Neuroscience and artificial intelligence (AI) share a long history of collaboration. Advances in neuroscience, alongside huge leaps in computer processing power over the last few decades, have given rise to a new generation of in silico neural networks inspired by the architecture of the brain. These AI systems are now capable of many of the advanced perceptual and cognitive abilities of biological systems, including object recognition and decision making. Moreover, AI is now increasingly being employed as a tool for neuroscience research and is transforming our understanding of brain functions. In particular, deep learning has been used to model how convolutional layers and recurrent connections in the brain’s cerebral cortex control important functions, including visual processing, memory, and motor control. Excitingly, the use of neuroscience-inspired AI also holds great promise for understanding how changes in brain networks result in psychopathologies, and could even be utilized in treatment regimes. Here we discuss recent advancements in four areas in which the relationship between neuroscience and AI has led to major advancements in the field: (1) AI models of working memory, (2) AI visual processing, (3) AI analysis of big neuroscience datasets, and (4) computational psychiatry.

© 2021 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Classically, our definition of intelligence has largely been based upon the capabilities of advanced biological entities, most notably humans. Accordingly, research into artificial intelligence (AI) has primarily focused on the creation of machines that can perceive, learn, and reason, with the overarching objective of creating an artificial general intelligence (AGI) system that can emulate human intelligence, so called Turing-powerful systems. Considering this aim, it is not surprising that scientists, mathematicians, and philosophers working on AI have taken inspiration from the mechanistic, structural, and functional properties of the brain.

Since at least the 1950s, attempts have been made to artificially model the information processing mechanisms of neurons. This primarily began with the development of perceptrons (Rosenblatt, 1958), a highly reductionist model of neuronal signaling, in which an individual node receiving weighted inputs could produce a binary output if the summation of inputs reached a threshold. Coinciding with the emergence of the cognitive revolution in the 50s and 60s, and extending until at least the 90s, there was initially much pushback against the development of artificial neural networks within the AI and cognitive science communities (Fodor & Pylyshyn, 1988; Mandler, 2002; Minsky & Papert, 1969). However, by the late 1980s, the development of multilayer neural networks and the popularization of backpropagation had solved many of the limitations of early perceptrons, including their inability to solve non-linear classification problems such as learning a simple boolean XOR function (Rumelhart, Hinton, & Williams, 1986). Neural networks were now able to dynamically modify their own connections by calculating error functions of the network and communicating them back through...
the constituent layers, giving rise to a new generation of AI capable of intelligent skills including image and speech recognition (Bengio, 1993; LeCun, Bengio, & Hinton, 2015; LeCun et al., 1989). To date, backpropagation is still commonly used to train deep neural networks (Lillicrap, Santoro, Marris, Akerman, & Hinton, 2020; Richards et al., 2019), and has been combined with reinforcement learning methods to create advanced learning systems capable of matching or outperforming humans in strategy-based games including Chess (Silver et al., 2018), Go (Silver et al., 2016, 2018), poker (Moravčík et al., 2017), and StarCraft II (Vinyals et al., 2019).

Despite the neuroscience-inspired origins of AI, the biological plausibility of modern AI is questionable. Indeed, there is little evidence that backpropagation of error underlies the modification of synaptic connections between neurons (Crick, 1989; Grossberg, 1987); although recent theories have suggested that an approximation of backpropagation may exist in the brain (Lillicrap et al., 2020; Whittington & Bogacz, 2019). While creating brain-like systems is clearly not necessary to achieve all goals of AI, as evidenced by the above-described accomplishments, a major advantage of biologically plausible AI is its usefulness for understanding and modeling information processing in the brain. Additionally, brain mechanisms can be thought of as an evolutionarily-validated template for intelligence, honed over millions of years for adaptability, speed, and energy efficiency. As such, increased integration of brain-inspired mechanisms may help to further improve the capabilities and efficiency of AI. These ideas have led to continued interest in the creation of neuroscience-inspired AI, and have further strengthened the partnership between AI and neuroscience research fields.

