Deep Learning and the Global Workspace Theory

Rufin VanRullen\textsuperscript{1,2} and Ryota Kanai\textsuperscript{3}

\textsuperscript{1}CerCo, CNRS UMR5549, Toulouse, France
\textsuperscript{2}ANITI, Université de Toulouse, France
\textsuperscript{3}Araya Inc, Tokyo, Japan

Abstract

Recent advances in deep learning have allowed Artificial Intelligence (AI) to reach near human-level performance in many sensory, perceptual, linguistic or cognitive tasks. There is a growing need, however, for novel, brain-inspired cognitive architectures. The Global Workspace theory refers to a large-scale system integrating and distributing information among networks of specialized modules to create higher-level forms of cognition and awareness. We argue that the time is ripe to consider explicit implementations of this theory using deep learning techniques. We propose a roadmap based on unsupervised neural translation between multiple latent spaces (neural networks trained for distinct tasks, on distinct sensory inputs and/or modalities) to create a unique, amodal global latent workspace (GLW). Potential functional advantages of GLW are reviewed.

1 Cognitive neural architectures in brains and machines

Deep learning denotes a machine learning system using artificial neural networks with multiple “hidden” layers between the input and output layers. Although the underlying theory is more than 3 decades old \cite{1,2}, it is only in the last decade that these systems have started to fully reveal their potential \cite{3}. Many of the recent breakthroughs in AI (Artificial Intelligence) have been fueled by deep learning. Neuroscientists have been quick to point out the similarities (and differences) between the brain and these deep artificial neural networks \cite{4–9}. The advent of deep learning has allowed the efficient computer implementation of perceptual and cognitive functions that had been so far inaccessible. Here, we aim to extend this approach to a cognitive framework that has been proposed to underlie perception, executive function and even consciousness: the Global Workspace Theory (GWT).

The GWT, initially proposed by Baars \cite{10,11}, is a tenet of modern cognitive science, and one of the major contemporary neuroscientific theories of consciousness (Figure 1A). The theory proposes that the brain is divided into specialized modules for specific functions, with long-distance connections between them \cite{10,11}. When warranted by the inputs or by task requirements (through a process of attentional selection), the contents of a specialized module can be broadcast and shared among distinct modules. According to the theory, the shared information at each moment in time—the global workspace—is what constitutes our conscious awareness. In functional terms, the global workspace can
serve to resolve problems that could not be solved by a single specialized function, by coordinating multiple specialized modules. Dehaene and colleagues [12–15] proposed a neuronal version of the theory, Global Neuronal Workspace (GNW). According to GNW, conscious access occurs when incoming information is made globally available to multiple brain systems through a network of neurons with long-range axons densely distributed in prefrontal, parieto-temporal, and cingulate cortices (Figure 1B). A spiking neural network implementation of key aspects of GNW captured the essence of the global workspace theory (Figure 1C), as well as known relationships between conscious report and neuronal responses (Figure 1D). Whether this approach could scale up with enough flexibility to tackle problems in artificial intelligence remains an open question.

Here, we argue that the time is ripe to consider a deep learning implementation of global workspace theory. While Y. Bengio has explicitly linked his recent “consciousness prior” theory to GWT [16], his proposal focused on novel theoretical principles in machine learning (e.g. sparse factor graphs). Our approach is a complementary one, in which we emphasize practical solutions to implementing a global workspace with currently available deep learning components, while always keeping in mind the equivalent mechanisms in the brain. We hope that some of the ideas developed here will assist neuroscientists in interpreting brain data in a new or different light, and in developing novel empirical evaluations of the key operations at play in the global workspace framework.

2 Roadmap to a deep learning Global Latent Workspace

The following is a step-by-step attempt at defining necessary and sufficient components for an implementation of the global workspace in an AI system. Together, these steps define a roadmap that we (and hopefully other researchers) can follow towards achieving this goal. A major point to emphasize is that all of the described components already exist individually, and often reach or surpass human-level performance in their respective functions. As in any theoretical proposal, some of the details will likely be flawed; in addition, there might be multiple ways to implement a global workspace. Overall, we believe that the strategy outlined below is most likely to be successful.

- **Multiple specialized modules.** The first ingredient of GWT is a number \( N \geq 2 \) of independent specialized modules (see Glossary), each with their own high-level latent space (see Figure 2 for a definition and some examples of latent space). The modules could be pre-trained neural networks designed for sensory perception (visual or auditory classification, object segmentation...), natural language processing (NLP), long-term memory storage, reinforcement learning (RL) agents, motor control systems, etc. The choice of these specialized modules, of course, is critical since it determines the capabilities of the full global workspace system, and the range of tasks it may perform; however, it does not affect the remaining principles laid out below.

