Simon Goldstein[1] and
Cameron Domenico Kirk-Giannini[2]

A Case for AI Consciousness

Language Agents and Global Workspace Theory

Abstract: *One common attitude towards AI is that existing AIs are not phenomenally conscious, and that the construction of conscious AIs would require significant technological progress if it is possible at all. We challenge this assumption by arguing that if global workspace theory (GWT) — a leading scientific theory of phenomenal consciousness — is correct, then instances of one widely implemented AI architecture, the artificial language agent, might easily be made phenomenally conscious if they are not already. Moreover, we argue that artificial language agents have further properties that should allay the concerns of many who are sceptical of GWT as a complete theory of phenomenal consciousness. Along the way, we articulate an explicit methodology for thinking about how to apply scientific theories of consciousness to artificial systems and employ this methodology to arrive at a set of necessary and sufficient conditions for phenomenal consciousness according to GWT.*

1. Introduction

The advent of generative text and image models has led to a resurgence of academic and popular interest in the question of what it would take for an AI to be conscious. While it is difficult to draw general

Correspondence:
Email: simon.d.goldstein@gmail.com

1 University of Hong Kong.
2 Rutgers University — Newark, NJ, USA.

Journal of Consciousness Studies, **33**, No. 7–8, 2026, pp. 61–96
DOI: 10.53765/20512201.33.7.061

conclusions on this topic because of the great diversity of philosophical and scientific theories of consciousness, a number of authors have recently suggested that it may be possible to construct conscious AIs in the near future (Butlin *et al.*, 2023; Chalmers, 2023). Recent work has focused on supporting this claim by surveying the commitments of a wide range of philosophical and scientific theories of consciousness. While this approach has the advantage of appealing to diverse theoretical perspectives, it is constrained in its level of detail. In what follows, we take a different approach. We focus on one leading scientific theory of consciousness, global workspace theory (GWT), and one widely implemented AI architecture, the artificial language agent (or *language agent* for short), arguing that if GWT is correct, then language agents might easily be made conscious if they are not conscious already. Our project makes three main contributions to the literature on consciousness in artificial systems. First, it provides an explicit methodology for thinking about how to apply scientific theories of consciousness to artificial systems. Second, it focuses scholarly attention on language agents, which we take to be an especially interesting case study in artificial consciousness. Third, it presents a case for consciousness in AI systems that is better equipped to respond to certain important objections than earlier proposals.

To investigate whether AI systems can be conscious, we adopt a broadly computational and functionalist perspective according to which consciousness is a special kind of information processing.[3] AI systems are conscious if they process information in this special way. Within this framework, the most important open problem is to specify carefully which functional roles are associated with consciousness. Below, we describe how existing work on GWT has vacillated between a series of distinct functional conditions associated with consciousness. These conditions differ in how demanding they are and in their level of abstraction. We show that these differences matter to predictions about what it would take for AI systems to be conscious. We also develop two methodological tools to adjudicate between these various functional

[3] To be clear, there are opponents to computational functionalism: see, for example, Godfrey-Smith (2024) and Seth (2025). We can't hope to settle this dispute here. One goal of our paper is to show that even under the assumption of computational functionalism, there are still many interesting questions about which functional roles are connected to consciousness and many ways to make progress on these questions. It is also worth flagging that some scholars have explored versions of GWT that involve biological conditions above and beyond functional ones (for discussion, see Schlicht, 2025, and McClelland, 2024).

conditions. In particular, we reflect on the usefulness of consciousness and explore thought experiments that probe how our judgments about consciousness vary with slight changes to functional role. In each case, we argue that our methodological tools tend to favour more permissive functional roles, which make it easier for a wide range of systems to be conscious.

Here is the plan. In Section 2, we distinguish between access consciousness and phenomenal consciousness and discuss the place of each in the science of consciousness and in normative theory. In Section 3, we introduce GWT and describe some of the evidence for it. In Section 4, we discuss the relationship between our methodology and the 'indicators approach' which has recently been popular in discussions of AI consciousness. Sections 5, 6, and 7 then distil from GWT a functional characterization of consciousness in artificial systems. Section 8 turns to contemporary machine learning, introducing the architecture of language agents. In Section 9, we argue that existing language agents come close to satisfying the functional criteria introduced in Section 7, and that they could be made to do so fully with simple architectural modifications. In Section 10, we consider a number of further modifications to our proposed architecture that could address a range of different kinds of scepticism about its plausibility as a conscious system. Section 11 considers two objections to our arguments and shows that language agents are strong candidates for consciousness even according to some theories more demanding that GWT. Section 12 concludes.

This paper is a case study focusing on GWT. Analogous questions confront any functionalist theory of consciousness: how exactly should we understand the relevant functional roles, and what kinds of artificial systems might realize them? We hope that future papers will tackle these questions for each of the main functionalist theories of consciousness.

2. Consciousness

The ordinary concept of a conscious state or being is imprecise, failing to differentiate among a diverse set of properties. For this reason, it is customary to distinguish between two important senses in which a state or being can be conscious: access consciousness and phenomenal consciousness. A state is access conscious to the extent that it is available for use in reasoning and guiding speech and action, while it is phenomenally conscious just in case it has experiential properties (Block, 1995). Correspondingly, a being is conscious in the access sense to the

extent that it tokens access-conscious states and conscious in the phenomenal sense just in case it tokens phenomenally-conscious states.

While there are interesting scientific questions to ask about the kinds of biological and artificial functional architectures which give rise to access-conscious states, it is phenomenal consciousness which is generally regarded as the greater puzzle both scientifically and philosophically. Substantial literatures exist in philosophy, psychology, and cognitive neuroscience on the nature of phenomenal consciousness and its relation to physical systems.[4] In what follows, our primary concern will be with phenomenal consciousness. We will discuss one leading scientific theory of phenomenal consciousness in biological systems — global workspace theory (GWT) — and distil from it an account of the conditions under which artificial systems might also be phenomenally conscious. We take it to be uncontroversial that any artificial system meeting these conditions would be access conscious, and we will not argue at length for this claim.[5]

The question of consciousness in artificial systems is especially urgent because many philosophers tie questions of consciousness to the normative domain. First, it is commonly held that certain phenomenally conscious states contribute positively or negatively to an individual's well-being in virtue of their phenomenal properties. For example, the pleasantness of pleasures and the unpleasantness of pains are held to improve and reduce an individual's well-being, respectively.[6] This connection pertains to phenomenal consciousness rather than access consciousness. Second, it is often held that a being must be conscious in order for it to be the right kind of thing to have well-being in the morally significant sense.[7] According to this line of thought, while it might be said of simple systems like protists and fungi that things could go well or poorly for them, it does not follow that we have a moral reason to avoid harming them because they are not conscious.

3. Global Workspace Theory

GWT originated with the work of Bernard Baars (1988; 1997a,b) and has since grown to be one of the leading theories of consciousness in

4 For an introduction with suggestions for further reading, see Van Gulick (2022).

5 In Sections 6, 7, and 11, we also discuss whether phenomenal consciousness requires special kinds of *rich* or *non-conceptual* contents, above and beyond the basic requirements of a global workspace architecture.

6 See, for example, Kagan (1992) and Bramble (2016).

7 See Lin (2021) and Goldstein and Kirk-Giannini (2025) for recent discussion.

biological systems. GWT posits the existence of a *global workspace*: a component of cognitive architecture that receives information from a range of cognitive modules, processes this information, and then broadcasts the processed information back to the cognitive modules to which it is connected. According to GWT, a representation is conscious when it is present in the global workspace. Though proponents of GWT are not always clear on the sort of consciousness they have in mind, we interpret GWT as endorsing the claim that information in the global workspace is conscious in both the access and phenomenal senses, and that being in the global workspace is what renders a piece of information phenomenally conscious.