In the last decade, several biologically plausible alternatives to backpropagation have been suggested, including predictive coding (Bastos et al., 2012; Millidge, Tschantz, & Buckley, 2020), feedback alignment (Lillicrap, Cownden, Tweed, & Akerman, 2016), equilibrium propagation (Scellier & Bengio, 2017), Hebbian-like learning rules (Krotov & Hopfield, 2019; Miconi, 2017), and zero-divergence inference learning (Salvatori, Song, Lukasiewicz, Bogacz, & Xu, 2021). Similarly, other recent efforts to bridge the gap between artificial and biological neural networks have led to the development of spiking neural networks capable of approximating stochastic potential-based communication between neurons (Pfeiffer & Pfeil, 2018), as well as attention-like mechanisms including transformer architectures (Vaswani et al., 2017).

The beneficial relationship between AI and neuroscience is reciprocal, and AI is now rapidly becoming an invaluable tool in neuroscience research. AI models designed to perform intelligence-based tasks are providing novel hypotheses for how the same processes are controlled within the brain. For example, work on distributional reinforcement learning in AI has recently resulted in the proposal of a new theory of dopaminergic signaling of probabilistic distributions (Dabney et al., 2020). Similarly, goal-driven deep learning models of visual processing have been used to estimate the organizational properties of the brain's visual system and accurately predict patterns of neural activity (Yamins & DiCarlo, 2016). Additionally, advances in deep learning algorithms and the processing power of computers now allow for high-throughput analysis of large-scale datasets, including that of whole-brain imaging in animals and humans, expediting the progress of neuroscience research (Thomas, Heekeren, Müller, & Samek, 2019; Todorov et al., 2020; Zhu et al., 2019). Deep learning models trained to decode neural imaging data can create accurate predictions of decision-making, action selection, and behavior, helping us to understand the functional role of neural activity, a key goal of cognitive neuroscience (Batty et al., 2019; Musall, Urai, Sussillo, & Churchland, 2019). Excitingly, machine learning and deep learning approaches are also now being applied to the emerging field of computational psychiatry to simulate normal and dysfunctional brain states, as well as to identify aberrant patterns of brain activity that could be used as robust classifiers for brain disorders (Cho, Yim, Choi, Ko, & Lee, 2019; Durstewitz, Koppe, & Meyer-Lindenberg, 2019; Koppe, Meyer-Lindenberg, & Durstewitz, 2021; Zhou et al., 2020).

In the last few years, several reviews have examined the long and complicated relationship between neuroscience and AI (see Hassabis, Kumaran, Summerfield, & Botvinick, 2017; Hasson, Nastase, & Goldstein, 2020; Kriegeskorte & Douglas, 2018; Richards et al., 2019; Ulman, 2019). Here, we aim to give a brief introduction into how the interplay between neuroscience and AI fields has stimulated progress in both areas, focusing on four important themes taken from talks presented at a symposium entitled “AI for neuroscience and neuromorphic technologies” at the 2020 International Symposium on Artificial Intelligence and Brain Science; (1) AI models of working memory, (2) AI visual processing, (3) AI analysis of neuroscience data, and (4) computational psychiatry. Specifically, we focus on how recent neuroscience-inspired approaches are resulting in AI that is increasingly brain-like and is not only able to achieve human-like feats of intelligence, but is also capable of decoding neural activity and accurately predicting behavior and the brain's mental contents. This includes spiking and recurrent neural network models of working memory inspired by the stochastic spiking properties of biological neurons and their sustained activation during memory retention, as well as neural network models of visual processing incorporating convolutional layers inspired by the architecture of the brain's visual ventral stream. Additionally, we discuss how AI is becoming an increasingly powerful tool for neuroscientists and clinicians, acting as diagnostic and even therapeutic aids, as well as potentially informing us about brain mechanisms, including information processing and memory.

2. Neuroscience-inspired artificial working memory

One of the major obstacles of creating brain-like AI systems has been the challenge of modeling working memory, an important component of intelligence. Today, most in silico systems utilize a form of working memory known as random-access memory (RAM) that acts as a cache for data required by the central processor and is separated from long-term memory storage in solid state or hard disk drives. However, this architecture differs considerably from the brain, where working and long-term memory appear to involve, at least partly, the same neural substrates, predominantly the neocortex (Baddeley, 2003; Blumenfeld & Ranganath, 2007; Rumelhart & McClelland, 1986; Shimamura, 1995) and hippocampus (Bird & Burgess, 2008; Eichenbaum, 2017). These findings suggest that within these regions, working memory is likely realized by specific brain mechanisms that allow for fast and short-term access to information.

