushima, K. (2003). Neocognitron for handwritten digit recognition. *Neurocomputing*, *51*, 161–180.

Fukushima, K., & Miyake, S. (1982). Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and cooperation in neural nets* (pp. 267–285). Springer.

Garmezy, N. (1977). The psychology and psychopathology of attention. *Schizophrenia Bulletin*, *3*(3), 360.

Gowen, E., & Hamilton, A. (2013). Motor abilities in autism: a review using a computational context. *Journal of Autism and Developmental Disorders*, *43*(2), 323–344.

Grasemann, U., Miikulainen, R., & Hoffman, R. (2007). A subsymbolic model of language pathology in schizophrenia. In *Proceedings of the annual meeting of the cognitive science society, Vol. 29*. (29).

Gustafsson, L. (1997). Inadequate cortical feature maps: A neural circuit theory of autism. *Biological Psychiatry*, *42*(12), 1138–1147.

Gustafsson, L., & Paplinski, A. P. (2004). Neural network modelling of autism. *Recent Developments in Autism Research*, 100–134.

Gustafsson, L., & Papliński, A. P. (2004). Self-organization of an artificial neural network subjected to attention shift impairments and familiarity preference, characteristics studied in autism. *Journal of Autism and Developmental Disorders*, *34*(2), 189–198.

Hahamy, A., Behrmann, M., & Malach, R. (2015). The idiosyncratic brain: distortion of spontaneous connectivity patterns in autism spectrum disorder. *Nature Neuroscience*, *18*(2), 302.

Happé, F. G. (1996). Studying weak central coherence at low levels: children with autism do not succumb to visual illusions. a research note. *Journal of Child Psychology and Psychiatry*, *37*(7), 873–877.

Happé, F., & Frith, U. (2006). The weak coherence account: detail-focused cognitive style in autism spectrum disorders. *Journal of autism and developmental disorders*, *36*(1), 5–25.

Hebb, D. O. (1949). *The organization of behavior: A neuropsychological theory*. John Wiley.

Henderson, C. M., & McClelland, J. L. (2011). A PDP model of the simultaneous perception of multiple objects. *Connection Science*, *23*(2), 161–172.

Henik, A., & Salo, R. (2004). Schizophrenia and the stroop effect. *Behavioral and Cognitive Neuroscience Reviews*, *3*(1), 42–59.

Hinz, N.-A., Lanillos, P., Mueller, H., & Cheng, G. (2018). Drifting perceptual patterns suggest prediction errors fusion rather than hypothesis selection: replicating the rubber-hand illusion on a robot. In *2018 Joint IEEE 8th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)* (pp. 125–132). http://dx.doi.org/10.1109/DEVLRN.2018.8761005.

Hoffman, R. E. (1987). Computer simulations of neural information processing and the schizophrenia-mania dichotomy. *Archives of General Psychiatry*, *44*(2), 178–188.

Hoffman, R. E., & Dobscha, S. K. (1989). Cortical pruning and the development of schizophrenia: a computer model. *Schizophrenia Bulletin*, *15*(3), 477–490.

Hoffman, R. E., Grasemann, U., Gueorguieva, R., Quinlan, D., Lane, D., & Miikkulainen, R. (2011). Using computational patients to evaluate illness mechanisms in schizophrenia. *Biological Psychiatry*, *69*(10), 997–1005.

Hoffman, R. E., & McGlashan, T. H. (1997). Synaptic elimination, neurodevelopment, and the mechanism of hallucinated "voices" in schizophrenia. *American Journal of Psychiatry*, *154*(12), 1683–1689.

Hoffman, R. E., & McGlashan, T. H. (2001). Book review: Neural network models of schizophrenia. *The Neuroscientist*, *7*(5), 441–454.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, *79*(8), 2554–2558.

Horn, D., & Ruppin, E. (1995). Compensatory mechanisms in an attractor neural network model of schizophrenia. *Neural Computation*, *7*(1), 182–205.

Huttenlocher, P. R., de Courten, C., Garey, L. J., & Van der Loos, H. (1982). Synaptogenesis in human visual cortex—evidence for synapse elimination during normal development. *Neuroscience Letters*, *33*(3), 247–252.

Huttenlocher, P. R., et al. (1979). Synaptic density in human frontal cortex – developmental changes and effects of aging. *Brain Research*, *163*(2), 195–205.

