Deep learning and the Global Workspace Theory

Rufin Vanrullen, Ryota Kanai

Deep Learning and the Global Workspace Theory

Rufin VanRullen[1, 2] and Ryota Kanai[3]

[1] *CerCo, CNRS UMR5549, Toulouse, France*
[2] *ANITI, Université de Toulouse, France*
[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, along with neuroscientific implications.

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 [1, 2], it is only in the last decade that these systems have started to fully reveal their potential [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 [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 [10, 11], is a key element of modern cognitive science (Figure 1A). The theory proposes that the brain is divided into specialized modules for specific functions, with long-distance connections between them [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 [12–16] proposed a neuronal version of the theory, Global Neuronal Workspace (GNW), which has become one of the major contemporary neuroscientific theories of consciousness. 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 neural signature of this global broadcast of information is the ignition property: an all-or-none activation of a broad network of brain regions, likely supported by long-range recurrent connections (Figure 1C).

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 [17], 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 towards achieving this goal, and highlight important issues and predictions for neuroscience research. 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. The value of our proposal is, therefore, to identify the appropriate components and the manner in which they should interact, so as to optimize functionality while remaining truthful to neuroscience findings. 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. Nonetheless, 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**. In deep learning, a latent space is a representation layer trained to encode the key elements of an input domain. This information corresponds to high-level conceptual representations such as visual object features, word meaning, chunks of action sequences, etc. (Figure 2). 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

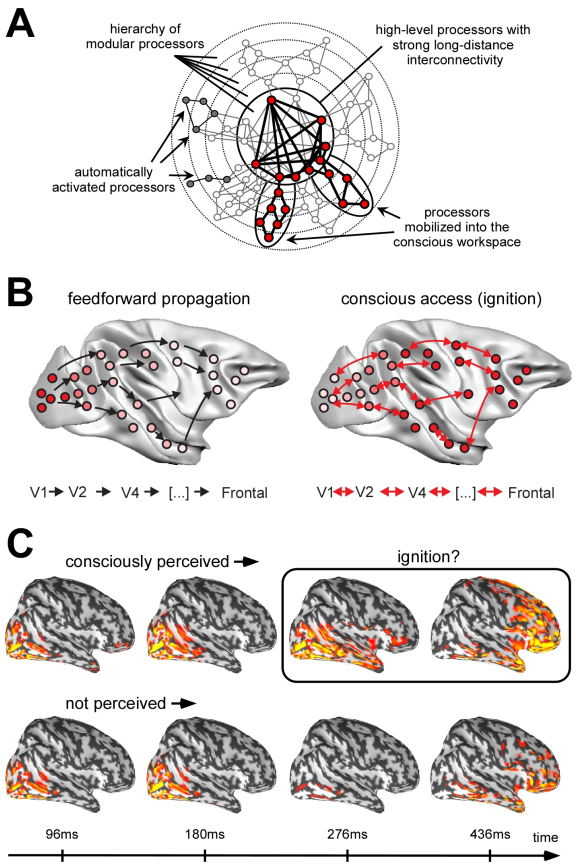

Fig. 1. Global workspace in the brain. A. Schematic illustration of GWT. Concentric circles depict peripheral (e.g. sensory inputs, motor outputs) vs. more central processes, with the global workspace at the center. Specialized modules process information independently from each other. Their outputs, 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. Redrawn from [10]. **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 (right), they activate local recurrent connections, resulting in "ignition" of the global workspace (a highly non-linear, all-or-none process, characterized by global recurrence across a network of long-range connections). Reproduced, with permission, from [15]. **C.** In certain experimental situations, the same sensory stimulus sometimes reaches consciousness (top row), and sometimes remains unconscious (bottom row). In human magneto-encephalography (MEG) recordings, the main signature of consciously perceived inputs is a late all-or-none activation (or "ignition") of frontal regions, accompanied by sustained activity in sensory regions. Adapted, with permission, from [18].

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 predic-

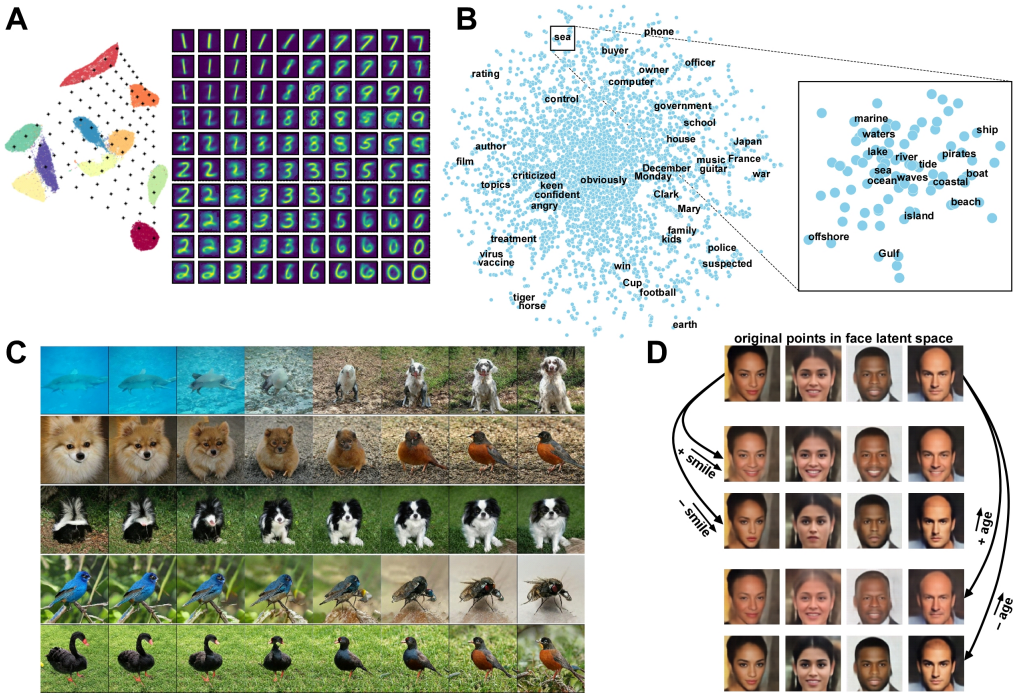

Fig. 2. Examples of deep learning latent spaces: a low-dimensional space that captures the relevant structure and topology of an input domain or task. In discriminative models, it is often considered to be the last feature layer, and the first layer for generative models. Examples (projected to 2D for 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 [19]). **B.** Word embedding space (Word2Vec algorithm [20]). 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 [21]. Each row samples different points along a single vector in the 256-D latent space. **D.** Face latent space from a VAE-GAN model [22]. In each column, a point is sampled from the latent space, then varying amounts of a pre-computed "smile" or "age" 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).