Working memory tasks performed in humans and non-human primates have indicated that elevated and persistent activity within cell assemblies of the prefrontal cortex, as well as other areas of the neocortex, hippocampus, and brainstem, may be critical for information retention within working memory (Boran et al., 2019; Christophel, Klink, Spitzer, Roelfsema, & Haynes, 2017; Fuster & Alexander, 1971; Goldman-Rakic, 1995; McFarland & Fuchs, 1992; Miller, Erickson, & Desimone, 1996; Watanabe & Niki, 1985). In response to these findings, several neural mechanisms have been proposed to account for this persistent activity (reviewed in Durstewitz, Seamans, & Sejnowski, 2000). These include recurrent excitatory connectivity between networks of neurons (Hopfield, 1982; O'Reilly, Braver, & Cohen, 1999), cellular bistability, where the intrinsic properties of neurons can produce a continuously spiking state (Lisman, Fellous,
continuous-rate RNN models (Kim, Li, & Sejnowski, 2019; Xue, similar performance in cognitive tasks to the above-described tic spiking properties of biological circuits and demonstrating current neural networks (SRNN), aiming to recreate the stochas-
tic coordinated bursts of spiking activity, rather than persistent activity across trials, potentially masking important intra-trial spiking dynamics (Lundqvist, Herman, & Miller, 2018). Interestingly, recent single-trial analyses of working memory tasks suggest that frontal cortex networks demonstrate sparse, transient coordinated bursts of spiking activity, rather than persistent activation (Bastos, Loonis, Kornblith, Lundqvist, & Miller, 2018; Lundqvist, Herman, Warden, Brincat, & Miller, 2018; Lundqvist et al., 2016). Such patterns of neural activity may be explained by models of transient spiking activity, such as the “synaptic attractor” model, where working memories are maintained by spike-induced Hebbian synaptic plasticity in between transient coordinated bursts of activity (Fiebig & Lansner, 2017; Huang & Wei, 2021; Lundqvist, Herman, & Miller, 2018; Mongillo, Barak, & Tsydyks, 2008; Sandberg, Tegnér, & Lansner, 2003). These models suggest that synaptic plasticity may allow working memory to be temporarily stored in an energy efficient manner that is also less susceptible to interference, while bursts of spiking may allow for fast reading of information when necessary (Huang & Wei, 2021; Lundqvist, Herman, & Miller, 2018). Further investigation of working memory in biological studies using single-trial analyses, as well as neuroscience-inspired AI models trained on working memory tasks, may help to elucidate precisely when and how these spiking and plasticity-based processes are utilized within the brain.

Here we have discussed how neuroscience findings over the last few decades inspired the creation of computational models of working memory in humans and non-human primates. These studies subsequently informed the creation of artificial neural networks designed to model the organization and function of brain networks, including the inclusion of recurrent connections between neurons and the introduction of spiking properties. The relationship between neuroscience and AI research has now come full circle, with recent SRNN models potentially informing about brain mechanisms underlying working memory (Kim & Sejnowski, 2021). In the next section we continue this examination of the benefits of the partnership between neuroscience and AI, discussing how brain architectures have inspired the design of artificial visual processing models and how human imaging data has been used to decode and inform how visual processing is controlled within the brain.

3. Decoding the brain’s visual system

The challenge of creating artificial systems capable of emulating biological visual processing is formidable. However, recent efforts to understand and reverse engineer the brain’s ventral visual stream, a series of interconnected cortical nuclei responsible for hierarchically processing and encoding of images into explicit neural representations, have shown great promise in the creation of robust AI systems capable of decoding and interpreting human visual processing, as well as performing complex visual intelligence skills including image recognition (Federer, Xu, Fyshe, & Zylberberg, 2020; Verschae & Ruiz-del-Solar, 2015), motion detection (Manchanda & Sharma, 2016; Wu, McGinley, Maguire, Cai, & Valderrama-Gonzalez, 2008), and object tracking (Luo et al., 2020; Soleimanitaleb, Keyvanrad, & Jafari, 2019; Zhang et al., 2021).