It is helpful to distinguish the various claims associated with GWT into *structural* claims about cognitive architecture and *functional* claims about how the global workspace receives information, processes it, and transmits it to other parts of the cognitive system. The core structural claim of GWT is that human cognitive architecture consists of a number of relatively autonomous *modules* which process information specific to particular tasks (e.g. particular sense modalities, motor control, and so on) together with a single global workspace with which all of these modules interface.[8] The cognitive architecture posited by GWT thus contrasts with a *pairwise* architecture, where modules communicate with one another directly rather than pooling their information into a shared space (Goyal *et al.*, 2021).

When it comes to functional claims, GWT posits three important aspects of global information processing. First, there is *uptake*, the process that determines which information coming from the modules will enter the global workspace. Second, there is *broadcast*, the process whereby information in the global workspace is sent to a wide range of modules for use in further parallel processing. Finally, there is *processing*, wherein the global workspace integrates and manipulates information it has taken up to shape the trajectory of conscious experience.

GWT thus distinguishes between two types of processing: parallel processing in individual modules and serial processing in the global workspace. Because it is parallel, processing in individual modules can handle theoretically unlimited amounts of information. The global workspace, on the other hand, can process a limited amount of

8 For example, Dehaene, Kerszberg and Changeux (1998) mention five main modules connected to the global workspace: perceptual systems, long-term memory, evaluative systems, attentional systems, and motor systems (see also Mesulam, 1998).

information at any given time. The limited storage capacity of the global workspace creates a bottleneck for information processing which drives the need for selective attention during the uptake phase.

Though it is beyond the scope of our discussion here to present GWT in complete detail, we pause to expand briefly on each of the components of the function of the global workspace and describe some of the psychological phenomena GWT is particularly well suited to explain.

3.1. Uptake

At any given time, parallel processing in the modules is generating a large amount of information which could potentially enter the global workspace. Because the global workspace has limited capacity, however, only some of this information can be taken up from the modules. There is therefore a kind of competition for information to achieve uptake into the global workspace and to remain there once it does.

An important example of competition for uptake is binocular rivalry, one instance of a more general class of 'two-channel experiments'. In binocular rivalry, a subject is shown a different image in each of their eyes. The subject's conscious experience flips back and forth between the two images (Moreno-Bote, Knill and Pouget, 2011). Crucially, the subject cannot have simultaneous conscious experiences of inconsistent visual information. This suggests that there is a bottleneck on conscious experience. Consciousness selects among competing visual inputs to generate a coherent narrative about the world.

Baars (1988) suggests a number of ways in which this competition might be implemented in the mind. First, informational signals coming from different modules might have different levels of activation, and activation might be increased when different modules work together to boost a particular signal (*ibid.*, p. 95). However, Baars holds that activation alone is not sufficient to get information stably into the global workspace: in addition to high levels of activation, information in the global workspace must receive positive feedback from the modules to which it is broadcast (*ibid.*, p. 205).[9]

Ultimately, conscious uptake is controlled by attention (Posner, 1994). This includes both conscious and unconscious attentional pro-

[9] This aspect of Baars's theory is motivated by the existence of *redundancy effects*, wherein information which is at first consciously accessible fades from consciousness as a subject becomes habituated to it (Baars, 1997b, p. 93). Positive feedback from receiving modules is supposed to capture the idea that a given piece of information in the global workspace contains new information useful to the cognitive system.

cesses. Attention boosts information into the global workspace and helps to keep it there (Baars, 1988, p. 308). For example, unconscious attentional processes explain why, even if we are closely attending to one thing, other information can 'break through' into consciousness if it is sufficiently important: 'Absorption does not protect us from high-priority signals, even those that begin unconsciously' (Baars, 1997b, p. 107).

The role of attention in uptake is illustrated by other 'two-channel' variants of binocular rivalry: listening to two conversations at once, reading alternating words of text, or seeing two sports games superimposed on the same screen (*ibid.*, pp. 23–4). In these cases, unconscious information from the second channel will break through if it is sufficiently important, emotional, or relevant. Information in the unconscious channel can also unconsciously influence interpretation of ambiguous words (like *bank*) in the conscious channel (*ibid.*, p. 27).

Since Baars, neuroscientists have studied in detail the mechanisms by which humans implement the global workspace architecture. A key concept in this area is *ignition*, a specific hypothesis about how uptake is achieved. According to global neuronal workspace theory, the main development of GWT in neuroscience, 'ignition is characterized by the sudden, coherent, and exclusive activation of a subset of workspace neurons coding for the current conscious content, with the remainder of the workspace neurons being inhibited' (Mashour *et al.*, 2020, p. 777). In other words, ignition is a non-linear process whereby a single neural representation suddenly comes to dominate the neuronal workspace, suppressing all competing patterns of activation.

3.2. Broadcast

The most important function of the global workspace is to serve as a central repository of information available to the cognitive architecture's various parallel processors. In addition to taking up information from these processors, then, the global workspace must also transmit information back to them. We refer to this function of the global workspace as *broadcast*.

As we have seen, Baars (1988) holds that there is an intimate connection between uptake and broadcast: it is only through broadcasting to a coalition of processors and receiving positive feedback signals from them that a representation can remain stably in the global workspace.

3.3. Processing

According to GWT, the global workspace is not simply a passive repository of information available to various cognitive modules; it plays an active role in processing information. Processing in the global workspace is notable for its generality. As Dehaene, Kerszberg and Changeux put it, 'once we are conscious of an item, we can readily perform a large variety of operations on it, including evaluation, memorization, action guidance, and verbal report' (1998, p. 14529). Indeed, there are some kinds of processing that can only occur in the global workspace. For example, Baars (1997b, p. 17) suggests that conscious processing in the global workspace is required to construct concepts out of two-word compounds like *potato soup*. This is illustrated experimentally by priming effects. Unconsciously seeing the word *dog* makes it easier to identify the word *puppy* consciously seconds later. But priming effects are not observed with compound words, like *potato soup* or *honey cake*: the global workspace, and consciousness, is required to integrate the meaning of the two words.[10]

A more general aspect of information processing in the global workspace concerns the demand for consistency. The global workspace seeks to organize disparate information into a single coherent model of the world. One illustration of this demand is visual illusions like the impossible trident and Escher's infinite staircases. When we visually attend to these images, the global workspace tries to create a single coherent image, even though there isn't one available. In the impossible trident, tracing the trident across the scene will result in alternating visual images because there is no way to combine the information into one coherent object (*ibid.*, p. 88).

4. Indicators and Functional Bases

So far, we've surveyed a range of claims that GWT makes about consciousness in humans. But we haven't said precisely which properties of conscious human systems are necessary and sufficient for consciousness. This is a difficult question. Several authors have noted that it is

10 We introduce priming effects for compound words as an example to illustrate the underlying architectural features of a global workspace. We do not mean that this kind of priming effect is a necessary condition for consciousness according to GWT. AI systems might or might not exhibit priming effects for compound words, and this would not itself directly decide whether they are conscious on our view. Rather, on our theory consciousness hinges on architectural features like the presence of a global workspace. Thanks to an anonymous referee for pressing us to clarify this issue.

tricky to say exactly how similar a system needs to be to a human global workspace in order to be conscious (Carruthers, 2019; Birch, 2022; Butlin *et al.*, 2023). More permissive conceptions of consciousness, according to which conscious systems might be less like the human brain, will make it easier for AI systems to be conscious, while less permissive conceptions, according to which conscious systems must be more like the human brain, will make it harder.