tive coding networks. This top-down pathway is trivially required if the global workspace is intended to influence the system's behavioral output. It is also necessary (though certainly not sufficient) in order to endow the system with creative or "imagination" abilities (e.g. generation of mental images), and more generally, to perform mental simulation, planning or "thinking" by iteratively conjuring up a possible future state or **counterfactual** state [23]. Finally, a recurrent top-down pathway may be key to account for the global **ignition** property observed in the brain, when an input reaches consciousness and the corresponding module is mobilized into the conscious global workspace (Figure 1B,C).

- **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 [24–38], here we emphasize **cycle consistency** as the major unsupervised training objective for neural translation (see Box 1). In brief, the translation system is optimized such that successive translation and back-translation (e.g. a cycle from language A to B, then back to A) always returns the original input. Using this strategy, 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 [39] 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 [40] (although attention and consciousness can be dissociated [41, 42]). Similarly, in the original GWT, attention selects the information that enters the workspace. In deep learning, attention has recently taken the spotlight [43], most particularly the transformer architecture used widely in NLP [44] and computer vision [45–48]. Although the term "attention" is the same, there are important differences between the neuroscience and machine learning usage of the notion [49]. 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 and accordingly, to break existing connections or create new ones. If the workspace includes a latent representation of the current task [50, 51], this signal can serve to emit a top-down attention *query*, compared against the current "key" vectors from all the candidate inputs to the workspace: whenever the latent space of an input module produces a matching *key*, the module is connected and 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 capture can prevail: in the above terminology, salient information has a "master key" that supersedes all queries, i.e., that can grant access to the workspace regardless of the current query. 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 a specific module is connected to the workspace as a result of attentional selection, its latent space activation vector is copied into the GLW. This **internal copy** serves the role of a bidirectional connection interface between the corresponding module and the GLW.

- **Broadcast.** The incoming 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 transferred 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 [52].

3 Global Latent Workspace in action

To clarify the inner workings of the proposed workspace, let us follow its step-by-step operations during a standard scenario (as illustrated also in Figure 3B). Before any stimulus appears, the prior state of the system, including the current task setting or instructions, can preset some modules to be connected to the GLW, while others remain disconnected (Step 0 in Figure 3B). "Connected" means that the latent space of the module is temporarily clamped, in a bidirectional way, to its internal copy in the workspace. If a stimulus appears in a disconnected module, it will not reach the workspace directly; but it may still affect the attentional system (not represented in Figure 3B), which may eventually result in the connection of the relevant module (either because the corresponding key matches the top-down attention query; or because it is a bottom-up "master key"). If a new stimulus appears in a connected module, the latent activity is immediately transferred to the corresponding "internal copy" inside the workspace (Step 1 in Figure 3B). Hence starts the broadcast, that is, an automatic translation to all other domains: via the shared latent space, each internal copy (no matter whether its module is connected or not) receives a translation of the new input in its own "language". In turn, the activation from each internal copy will return to the shared latent space and potentially modify it, reverberating ad infinitum inside the GLW. This might be what "ignition" means (Figure 1C): long-range and long-lasting recurrent interactions between the latent spaces of the different modules. The shared latent space can use the translations and corresponding "back-translations" from all modules to compute its "cycle-consistency" error, required to train or fine-tune the unsupervised neural translation system (e.g. via error backpropagation).

What will modules do with the broadcasted information they receive on their internal copy (Step 2 in Figure 3B)? If the module is disconnected, the broadcast only reaches the internal copy, but not the actual module. Still, this can be helpful for grounding and affordance, as described more fully in the next section. If the module is connected, that is, if its latent space is clamped to the internal copy, the broadcasted information reaching the internal copy will also modify the latent space and potentially the inner layers of the module (for a generative module). For a language network, speech may be produced; for a movement network, an evasive action or a break-dance move may be launched; for a generative visual network, an image reflecting the contents of the workspace may be summoned, etc. This is what it means for a network to be "recruited" in the workspace: because of the bidirectional connection, the inner network activity directly affects the GLW, but is also directly affected by activity changes in the workspace.

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), their associated semantic representation spaces are likely to share a similar topology: for example, the words "cat" and "dog" are likely to be found close together, while the word "machine" would be more distant (see Figure I). Therefore, it is theoretically possible to learn to align linguistic representations in two (or more) languages based solely on the geometry of their semantic representation spaces, 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 [39, 53, 54]. Similar methods have been applied to neural translation between varied domains, e.g. unpaired image-to-image translation [55–57], **text-to-image** translation [28, 31, 58] or **touch-to-image** translation [36]. 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 [59–64].
We suggest that the core challenge for any artificial system based on GWT is in fact a problem of unsupervised neural translation: learning and retrieving appropriate correspondences between elements of distinct domains or modalities, which may not always directly co-occur in the environment. Accordingly, our framework places a strong emphasis on cycle-consistency as an objective function for training the translation mechanism at the heart of the Global Latent Workspace.