Huys, Q. J., Moutoussis, M., & Williams, J. (2011). Are computational models of any use to psychiatry?. *Neural Networks*, *24*(6), 544–551.

Ichinose, K., Park, J., Kawai, Y., Suzuki, J., Asada, M., & Mori, H. (2017). Local overconnectivity reduces the complexity of neural activity: toward a constructive understanding of brain networks in patients with autism spectrum disorder. In *2017 joint IEEE international conference on development and learning and epigenetic robotics (ICDL-EpiRob)* (pp. 233–238). http://dx.doi.org/10.1109/DEVLRN.2017.8329813.

Idei, H., Murata, S., Chen, Y., Yamashita, Y., Tani, J., & Ogata, T. (2017). Reduced behavioral flexibility by aberrant sensory precision in autism spectrum disorder: A neurorobotics experiment. In *Development and learning and epigenetic robotics (ICDL-EpiRob), 2017 joint IEEE international conference on* (pp. 271–276). IEEE.

Idei, H., Murata, S., Chen, Y., Yamashita, Y., Tani, J., & Ogata, T. (2018). A neurorobotics simulation of autistic behavior induced by unusual sensory precision. *Computational Psychiatry*, *2*, 164–182.

Izhikevich, E. M. (2003). Simple model of spiking neurons. *IEEE Transactions on Neural Networks*, *14*(6), 1569–1572.

Izhikevich, E. M., & Desai, N. S. (2003). Relating stdp to bcm. *Neural Computation*, *15*(7), 1511–1523.

Jardri, R., & Deneve, S. (2013). Circular inferences in schizophrenia. *Brain*, *136*(11), 3227–3241.

Jardri, R., Duverne, S., Litvinova, A. S., & Denève, S. (2017). Experimental evidence for circular inference in schizophrenia. *Nature Communications*, *8*, 14218.

Jardri, R., Hugdahl, K., Hughes, M., Brunelin, J., Waters, F., Alderson-Day, B., et al. (2016). Are hallucinations due to an imbalance between excitatory and inhibitory influences on the brain?. *Schizophrenia Bulletin*, *42*(5), 1124–1134.

Just, M. A., Cherkassky, V. L., Keller, T. A., & Minshew, N. J. (2004). Cortical activation and synchronization during sentence comprehension in high-functioning autism: evidence of underconnectivity. *Brain*, *127*(8), 1811–1821.

Kanner, L., et al. (1943). Autistic disturbances of affective contact. *Nervous Child*, *2*(3), 217–250.

Karvelis, P., Seitz, A. R., Lawrie, S. M., & Seriès, P. (2018). Autistic traits, but not schizotypy, predict increased weighting of sensory information in Bayesian visual integration. *eLife*, *7*, e34115.

Keown, C. L., Shih, P., Nair, A., Peterson, N., Mulvey, M. E., & Müller, R.-A. (2013). Local functional overconnectivity in posterior brain regions is associated with symptom severity in autism spectrum disorders. *Cell Reports*, *5*(3), 567–572.

Keshavan, M. S., Anderson, S., & Pettergrew, J. W. (1994). Is schizophrenia due to excessive synaptic pruning in the prefrontal cortex? the feinberg hypothesis revisited. *Journal of Psychiatric Research*, *28*(3), 239–265.

King, B. H., & Lord, C. (2011). Is schizophrenia on the autism spectrum?. *Brain Research*, *1380*, 34–41.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.

Krogh, A., & Hertz, J. A. (1992). A simple weight decay can improve generalization. In *Advances in neural information processing systems* (pp. 950–957).

Lang, P. J., & Buss, A. H. (1965). Psychological deficit in schizophrenia: II. Interference and activation.. *Journal of Abnormal Psychology*, *70*(2), 77.

Lanillos, P., & Cheng, G. (2018). Adaptive robot body learning and estimation through predictive coding. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 4083–4090). IEEE.

Lanillos, P., Dean-Leon, E., & Cheng, G. (2017). Enactive self: a study of engineering perspectives to obtain the sensorimotor self through enaction. In *2017 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)* (pp. 72–78). IEEE.

Lawson, R. P., Mathys, C., & Rees, G. (2017). Adults with autism overestimate the volatility of the sensory environment. *Nature Neuroscience*, *20*(9), 1293.

Lawson, R. P., Rees, G., & Friston, K. J. (2014). An aberrant precision account of autism. *Frontiers in Human Neuroscience*, *8*, 302.