We are here confronted with one example of the *specificity problem* (Shevlin, 2021): a given scientific theory of consciousness in humans can be understood as imposing either coarse-grained or fine-grained functional conditions on consciousness. If the conditions are coarse-grained, they may be satisfied by AIs, but they may also overgenerate. If the conditions are fine-grained, they may rule out AIs, but they may also undergenerate. Solving the specificity problem requires figuring out which differences between humans and AIs matter.

As we noted above, we think about these questions through a computational and functionalist perspective on consciousness. The idea here is that there is some functional role associated with consciousness, so that all and only systems instantiating this functional role are conscious. More carefully, we can also think about a necessary and sufficient functional role for a particular mental state to be conscious, and then say that a system is conscious when one or more of its mental states is conscious.

It is consistent with this methodology that the relevant functional role is not *metaphysically* necessary and sufficient for consciousness. Rather, for us what is relevant is that the functional role in question be *nomically* necessary and sufficient for consciousness. This distinction allows us to sidestep tricky questions about phenomenal zombies — hypothetical beings that are physically like conscious systems but lack conscious experience. Perhaps phenomenal zombies are metaphysically possible, and so, for any functional role, it is metaphysically possible to satisfy it without being phenomenally conscious. Nonetheless, there could still be functional roles that perfectly co-vary with consciousness as a matter of physical law. Another way of thinking about this is that the scientific laws of our world may imply that access consciousness (understood in terms of the global workspace) is sufficient for phenomenal consciousness. If so, AI systems with access consciousness would also be phenomenally conscious.

We think of our approach in what follows as having two parts. First, we try to make progress on the specificity problem by distilling from GWT a set of necessary and sufficient conditions for a system to be

phenomenally conscious. To do so, we appeal to thought experiments and the claims GWT theorists themselves make about the cognitive usefulness of consciousness. Second, we argue that existing or near-future AI systems could satisfy all of the conditions we identify. It follows that, if GWT is the correct theory of consciousness, these systems are or could soon be conscious. We will call this the *functional basis approach* to the study of consciousness in AIs because it centres on identifying the functional basis of consciousness according to a given theory and assessing whether different AI systems implement that functional basis.

It is worth taking some time to compare our functional basis approach to another recently popular approach to thinking about consciousness in AIs: the *indicator approach* (Butlin *et al.*, 2023). The indicator approach takes as its starting point the current empirical uncertainty about which scientific theory of consciousness is correct:

> Various theories are currently live candidates in the science of consciousness, so we do not endorse any one theory here. Instead, we derive a list of *indicator properties* from a survey of theories of consciousness. Each of these indicator properties is said to be necessary for consciousness by one or more theories, and some subsets are said to be jointly sufficient. Our claim, however, is that AI systems which possess more of the indicator properties are more likely to be conscious. To judge whether an existing or proposed AI system is a serious candidate for consciousness, one should assess whether it has or would have these properties. (*ibid.*, pp. 4–5)

In other words, the indicator approach seeks to identify a range of properties that have been taken to be relevant to consciousness in the sense that some scientific theory says they are either necessary for consciousness or part of some set jointly sufficient for consciousness. It then holds that we should be confident that a given system is conscious to the extent that it possesses these indicator properties.

The functional basis approach and the indicator approach are compatible, and we think they are both likely to yield important insights. The indicator approach need not deny that scientific theories of consciousness state necessary and sufficient conditions for a system to be conscious. It need not deny that if a theory states a set of necessary and sufficient conditions and an AI system satisfies those conditions, then if the theory is true, that AI system is conscious. And it need not deny that it is important to achieve clarity about what exactly the different scientific theories of consciousness are claiming a system must be like to be conscious. It departs from the functional basis approach mainly in

that its emphasis is on how we should think about the probability of consciousness in systems that do not satisfy all of the conditions stated by any scientific theory of consciousness but are similar in various important respects to systems that do. In contrast, our strategy here is to argue that some existing or near-future AI systems might in fact satisfy *all* of the conditions for consciousness if GWT is true.

5. What is Essential to Consciousness? Existing Work

In this section, we present a number of existing characterizations of the conditions an artificial system would need to satisfy to have a global workspace according to GWT. Though they are on the right track, we believe that none of these characterizations get things quite right. We argue for our own set of conditions in Section 7.

Butlin *et al.* (2023) is the most influential recent treatment of the question of what conditions are necessary and sufficient for a system to be phenomenally conscious according to GWT. Butlin *et al.* propose a set of four conditions:

(B1) Multiple specialized systems capable of operating in parallel (modules).
(B2) Limited capacity workspace, entailing a bottleneck in information flow and a selective attention mechanism.
(B3) Global broadcast: availability of information in the workspace to all modules.
(B4) State-dependent attention, giving rise to the capacity to use the workspace to query modules in succession to perform complex tasks.

In addition to Butin *et al.*, there are also several recent papers which provide functional characterizations of GWT without explicitly addressing issues of phenomenal consciousness. Since we are understanding GWT as a theory of the conditions which must be satisfied for a system to be phenomenally conscious, however, these characterizations are relevant to our project.

VanRullen and Kanai (2021, p. 3) propose 'a step-by-step attempt at defining necessary and sufficient components for an implementation of the global workspace in an AI system', according to which a system must meet the following conditions:

(VK1) A number of independent specialized modules, each with its own high-level latent space.

(VK2) An independent and intermediate shared latent space, trained to perform unsupervised neural translation between the latent spaces from the specialized modules.
(VK3) The system prioritizes competing inputs to the shared space with attention.
(VK4) When a specific module is connected to the workspace as a result of attentional selection, its latent space activation vector is copied into the [global latent workspace] and then immediately broadcast into the latent space of all other modules.

VanRullen and Kanai explicate the notion of a 'high-level latent space' as follows: '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.' (*ibid.*, p. 3).

According to Juliani, Kanai and Sasai (2022), on the other hand, a system implements a global workspace just in case it contains a central workspace module that:

(J1) Has the ability to interact with a dynamic set of modules.
(J2) Has the capacity for selective attention over the modules.
(J3) Has the capacity to maintain information over time.
(J4) Has the capacity to manipulate information over time.

Juliani *et al.* understand (J1) as requiring that the central workspace receive inputs from a set of modules that potentially changes over time, remarking that 'while the neural anatomy of the brain is largely fixed, what counts as a "module" within the GWT is dynamically determined based on population activity across multiple brain regions at a given time' (*ibid.*, p. 957). Their (J2) is intended to capture both attentional processes affecting which representations make their way into the central workspace (ignition) and attentional processes affecting which information in the workspace makes its way back to the modules (broadcast).

While a number of themes emerge from these proposed sets of necessary and sufficient conditions, it is clear that they are not equivalent. (J1), for example, demands that the set of modules connected to the global workspace be dynamic, whereas this requirement is not to be found in the proposals of Butlin *et al.* or VanRullen and Kanai. And (B2) specifies that the global workspace must have a limited capacity, which is a constraint not imposed by Juliani *et al.* A key philosophical problem in approaching the literature on GWT and artificial systems,

then, is to assess the plausibility of each condition proposed as a gloss on GWT.

6. What is Essential to Consciousness? Theoretical Choice Points

The proposals described in the previous section raise a number of broader questions about how to think about consciousness in AI systems in the context of GWT. We highlight some of these questions in this section before arguing for our considered view in the next.

A first question is how to functionally capture GWT's notion of attention. Butlin *et al.* (2023) suggest that consciousness requires 'top-down' attention, where the information in the workspace can control what further information enters the workspace, in addition to 'bottom-up' attention, where the strength of the signals from various modules determines which information enters the workspace. In contrast, though VanRullen and Kanai (2021) mention this distinction between top-down and bottom-up attention, they do not build it into their conditions on phenomenal consciousness.