In an effort to understand and measure human visual perception, machine learning models, including support-vector networks, have been trained to decode stimulus-induced fMRI activity patterns in the human V1 cortical area, and were able to visually reconstruct the local contrast of presented and internal mental images (Kamitani & Tong, 2005; Miyawaki et al., 2008). Similarly, those trained to decode stimulus-induced activity in higher visual cortical areas were able to identify the semantic contents of dream imagery (Horikawa, Tamaki, Miyawaki, & Kamitani, 2013). These findings indicate that the visual features of both perceived and mental images are represented in the same neural substrates (lower and higher visual areas for low-level perceptual and high-level semantic features, respectively), supporting previous evidence from human PET imaging studies revealing mental imagery to activate the primary visual cortex, an...
area necessary for visual perception (Kosslyn et al., 1993; Kosslyn, Thompson, Klm, & Alpert, 1995). Additionally, these studies add to a growing literature demonstrating the utility of AI for decoding of brain imaging data for objective measurement of human visual experience (Kamitani & Tong, 2005; Nishimoto et al., 2011).

Finally, beyond machine learning methods, the incorporation of a deep generator network (DGN) to a very deep convolutional neural network (CNN) image reconstruction method, allowing CNN hierarchical processing layers to be fully utilized in a manner similar to that of the human visual system, has recently been demonstrated to improve the quality of visual reconstructions of perceived or mental images compared with the same CNN without the DGN (Shen, Horikawa, Majima, & Kamitani, 2019).

Interestingly, neural networks trained to perform visual tasks have often been reported to acquire similar representations to regions of the brain's visual system required for the same tasks (Nonaka, Majima, Aoki, & Kamitani, 2020; Yamins & DiCarlo, 2016). CNNs incorporating hierarchical processing layers similar to that of the visual ventral stream and trained on image recognition tasks have been reported to be able to accurately predict neural responses in the inferior temporal (IT) cortex, the recognition tasks have been reported to accurately predict neural responses in the inferior temporal (IT) cortex, the highest area of the ventral visual stream, of primates (Cadiou et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014). What is more, high-throughput computational evaluation of candidate CNN models revealed a strong correlation between a model's object recognition capability and its ability to predict IT cortex neural activity (Yamins et al., 2014). Accordingly, recent evidence revealed that the inclusion of components that closely predict the activity of the front-end of the visual stream (V1 area) improve the accuracy of CNNs by reducing their susceptibility to errors resulting from image perturbations, so-called white box adversarial attacks (Dapello et al., 2020).

While these studies appear to suggest the merit of "brain-like" AI systems for visual processing, until recently there has been no method to objectively measure how "brain-like" an AI visual processing model is. In response to this concern, two novel metrics, the Brain-Score (BS) (Schrömpf et al., 2020) and the brain hierarchy (BH) score (Nonaka et al., 2020), have been created to assess the functional similarity between AI models and the human visual system. Specifically, the BS measures the ability of models to predict brain activity and behavior, whereas the BH is designed to evaluate the hierarchical homology across layers/areas between neural networks and the brain (Nonaka et al., 2020; Schrömpf et al., 2020). Interestingly, while evaluation of several commonly used AI visual processing models found a positive correlation between the BS and the accuracy for image recognition (i.e., brain-like neural networks performed better), the opposite result was found when the BH was used (Nonaka et al., 2020; Schrömpf et al., 2020). Although these findings appear to contradict each other, more recently developed high-performance neural networks tended to have a lower BS, suggesting that AI vision may now be diverging from human vision (Schrömpf et al., 2020). Importantly, and particularly for the BS, it should be considered that while the ability of a model to predict brain activity may indicate its functional similarity, it does not necessarily mean that the model is emulating actual brain mechanisms. In fact, statisticians have long stressed this importance of the distinction between explanatory and predictive modeling (Shmueli, 2010). Thus, if we intend to use AI systems to model and possibly inform our understanding of visual processing in the brain, it is important that we continue to increase the structural and mechanistic correspondence of AI models to their counterpart systems within the brain, as well as strengthening the ability of metrics to measure such correspondence. Indeed, considering the known complexity of the brain's visual system, including the existence of multiple cell types (Gonchar, Wang, & Burkhalter, 2008; Pfeffer, Xue, He, Huang, & Scanziani, 2013) that are modulated by various neurotransmitters (Azimi et al., 2020; Noudoost & Moore, 2011), it is likely that comparatively simplistic artificial neural networks do not yet come close to fully modeling the myriad of processes contributing to biological visual processing.