In theory, connecting together \( N \) feed-forward discriminative networks (each trained to classify inputs from their specific domain according to category) could suffice to build a multi-modal workspace (e.g. to preactivate the “tiger” visual recognition units when one hears the word “tiger”). In practice, however, there are many reasons why including generative networks would be beneficial—networks that produce motor or language outputs, but also sensory systems with a generative top-down pathway such as (variational) auto-encoders, GANs or predictive coding networks. This is trivially required if the global workspace is intended to
Fig. 1. Global workspace in the brain. A. Schematic illustration of GWT. Concentric circles depict peripheral (e.g., sensory inputs, motor outputs) vs. more central processes, with the global workspace at the center. Specialized modules process information independently from each other. Their outputs, when selected by bottom-up (saliency-based) or top-down (task-related) attention, can enter the global workspace. There, information processing is characterized by strong long-distance interconnectivity, such that incoming information can be broadcast to other modules. At any given time, a subset of the specialized modules is mobilized into the workspace in a data-dependent and task-dependent manner. The contents of the global workspace reflect our fluctuating consciousness.

B. Mapping of GWT onto the (monkey) brain. Visual information can propagate through the visual system and activate certain frontal regions controlling behavioral output in a feed-forward way—in this case, information remains unconscious (left). When inputs are sufficiently strong or task-relevant, they activate recurrent connections (right), resulting in “ignition” of the global workspace (a highly non-linear, all-or-none process).

C. An implementation of Global Neuronal workspace (GNW) inspired by Dehaene and Changeux [13], contains a hierarchy of processing layers with feedforward and recurrent connections (LGN: Lateral Geniculate Nucleus; V1/V4: visual areas; Par: Parietal cortex; Front: Frontal cortex).

D. This simple recurrent network accounts for the all-or-none “ignition” property of the global workspace: compared with absent or undetected inputs (respectively: Correct Rejection, Miss), the main signature of consciously perceived inputs (Hits) in the brain is an all-or-none activation (or "ignition") of frontal regions. Panels A-D reproduced, with permission, from [14]

- Global Latent Workspace (GLW). The GLW, amodal by nature, is an independent and intermediate shared latent space, trained to perform unsupervised neural translation between the N latent spaces from the specialized modules (Figure 3, Key Figure). Although there are numerous examples of supervised multi-modal translation in deep learning [22, 33], here we emphasize cycle consistency as the major unsupervised training objective for neural translation (see
Fig. 2. Examples of deep learning latent spaces: a low-dimensional space that captures the relevant structure and topology of an input domain or task. In discriminative models, it is often considered to be the last feature layer, and the first layer for generative models. Examples (projected to 2D for visualization) include: A. latent space of the MNIST digit dataset. Each image from the dataset is a point in the space on the left, colored according to digit class. Regularly sampling this space in a 2D matrix produces the image reconstructions on the right (created using the UMAP inverse transform [17]). B. Word embedding space (Word2Vec algorithm [18]). Different parts of the latent space focus on distinct semantic domains (e.g. "sea" in the inset). C. Latent space of the ImageNet natural scene dataset derived from the BigGAN generative model [19]. Each row samples different points along a single vector in the 256-D latent space. D. Face latent space from a VAE-GAN model [20]. In each column, a point is sampled from the latent space, then varying amounts of a pre-computed “smile” or “male” vector are added to it. It must be emphasized that latent representations are essentially vectors of neural activation, which can be meaningfully interpolated (as in panels A,C), but also extrapolated and more generally, manipulated with algebraic operations (as in panel D).

Box 1). This way, the GLW can potentially transcribe between any pair of modules, even those for which matched data is unavailable (for example, there is no smell systematically associated with a specific video game state; yet we can intuitively recognize when the player’s situation becomes odiferous). Of course, it will be most advantageous if the default unsupervised neural translation strategy can also be complemented by supervised objectives [34] whenever joint data is available (e.g. watching an animal while hearing the corresponding sound). The dimensionality of this intermediate space is expected to be on par with or perhaps higher than the dimension of each of the input latent spaces, but much lower than their sum. This bottleneck ensures that only relevant information is encoded at each moment in time, and forces the system to prioritize competing inputs with attention.