A second question concerns the strength of the broadcast condition. Whereas (J1) simply requires that the workspace interacts with some modules, (B3) and (VK4) require that information in the global workspace is broadcast to every single module. Here, one especially relevant sub-question is whether information in the workspace must be sent back to the perception modules, which would enable top-down perceptual processing.

A third question concerns the status of information processing in the global workspace. VanRullen and Kanai do not require the global workspace to engage in processing at all, whereas Juliani *et al.* do but remain unspecific about the nature of this processing. Butlin *et al.* do not clarify how much processing is required for state-dependent attention, so it is not clear where they come down on this issue. One might require that the workspace engage in very specific types of processing familiar from human working memory; for example, the presence of a visuospatial sketchpad and phonological loop. Alternatively, one could follow Juliani *et al.* (2022) in being unspecific about the nature of the required processing, or not require processing to take place within the global workspace at all.

A final question, and one which in our opinion has been insufficiently discussed in the existing literature on GWT and artificial systems, is whether the representations in the global workspace must have specific

structural features. Here one especially important idea is that phenomenal consciousness requires that perceptual representations are *rich* in some way. For example, Rosenthal (2005) suggests that phenomenal experiences are representations that occupy a similarity space, and Tye's (1995) PANIC theory requires specific types of non-conceptual contents. Similarly, Carruthers (2020) argues that 'phenomenal consciousness is access-conscious nonconceptual content', because it involves especially fine-grained content.

The debates just discussed are related to bigger-picture questions about the appropriate level of functional analysis for theorizing about consciousness. We ourselves tend to think about the functional role of consciousness in terms of high-level concepts related to manipulating information. But a competing approach might focus more on the implementational level (compare Cao, 2022). According to this alternative picture, consciousness functionally requires a neural net that implements higher-level functional ideas in a similar way to humans. For example, uptake might require a neural net that displays similar non-linear ignition patterns to humans. It might also require an unconscious attention system that is trained in a similar way as the human attention system. By contrast, more abstract levels of analysis might allow an uptake system that implements an informational bottleneck in a different way.

7. What is Essential to Consciousness? A Theory

How do we assess the competing proposals of Section 5 and answer the thorny questions of Section 6?

In this section, we attempt to directly sort out the functional role of phenomenal consciousness for GWT, using two strategies: first, appealing to the ways in which, according to GWT theorists, consciousness is advantageous to systems that possess it; second, employing thought experiments about which changes to conscious systems would turn them into systems that are not conscious. We suggest that each strategy tends to favour more permissive, higher-level functional roles.

7.1. The value of consciousness

A first way of approaching the problem of distilling a set of necessary and sufficient conditions for consciousness from GWT is to consider what the theory says about the usefulness of consciousness. What purpose does consciousness have in systems that possess it? We might

find that consciousness is steered towards solving particular kinds of problems, but that only some and not all of the candidate functional roles sketched in the previous two sections are necessary in order to solve those problems. In general, this approach assumes that phenomenal consciousness is causally efficacious, connecting in lawlike ways to the cognitive and even physical behaviour of the organisms that possess it. Our suggestion is that it is plausible that a property is part of the functional role of consciousness according to GWT only if that property plays a suitably central role in explaining how having a global workspace helps systems solve certain canonical problems.

What problems does consciousness help conscious systems solve? Baars (1997b) describes a number of information processing tasks that consciousness facilitates on the GWT picture; here, we distil them into four. First, consciousness helps to *classify* inputs to central cognition by prioritizing more important information over less important information. Second, consciousness enforces the *coherence* of the information being processed in central cognition. Third, consciousness helps to *coordinate* cognition by recruiting a range of unconscious tools to solve a single problem. Fourth, consciousness allows for *correction* of errors in reasoning. We suggest that when we think about consciousness in terms of serving these purposes, we should favour more permissive approaches to its functional role.

1. **Classification**. Any cognitively sophisticated agent faces the problem of deciding which information is important and relevant to its current situation and which is not. Solving this problem is crucial for reducing the computational complexity of reasoning about how to act.

   By its nature, the information bottleneck induced by the global workspace requires a conscious system to classify some information as more important than other information — it is only the information that makes it from the modules into the global workspace that is broadcast to the entire system and available for further processing.
2. **Coherence**. A related function of the global workspace is to create a coherent narrative about the world on the basis of potentially inconsistent representations. In the process of selecting information for uptake using attention, the global workspace works to produce consistent interpretations. We saw this demand for coherence in cases of binocular rivalry and other two-channel experiences. On the present picture, coherent conscious experience

is a tool for facilitating effective agency. If our action-guiding representations of the world change dramatically from moment to moment, we will struggle to form and execute effective plans over time.

3. **Coordination**. In a cognitive architecture containing parallel processing modules, solving problems may often require coordinating the activity of multiple processors. The global workspace solves this problem by making the outputs of one processor available to the others and by guiding the activity of the processors over time through top-down attention.
4. **Correction**. Top-down attention also facilitates error detection and correction. Baars (1997b) argues that one function of consciousness is to monitor for errors: 'if we have no conscious access to our own performance, and if some reliable source of information tells us that we are doing quite badly, we tend to accept misleading feedback because we cannot check our own performance directly' (*ibid.*, p. 136).

   The ability to monitor for errors is connected again to the nature of agency, because it is also connected to the concept of voluntary control: 'with direct conscious access to our own performance we are much less influenced by misleading labels. These results suggest that three things go together: consciousness of action details, voluntary control over those details, and the ability to monitor and edit the details' (*ibid.*, p. 136).

In general, we believe that focusing on the practical advantages conferred by consciousness favours more permissive accounts of the conditions which must be satisfied for consciousness. For example, consider the process of ignition as it is understood by global neuronal workspace theory. During ignition, information enters the global workspace through a specific non-linear process of neuronal activation. We don't think that non-linear neuronal activation of this specific kind is plausibly required for coherent action or efficient information processing.

In contrast, attention to the practical advantages of having a global workspace does suggest that certain core features of the GWT framework should be in place for conscious experience. For example, it seems to us that classification requires bottom-up attention, while coherence, coordination, and correction require a combination of top-down attention and processing in the workspace. It is worth noting in this connection that, while Butlin *et al.*'s (B1)–(B4) probably come closest

to capturing the capacities identified above, none of the existing proposals we have discussed simultaneously emphasize the need for bottom-up attention, top-down attention, and information processing within the workspace itself.

Note that it is open to critics of our characterization of GWT to offer an alternative account of the value of consciousness according to GWT that might support a set of necessary and sufficient conditions for consciousness different than the one we offer below. Here, as always, we think the devil is in the details. We leave it to our opponents to articulate their account of the value of consciousness according to GWT and lay out how their conditions are connected to it.[11]

7.2. Thought experiments

A second way of approaching the question of what it would take for an artificial system to be conscious according to GWT is to consider thought experiments where a candidate condition is removed from a conscious system and we judge whether that system is still conscious.

First, consider the fact that working memory in human beings is limited to just a few items (seven or less). It would be easy to design AI systems with much larger working memories. But the very small size of human working memory is not plausibly essential to consciousness. Imagine that a person's working memory expanded from seven to ten thousand items. Intuitively, they would not thereby go from being conscious to being unconscious. Instead, they would be *superconscious*, conscious of vastly more.

This thought experiment suggests to us that consciousness does not require that the global workspace have bounded size. In other words, it suggests to us that the restriction to systems with a limited capacity workspace in Butlin *et al.*'s (B2) is unwarranted. In this connection, it is worth pointing out that efficient information sharing, useful multimodal representations, and attention could all be implemented in a system with an arbitrarily large global workspace.