Finally, in addition to their utility for image recognition, brain-inspired neural networks are now beginning to be applied to innovative and practical uses within the field of visual neuroscience research. One example of this is the recent use of an artificial neural network to design precise visual patterns that can be projected directly onto the retina of primates to accurately control the activity of individual or groups of ventral stream (V4 area) neurons (Bashivan, Kair, & DiCarlo, 2019). These findings indicate the potential of this method for non-invasive control of neural activity in the visual cortex, creating a powerful tool for neuroscientists. In the next section, we further describe how AI is now increasingly being utilized for the advancement of neuroscience research, including in the objective analysis of animal behavior and its neural basis.

4. AI for analyzing behavior and its neural correlates

Understanding the relationship between neural activity and behavior is a critical goal of neuroscience. Recently developed large-scale neural imaging techniques have now enabled huge quantities of data to be collected during behavioral tasks in animals (Ahrens & Engert, 2015; Cardin, Craig, & Higley, 2020; Weisenburger & Vaziri, 2016; Yang & Yuste, 2017). However, given the quantity and speed of individual movements animals perform during behavioral tasks, as well as the difficulty in identifying individual neurons among large and crowded neural imaging datasets, it has been challenging for researchers to effectively and objectively analyze animal behavior and its precise neural correlates (Berman, 2018; Giovannucci et al., 2019; von Ziegler, Sturman, & Bohacek, 2021).

To address difficulties in human labeling of animal behavior, researchers have turned to AI for help. Over the last few years, several open-source, deep learning-based software toolboxes have been developed for 3D markerless pose estimation across several species and types of behaviors (Aarc, Zhao, Dobkin, Carmichael, & Golshani, 2019; Forys, Xiao, Gupta, & Murphy, 2020; Graving et al., 2019; Günel et al., 2019; Mathis et al., 2018; Nath et al., 2019; Pereira et al., 2019). Perhaps the most widely used of these has been DeepLabCut, a deep neural network that incorporates the feature detectors from DeeperCut, a multi-person pose estimation model, and is able to accurately estimate the pose of several commonly used laboratory animals with minimal training (Lauer et al., 2021; Mathis et al., 2018; Nath et al., 2019). This pose estimation data can then be combined with various supervised machine learning tools, including JAABA (Kabra, Robie, Rivera-Alba, Branson, & Branson, 2013) and SimBA (Nilsson et al., 2020) that allow for automated identification of specific behaviors labeled by humans, such as grooming, freezing, and various social behaviors. This combined use of such tools has been shown to be able to match human ability for accurate quantification of several types of behaviors, and can outperform commercially available animal-tracking software packages (Sturman et al., 2020). In addition to supervised machine learning analysis of animal behavioral data, several unsupervised machine learning tools have been developed, including MotionMapper (Berman, Choi, Bialek, & Saezvitz, 2014), MoSeq (Wiltshire et al., 2015), and more recently, uBAM (Brattoli et al., 2021). These unsupervised approaches allow objective classification of the full repertoire of animal behavior and can potentially uncover subtle behavioral traits that might be missed by humans (Kwok, 2019).
As with animal behavioral data, human annotation of animal neural imaging datasets acquired from large-scale recording of neural activity, such as that acquired using in-vivo imaging of neural activity markers such as calcium indicators, is time consuming and suffers from a large degree of variability between annotators in the segmentation of individual neurons (Giovannucci et al., 2019). In the last decade, several tools utilizing classic machine learning and deep learning approaches have been developed to assist the analysis of animal calcium imaging data through the automated detection and quantification of individual neuron activity, known as source extraction (Pnevmatikakis, 2019). The most widely used of these have been unsupervised machine learning approaches employing activity-based segmentation algorithms, including principal component and independent component analysis (PCA/ICA) (Mukamel, Nimimerjahn, & Schnitzer, 2009), variations of constrained non-negative matrix factorization (CNMF) (Friedrich, Giovannucci, & Pnevmatikakis, 2021; Guan et al., 2018; Pnevmatikakis et al., 2016; Zhou et al., 2018), and dictionary learning (Giovannucci et al., 2017; Petersen, Simon, & Witten, 2017), to extract signals of neuron-like regions of interest from the background. While these techniques offer the benefit that they require no training and thus can be applied to analysis of various cell types and even dendritic imaging, they often suffer from false positives and are unable to identify low-activity neurons, making it difficult to longitudinally track the activity of neurons that may be temporally inactive in certain contexts (Lu et al., 2018). To address this limitation, several supervised deep learning approaches that segment neurons based upon features learned from human-labeled calcium imaging datasets have been developed (Aphorpe et al., 2016; Denis, Dard, Quiroli, Cossart, & Picardo, 2020; Giovannucci et al., 2019; Klibisz, Rose, Eicholtz, Blundon, & Zakharenko, 2017; Soltanian-Zadeh, Sahingur, Blau, Gong, & Farsi, 2019; Xu, Su, Zhut, Guan, & Zhangt, 2016). Many of these tools, including U-Net2DS (Klibisz et al., 2017), STNeuroNet (Soltanian-Zadeh et al., 2019), and DeepCINAC (Denis et al., 2020), train a CNN to segment neurons in either 2D or 3D space and have been demonstrated to be able to detect neurons with near-human accuracy, and outperform other techniques including PCA/ICA, allowing for accurate, fast, and reproducible neural network detection and classification (Aphorpe et al., 2016; Giovannucci et al., 2019; Mukamel et al., 2009).