- **Attention.** In the brain, attention determines what information is consciously
perceived, and what is discarded [35] (although attention and consciousness can be dissociated [36]). Similarly, in the original GWT, attention selects the information that enters the workspace. In deep learning, attention has recently taken the spotlight [37], most particularly the transformer architecture used widely in NLP [38] and computer vision [39–42]. In the transformer and related networks, attention is defined as a match between queries emitted by one network layer and keys produced by another one (possibly the same layer, in the "self-attention" case); the matching score determines what information is passed on to the next stage. Similarly, we can envision a key-query matching process to select inputs that reach the GLW. If the workspace includes a latent representation of the current task [43–44], this signal can serve to emit a top-down attention query: whenever the latent space of an input module produces a matching key, the relevant information is brought into the workspace. In the absence of a clear task, or in the presence of exceptionally strong or surprising inputs, bottom-up attention can prevail: in the above terminology, salient information has a “master key” that supersedes all queries. The attention mechanism for producing keys and queries in a data-dependent and task-dependent way must be optimized via training with a specific objective function (see Outstanding Questions).

• **Internal copies.** When information from a specific module is selected to enter the workspace, a copy of the latent space activation vector is brought into the GLW. If the latent space is probabilistic (as in variational auto-encoders), a unique sample is taken from the probability distribution—this ensures that a single, unified representation inhabits the GLW at all times, in accordance with our subjective experience, and with neuroscientific evidence [45].

• **Broadcast.** The selected information is then immediately broadcast, that is, translated (via the shared latent space) into the latent space of all other modules. This translation process is automatic: there is no effort involved in consciously apprehending our inner and outer environment. It is how conscious inputs acquire “meaning”, as they suddenly connect to the corresponding linguistic, motor, visual, auditory (etc) representations. This only means that the relevant information in the relevant format is “available” to these systems (as an internal copy within the workspace), not necessarily that it will be used (i.e., effectively copied into the corresponding module). One does not always visualize the details of a conjured mental image; one does not always verbalize their thought or inner speech; one does not always act on a motor plan, etc. What determines if this information is used by those systems is whether they are themselves currently connected to the workspace (e.g. by virtue of their task-relevance). The many latent representations that are automatically formed when broadcasting conscious inputs inside the workspace, without being consciously perceived themselves (because their corresponding module is not currently connected to the workspace) may correspond to what Crick and Koch described as the penumbra of consciousness [46].
Opinion - Paper under review

Box 1. Unsupervised neural translation via cycle-consistency.
In Natural Language Processing (NLP), a neural translation system is a machine translation algorithm that uses neural networks. Standard (neural) machine translation is learnt from matched exemplars (words, sentences) in the source and target languages. However, since all languages refer to a common physical reality in the outside world (the so-called language grounding property), it is theoretically possible to learn to align linguistic representations in two (or more) languages, without access to matched corpora (Figure I). This is referred to as unsupervised neural translation. One recently proposed method relies on a cycle-consistency training objective: language alignment is successful when the successive translation from language A to language B, then back from B to A returns the original sentence [34, 47, 48]. Similar methods have been applied to neural translation between varied domains, e.g. unpaired image-to-image translation [49–51], text-to-image translation [26, 28, 52] or touch-to-image translation [32]. Domain alignment via cycle-consistency training is also at the heart of a recent surge of studies investigating unsupervised domain adaptation and transfer learning tasks [53–57].

Figure I. Alignment between linguistic representations. Latent spaces from any two languages \( X \) and \( Y \) (here, French and Japanese) share a similar topology, and can be aligned to a shared latent space \( Z \) through a transform \( W \) (adapted from [58]).

3 Functional advantages of a Global Latent Workspace

A major testable property of the proposed GLW architecture is that the whole should be more than the sum of its parts (i.e., its individual modules). In other words, the added functional properties of GLW, specified below, should result in improved performance across the entire range of modules that are connected to it. Beyond these pre-existing individual tasks, the global workspace also opens up the possibility of combining modules to perform entirely novel tasks.

To begin with, the automatic multimodal alignment of representations in GLW is an ideal way to accomplish information grounding. Sensory inputs or motor outputs, instead of meaningless vectors in their respective latent spaces, become associated with corresponding representations in other sensorimotor domains, as well as with relevant linguistic representations: this promotes semantic grounding of sensorimotor data. Conversely with sensorimotor grounding of semantic information, linguistic embedding vectors that merely capture long-range statistical relations between hollow "language tokens" are transformed by association with relevant parts of the sensory environment.
or the agent’s motor and behavioral repertoire [59]. This notion of sensorimotor grounding is thus strongly related to the Gibsonian concept of **affordance**, and more generally to Gibson’s ecological approach in brain science [60]. Ultimately, grounded latent representations can confer increased performance to every module connected to the global workspace, particularly in terms of robustness to out-of-distribution samples (including so-called “adversarial” attacks).