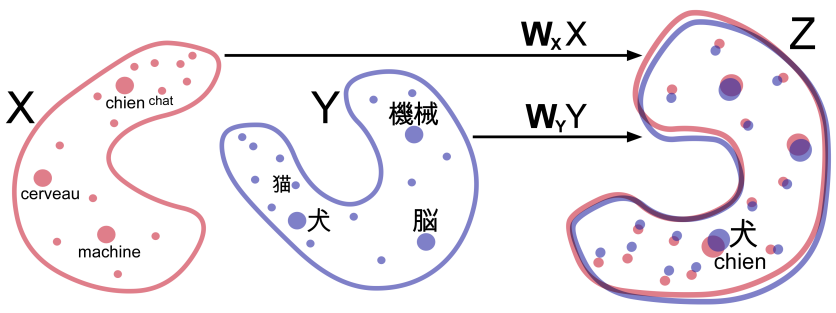

Figure I. Alignment between linguistic representations. Latent spaces from any two languages X and Y (here, French and Japanese) share a similar topology, and can be aligned to a shared latent space Z through a transform W (adapted from [65]).

4 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 (and leaving aside the possible emergence of conscious experience, which

we address in the next section), 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 [66]. 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 [67]. Ultimately, grounded latent representations can confer increased performance to every module connected to the global workspace. We thus predict that GLW should result in performance improvements, particularly in terms of robustness to out-of-distribution samples (including so-called "adversarial" attacks [68]).

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 [69, 70]. 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 [71]. This is what Kahneman, and after him Bengio, have dubbed **system-2** cognition [17, 72].

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. In this context, a particularly useful module could be a "world model". This is an internal model of how the environment reacts to one's actions, which can be queried iteratively as a "forward model" to predict future states of the world given an initial state and possible action [73, 74]. This function is at the core of many emblematic attributes of high-level cognition: imagination and creativity, planning, mental simulation, iterative reasoning about possible future states [23].

Arguably, 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.

5 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

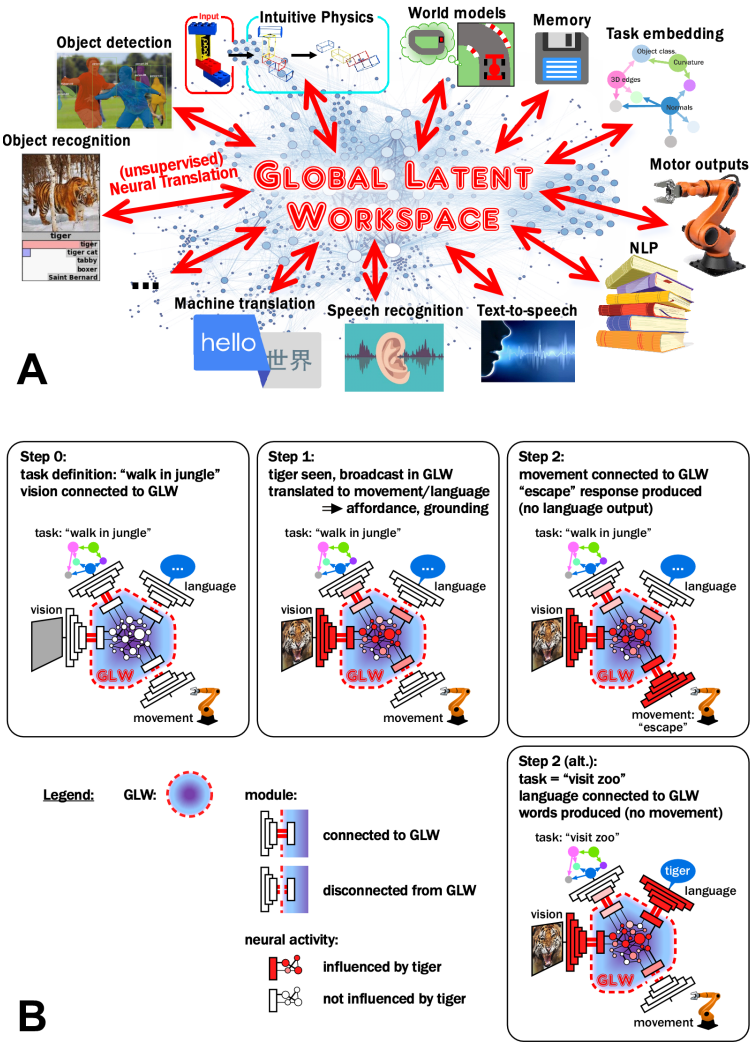

Fig. 3. Schematic of a deep learning "Global Latent Workspace" (A) and its operation (B). A. Specialized modules are arranged in the periphery. These can be pre-trained networks for any variety of tasks: sensory (object recognition, detection, segmentation, speech recognition...), motor (robotic arm 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, world model...). Each module is connected to the GLW (schematically represented at the center) via an internal copy of the module's relevant latent space, effectively acting as a connection interface.. 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). **B**. When bottom-up or top-down attention (not represented here) selects inputs from one module (Step 0), its latent space activation is copied into the GLW, and immediately translated into representations suitable for each of the other modules (Step 1). 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 (Step 1); but the flight would only be initiated (Step 2), or the word "tiger" pronounced (Step 2alt.), if the corresponding module (motor output, text-to-speech) was effectively recruited in the workspace at this instant.

necessarily express (a minimal form of) consciousness. In philosophy of mind and in related neuroscientific theories of consciousness, two aspects of consciousness are usually distinguished [75]: **phenomenal consciousness** is the immediate subjective experience of sensations, perceptions, thoughts, wants and emotions; **access consciousness** requires further consolidation, and is used for reasoning and executive control of actions, including language. GWT does not explicitly distinguish between these two forms of consciousness, but other authors have suggested that local recurrence could be sufficient for phenomenal awareness, while global recurrence is a hallmark of access consciousness [41, 76]. In this view, the global workspace ignition that produces global recurrence of brain activity would more naturally map to access consciousness. Indeed, the functional advantages that we highlight in terms of flexible cognitive control seem in line with the definition of access consciousness, and do not critically depend on the emergence of phenomenal consciousness. Still, there are two aspects of GLW that may be conducive to a form of phenomenal consciousness. First, the grounding and affordance properties could account for the emergence of implicit associations between different sensorimotor properties of an object as well as the associated declarative knowledge (e.g., the word that comes on the tip of the tongue, the different ways we know that we could grasp an object if we decided to, etc). Second, the recruitment of a sensory module into the workspace could explain the vivid and detailed nature of our sensory phenomenal experience: as the connection between the module and its internal copy is bidirectional, the workspace can access sensory information but can also modify it and enrich it with semantically grounded information. Thus, on the one hand, GLW could reasonably be viewed as a way to endow an artificial system with phenomenal consciousness. On the other hand, 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 [77], non-trivial information closure [78], synergistic mutual information [79], etc.