Lord, C., Rutter, M., DiLavore, P. C., Risi, S., Gotham, K., Bishop, S., et al. (2012). *Autism diagnostic observation schedule: ADOS*. Los Angeles, CA: Western Psychological Services.

Lucker, J. R. (2013). Auditory hypersensitivity in children with autism spectrum disorders. *Focus on Autism and Other Developmental Disabilities*, *28*(3), 184–191.

Lynall, M.-E., Bassett, D. S., Kerwin, R., McKenna, P. J., Kitzbichler, M., Muller, U., et al. (2010). Functional connectivity and brain networks in schizophrenia. *Journal of Neuroscience*, *30*(28), 9477–9487.

Margolis, R. L., Chuang, D.-M., & Post, R. M. (1994). Programmed cell death: implications for neuropsychiatric disorders. *Biological Psychiatry*, *35*(12), 946–956.

McClelland, J. L. (2000). The basis of hyperspecificity in autism: A preliminary suggestion based on properties of neural nets. *Journal of Autism and Developmental Disorders*, *30*(5), 497–502.

McClelland, J. L., McNaughton, B. L., & 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.. *Psychological Review*, *102*(3), 419.

Miikkulainen, R. (1993). *Subsymbolic natural language processing: An integrated model of scripts, lexicon, and memory*. MIT press.

Miikkulainen, R., & Dyer, M. G. (1991). Natural language processing with modular PDP networks and distributed lexicon. *Cognitive Science*, *15*(3), 343–399.

Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. *Trends in cognitive sciences*, *16*(1), 72–80.

Morrens, M., Hulstijn, W., Lewi, P. J., De Hert, M., & Sabbe, B. G. (2006). Stereotypy in schizophrenia. *Schizophrenia Research*, *84*(2–3), 397–404.

Mountcastle, V. B. (1957). Modality and topographic properties of single neurons of cat's somatic sensory cortex. *Journal of Neurophysiology*, *20*(4), 408–434.

Moustafa, A. A., Misiak, B., & Frydecka, D. (2017). Neurocomputational models of schizophrenia. *Computational Models of Brain and Behavior*, 73.

Murata, S., Namikawa, J., Arie, H., Sugano, S., & Tani, J. (2013). Learning to reproduce fluctuating time series by inferring their time-dependent stochastic properties: Application in robot learning via tutoring. *IEEE Transactions on Autonomous Mental Development*, *5*(4), 298–310.

Nagai, Y. (2019). Predictive learning: its key role in early cognitive development. *Philosophical Transactions of the Royal Society B*, *374*(1771), 20180030.

Nagai, Y., Moriwaki, T., & Asada, M. (2015). Influence of excitation/inhibition imbalance on local processing bias in autism spectrum disorder. In *Proc. of the 37th annual meeting of the cognitive science society* (pp. 1685–1690).

Noel, J.-P., Cascio, C. J., Wallace, M. T., & Park, S. (2017). The spatial self in schizophrenia and autism spectrum disorder. *Schizophrenia Research*, *179*, 8–12.

Nordby, H., Hammerborg, D., Roth, W. T., & Hugdahl, K. (1994). ERPS for infrequent omissions and inclusions of stimulus elements. *Psychophysiology*, *31*(6), 544–552.

Noriega, G. (2007). Self-organizing maps as a model of brain mechanisms potentially linked to autism. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *15*(2), 217–226.

Noriega, G. (2008). Modeling propagation delays in the development of SOMs—A parallel with abnormal brain growth in autism. *Neural Networks*, *21*(2–3), 130–139.

Notredame, C.-E., Pins, D., Deneve, S., & Jardri, R. (2014). What visual illusions teach us about schizophrenia. *Frontiers in integrative neuroscience*, *8*, 63.

OLaughlin, C., & Thagard, P. (2000). Autism and coherence: A computational model. *Mind & Language*, *15*(4), 375–392.

Oliver, G., Lanillos, P., & Cheng, G. (2019). Active inference body perception and action for humanoid robots. arXiv preprint arXiv:1906.03022.

Park, J., Ichinose, K., Kawai, Y., Suzuki, J., Asada, M., & Mori, H. (2019). Macroscopic cluster organizations change the complexity of neural activity. *Entropy*, *21*(2), 214.