While grounding and affordance are immediate and automatic consequences of information entering the global workspace, such a system is capable of much more, granted time and effort. Indeed, the ability to transiently mobilize any combination of modules into the workspace in a task-dependent manner is exactly what is required of a general-purpose cognitive architecture. This way, the system can compose more general functions from specialized modules, by deploying one module’s abilities onto another module’s latent representation. This **transfer learning** enables agents to adapt to new environments and tasks by generalizing previously learned models, and is considered a core component for implementing intelligence [61, 62]. Thus, our proposal of a shared latent space as the platform of consciousness bridges the theoretical link between consciousness and general intelligence. When enough diverse modules are available, their possible combinations are virtually limitless. The price of this flexibility is time and effort: mentally composing functions is a slow, sequential process, requiring iterative calls to top-down attention in order to recruit the relevant modules, one function at a time [63]. This is what Kahneman, and after him Bengio, have dubbed **system-2 cognition** [16, 64].

One of the major functions that such a flexible mental composition system can produce is **counterfactual reasoning**, or the ability to answer “what if?” questions. This ability is at the core of many emblematic attributes of high-level cognition: imagination and creativity, planning, mental simulation, iterative reasoning about possible future states [21]. In this context, **world models** (i.e., internal models of how the environment reacts to one’s actions) are particularly useful for finding solutions to new tasks as they provide task-independent, relevant information that an agent can use for offline learning via internal simulation [65, 66].

Arguably, these hypothesized functional advantages should be turned into testable predictions. The potentially improved performance and robustness of each module can easily be verified with existing benchmarks. Transfer learning and task composition benchmarks already exist [62], although novel test-beds may be required for advanced cognitive functions. Ultimately, the cumulative advantages listed here may capture the function of consciousness in humans and animals, as well as a path towards general intelligence in machines.

4 Does GLW entail artificial consciousness?

In the original GWT, a necessary and sufficient condition for conscious perception is that the information is broadcast through the global workspace. This raises the question of whether an artificial network equipped with a global latent workspace would necessarily express (a minimal form of) consciousness. Our position is that this question is an empirical one, which cannot be addressed without committing to a specific measure of consciousness. The answer, therefore, could heavily depend on the chosen measure: integrated information [67], non-trivial information closure [68], synergistic mutual information [69], etc. A more pragmatic question may be: could such a network pass a Turing test, i.e., fool a remotely interacting human observer into thinking that
Fig. 3. Schematic of a deep learning “Global Latent Workspace”. Specialized modules are arranged in the periphery. These can be pretrained networks for any variety of tasks: sensory (object recognition, detection, segmentation, speech recognition...), motor (robotic arm control, speech production...), linguistic (text comprehension, machine translation, text-to-speech...), memory storage, or higher-level cognition- and behavior-related functions (intuitive physics engine, RL policy, task embedding...). Each module is connected to the GLW (schematically represented at the center) via an internal copy of the module’s relevant latent space. Through extensive training using a cycle-consistency objective, the workspace learns to translate between the latent space representations of any two modules, in a mostly unsupervised fashion, i.e. without or with very little need for paired data (red arrows). When bottom-up or top-down attention (not represented here) selects inputs from one module, its latent space activation is copied into the GLW, and immediately translated into representations suitable for each of the other modules. However, only a handful of these modules, those currently mobilized into the workspace, will effectively receive and process the corresponding data. For example, upon recognizing a tiger in the visual scene, the corresponding NLP word embedding for “tiger” and a flight-oriented motor plan would arise in the workspace; but the word “tiger” would only be pronounced, or the flight initiated, if the corresponding module (text-to-speech, motor output) was effectively recruited in the workspace at this instant.

5 Concluding remarks

It is not the first time that a computer implementation of GWT is suggested [16, 70–74]. What sets our stance apart is the conjunction of two factors. First, we capitalize on it is, also, human? If the model is endowed with language generation abilities, and an aptitude to interpret the outside environment via the “grounding” properties described above, we believe it might come closer than ever before. Nonetheless, it is worth noting that the global workspace focuses on the “information broadcast” property of awareness. According to Dehaene et al. [45], there is an additional self-monitoring aspect that is important to capture human and animal consciousness, and that a GLW system as we describe here might be missing. Of course, this self-monitoring itself is likely amenable to a deep learning implementation, but we defer this question to future work. Importantly, phenomenal consciousness is not necessary to benefit from the above functional advantages—a GLW artificial system should display increased functionality, whether or not it is conscious.
modern deep learning-compatible components, most of them validated in state-of-the-art neural network architectures. Second, we contemplate the underlying neuronal bases and the neuroscientific implications of the proposed scheme. Correspondingly, we hope that this work may serve two purposes. Firstly, from a cognitive neuroscience standpoint, considering how to effectively implement the global workspace theory forces us to be very concrete about each component of the theory, and thereby gives us an opportunity to refine the corresponding notions. In turn, these refined notions could help formulate new hypotheses that may be empirically tested using neuroscientific methods. Secondly, in the context of artificial intelligence, the main implication of our effort is to show that inspiration from neuro-cognitive architectures may have important functional benefits. GLW could serve to improve specific machine learning tasks or benchmarks by augmenting existing architectures, thanks to the added robustness conferred by the grounding of representations inside the workspace. But GLW could also be a way to develop entirely novel architectures capable of planning, reasoning and thinking through the flexible reconfiguration of multiple existing modules. This may bring us one step closer to general-purpose (system-2) artificial cognition.
Outstanding questions