Finally, efforts are now being made to combine AI analysis of animal behavior and neural imaging data, not only for automated mapping of behavior to its neural correlates, but in order to predict and model animal behavior based on analyzed neural activity data. One such recently developed system is BehaveNet, a probabilistic framework for the unsupervised analysis of behavioral video, with semi-supervised decoding of neural activity (Batty et al., 2019). The resulting generative models of this framework are able to decode animal neural activity data and create probabilistic full-resolution video simulations of behavior (Batty et al., 2019). Further development of technologies designed to automate the mapping of neural activity patterns with behavioral motifs may help to elucidate how discrete patterns of neural activity are related to specific movements (Musall et al., 2019).

While the studies here describe approaches to analyzing and modeling healthy behavior and brain activity in animals, efforts have also been made to use AI to understand and identify abnormal brain functioning. In the next section, we discuss AI-based approaches to objective classification of psychiatric disorders, and how deep learning approaches have been utilized for modeling such disorders in artificial neural networks.

5. The interface between AI and psychiatry

Despite the adoption of standardized diagnostic criteria in clinical manuals such as the Diagnostic and Statistical Manual of Mental Disorders (DSM) and the International Classification of Disease (ICD), psychiatric and developmental disorders are still primarily identified based upon a patient’s subjective behavioral symptoms and self-report measures. Not only is this method often unreliable due to its subjectivity (Wakefield, 2016), but it also leads to an explanatory gap between phenomenology and neurobiology. However, in the last few decades, huge advancements in the power of computing, alongside the collection of large neuroimaging datasets, have allowed researchers to begin to bridge this gap by using AI to identify, model, and potentially even treat psychiatric and developmental disorders.