Pellicano, E., & Burr, D. (2012). When the world becomes too real: a bayesian explanation of autistic perception. *Trends in Cognitive Sciences*, *16*(10), 504–510.

Pfeifer, R., & Bongard, J. (2006). *How the body shapes the way we think: a new view of intelligence*. MIT press.

Philippsen, A., & Nagai, Y. (2018). Understanding the cognitive mechanisms underlying autistic behavior: a recurrent neural network study. In *Development and learning and epigenetic robotics (ICDL-EpiRob), 2018 joint IEEE international conference on* (pp. 84–90). IEEE.

Pinkham, A. E., Hopfinger, J. B., Pelphrey, K. A., Piven, J., & Penn, D. L. (2008). Neural bases for impaired social cognition in schizophrenia and autism spectrum disorders. *Schizophrenia Research*, *99*(1–3), 164–175.

Powers III, A. R., Kelley, M., & Corlett, P. R. (2016). Hallucinations as top-down effects on perception. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *1*(5), 393–400.

C.-D. G. of the Psychiatric Genomics Consortium, et al. (2013). Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *The Lancet*, *381*(9875), 1371–1379.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79.

Rapoport, J., Chavez, A., Greenstein, D., Addington, A., & Gogtay, N. (2009). Autism spectrum disorders and childhood-onset schizophrenia: clinical and biological contributions to a relation revisited. *Journal of the American Academy of Child & Adolescent Psychiatry*, *48*(1), 10–18.

Redish, D., & Gordon, J. (2016). *Computational psychiatry: New perspectives on mental illness (strungmann forum reports)*. Cambridge, MA: MIT Press.

Reggia, J. A., Ruppin, E., & Berndt, R. S. (1996). *Neural modeling of brain and cognitive disorders, Vol. 6*. World Scientific.

Robbins, M. (1993). *Experiences of schizophrenia: An integration of the personal, scientific, and therapeutic*. Guilford Press.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain.. *Psychological Review*, *65*(6), 386.

Rosvold, H. E., Mirsky, A. F., Sarason, I., Bransome Jr, E. D., & Beck, L. H. (1956). A continuous performance test of brain damage.. *Journal of Consulting Psychology*, *20*(5), 343.

Rubenstein, J., & Merzenich, M. M. (2003). Model of autism: increased ratio of excitation/inhibition in key neural systems. *Genes, Brain and Behavior*, *2*(5), 255–267.

Ruppin, E., Reggia, J. A., & Horn, D. (1995). A neural model of delusions and hallucinations in schizophrenia. In *Advances in neural information processing systems* (pp. 149–156).

Ruppin, E., Reggia, J. A., & Horn, D. (1996). Pathogenesis of schizophrenic delusions and hallucinations: a neural model. *Schizophrenia Bulletin*, *22*(1), 105–121.

Saha, S., Chant, D., Welham, J., & McGrath, J. (2005). A systematic review of the prevalence of schizophrenia. *PLoS Medicine*, *2*(5), e141.

Sandin, S., Lichtenstein, P., Kuja-Halkola, R., Hultman, C., Larsson, H., & Reichenberg, A. (2017). The heritability of autism spectrum disorder. *Jama*, *318*(12), 1182–1184.

Schaal, S., Kotosaka, S., & Sternad, D. (2000). Nonlinear dynamical systems as movement primitives. In *IEEE international conference on humanoid robotics* (pp. 1–11).

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117.

Selfe, L. (2012). *Nadia revisited: A longitudinal study of an autistic savant*. Psychology Press.

Series, P., Reichert, D. P., & Storkey, A. J. (2010). Hallucinations in charles bonnet syndrome induced by homeostasis: a deep Boltzmann machine model. In *Advances in Neural Information Processing Systems* (pp. 2020–2028).

Sims, A. (1988). *Symptoms in the mind: An introduction to descriptive psychopathology*.. Bailliere Tindall Publishers.

Snijders, T. M., Milivojevic, B., & Kemner, C. (2013). Atypical excitation–inhibition balance in autism captured by the gamma response to contextual modulation. *NeuroImage: Clinical*, *3*, 65–72.

Spitzer, M. (1995). A neurocomputational approach to delusions. *Comprehensive Psychiatry*, *36*.

Stanghellini, G. (2009). Embodiment and schizophrenia. *World Psychiatry*, *8*(1), 56–59.