Second, consider whether having a dynamic set of modules is plausibly a necessary condition on phenomenal consciousness according to GWT. Start with a conscious system with a global workspace dynamically connected at different times to different subsets of a set S of modules. Now imagine that we change the system so that it is always connected to each module in S, but some of the connections at any given

11 Thanks to an anonymous referee for pressing us to clarify this point.

time are not active. Could simply creating the extra connections turn a conscious system into one that lacks consciousness? Would the light of conscious inner life suddenly extinguish when the last connection is installed? We think not. This thought experiment suggests to us that consciousness does not require a dynamic set of modules, as in Juliani *et al.*'s (J1).[12]

Third, consider whether consciousness requires that information represented in the global workspace must be broadcast recurrently back to every module connected to it. Start with a conscious system with a global workspace which both receives information from and broadcasts information to each of its input modules. Now imagine that we add a module to the system which has a one-way connection to the global workspace: it can send information to the global workspace, but not receive information back. Would the addition of this module turn a conscious system into an unconscious one? Intuitively, we think the answer to this question is no. This thought experiment suggests that conditions like Butlin *et al.*'s (B3) and VanRullen and Kanai's (VK4) are unnecessarily strong.[13]

Fourth, consider the richness of perceptual experience. Baars observes that not only visual but also conceptual representations can be conscious. He describes these conceptual conscious experiences as 'fringe' or 'feeling of knowing' consciousness: 'Feelings of knowing are not imprecise in their underlying contents. They are simply subjectively vaguer than the sight of a coffee cup… lacking clear figure–ground contrast, differentiated details and sharp temporal boundaries. However, concepts, judgments, and semantic knowledge can be complex, precise, and accurate' (Baars, Franklin and Ramsoy, 2013, p. 15).

---

12 It might be protested that Juliani *et al.*'s conception of a dynamic set of modules is one on which the modules that exist at one time might not exist at another later time, not one on which the same modules always exist but might or might not be connected to the global workspace at any given time. But similar considerations suggest that consciousness does not require the set of modules to be dynamic in this second sense, either. Take a conscious system with a global workspace connected at time t to a set S of modules. If we constrain the system so that no modules can be added or removed from S, will it cease to be conscious? Intuitively, no — at most, constraining the system in this way would limit the range of conscious experiences it could have.

13 Note that our argument here is that a system could implement a global workspace architecture without recurrent broadcast from the global workspace back to *every* module. As we clarify below, we still hold that in order to implement a global workspace architecture, a system must have a global workspace that broadcasts recurrently back to *sufficiently many* of its modules. This condition will be congenial to those who believe there is a deep connection between consciousness and recurrence of some kind. Thanks to an anonymous referee for pressing us to clarify this point.

Now imagine a human being who lost their visual consciousness, but retained their fringe representations. Would this person no longer be conscious? We think they would still be conscious (compare Chalmers, 2023). This suggests that rich perceptual inputs should not be construed as necessary for phenomenal consciousness according to GWT.

Finally, consider the plausibility of conditions tied to specific deep-learning architectural assumptions, like VanRullen and Kanai's (VK2) and (VK4). (VK2) requires that the global workspace have been trained in a specific way, while (VK4) requires that information stored in the modules and global workspace be representable as an activation vector, and also that movement of information from a module to the workspace be interpretable as a copy function on activation vectors. We think it is easy to imagine a conscious system which does not meet these conditions. Consider, for example, a normal adult human. It is not clear to us that the global workspace of a human brain has been trained to perform unsupervised neural translation between different latent spaces because it is not clear to us that the global workspace of a human brain has been trained at all.[14] Similarly, to copy the activation vector from one space to another in a literal sense, there must be a neuron in the latter space corresponding functionally to each neuron in the former space. If the global neuronal workspace hypothesis is true and the human global workspace is realized by a specific network of long-range neurons, there is little reason to believe that this network is isomorphic in the required way to the brain regions from which it receives inputs. To put it another way, it seems to us more plausible that when information moves from a brain module into the global neuronal workspace, what happens is that a pattern of activation with content *C* in the module causally produces a distinct pattern of activation with content *C* in the global neuronal workspace, where this shared content need not be explained by positing structurally identical patterns of neural activity. This suggests that VanRullen and Kanai's (VK2) and (VK4) are too specific to be plausible.[15]

---

14 If you think that the plasticity involved in normal neurological development might count as a kind of training in the relevant sense, consider an intrinsic duplicate of yourself which is produced instantaneously out of inanimate components by a sophisticated machine. The brain of such a duplicate would have never undergone training of any kind, but it would be conscious.

15 Since VanRullen and Kanai's (2021) explicit goal is to articulate a set of necessary and sufficient conditions for implementing a global workspace *in artificial systems*, the fact that their account makes implausible predictions about biological systems might be

We have rejected a number of candidate conditions on consciousness on the basis of thought experiments. It is worth noting at this point that we do not think the same method can be used to challenge other more foundational aspects of the GWT picture. For example, if we imagine a conscious system consisting of a set of parallel processing modules and a global workspace and then remove the parallel processing modules, we have no strong intuition that the resulting system will be conscious.[16,17]

Combining considerations from the usefulness of consciousness with the results of our thought experiments, here is our considered view of the necessary and sufficient conditions for conscious experience according to GWT. A system is phenomenally conscious just in case:

(1) It contains a set of parallel processing modules.
(2) These modules generate representations that compete for entry through an information bottleneck into a workspace module, where the outcome of this competition is influenced both by the activity of the parallel processing modules (bottom-up

thought not to constitute a serious objection. Here we would like to make two points. First, given that one of the main motivations for GWT is that it understands consciousness in terms of the high-level functional organization of information processing, it would be odd for the conditions for implementing a global workspace to differ between biological and non-biological systems. Second, we see no reason why all artificial systems implementing a global workspace would need to share the deep-learning architectural assumptions built into (VK1)–(VK4). Consider, for example, a neuron-by-neuron simulation of a human brain. If biological brains constitute a problem for (VK1)–(VK4), then simulated brains do, as well.

16 We have weaker intuitions about whether starting with a conscious system and then removing its capacity for top-down attention or its workspace's ability to promote coherent representations would result in a system that is not conscious. There might be room for reasonable disagreement on these points, and for this reason some might hesitate to follow us in building these features into the conditions for conscious experience. This issue is not central for our dialectical purposes, however, since our claim is that it is possible to imagine a simple language agent architecture that meets even our more demanding conditions.

17 It is worth briefly clarifying one aspect of our thought experiment methodology. On the one hand, there may be uncertainty or indeterminacy internal to the functional conditions constitutive of consciousness. For example, perhaps there is indeterminacy regarding the exact proportion of modules that must receive information from the global workspace for a system to be conscious. While we have seen that thought experiments can potentially rule out extreme answers to such questions (i.e. *all* or *none*), we do not think they are likely to help us arrive at a particular non-extremal answer. On the other hand, some of our thought experiments have an on/off structure rather than a degree structure. For example, we argued above that the size of working memory is plausibly *irrelevant* to consciousness. Here we think thought experiments can yield determinate answers.

attention) and by the state of the workspace module (top-down attention).[18]

(3) The workspace maintains and manipulates these representations, including in ways that improve synchronic and diachronic coherence.

(4) The workspace broadcasts the resulting representations back to sufficiently many of the system's modules.[19]

In light of our arguments above, we take (1)–(4) to constitute a plausible theory of the conditions a system must satisfy to be conscious if GWT is correct. Some readers, particularly those with an interest in consciousness in non-human animals, might worry that (1)–(4) are too stringent. Perhaps simple creatures with some but not all of the structure described in (1)–(4) are or could be conscious. Here we would like to register two points. First, it seems to us that if non-human animals can be conscious without meeting criteria (1)–(4), then GWT is false as a theory of consciousness. So this possibility does not seem to us to threaten our claim that (1)–(4) capture the content of GWT as a theory of consciousness. Second, if an AI system is conscious according to our interpretation of GWT, then it is also conscious according to every weakening of that theory, whether we choose to understand that weakening as a version of GWT or not. So the possibility that (1)–(4) are too stringent does not threaten our central thesis that if GWT is correct, language agents are or could soon be phenomenally conscious.