Finally, it is worth noting that the global workspace focuses on the "information broadcast" property of awareness. According to Dehaene et al. [80], 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.

6 Implications for Neuroscience

A global workspace using unsupervised neural translation to broadcast information between internal copies of every modality-specific latent space—if it exists in the brain—should have a number of telltale signatures that could be explored by neuroscientists. An internal copy, for example, would correspond to a population of neurons having a clear preference for a specific sensorimotor domain or modality, but whose response is heavily influenced by high-level, semantic or multimodal information (the grounding and affordance properties). While there are many candidate high-level or multimodal regions in the brain, the concept of internal copy further implies that the activation of this neural population (i) could happen without stimulation of its preferred modality, yet (ii) would be systematically coupled to a global ignition of the entire workspace.

Training the translation mechanism by optimizing cycle-consistency may be relatively straightforward to implement with biological neurons, by making the networks mutually predictive of each other. In this sense, cycle-consistency could be envisioned as a form of "predictive coding", a well-studied framework in neuroscience [81, 82]. Broadcast

implies a recurrent loop between translations, back-translations and error estimations, resembling the prediction error mimimization objective of predictive coding. As this sort of error minimization loop is also known to be a source of brain oscillations [82, 83], we further suggest that internal copy neurons in the brain could be characterized by oscillatory responses at a specific frequency.

A dedicated attention system is required to control the workspace inputs and outputs. In our framework (inspired by the deep learning transformer architecture [44]), the workspace constantly emits context-dependent attention queries, each module emits attention keys, and the match between keys and queries determines the module's connection status. In the brain, this would correspond to endogenous attention systems, particularly the dorsal part of the frontoparietal network responsible for top-down attention control [84]. As we explained, it would be advantageous if the modules with especially salient inputs had the ability to emit "master keys" to force their recruitment into the workspace, regardless of the current query. This is a form of bottom-up attention capture, reminiscent of the "circuit-breaking" property of exogenous attention in the brain [84, 85]. Finally, while the workspace requires a dedicated and unified attention system, this does not preclude the existence of other independent attention systems within each module. Similarly in the brain, there are global forms of attention to select entire modalities while inhibiting others, but also more "local" forms of attention operating within each modality, e.g. to highlight one object among others [86]. For optimal performance, these multiple attention systems should be allowed to interact [87], for instance by sharing queries. The resulting widespread network of within- and between-modality attention systems could correspond to the so-called frontoparietal attention network [84, 88].

Common neuronal or cognitive phenomena may be revisited in the light of our proposed framework. For instance, the suppression of consciousness during general anesthesia has been linked to a specific impairment of long-range connections [16, 89], which are crucial for the normal operation of the global workspace—specifically for ignition, broadcast and translation. A model like the one we propose could serve to perform pre-clinical studies, e.g. to evaluate how various aspects of consciousness depend on certain anesthetic drug targets. Just like anesthesia may reflect impaired translation mechanisms, synesthesia could be related to hyperactive translation. Humans show the ability to discover patterns through analogical reasoning, as well as a natural tendency to connect seemingly unrelated stimuli in a consistent manner–a tendency that culminates in the arbitrary and mandatory cross-modal associations of synesthesia [90]. Yet neuronal mechanisms for such phenomena have been elusive. The unsupervised neural translation we discussed here offers a possible algorithmic method to establish such alignment of high-level representations across modalities, and could thus help understand the origins of synesthesia.

7 Concluding remarks

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. Of course, it is not the first time that a computer implementation of GWT is suggested [17, 91–95]. What sets our stance apart is the conjunction of two factors. First, we capitalize on 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

- A global workspace serves to flexibly connect neural representations arising in multiple separate modules. Is there a minimal number of modules feeding into the workspace? When does bimodal, trimodal, multimodal integration become a "global workspace"?
- 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.
- Is cycle-consistency implemented in the brain? If yes, does it correspond to a form of predictive coding?
- 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?
- How does attention learn to select the relevant information to enter the GLW? What is the corresponding objective function? Many candidates exist and could be tested: self-prediction, free energy, survival, reward of a RL agent, metalearning (learning progress), etc.
- 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.

Glossary

Our terminology is borrowed from different fields, with the same term sometimes taking distinct meanings across the fields. To alleviate any confusion, we begin each definition by indicating whether the term is employed in a way traditionally associated with Cognitive Neuroscience *(Neuro)* or AI *(AI)*.

- **affordance:** *(Neuro)* 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:** *(AI)* bottom-up or top-down selection of information to enter the workspace, by means of matching *query* and *key* vectors

- **broadcast:** *(AI)* automatic translation of incoming information from one selected module into a format suitable for the latent space of all other modules
- **counterfactual:** *(Neuro/AI)* resulting from simulation of possible situations, without a direct connection to reality or facts
- **cycle-consistency:** *(AI)* 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:** *(AI)* 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:** *(Neuro)* how representations from one domain acquire "meaning", by associating them with other related (and possibly unrelated) domains
- **internal copy:** *(AI)* 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:** *(AI)* low-dimensional space that captures the structure and topology of an input and/or output domain (for discriminative or generative networks, respectively)
- **module:** *(AI)* 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:** *(AI)* machine translation algorithm that uses neural networks
- **objective function:** *(AI)* the measure that a network aims to optimize via training
- **penumbra:** *(Neuro)* according to Crick and Koch, the ensemble of neural activity produced by the current conscious state, yet not strictly part of it
- **phenomenal/access consciousness:** *(Neuro)* the immediate subjective experience of sensations, emotions, thoughts (etc.) is called *phenomenal* consciousness; *access* consciousness denotes information used for reasoning and executive control of actions, including language
- **supervised/unsupervised learning:** *(AI)* training a network with/without a desired output corresponding to each input
- **system-2:** *(Neuro/AI)* 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:** *(AI)* application of a model trained on one problem to a distinct but related problem. Domain adaptation tasks are a subset of transfer learning