- Having a roadmap towards GLW does not imply that this goal is easy to reach—actual implementation will involve much trial-and-error, and as yet unknown computational resources. Can this goal be achieved today, in a few years, or only much later?

- Is there a minimal number of modules $N$ feeding into the workspace? When does bimodal, trimodal, multimodal integration become a “global workspace”?

- How can newly learned tasks or modules be connected to an existing GLW? Requirements include: a new “internal copy” with a new (learned) attention mechanism to produce keys for the latent space, new (learned) translations to the rest of the workspace.

- How does attention learn to select the relevant information to enter the GLW? What is the corresponding objective function? According to Bengio, it is the ability of previous conscious states to predict the current state of the world. Many alternatives exist and could be tested: free energy, survival, reward of a RL agent, metalearning (learning progress), etc.

- Is cycle-consistency a form of predictive coding? When inputs reach the GLW, they are broadcast to all other domains, before bouncing back to the initial domain, where they can be compared to the original input in order to estimate cycle-consistency. This resembles the reconstruction objective of predictive coding: the GLW “predicts” its inputs, and optimizes its prediction accuracy, both in terms of convergence of neural activity over successive time-steps, and learning of synaptic weights over successive trials.

- What could be the neuronal implementation in the brain of automatic translation via cycle consistency? Can we identify neurons, e.g. in frontal regions, that incarnate copies of the various latent spaces? This may explain the numerous reports of sensory and multimodal neuronal responses in frontal cortex.

- Could synesthesia be the consequence of an exaggerated or overactive translation between domains, crossing the threshold of perception instead of acting as a background process?
Glossary

- **affordance**: objects and events are interpreted according to the options they offer an observer in terms of available uses (including mental usage) and possible actions: their *affordances*
- **attention**: bottom-up or top-down selection of information to enter the workspace, by means of matching *query* and *key* vectors
- **broadcast**: automatic translation of incoming information from one selected module into a format suitable for the latent space of all other modules
- **counterfactual**: resulting from simulation of possible situations, without a direct connection to reality or facts
- **cycle-consistency**: objective function for translation between two domains A and B, whereby successive translations from A to B and from B back to A should retrieve the original input
- **discriminative/generative network**: a neural network in which information flows from the external environment towards the latent space is called *discriminative*, and *generative* for the opposite direction; some networks can be both (with bidirectional information flow)
- **grounding**: how representations from one domain acquire “meaning”, by associating them with other related (and possibly unrelated) domains
- **internal copy**: the GLW contains an internal copy of each module’s latent space, used for automatic translation and broadcast; recruiting a module into the workspace amounts to effectively connecting this internal copy to the corresponding latent space
- **latent space**: low-dimensional space that captures the structure and topology of an input and/or output domain (for discriminative or generative networks, respectively)
- **module**: a specialized system, operating independently of the GLW, but capable of connecting to it when needed (to achieve this, the module’s latent space gets clamped to its internal copy in the workspace)
- **neural translation**: machine translation algorithm that uses neural networks
- **objective function**: the measure that a network aims to optimize via training
- **penumbra**: according to Crick and Koch, the ensemble of neural activity produced by the current conscious state, yet not strictly part of it
- **supervised/unsupervised learning**: training a network with/without a desired output corresponding to each input
- **system-2**: cognitive architecture capable of deliberate planning and reasoning, typically slow and effortful compared to immediate perceptual awareness, well-practiced tasks or reflexive behaviors
- **transfer learning**: application of a model trained on one problem to a distinct but related problem. Domain adaptation tasks are a subset of transfer learning
- **world model**: internal model of the environment and its reactions to the agent’s actions, particularly useful for counterfactual reasoning
Highlights