One area of particular promise has been the use of AI for the objective identification of brain disorders. Using machine learning methods, classifiers have been built to predict diagnostic labels of psychiatric and developmental disorders (see Bzdok & Meyer-Lindenberg, 2017; Cho et al., 2019; Zhou et al., 2020 for review). The scores produced by these probabilistic classifiers provide a degree of classification certainty that can be interpreted as a neural liability for the disorder and represent new biological dimensions of the disorders. However, while many of these classifiers, including those for schizophrenia (Greenstein, Malley, Weisinger, Clasen, & Gogtay, 2012; Orrù, Pettersson-Yeo, Marquand, Sartori, & Mechelli, 2012; Yassin et al., 2020) and ASD (Eslami, Almuqhim, Raiker, & Saeed, 2021; Yassin et al., 2020), are able to accurately identify the intended disorder, a major criticism has been that they are often only validated in a single sample cohort. To address this issue, recent attempts have been made to establish robust classifiers using larger and more varied sample data. This has led to the identification of classifiers for ASD and schizophrenia that could be generalized to independent cohorts, regardless of ethnicity, country, and MRI vendor, and still demonstrated classification accuracies of between 61%–76% (Yamada et al., 2017; Yoshihara et al., 2020). Aside from machine learning, deep learning approaches have also been applied to the classification of psychiatric and developmental disorders (see Durstewitz et al., 2019; Koppe et al., 2021, for review). A major advantage of deep neural networks is that their multi-layered design makes them particularly suited to learning high-level representations from complex raw data, allowing features to be extracted from neuroimaging data with far less parameters than machine learning architectures (Durstewitz et al., 2019; Jang, Plis, Calhoun, & Lee, 2017; Koppe et al., 2021; Plis et al., 2014; Schmidhuber, 2015). Accordingly, in the last few years, several deep neural networks have been reported to effectively classify brain disorders from neuroimaging data, including schizophrenia (Oh, Oh, Lee, Chae, & Yun, 2020; Sun et al., 2021; Yan et al., 2019; Zeng et al., 2018), autism (Guo et al., 2017; Heinsfeld, Franco, Craddock, Buchweitz, & Meneguzzi, 2018; Misman et al., 2019; Raj & Masood, 2020), ADHD (Chen, Li, et al., 2019; Chen, Song, & Li, 2019; Dubreuil-Vall, Ruffini, & Camprodon, 2020), and depression (Li, La, Wang, Hu, & Zhang, 2020; Uyulan et al., 2020). Further development of AI models for data-driven, dimensional psychiatry will likely help to address the current discontent surrounding categorical diagnostic criteria.

In addition to classification of disorders based on neuroimaging data, AI is also increasingly being used to model various psychiatric and developmental disorders (see Lanillos et al., 2020 for in-depth reviews). This largely began in the 1980s and 90s with studies modeling schizophrenia and ASD using artificial neural networks (Cohen, 1994; Cohen & Servan-Schreiber, 1992; Hoffman, 1987; Horn & Ruppin, 1995). Many of these models were inspired by biological evidence of structural and synaptic abnormalities associated with particular psychiatric disorder
symptoms. For example, evidence of reduced metabolism in the frontal cortex (Feinberg, 1983; Feinberg, Koresso, & Gottlieb, 1965; Feinberg, Koresso, Gottlieb, & Wender, 1964) and aberrant synaptic regeneration of abnormal brain structures (Stevens, 1992) have prompted the creation of neural networks designed to simulate how synaptic pruning (Hoffman & Dobscha, 1989) and reactive synaptic reorganization (Horn & Ruppin, 1995; Ruppin, Reggia, & Horn, 1996) may explain delusions and hallucinations in schizophrenia patients. Similarly, neural network models of excessive or reduced neuronal connections (Cohen, 1994, 1998; Thomas, Knowland, & Karmiloff-Smith, 2011) were generated to model biological observations of abnormal neuronal density in cortical, limbic and cerebellar regions (Bailey et al., 1998; Baum, 1991; Bauman & Kemper, 1985) hypothesized to contribute to developmental regression in ASD. Excitingly, more recently, deep learning models, including high-dimensional RNN models of schizophrenia (Yamashita & Tani, 2012) and ASD (Idee et al., 2017, 2018), have begun to be implemented into robots, allowing direct observation and comparison of modeled behavior with that seen in patients.

Finally, in the near future, AI could begin to play an important role in the treatment of psychiatric and developmental disorders. Computer-assisted therapy (CAT), including AI chatbots delivering cognitive behavioral therapies, is beginning to be tested for treatment of psychiatric disorders including depression and anxiety (Carroll & Rounsaville, 2010; Fitzpatrick, Darcy, & Vierhile, 2017; Fulmer, Joerin, Gentile, Lakerink, & Rauws, 2018). While still in their infancy, these CATs offer distinct advantages to human-led therapies in terms of price and accessibility, although their effectiveness in comparison to currently used therapeutic methods has yet to be robustly measured. Additionally, the identification of neuroimaging-based classifiers for psychiatric disorders (described above) has inspired the launch of a real-time fMRI-based neurofeedback projects in which patients attempt to normalize their own brain connectivity pattern through neurofeedback. Meta-analyses of such studies have indicated neurofeedback treatments to result in significant amelioration of the symptoms of several disorders including schizophrenia, depression, anxiety disorder, and ASD, suggesting the potential benefit of further use of such digital treatments (Dudek & Dodell-Feder, 2020; Schoenberg & David, 2014).