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, possibly inspired by Neuroscience, 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
- The Global Workspace accounts for conscious information processing in the human brain, but its associated functional advantages could generalize to artificial systems
- In turn, considering an artificial global workspace can help constrain neuroscientific investigations of brain function and consciousness

Acknowledgments

RV is supported by an ANITI (Artificial and Natural Intelligence Toulouse Institute) Research Chair (grant ANR-19-PI3A-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. Sergent, C., & Dehaene, S. (2004). Neural processes underlying conscious perception: Experimental findings and a global neuronal workspace framework. *Journal of Physiology-Paris*, *98*(4-6), 374–384.

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

15. 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.

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

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

18. Sergent, C., Baillet, S., & Dehaene, S. (2005). Timing of the brain events underlying access to consciousness during the attentional blink. *Nature neuroscience*, *8*(10), 1391–1400.

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

20. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality, In *Advances in neural information processing systems*.

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

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

23. 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.

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

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

26. Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions, In *Proceedings of the ieee conference on computer vision and pattern recognition*.

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

28. Qiao, T., Zhang, J., Xu, D., & Tao, D. (2019). Mirrorgan: Learning text-to-image generation by redescription, In *Proceedings of the ieee conference on computer vision and pattern recognition*.

29. Kim, E., Hannan, D., & Kenyon, G. (2018). Deep sparse coding for invariant multimodal halle berry neurons, In *Proceedings of the ieee conference on computer vision and pattern recognition*.

30. Gorti, S. K., & Ma, J. (2018). Text-to-image-to-text translation using cycle consistent adversarial networks. *arXiv preprint arXiv:1808.04538*.

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

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

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

34. Chen, Y.-C., Li, L., Yu, L., Kholy, A. E., Ahmed, F., Gan, Z., Cheng, Y., & Liu, J. (2019). Uniter: Learning universal image-text representations. *arXiv preprint arXiv:1909.11740*.

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

36. Li, Y., Zhu, J.-Y., Tedrake, R., & Torralba, A. (2019). Connecting touch and vision via cross-modal prediction, In *Proceedings of the ieee conference on computer vision and pattern recognition*.

37. Wang, X., Ye, Y., & Gupta, A. (2018). Zero-shot recognition via semantic embeddings and knowledge graphs, In *Proceedings of the ieee conference on computer vision and pattern recognition*.

38. Pham, H., Liang, P. P., Manzini, T., Morency, L.-P., & Paczos, 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

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

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

41. Lamme, V. A. (2003). Why visual attention and awareness are different. *Trends in cognitive sciences*, *7*(1), 12–18.

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

43. 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.

44. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need, In *Advances in neural information processing systems*.

45. Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., & Tang, X. (2017). Residual attention network for image classification, In *Proceedings of the ieee conference on computer vision and pattern recognition*.

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

47. Bello, I., Zoph, B., Vaswani, A., Shlens, J., & Le, Q. V. (2019). Attention augmented convolutional networks, In *Proceedings of the ieee international conference on computer vision*.

48. Zhao, H., Jia, J., & Koltun, V. (2020). Exploring self-attention for image recognition, In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*.

49. Lindsay, G. W. (2020). Attention in psychology, neuroscience, and machine learning. *Frontiers in Computational Neuroscience*, *14*, 29. https://doi.org/10.3389/fncom.2020.00029

50. Zamir, A. R., Sax, A., Shen, W., Guibas, L. J., Malik, J., & Savarese, S. (2018). Taskonomy: Disentangling task transfer learning, In *Proceedings of the ieee conference on computer vision and pattern recognition*.

51. Achille, A., Lam, M., Tewari, R., Ravichandran, A., Maji, S., Fowlkes, C. C., Soatto, S., & Perona, P. (2019). Task2vec: Task embedding for meta-learning, In *Proceedings of the ieee international conference on computer vision*.

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

53. He, D., Xia, Y., Qin, T., Wang, L., Yu, N., Liu, T.-Y., & Ma, W.-Y. (2016). Dual learning for machine translation, In *Advances in neural information processing systems*.

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

55. Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks, In *Proceedings of the ieee international conference on computer vision*.

56. Liu, M.-Y., Breuel, T., & Kautz, J. (2017). Unsupervised image-to-image translation networks, In *Advances in neural information processing systems*.

57. Yi, Z., Zhang, H., Tan, P., & Gong, M. (2017). Dualgan: Unsupervised dual learning for image-to-image translation, In *Proceedings of the ieee international conference on computer vision*.

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

59. Hoffman, J., Tzeng, E., Park, T., Zhu, J.-Y., Isola, P., Saenko, K., Efros, A., & Darrell, T. (2018). CyCADA: Cycle-consistent adversarial domain adaptation (J. Dy & A. Krause, Eds.). In J. Dy & A. Krause (Eds.), Stockholmsmässan, Stockholm Sweden, PMLR. http://proceedings.mlr.press/v80/hoffman18a.html

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

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

62. Hosseini-Asl, E., Zhou, Y., Xiong, C., & Socher, R. (2018). Augmented cyclic adversarial learning for low resource domain adaptation, In *International conference on learning representations*.

63. Tian, Y., & Engel, J. (2019). Latent translation: Crossing modalities by bridging generative models. *arXiv preprint arXiv:1902.08261*.

64. Chen, Y.-C., Lin, Y.-Y., Yang, M.-H., & Huang, J.-B. (2019). Crdoco: Pixel-level domain transfer with cross-domain consistency, In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr)*.