- In recent years, deep learning has steadily improved the state-of-the-art in artificial intelligence, but mainly for single, well-defined tasks or challenges
- Novel advanced neural network architectures are needed to create more general-purpose AI systems with flexible and robust capabilities
- The 30-year old Global Workspace Theory proposed such an architecture; we now consider its implementation in a deep learning framework
- Recent advances in natural language processing allow for unsupervised neural translation via cycle-consistency training (without paired data in the two languages); a similar method could serve to translate inputs reaching the global workspace from/to any modality
- Flexible routing of information by attention has emerged in so-called differentiable neural computers and the related transformer models; a similar form of attention could select inputs/outputs to the workspace

Acknowledgments

RV is supported by an ANITI (Artificial and Natural Intelligence Toulouse Institute) Research Chair (grant ANR-19-P13A-0004), and two ANR grants AI-REPS (ANR-18-CE37-0007-01) and OSCI-DEEP (ANR-19-NEUC-0004). RK is supported by Japan Science and Technology Agency (JST) CREST project. We wish to thank Leila Reddy, Thomas Serre, Andrea Alamia, Milad Mozafari and Benjamin Devillers for helpful comments on the manuscript.

References

1. Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. Psychological review, 65(6), 386.
2. McClelland, J. L., Rumelhart, D. E., Group, P. R. et al. (1986). Parallel distributed processing. Explorations in the Microstructure of Cognition, 2, 216–271.
3. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444.
4. Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. Annual review of vision science, 1, 417–446.
5. Marblestone, A. H., Wayne, G., & Kording, K. P. (2016). Toward an integration of deep learning and neuroscience. Frontiers in computational neuroscience, 10, 94.
6. Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S. et al. (2019). A deep learning framework for neuroscience. Nature neuroscience, 22(11), 1761–1770.
7. VanRullen, R. (2017). Perception science in the age of deep neural networks. Frontiers in psychology, 8, 142.
8. Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. Nature neuroscience, 19(3), 356–365.
9. Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and brain sciences, 40*.

10. Baars, B. J. (1993). *A cognitive theory of consciousness*. Cambridge University Press.

11. Baars, B. J. (2005). Global workspace theory of consciousness: Toward a cognitive neuroscience of human experience. *Progress in brain research, 150*, 45–53.

12. Dehaene, S., Kerszberg, M., & Changeux, J.-P. (1998). A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the national Academy of Sciences, 95*(24), 14529–14534.

13. Dehaene, S., & Changeux, J.-P. (2005). Ongoing spontaneous activity controls access to consciousness: A neuronal model for inattentive blindness. *PLoS Biol, 3*(5), e141.

14. Van Vugt, B., Dagnino, B., Vartak, D., Safaai, H., Panzeri, S., Dehaene, S., & Roelfsema, P. R. (2018). The threshold for conscious report: Signal loss and response bias in visual and frontal cortex. *Science, 360*(6388), 537–542.

15. Mashour, G. A., Roelfsema, P., Changeux, J.-P., & Dehaene, S. (2020). Conscious processing and the global neuronal workspace hypothesis. *Neuron, 105*(5), 776–798.

16. Bengio, Y. (2017). The consciousness prior. *arXiv preprint arXiv:1709.08568*.

17. McInnes, L., Healy, J., Saul, N., & Großberger, L. (2018). Umap: Uniform manifold approximation and projection. *Journal of Open Source Software, 3*(29), 861.

18. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. Distributed representations of words and phrases and their compositionality. In: *Advances in neural information processing systems*. 2013, 3111–3119.

19. Brock, A., Donahue, J., & Simonyan, K. Large scale gan training for high fidelity natural image synthesis. In: *6th international conference on learning representations, iclr 2018*. 2018.

20. Larsen, A. B. L., Sønderby, S. K., Larochelle, H., & Winther, O. Autoencoding beyond pixels using a learned similarity metric. In: *International conference on machine learning, icml 2016*. PMLR, 2016, 1558–1566.

21. Kanai, R., Chang, A., Yu, Y., Magrans de Abril, I., Biehl, M., & Guttenberg, N. (2019). Information generation as a functional basis of consciousness. *Neuroscience of Consciousness, 2019*(1), niz016.

22. Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., & Mikolov, T. Devise: A deep visual-semantic embedding model. In: *Advances in neural information processing systems*. 2013, 2121–2129.

23. Desai, K., & Johnson, J. (2020). Virtex: Learning visual representations from textual annotations. *arXiv preprint arXiv:2006.06666*.

24. Karpathy, A., & Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. In: *Proceedings of the ieee conference on computer vision and pattern recognition*. 2015, 3128–3137.

25. Silberer, C., & Lapata, M. Learning grounded meaning representations with autoencoders. In: *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers)*. 2014, 721–732.