6. Conclusions

Since the inception of AI research midway through the last century, the brain has served as the primary source of inspiration for the creation of artificial systems of intelligence. This is largely based upon the reasoning that the brain is proof of concept of a comprehensive intelligence system capable of perception, planning, and decision making, and therefore offers an attractive template for the design of AI. In this review, based upon topics presented at the 2020 International Symposium on Artificial Intelligence and Brain Science, we have discussed how brain-inspired mechanistic, structural, and functional elements are being utilized to create novel, and optimize existing, AI systems. In particular, this has led to the development of high-dimensional deep neural networks often incorporating hierarchical architectures inspired by those found in the brain and capable of feats of intelligence including visual object recognition and memory-based cognitive tasks. Advancements in AI have also helped to foster progress within the field of neuroscience. Here we have described how the use of machine learning and neural networks for automated analysis of big data has revolutionized the analysis of animal behavioral and neuroimaging studies, as well as being utilized for objective classification of psychiatric and developmental disorders.

Importantly, while it has not been discussed in great detail in the current review, it should be considered that the relationship between AI and neuroscience is not simply two-way, but rather also includes the field of cognitive science (see Battleday, Peterson, & Griffiths, 2021; Cichy & Kaiser, 2019; Forbus, 2010; Kriegeskorte & Douglas, 2018, for review). Indeed, over the years, much of AI research has been guided by theories of brain functioning established by cognitive scientists (Elman, 1990; Hebb, 1949; Rumelhart & McClelland, 1986). For example, the convolutional neural networks discussed earlier in this review (in the section on visual processing) were inspired in part by computational models of cognition within the brain, including principles such as nonlinear feature maps and pooling of inputs, which were themselves derived from observations from neurophysiological studies in animals (Battleday et al., 2021; Fukushima, 1980; Hubel & Wiesel, 1962; Mozer, 1987; Riesenhuber & Poggio, 1999). In turn, neural networks have been used to guide new cognitive models of intellectual abilities, including perception, memory, and language, giving rise to the connectionism movement within cognitive science (Barrow, 1996; Fodor & Pylyshyn, 1988; Mayor, Gomez, Chang, & Lupyan, 2014; Yamins & DiCarlo, 2016). If we are to use AI to model and potentially elucidate brain functioning, the primary focus of cognitive science, it is important that we continue to not only use biological data from neuroscience studies, but also cognitive models, to inspire the architectural, mechanistic, and algorithmic design of artificial neural networks.

Despite their accomplishments and apparent complexity, current AI systems are still remarkably simplistic in comparison to brain networks and in many cases still lack the ability to accurately model brain functions (Bae, Kim, & Kim, 2021; Barrett, Morris, & Macke, 2019; Hasson et al., 2020; Pulvermüller, Tomasello, Henningsen-Schomers, & Wennekers, 2021; Tang et al., 2019). A major limitation is that, in general, current models are still not able to model the brain at multiple levels; from synaptic reorganization and the influence of neurotransmitter and hormone neuromodulation of neuronal excitability at the microlevel, to large-scale synchronization of spiking activity and global connectivity at the macrolevel. In fact, integration of various AI models of brain functioning, including the models of the cerebral cortex described in this review, as well as models of other brain regions, including limbic and motor control regions (Kowalczyk & Czubenko, 2016; Merel, Botvinick, & Wayne, 2019; Parsapooy, 2016), remains one of the greatest challenges in the creation of an AGI system capable of modeling the entire brain. In spite of these difficulties, it is clear that the continued interaction between neuroscience and AI will undoubtedly expedite progress in both areas.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This is supported by MEXT Grant-in-Aid Scientific Research on Innovative Areas “Correspondence and Fusion of Artificial Intelligence and Brain Science” (JP16H06568 to TH and TM, JP16H06572 to HT).

References