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

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

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

68. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

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

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

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

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

73. Ha, D., & Schmidhuber, J. (2018). Recurrent world models facilitate policy evolution, In *Advances in neural information processing systems*.

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

75. Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and brain sciences*, *18*(2), 227–247.

76. Lamme, V. A. (2018). Challenges for theories of consciousness: Seeing or knowing, the missing ingredient and how to deal with panpsychism. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1755), 20170344.

77. 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.

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

79. Griffith, V., & Koch, C. (2014). Quantifying synergistic mutual information, In *Guided self-organization: Inception*. Springer.

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

81. 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), 79–87.

82. Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., & Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron*, *76*(4), 695–711.

83. Alamia, A., & VanRullen, R. (2019). Alpha oscillations and traveling waves: Signatures of predictive coding? *PLoS Biology*, *17*(10), e3000487.

84. Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature reviews neuroscience*, *3*(3), 201–215.

85. Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature reviews neuroscience*, *2*(3), 194–203.

86. Macaluso, E., Frith, C., & Driver, J. (2002). Directing attention to locations and to sensory modalities: Multiple levels of selective processing revealed with pet. *Cerebral Cortex*, *12*(4), 357–368.

87. Driver, J., & Spence, C. (1998). Attention and the crossmodal construction of space. *Trends in cognitive sciences*, *2*(7), 254–262.

88. Szczepanski, S. M., Pinsk, M. A., Douglas, M. M., Kastner, S., & Saalmann, Y. B. (2013). Functional and structural architecture of the human dorsal frontoparietal attention network. *Proceedings of the National Academy of Sciences*, *110*(39), 15806–15811.

89. Mashour, G. A. (2013). Cognitive unbinding: A neuroscientific paradigm of general anesthesia and related states of unconsciousness. *Neuroscience & Biobehavioral Reviews*, *37*(10), 2751–2759.

90. Hubbard, E. M., & Ramachandran, V. S. (2005). Neurocognitive mechanisms of synesthesia. *Neuron*, *48*(3), 509–520.

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

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

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

94. 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

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

Thoughts

(this section will be deleted before submission, let's use it as a notepad).
Potential [readers]/reviewers: [A. Seth, N. Tsuchyia, T. Serre, Sid Kouider, D. Ha], P. Roelfsema, A. Torralba, A. Cleeremans
To mention somewhere:

- What's the dimensionality of the GLW? 1,000-2,000 dims sounds about right.
- As in most modern deep learning approaches, we focus here on functional objectives that can be optimized via back-propagation. This learning strategy is efficient, but its biological plausibility has been questioned (refs).
- We could also cite the very recent paper by Hill et al (Deepmind) on "Grounded langugage learning fast and slow", the architecture has a lot of similarity with GWT: text+visual encoders, joint latent embedding, separate decoders with reconstruction loss, etc. (although they don't mention GWT at all). Maybe in the section on grounding? On the other hand, it's on arXiv, no proper reference yet.
- Kaiser et al (2017): one model to learn them all. I had completely forgotten about this model. It's definitely relevant...

Resolved comments Ryoya's comments are in blue Rufin's comments are in purple
In this opinion paper, we employ the notion of latent space in deep learning literature and adapt it to interpret the function of global workspace. Specifically, we view the global workspace as a latent space shared across functionally specialized neural networks. With a shared latent space (e.g. modalities such as vision and audition in the brain), information coming from various specialised networks are interpreted in the same format and can be used for solving novel problems via transfer learning or combining existing neural networks. This view offers more concrete ideas as to how the function of a global workspace is implemented by modern deep learning methods.

There are at least two ways to implement global workspace. One is to train multiple modules simultaneously to obtain joint distributions of events and instances from those specialized modules. The other is to align the relational structures of events encoded by distinct modules to match each other (e.g. [39]). The first one requires parallel/matched data to estimate the joint distribution, the second one doesn't. However, there is no reason to choose, and both of them can happen in the same system, depending on the availability of joint data. Similarly, the "unsupervised neural translation" approaches do work without parallel data, but they work much better when there is (a little) supervision data. The cycle consistency objective (relational structure alignment) is what you do by default, when you don't have joint data.

Shared latent space within the same modality (e.g. image-to-image) and across modalities (e.g. text-to-image). Is there such a representation between tasks and sensor information? Could you expand/clarify? I'm not sure to understand.It's about the task embeddings.

Do the systems need to be generative, e.g. auto-encoders, predictive coding, etc., or can discriminative networks suffice? I think in the original GWT any architecture could do. But to explain the "ignition" properties of GNWT, recurrence seems necessary, and therefore we should favor generative models? Also, how does the GW influence the specialized systems if there is no top-down route? Maybe this is a point to discuss: if you just want to create a multimodal workspace, e.g. to pre-activate the "tiger" visual recognition units when you hear the word "tiger", then discriminative models are sufficient; but if you want this activation to "recruit" the entire visual recognition hierarchy (something like the "blackboard" idea of Lamme, Bullier etc), then a generative/recurrent model seems necessary
RK:This is a difficult question and depends on how we interpret ignition. My interpretation is that ignition as the process of entering the global workspace corresponds to attention gating, i.e. key-query matching in the transformer network. Regarding how generative models are used in the context of global workspace, I interpret generative models as a simulator. The state vector from the global workspace can be used as an input to a generative model, which generates a specific image (generated data in sensor space) and this is re-encoded into the global workspace. This can be useful when we perform "thinking" by obtaining a possible future state (or counterfactual state) via iteration.
1. Agreed, ignition in our model doesn't need recurrence, it's just what happens when inputs enter the workspace (because they were selected by attention). I meant that in the brain, it seems that recurrence is the key factor that determines whether ignition takes place or inputs are "lost". At least that's the standard view of Dehaene, Lamme, etc. It's also how they obtain ignition in their GNWT implementation (using the recurrent connections). But you're right, this should not get confused with the question of feedback generative circuits in our model.