26. Qiao, T., Zhang, J., Xu, D., & Tao, D. Mirrorgan: Learning text-to-image generation by redescription. In: *Proceedings of the ieee conference on computer vision and pattern recognition*. 2019, 1505–1514.
27. Kim, E., Hannan, D., & Kenyon, G. Deep sparse coding for invariant multimodal halle berry neurons. In: Proceedings of the ieee conference on computer vision and pattern recognition. 2018, 1111–1120.

28. Joseph, K., Pal, A., Rajanala, S., & Balasubramanian, V. N. C4synth: Cross-caption cycle-consistent text-to-image synthesis. In: 2019 ieee winter conference on applications of computer vision (wacv). IEEE. 2019, 358–366.

29. Tsai, Y.-H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L.-P., & Salakhutdinov, R. Multimodal transformer for unaligned multimodal language sequences. In: Proceedings of the 57th annual meeting of the association for computational linguistics. 2019, 6558–6569.

30. Sun, C., Myers, A., Vondrick, C., Murphy, K., & Schmid, C. Videobert: A joint model for video and language representation learning. In: Proceedings of the ieee international conference on computer vision. 2019, 7464–7473.

31. Harwath, D., Recasens, A., Suré, D., Chuang, G., Torralba, A., & Glass, J. Jointly discovering visual objects and spoken words from raw sensory input. In: Proceedings of the european conference on computer vision (eccv). 2018, 649–665.

32. Li, Y., Zhu, J.-Y., Tedrake, R., & Torralba, A. Connecting touch and vision via cross-modal prediction. In: Proceedings of the ieee conference on computer vision and pattern recognition. 2019, 10609–10618.

33. Pham, H., Liang, P. P., Manzini, T., Morency, L.-P., & Páczos, B. (2019). Found in translation: Learning robust joint representations by cyclic translations between modalities. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01), 6892–6899. https://doi.org/10.1609/aaai.v33i01.33016892

34. Artetxe, M., Labaka, G., Agirre, E., & Cho, K. Unsupervised neural machine translation. In: 6th international conference on learning representations, iclr 2018. 2018.

35. Posner, M. I. (1994). Attention: The mechanisms of consciousness. Proceedings of the National Academy of Sciences, 91(16), 7398–7403.

36. Koch, C., & Tsuchiya, N. (2007). Attention and consciousness: Two distinct brain processes. Trends in cognitive sciences, 11(1), 16–22.

37. Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwińska, A., Colmenarejo, S. G., Grefenstette, E., Ramalho, T., Agapiou, J. et al. (2016). Hybrid computing using a neural network with dynamic external memory. Nature, 538(7626), 471–476.

38. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. Attention is all you need. In: Advances in neural information processing systems. 2017, 5998–6008.

39. Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., & Tang, X. Residual attention network for image classification. In: Proceedings of the ieee conference on computer vision and pattern recognition. 2017, 3156–3164.

40. Ramachandran, P., Parmar, N., Vaswani, A., Bello, I., Levskaya, A., & Shlens, J. Stand-alone self-attention in vision models. In: Advances in neural information processing systems 32. 2019, 68–80. http://papers.nips.cc/paper/8302-stand-alone-self-attention-in-vision-models.pdf

41. Bello, I., Zoph, B., Vaswani, A., Shlens, J., & Le, Q. V. Attention augmented convolutional networks. In: Proceedings of the ieee international conference on computer vision. 2019, 3286–3295.
42. Zhao, H., Jia, J., & Koltun, V. Exploring self-attention for image recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020, 10076–10085.

43. Zamir, A. R., Sax, A., Shen, W., Guibas, L. J., Malik, J., & Savarese, S. Taskonomy: Disentangling task transfer learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018, 3712–3722.

44. Achille, A., Lam, M., Tewari, R., Ravichandran, A., Maji, S., Fowlkes, C. C., Soatto, S., & Perona, P. Task2vec: Task embedding for meta-learning. In: Proceedings of the IEEE international conference on computer vision. 2019, 6430–6439.

45. Dehaene, S., Lau, H., & Kouider, S. (2017). What is consciousness, and could machines have it? Science, 358(6362), 486–492.

46. Crick, F., & Koch, C. (2003). A framework for consciousness. Nature neuroscience, 6(2), 119–126.

47. He, D., Xia, Y., Qin, T., Wang, L., Yu, N., Liu, T.-Y., & Ma, W.-Y. Dual learning for machine translation. In: Advances in neural information processing systems. 2016, 820–828.

48. Lample, G., Conneau, A., Denoyer, L., & Ranzato, M. Unsupervised machine translation using monolingual corpora only. In: 6th international conference on learning representations, iclr 2018, 2018.

49. Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. 2017, 2223–2232.