Sterzer, P., Adams, R. A., Fletcher, P., Frith, C., Lawrie, S. M., Muckli, L., et al. (2018). The predictive coding account of psychosis. *Biological Psychiatry*, *84*(9), 634–643.

Stevens, J. R. (1992). Abnormal reinnervation as a basis for schizophrenia: A hypothesis. *Archives of General Psychiatry*, *49*, 238–243.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions.. *Journal of Experimental Psychology*, *18*(6), 643.

Sun, L., Grützner, C., Bölte, S., Wibral, M., Tozman, T., Schlitt, S., et al. (2012). Impaired gamma-band activity during perceptual organization in adults with autism spectrum disorders: evidence for dysfunctional network activity in frontal-posterior cortices. *Journal of Neuroscience*, *32*(28), 9563–9573.

Supekar, K., Uddin, L. Q., Khouzam, A., Phillips, J., Gaillard, W. D., Kenworthy, L. E., et al. (2013). Brain hyperconnectivity in children with autism and its links to social deficits. *Cell Reports*, *5*(3), 738–747.

Tani, J. (2003). Learning to generate articulated behavior through the bottom-up and the top-down interaction processes. *Neural Networks*, *16*(1), 11–23.

Thakkar, K. N., Nichols, H. S., McIntosh, L. G., & Park, S. (2011). Disturbances in body ownership in schizophrenia: evidence from the rubber hand illusion and case study of a spontaneous out-of-body experience. *PLoS One*, *6*(10), e27089.

Thompson, P. (1980). Margaret thatcher: a new illusion.. *Perception*, *9*(4), 483–484.

Treffert, D. A. (2009). The savant syndrome: an extraordinary condition. a synopsis: past, present, future. *Philosophical Transactions of the Royal Society, Series B (Biological Sciences)*, *364*(1522), 1351–1357.

Tsodyks, M. (1988). Associative memory in asymmetric diluted network with low level of activity. *EPL (Europhysics Letters)*, *7*(3), 203.

Tsodyks, M. V., & Feigel'man, M. V. (1988). The enhanced storage capacity in neural networks with low activity level. *EPL (Europhysics Letters)*, *6*(2), 101.

Valton, V., Romaniuk, L., Steele, J. D., Lawrie, S., & Seriès, P. (2017). Comprehensive review: computational modelling of schizophrenia. *Neuroscience & Biobehavioral Reviews*, *83*, 631–646.

Vladusich, T., Olu-Lafe, O., Kim, D.-S., Tager-Flusberg, H., & Grossberg, S. (2010). Prototypical category learning in high-functioning autism. *Autism Research*, *3*(5), 226–236.

Von Helmholtz, H. (1867). *Handbuch der physiologischen Optik, Vol. 9*. Voss.

Wang, J., Barstein, J., Ethridge, L. E., Mosconi, M. W., Takarae, Y., & Sweeney, J. A. (2013). Resting state EEG abnormalities in autism spectrum disorders. *Journal of Neurodevelopmental Disorders*, *5*(1), 24.

Wang, X.-J., & Krystal, J. H. (2014). Computational psychiatry. *Neuron*, *84*(3), 638–654.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world'networks. *Nature*, *393*(6684), 440.

van der Weiden, A., Prikken, M., & van Haren, N. E. (2015). Self–other integration and distinction in schizophrenia: A theoretical analysis and a review of the evidence. *Neuroscience & Biobehavioral Reviews*, *57*, 220–237.

Wood, S. J. (2017). Autism and schizophrenia: one, two or many disorders?. *The British Journal of Psychiatry*, *210*(4), 241–242.

Yamada, Y., Kanazawa, H., Iwasaki, S., Tsukahara, Y., Iwata, O., Yamada, S., et al. (2016). An embodied brain model of the human foetus. *Scientific Reports*, *6*, 27893.

Yamashita, Y., & Tani, J. (2008). Emergence of functional hierarchy in a multiple timescale neural network model: a humanoid robot experiment. *PLoS Computational Biology*, *4*(11), e1000220.

Yamashita, Y., & Tani, J. (2012). Spontaneous prediction error generation in schizophrenia. *PLoS One*, *7*(5), e37843.

Yizhar, O., Fenno, L. E., Prigge, M., Schneider, F., Davidson, T. J., O'shea, D. J., et al. (2011). Neocortical excitation/inhibition balance in information processing and social dysfunction. *Nature*, *477*(7363), 171.