2. Regarding the "simulation" aspect: does it make sense to say (as I will argue a bit later) that the global workspace "prepares" a translation of conscious inputs from one modality into the latent space of every other modality; but this only gets "used" if the corresponding modality or network gets recruited into the workspace (via attention). In that case, the prepared latent vector is copied into the corresponding network; in the case of the generative network, this results in a "simulation" in sensor space, as you suggest. So, a generative model is not necessary here, but it presents additional advantages (simulation, thinking, counterfactual reasoning, etc.)?

RK: I think there are two ways to align embedding spaces. One is from co-occurrences such as visual and auditory events. The other method is unsupervised neural translation, which allows mapping between two specialized modules that learned to encode inputs independently. The former seems to occur automatically both in deep learning and in the brain, but the latter seems to require additional optimization processes. This should be a slow process, and may correspond to cognitive processes that require some mental effort (i.e. system 2). For the 2 methods, see my response above: they are not mutually exclusive, but rather complementary, depending on the availability of joint data. For the automatic/effortful aspect of the translation, see my response below: I want to argue that both methods can be automatic. (Consciousness happens without effort. What is "effortful" in the sense of Bengio's or Kahneman's system 2 is recruiting the right networks/task modalities into the workspace in a flexible way so that novel tasks can be performed.) effortful <= iterations

RK: Another question is whether the unsupervised neural translation needs to be performed on an ad hoc basis or is continuously performed as part of learning. (Perhaps this is for the outstanding questions) I agree this is a fundamental question, and we should probably be on the same page on this. My intuition is that the translation does/should happen all the time, but in the background: that's how conscious inputs acquire "meaning", because they suddenly "connect" to the corresponding language representation, motor representation, audio representation etc. However, this "automatic" translation only happens for consciously perceived inputs (those that enter the global workspace because they were selected by attention); and it is only a translation between high-level (unimodal) latent representations, not all the way to the (sensory) input level.
To be practical (at the risk of being wrong): If you think of the architecture in our Figure 3, each separate "modality" (object recognition, speech recognition, NLP, etc.) has its own high-level latent space (e.g. the highest feature layer of a ResNet, the output layer of a BERT, etc.). The global workspace "automatically" translates between these high-level spaces (using the cycle-consistency objective). If the visual input is selected by attention, (a copy of) its ResNet activation vector enters the workspace, and gets immediately translated, via a shared latent space, into (copies of) all the other high-level latent vectors. These don't actually connect into the corresponding modalities, unless the modalities are themselves recruited in the workspace. So, for example, my visual feature vector gets translated into a high-level latent speech representation, but that's just a copy of the text-to-speech latent space, so no speech needs to be generated. On the other hand, if the text-to-speech network is recruited in the global workspace, then the latent speech "translation" from the visual image is actually copied into the text-to-speech network, and speech is effectively produced.

Is embedding mapping (i.e., unsupervised translation of one representation to another) performed in the brain? If so, what might be the neural mechanism? I'm confused by the question. If we assume that this translation IS what makes the global latent

workspace happen, then aren't we also proposing that it does happen in the brain? Doesn't the question contradict the entire paper?What would be the brain equivalent/neuronal implementation of translation via cycle consistency? Can we identify neurons, e.g. in frontal regions, that represent copies of the latent space? (there are sensory and multimodal neurons in frontal cortex, maybe that's what they are for?).

Why haven't we built this yet? What more do we need? Just compute?
"what we need is a set of criteria that an artificial global workspace should satisfy. Demonstration of flexibility is one thing. But if we could find a few other key properties that an artificial global workspace should have, that would be great"

Comments from Andrea ALAMIA (post-doc in my group)
Overall congrats, really cool paper! it reads very easily, and it's definitely clear in its overal goal and ideas. yet there are few specifics points that I didn't get completely. I listed them below more or less in order of appearance in the text.

1. It is unclear whether the GWS speaks its own language, or whether it acts simply as a translator (or a bridge) between modules. From the text, I would understand you intend the second case, as suggested by the cycle consistency. However, it's unclear how it would be able to generate top down predictions (such as the creative or imagination abilities that u mention in the 'multiple specialized module' paragraph) if it's just a 'bridge' between modules? From which language -if not its own- it's translating its own internal top-down predictions down to the modules? This part I found it quite unclear. (this gets even more confusing in the 5th outstanding question, where you link -interestingly- predictive coding and cycle consistency. I'm lost there).

2. the query-key system is intruiguing. However, I don't understand whether the query emitted by the GLW has to match all the keys from all the lower modules at any given moment in time (or whenever a new key is produced, which may be very often given a rapid stimulation). Also, when would the GLW emit queries? Only when is actively doing something (a task), or all the time? Are these related with top-down activity, such as imagination/dreaming?

3. "a copy of the latent space activation vector is brought into the GLW." in practice what does it mean? that it is translated in the GLW language? (but does it have one?)

4. brodcast. "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)." but aren't all modules always connected? you rather mean whether their keys are matching the current task-related queries? I found some discrepancy with Deahene 2006 paper. In my possibly wrong understanding of his paper, things are framed a bit differently than as you write. In your case -if I got it- one is conscious of things that are actually used by the module (effectively copied in the module). But if I understood Dehaene paper correctly (paragraph "Distinguishing accessibility from access"), being 'accessible' (i.e. in the workspace) is already enough to be conscious, the information doesn't need to actually be accessed by any specific module. Differently, he defined a 'preconscious' state (distinct from the subliminal) in which the activation has enough strength to enter the workspace, but doesn't get top-down amplification (you could say, in your words, that the key doesn't match the query).

5. Functional advantages: I found great the link to Gibson and the idea of 'affordance', and the counterfactual reasoning argument in line with the slow system 2. But I personally found the bottom-line -as described at the very beginning- a little weak: ".. should result in improved performance across the entire range of modules that are connected to it." It's not very convincing to me. It reads as if the goal of the GWS is eventually just to perform better at each individual module, by sharing information together. This is most likely true, but I would have put the emphasis on this question from a different perspective (guess what.. predictive coding!). That is, an integrated workspace does a better job at representing a coherent and multimodal model of the world, from which it's possible to make predictions and -at the end of the day- maximize your chances of survival. So for me it wouldn't be about improve the performance of each module individually, but rather improve the behavior of the whole system/agent.