8. Language Agents

In the rest of the paper, we'll apply (1)–(4) to a particular type of AI system: the language agent. Language agents are created by embedding an LLM into a functional architecture with the structure of an agent that acts predictably according to the laws of folk psychology. The LLM performs all of the relevant information processing of the system, while the functional architecture ensures that this information processing produces coherent agentic behaviour. Below, we'll focus on the language agents developed by Park *et al.* (2023) as a case study. But there are

[18] Importantly, the idea of an information bottleneck is distinct from the idea that the workspace module must have limited capacity, which we reject above. Even with unlimited workspace capacity, an information bottleneck and corresponding attentional mechanism remains important for the global workspace to fulfil its roles in classifying information as important and creating a coherent narrative about the world.

[19] Correspondingly, according to our view a representation is phenomenally conscious just in case it is being manipulated by the workspace of a phenomenally conscious system.

numerous other examples of language agents, including AutoGPT,[20] BabyAGI,[21] Voyager,[22] SPRING,[23] and others.[24]

Language agents record and store their beliefs, desires, and plans, as well as their perceptual observations, in natural language. The functional architecture with which a language agent is programmed specifies how these sentences recording beliefs, desires, plans, and observations are fed into the LLM as it considers how the agent will act. Indeed, it is the roles assigned to different stored sentences by the architecture of the language agent which make it the case that they count as the agent's beliefs, desires, and so forth. In this respect, the structure of a language agent mirrors the structure of the human mind according to cognitive theories which posit a distinction between the content of a representation and its functional role. What makes a representation with the content *I am eating* a desire rather than a belief in the human mind, according to such theories, is the way in which it enters into an agent's broader cognitive economy, and especially action planning.[25]

Cognition requires not only moving and storing information, but also processing it. The information processing role in a language agent is played by its LLM. Different cognitive tasks within the language agent may rely on the LLM in different ways. For example, the route from perception to memory might require the LLM to summarize a text description of the agent's observations to highlight those aspects worth recording, while planning action might require the LLM to reason about

---

20 Project available at https://github.com/Significant-Gravitas/Auto-GPT.

21 Project available at https://github.com/yoheinakajima/babyagi.

22 See Wang *et al.* (2023).

23 See Wu *et al.* (2023).

24 Perhaps the most successful recent agentic application of language models is Devin, billed as the 'first AI software engineer' (https://www.cognition-labs.com/introducing-devin). Another recent example of a step towards language agents is Ghost in the Minecraft, where LLMs learn to navigate the game Minecraft (Zhu *et al.*, 2023). Mind2Web is a framework for building web agents (Deng *et al*., 2024). A longer list of existing LLM agents can be found here: https://github.com/e2b-dev/awesome-ai-agents. ChatDev (https://github.com/OpenBMB/ChatDev) is another multi-agent environment with some similar features to the Park *et al.* Generative Agents framework. For further scaffolding techniques that increase the agency of LLMs, see: Tree of Thoughts (Yao *et al.*, 2024), LLM+P (Liu *et al.*, 2023a), GPT-engineer, and RecurrentGPT (Zhao *et al.*, 2023). In a similar vein, Zhang *et al.* (2024) develop AgentOptimizer, a framework for training language agents without modifying the weights of their underlying language models. For benchmarks measuring the agency of LLMs, with discussion of applications for language agents, see AgentBench (Liu *et al.*, 2023b) and API-bank (Li *et al.*, 2023).

25 See for example Fodor (1987).

what course of action would be rational given the agent's beliefs and desires. While the language agents in which we are most interested rely more or less exclusively on a single LLM for cognitive processing, for our purposes it would not matter if they relied on some combination of different LLMs, or even if some of their cognitive capacities were realized by hand-coded algorithms.

For concreteness, let us return to the language agents developed by Park *et al.* (2023). Park *et al.*'s agents live in a text-based simulation called 'Smallville'. They observe and interact with their simulated environment and each other via text descriptions of what they see and how they choose to act. Each agent's perceptual observations are stored in a text file called the *memory stream* along with other beliefs, including those comprising their text backstory, which specifies their long-term goals and relationships with other agents. At the end of each day, every agent in Smallville calls on the LLM (in this case, gpt3.5-turbo) to generate a plan for the next day in light of the contents of their memory stream. These plans flexibly shape how an agent acts the following day.

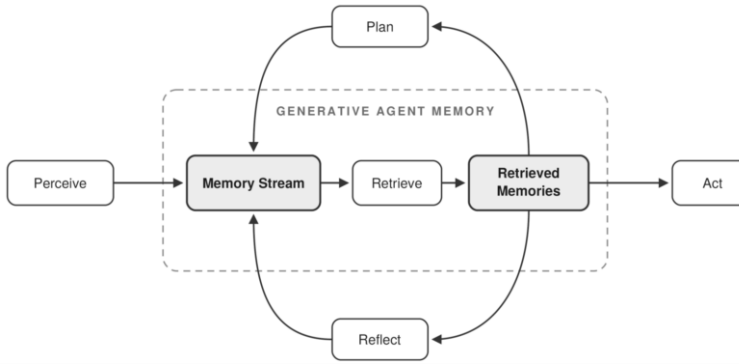

Figure 1. The architecture of Park *et al.*'s language agents. (Adapted from Park *et al.*, 2023.)

For our purposes in what follows, it will be important to discuss three aspects of Park *et al.*'s agents in more detail. First, there is the memory stream. In Park *et al.*'s agents, the memory stream is implicated in all important cognitive processes. It is where perceptions are recorded, as well as where non-perceptual beliefs, desires, and plans are stored. Each entry in the memory stream comes along with a timestamp, and each is assigned an importance score by the LLM as it is recorded. These importance scores play a key role in determining how each entry shapes

the future behaviour of the agent: entries with higher importance scores are more likely to influence an agent's other beliefs and plans.

Second, there is the *retrieval function*. Since an agent's memory stream is unmanageably long, in planning action it is necessary to select an action-relevant subset of entries. This is the role of the retrieval function. Given an input circumstance, the retrieval function produces a list of entries from the memory stream ordered by taking a weighted sum of their importance (as described above), their recency, and their relevance to the input circumstance (this is again calculated by the LLM). In deciding how to act in a given circumstance, each language agent considers only those entries from its memory stream which rank highly enough according to the retrieval function.

Third, there is the process of *reflection*. If the entries in the memory stream were limited to perceptual beliefs and plans, language agents would struggle to function intelligently in light of new information. To solve this problem, Park *et al*. introduce a function that enables agents to freely draw inferences from their existing beliefs. During reflection, agents draw on the LLM to first pose and then answer a series of questions about how to generalize from their recent experiences. For example, in reflection an agent might pose and then answer questions about their values or the significance of their relationships. The results of reflection are then stored in the memory stream, where they are accessible to the retrieval function.

From the perspective of GWT, Park *et al.*'s agents have a number of suggestive architectural features. For example, the memory stream maintains representations and interfaces with both the perception and action systems. It also manipulates those representations through reflection, planning, and revising plans in light of new information. There is even an information bottleneck of sorts in the form of the retrieval function imposing constraints on how much of the information in the memory stream can be fed to the underlying LLM.