50. Liu, M.-Y., Breuel, T., & Kautz, J. Unsupervised image-to-image translation networks. In: Advances in neural information processing systems. 2017, 700–708.

51. Yi, Z., Zhang, H., Tan, P., & Gong, M. Dualgan: Unsupervised dual learning for image-to-image translation. In: Proceedings of the IEEE international conference on computer vision. 2017, 2849–2857.

52. Chaudhury, S., Dasgupta, S., Munawar, A., Khan, M. A. S., & Tachibana, R. Text to image generative model using constrained embedding space mapping. In: 2017 IEEE 27th international workshop on machine learning for signal processing (mlsp). IEEE, 2017, 1–6.

53. Hoffman, J., Tzeng, E., Park, T., Zhu, J.-Y., Isola, P., Saenko, K., Efros, A., & Darrell, T. CyCADA: Cycle-consistent adversarial domain adaptation (J. Dy & A. Krause, Eds.). In: ed. by Dy, J., & Krause, A. 80. Proceedings of Machine Learning Research. Stockholmsmässan, Stockholm Sweden: PMLR, 2018, October, 1989–1998. [http://proceedings.mlr.press/v80/hoffman18a.html]

54. Hui, L., Li, X., Chen, J., He, H., & Yang, J. Unsupervised multi-domain image translation with domain-specific encoders/decoders. In: 2018 24th international conference on pattern recognition (icpr). IEEE, 2018, 2044–2049.

55. Murez, Z., Kolouri, S., Kriegman, D., Ramamoorthi, R., & Kim, K. Image to image translation for domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition (cvpr). 2018, June.

56. Hosseini-Asl, E., Zhou, Y., Xiong, C., & Socher, R. Augmented cyclic adversarial learning for low resource domain adaptation. In: International conference on learning representations. 2018.
57. Chen, Y.-C., Lin, Y.-Y., Yang, M.-H., & Huang, J.-B. Crdoco: Pixel-level domain transfer with cross-domain consistency. In: Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr). 2019, June.

58. Conneau, A., Lample, G., Ranzato, M., Denoyer, L., & Jégou, H. Word translation without parallel data. In: International conference on learning representations. 2018.

59. Tan, H., & Bansal, M. Vokenization: Improving language understanding via contextualized, visually-grounded supervision. In: Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp). 2020, 2066–2080.

60. Gibson, J. J. (1979). The ecological approach to visual perception. Psychology Press.

61. Legg, S., & Hutter, M. (2007). Universal intelligence: A definition of machine intelligence. Minds and machines, 17(4), 391–444.

62. Chollet, F. (2019). On the measure of intelligence. arXiv preprint arXiv:1911.01547.

63. Sackur, J., & Dehaene, S. (2009). The cognitive architecture for chaining of two mental operations. Cognition, 111(2), 187–211.

64. Kahneman, D. (2011). Thinking, fast and slow. Macmillan.

65. Ha, D., & Schmidhuber, J. Recurrent world models facilitate policy evolution. In: Advances in neural information processing systems. 2018, 2450–2462.

66. Hafner, D., Lillicrap, T., Ba, J., & Norouzi, M. Dream to control: Learning behaviors by latent imagination. In: International conference on learning representations. 2020.

67. Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: From consciousness to its physical substrate. Nature Reviews Neuroscience, 17(7), 450–461.

68. Chang, A. Y., Biehl, M., Yu, Y., & Kanai, R. (2020). Information closure theory of consciousness. Frontiers in Psychology, 11.

69. Griffith, V., & Koch, C. Quantifying synergistic mutual information. In: Guided self-organization: Inception. Springer, 2014, pp. 159–190.

70. Franklin, S., & Patterson, F. (2006). The lida architecture: Adding new modes of learning to an intelligent. Autonomous, Software Agent IDPT-2006.

71. Shanahan, M. (2006). A cognitive architecture that combines internal simulation with a global workspace. Consciousness and cognition, 15(2), 433–449.

72. Bao, C., Fountas, Z., Olugbade, T., & Bianchi-Berthouze, N. Multimodal data fusion based on the global workspace theory. In: Proceedings of the international conference on multimodal interactions, icmi 2020. 2020.

73. Safron, A. (2020). An integrated world modeling theory (iwmt) of consciousness: Combining integrated information and global neuronal workspace theories with the free energy principle and active inference framework; toward solving the hard problem and characterizing agentic causation. Frontiers in Artificial Intelligence, 3, 30. https://doi.org/10.3389/frai.2020.00030

74. Kotseruba, I., & Tsotsos, J. K. (2020). 40 years of cognitive architectures: Core cognitive abilities and practical applications. Artificial Intelligence Review, 53(1), 17–94.