6. In the bit about artificial consciousness, which I endorse fully in both ideas and cautiousness, it would be cool to mention that a self monitoring system could be simply a model of the world complex enough that it has included itself in it (Higher Order Theory of consciousness - Rosenthal 2002). Btw, this intuition is beautifully suitable for a predictive coding system..

7. In teh concluding remarks you mentione that you "contemplate the underlying neuronal bases and the neuroscientific implications of the proposed scheme". I'm not sure I see where exactly you discuss -or mention- the neuronal bases, if not briefly in describing the original idea of the GWS (like in figure 1). Quite minor, but when discussing the need for recursion ('Multiple specialized modules') an additional interesting point could be found in the information integration theory, which indeed states that a FF network can't be conscious (as measured by phi). Maybe an IIT-friendly reviewer would appreciate the pointer.

8. peanuts details: figure 1C why is the FB arrow bigger from frontal to Parietal than anywhere else? I'm not very clear about figure 1C but mostly figure 1D, which is not well explained nor contextualized in the main text/legend.

Comments from Benjamin DEVILLERS (PhD in my group)
As Andrea I found the article very interesting and easy to read. I think the stance makes perfect sense, and I particularly agree with the importance of cycle consistency. It also gave me some clear context on my current work and where it could potentially lead.

1. Besides, I would like to compare your stance in section 3 of composing general function from modules with François Chollet's opinion on the architecture of an intelligent system (https://arxiv.org/pdf/1911.01547.pdf, p.28 - II.2.1 Intelligence as skill-acquisition efficiency). With the generation of a "skill program" to solve a particular task. His paper is not at all on how to implement but rather on how to measure the capabilities of an intelligent system. Also, it only focuses on the tasks and not on how to gather and merge the different sensory information.

2. Finally, I have 2 questions: On grounding. I'm not sure I fully understand how grounding is working. I initially thought that the grounded vectors are the amodal representation of the GW, however you say "Ultimately, grounded latent representations can confer increased performance to every module connected to the global-Workspace". Does it mean that the pre-trained expert modules continue to learn concurrently with the GW and that the modules' representations are grounded by means of cycle-consistency?

3. On temporality / memory. From such a system, it seems that 2 timelines arise. One is the world's time, i.e. the continuous flow of input. The other one comes from the current task that is being processed, that could need information from what the expert modules were processing in the past. Should the GW have access to some sort of short-term memory? For example, can a several attentional query vectors be applied to the same set of keys?

Comments from Milad MOZAFARI (Postdoc in my group)
Just finished reading the paper. Easy to follow and very interesting ideas that truly make sense. Andrea's first and third points were also my questions specially the section that talks about the "copy of latent activation".

- But my own concern (in my opinion) is that reinforcement learning or a reward system is a crucial part to learn broadcasting/gating strategies while dealing with tasks like planning and decision making. I am saying this based on my early-phd research on the working memory and the gating role of the basal ganglia in information flow across the cortex. I acknowledge that my knowledge is not up-to-date anymore but at that time the literature mostly agreed on the role of brain reward system in learning gating strategies. In the beginning of your roadmap, I can see you have mentioned RL but I think it is worth elaborating more on it (maybe in a similar way that you talked about the discriminative and generative networks). The attention mechanism that you have mentioned might be the substitution for my point however I believe reinforcement learning should be involved to make it a general purpose attention mechanism.
  All in all, in my opinion working memory is important in consciousness and coordination/regulation of modalities and brain reward system is important for making working memory work! Again, thank you for sharing this interesting work with us.
- PS: I just saw you are posing my point as question in the end.

Comments from Thomas SERRE
I enjoyed reading the article. To be honest I had never heard of GWT and the article nicely introduced it for me in a clear and concise way. I dont have any major feedback — just a few small points.

I know that Tsotsos has done a lot of work including recent reviews on cognitive architectures. Below are 2 links that I found after a quick googling. Not necessary but might not hurt to add at least 1-2 citations for reviews on alternative architectures?

https://www.frontiersin.org/articles/10.3389/fpsyg.2014.01260/full

https://link.springer.com/article/10.1007/s10462-018-9646-y

I thought Fig 1 which is key could be improved. No idea what Par is and I did not have the faintest idea what panel D shows.

My main comment would be that there is a lot of work on multi-modal learning these days and it might be helpful to add a couple of FAQs at the end addressing briefly 1) how different this proposal is from other cognitive architectures or 2) work on multi-modal learning in ML. 3) possibly address how you could even start evaluating the proposal from an AI perspective... is it that you would already expect that the ind modules would start outperforming specialized approaches simply because of the richness of training or we would need to come up with new challenges to evaluate progress (in which case do you have any such proposal?)

4012 words 22503 characters (not including spaces)

```
File: main.tex
Encoding: utf8
Sum count: 4012
Words in text: 3965
Words in headers: 44
Words outside text (captions, etc.): 0
Number of headers: 8
Number of floats/tables/figures: 0
Number of math inlines: 3
Number of math displayed: 0
Subcounts:
  text+headers+captions (#headers/#floats/#inlines/#displayed)
  8+7+0 (1/0/0/0) _top_
  453+7+0 (1/0/0/0) Section: Cognitive neural architectures in brains and machines
  1214+8+0 (1/0/3/0) Section: Roadmap to a deep learning Global Latent Workspace
  461+5+0 (1/0/0/0) Section: \textcolor{red}{Global Latent Workspace in action}
  532+7+0 (1/0/0/0) Section: Functional advantages of a Global Latent Workspace
  439+5+0 (1/0/0/0) Section: Does GLW entail artificial consciousness ?
  613+3+0 (1/0/0/0) Section: \textcolor{red}{Implications for Neuroscience}
  245+2+0 (1/0/0/0) Section: Concluding remarks
```