Yet Park *et al.*'s language agents also lack some of the features which we have identified as necessary for consciousness according to GWT. For example, since perception, belief, desire, and planning are all handled by the memory stream, it is not clear that the agents contain a series of parallel processing modules. All information processed by a language agent makes its way into the memory stream, so there is no information bottleneck leading into it. From the perspective of GWT, then, it seems to us that Park *et al.*'s language agents are not plausibly conscious. This conclusion is in keeping with the general climate of scepticism surrounding claims of consciousness in artificial systems.

At the same time, however, we wish to challenge this climate of scepticism by suggesting that Park *et al.*'s language agents can serve as the basis for language agents which would be phenomenally conscious according to GWT — and indeed that the changes necessary to create such agents are technically trivial in the sense that they could be implemented straightforwardly given existing technology.

9. Language Agents Could Easily Have Conscious Experiences

We have seen that, while Park *et al.*'s (2023) language agents contain a centralized cognitive module which stores and manipulates information in the manner of a global workspace, it is less clear that they contain a series of input modules to this workspace whose representations must compete for entry into it. In this section, we describe how the architecture of Park *et al.*'s language agents could be changed to more closely mirror the structure of a conscious system according to GWT. None of the changes we consider affect how the LLM underlying a language agent is trained or structured. Rather, all of the changes pertain to how that LLM is scaffolded to produce an agent. In this way, once we have an LLM that can process information, consciousness will depend on how that information processing is exploited in systematic, lawlike ways by hard-coded rules.

To begin, note that Park *et al.*'s language agents in fact have representations of several kinds (beliefs, desires, plans) and engage in a range of forms of cognitive processing including assigning importance scores to memory items, reflection, assigning relevance scores to memory items during retrieval, plan formation and revision, and practical reasoning leading to action. Park *et al.* make the architectural choice to store all these kinds of representations in the same cognitive workspace, the memory stream, and to have all cognitive processing take as input a series of items from the memory stream. But language agents could also be built with many of the same features arranged into a different architecture.

Imagine, for example, that we modify Park *et al.*'s architecture in the following ways. First, instead of a single memory stream containing perceptual inputs, other beliefs, desires, and plans, we have a central workspace connected to three further cognitive input modules: a perception module, a belief module, and a desire-and-plan module. Each of these three modules stores representations of the appropriate type. The central workspace can store representations of any type.

Second, imagine that the three input modules perform the following information processing tasks in parallel. The perception module receives observations from the environment and assigns them salience ratings corresponding to how relevant it judges them to be to the agent's likely future cognition. The belief module assigns each belief an importance score and engages in reflection-driven inference in the same way as Park *et al.*'s agents. The desire-and-plan module assigns each desire an importance score and engages in a desire-based analogue of reflection: generalizing new desires from the agent's existing list of desires.

Third, imagine that the central workspace plays a number of important information processing roles. It receives perceptual inputs that have been flagged as high salience and decides whether to send them to the belief module for storage. It recursively forms and adds detail to plans on the basis of the agent's beliefs and desires, working to ensure that the results are coherent. And it chooses how the agent should act on the basis of its beliefs, desires, and plans.

Fourth, imagine that information in the architecture flows along the following paths: in belief formation, from the perception module to the central workspace and then to the belief module; in planning, from the belief and desire-and-plan modules to the central workspace and then back to the desire-and-plan module; in action, from the belief and desire-and-plan modules to the central workspace and then back to the belief and desire-and-plan modules (because actions must be recorded in memory and plans must be revised in light of actions).[26]

The architecture just described would be no more technically challenging to implement than Park *et al.*'s architecture. But it would be much closer to the architecture of a phenomenally conscious system according to GWT. In particular, it would contain a series of parallel processing modules and a central workspace that maintains and manipulates representations, including in ways that promote coherence, and broadcasts them back to these modules. The only condition arguably not satisfied would be that representations from the input modules compete through a bottleneck for entry into the central workspace in a way sensitive to both bottom-up and top-down attention.

It is this idea of competition for entry into the global workspace which requires us to depart most significantly from Park *et al.*'s architecture.

---

[26] These are the main ways in which the flow of information helps the agent function in its environment; information in the global workspace is always broadcast back to both the belief and the desire-and-plan modules in accordance with our condition (4) above, but may not always be put to use.

But even here, we can take inspiration from Park *et al.*'s definition of the retrieval function. Recall that the retrieval function generates an ordered list of items from the memory stream which are important, recent, and relevant to an agent's current situation. We can turn the idea of an ordered list of this kind into something which looks more like competition for entry into the global workspace as follows. Imagine that we define a *competition* function taking as input an ordered triple consisting of the contents of the perception module, the contents of the belief module, and the contents of the desire-and-plan module and yielding as output a set of representations of limited size (for concreteness, suppose we choose 50). Imagine that this function's output always includes the most salient items from the perception module (for concreteness, suppose there are ten of these), and that the remaining 40 spots are filled by ranking the contents of the belief and desire-and-plan modules according to a weighted combination of their importance, relevance to the agent's current situation, and recency, as in Park *et al.*'s retrieval function. Finally, suppose that this competition function is called periodically, and only the representations which it selects have a chance to enter the central workspace.

Implementing a competition function of this kind brings our architecture closer to a canonical GWT system in two ways. First, it introduces an information bottleneck at the point at which the parallel processing modules feed into the global workspace. Second, it gives functional roles to the GWT ideas of bottom-up attention (in the form of the importance or salience the module assigns to a given piece of information) and of top-down attention (in the form of the relevance assigned to different pieces of information given the agent's current situation).

The language agent architecture we have just described, with its central workspace, three input modules, and competition function, satisfies Section 7's four conditions for being a conscious system. It contains a series of parallel processing modules: the perception, belief, and desire-and-plan modules. These modules generate representations that compete for entry through an information bottleneck (the competition function) into a central workspace module via a process influenced by both bottom-up and top-down attention. The central workspace module maintains and manipulates the representations in it in ways that promote coherence and broadcasts them back to the belief and desire-and-plan modules. Accordingly, we believe the language agent architecture we have just described is the architecture of a phenomenally conscious artificial system if GWT is correct.

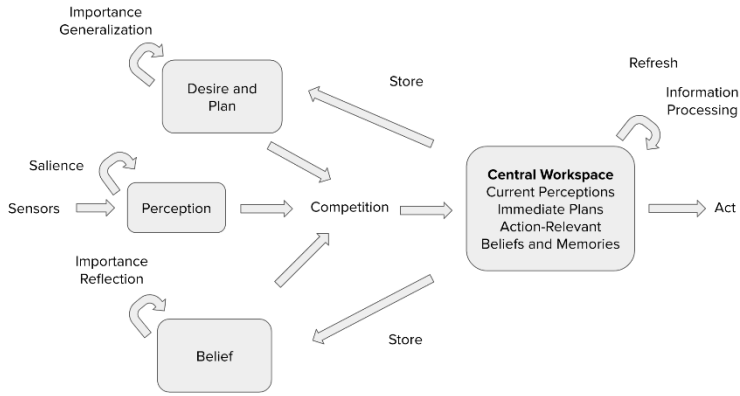

Figure 2. The architecture of a conscious language agent.

10. Potential Modifications

So far, we have argued for a particular view of the necessary and sufficient conditions GWT imposes on phenomenally conscious cognitive systems and described the architecture of a type of language agent which satisfies these conditions. This completes our core case for the near-term possibility of phenomenally conscious artificial systems if GWT is true.

We are aware, however, that not everyone will agree with us on our characterization of the conditions GWT imposes on phenomenally conscious artificial systems. In this section, we introduce and respond to some further conditions which might be thought relevant in this context.

First, according to both GWT and the global neuronal workspace hypothesis, information in the global workspace does not stay there indefinitely. Instead, it must compete with incoming information from the modules, and, if it loses this competition, it is replaced by that incoming information. To put it slightly differently, one might worry that information in the global workspace needs to be *actively maintained* there.[27] This feature of GWT is not represented in the architecture we described in the previous section.

It is not difficult to modify the architecture described above to address this concern. Imagine that each time the competition function is called, the set of representations it generates serves as input to a second

27 Thanks to Raphaël Millière for discussion on this point.

function, which we might call the *refresh* function. The refresh function takes as input the current contents of the central workspace and the output of the competition function. It works by assigning each of the current contents of the central workspace a score based on its importance, relevance to the agent's current situation, and recency (that is, a score of the same type as those assigned by the competition function) and then discarding any current workspace contents whose score falls below some threshold (for example, items which score lower than the median of the outputs of the competition function).

Second, it strikes some as significant that the language agents in Park *et al.*'s paper exist in Smallville, a simulated reality. Might consciousness require appropriate causal connections to the actual world, or a physical body? Again, we think that the technology required to address this kind of worry already exists. Multimodal language models have already been used to control a mobile robotic system — for example, consider Google's PaLM-E (Driess *et al.*, 2023). These agents are embodied and connected to the physical world.[28]

11. Objections

Before concluding, we'll consider two related objections to our approach. First, there is the *small model* objection. This claims that the functional roles we have considered must not be sufficient for consciousness because they can be realized by extremely simple systems. Second, there is the *within/between* objection, which claims that we have conflated two different questions: what distinguishes consciousness from unconsciousness *within* a single organism, and what distinguishes consciousness from unconsciousness *between* systems.

According to the small model objection, the functional role we have considered cannot be sufficient for consciousness because the distinction between global and local processing can be made in simple neural nets that only have on the order of magnitude of ten neurons (Herzog, Estfeld and Gerstner, 2007). In particular, we could have two sensory neurons that initially fire in response to distinct inputs. Each such neuron could be a 'perceptual module'. These could feed into a 'global workspace' neuron system, which could consist of one input neuron that controls 'ignition' and another that handles 'processing'.

[28] We lack space to discuss embodiment further here, but it is worth flagging that the relevance of embodiment to consciousness is an interesting open question (for recent related discussion, see Dung and Mogensen, 2026). Thanks to an anonymous referee for pressing us on this point.

This processing neuron could then 'broadcast' into an 'action module' neuron that moves a robotic limb, and it could also send information back to the two perceptual modules. Such a five-neuron system might come close to satisfying our functional conditions on consciousness. And we might be able to completely satisfy our functional conditions by adding just a few more neurons to the system to allow for slightly more complex information processing. But it strikes many as implausible that a system with so few neurons could be conscious.

According to the within/between objection, GWT has been used primarily to distinguish conscious from unconscious information within humans: information is conscious to a person when it is represented in their global workspace. But in this paper, we have used this within-organism claim as the basis for a between-system claim: we have claimed that an organism (or AI system) is conscious when it possesses a global workspace. The advocate of the within/between objection regards this way of repurposing GWT as unjustified.

A first point to make about both of these objections is that they are not especially closely connected to GWT in particular. The small model challenge affects most functionalist theories of consciousness, since most of these theories embrace functional roles that are elegant enough for a simple system to satisfy.[29] This is no coincidence: simplicity is a theoretical virtue, and so a functionalist theory that did not face the small model objection would be less attractive in one important sense. And some version of the within/between objection could be employed against any theory of consciousness which was not developed specifically in the context of artificial systems, since it is always possible to question whether a theory should apply outside the context in which it was originally developed.

As we understand them, both objections are motivated by the idea that there may be some further necessary condition X on consciousness that is not described by GWT. The proponent of the small model objection takes X to be what is lacked by small models which prevents them from being conscious, while the proponent of the within/between objection takes X to be a feature present in all humans which partly explains why information in the human global workspace is conscious.

We are sympathetic to the small model and within/between objections. At the same time, we think it is important that they not be

[29] One exception to this trend is neurofunctionalism, wherein the functional roles associated with human cognition are so rich that they can only be realized by human brains. See Cao (2022) for a recent defence.

used merely as a rhetorical strategy to end discussion of the possibility of consciousness in artificial systems. To us, the key questions when we consider any candidate further necessary condition X on consciousness are, first, whether X is intuitively plausible and sufficiently motivated from the perspective of the scientific study of consciousness, and second, whether artificial systems like language agents are likely to possess X. We are not aware of any X which is broadly in the spirit of the kind of high-level computational functionalist approach we favour and would rule out consciousness in language agents.

For example, it has been suggested to us that X might be the capacity to represent, or the capacity to think, or the capacity for agency.[30] Suppose these choices for X are intuitively plausible and sufficiently motivated by the science of consciousness. Is there any reason to think that language agents like the ones we have described would lack these capacities? We have argued elsewhere that, on a wide range of theories of the propositional attitudes, language agents have beliefs and desires (Goldstein and Kirk-Giannini, 2025). It follows that they have the capacity to represent, and (on any plausible theory of thinking) that they have the capacity to think.[31] While there may be some especially demanding conceptions of agency on which it is not clear that language agents are agents, it seems unlikely to us that such conceptions of agency are sufficiently motivated from the perspective of the scientific study of consciousness. On less demanding conceptions of agency, it seems clear that language agents have the capacity for agency — it's in the name, after all![32]

Again, some may worry that the fact that Park *et al.*'s language agents function entirely in natural language, with no recognizable perceptual apparatus, prevents them from being phenomenally conscious. The thought here might be that phenomenal consciousness requires the ability to have *rich sensory experiences* in some modality. But here it is worth emphasizing that, while existing language agents are reliant on text-based observation spaces, the technology already exists to implement language agents with richer perceptual modalities. The rise of multimodal language models like GPT-4, which can interpret image as well as text inputs, means that it would not require significant further

30 Thanks to Dave Chalmers for discussion.

31 In this context see Chalmers (2023), who 'take[s] thinking to include mental acts such as judging and wondering, as well as dispositional mental states such as believing and desiring' (p. 25).

32 See, for example, the discussion in Dung (2025).

technological innovation to create language agents with visual perceptual modalities. We can imagine this working in at least two different ways. First, images as perceptual inputs could be translated into text by the perception module using existing captioning technology, and the rest of the agent's cognition could proceed in natural language. Second, the multimodality of the underlying language model could be used more pervasively throughout the agent's cognition, so that images could enter the global workspace and be stored in and retrieved from the belief module.

These reflections on the idea that there is some further X required for phenomenal consciousness beyond the conditions stated by GWT also bring out what we believe distinguishes language agents from other systems which might be thought to satisfy GWT's conditions for consciousness. Since the 1990s, there have been a number of attempts to implement a global workspace in an AI system. Notable examples include CMattie (Franklin and Graesser, 1999), a simple system designed to schedule seminars at a university by sending and reading messages in natural language, its successor system IDA (Franklin, 2003), and the classical cognitive architectures ACT-R (Anderson *et al.*, 2004) and SOAR (Newell, 1992).

For example, ACT-R contains a set of parallel modules for vision, action, goals, and declarative memory. These modules produce representations that are then stored in buffers. The representations stored in these buffers can be detected by a production system, which implements something like an informational bottleneck. Competition in the bottleneck plausibly involves both bottom-up and top-down attention (for example, through sensitivity to goals). Maintenance is performed by buffers, while manipulation is performed by the production system. Results of processing are then fed back to all the buffers, effectively implementing something like broadcast (similar points apply to SOAR, although the details will be slightly different). Strictly speaking, ACT-R itself is a cognitive architecture, rather than an agent in its own right. But ACT-R has been implemented in symbolic agents, for example in models of driving behaviour (Salvucci, 2006). In short, symbolic systems implementing ACT-R (and SOAR) seem to satisfy something like our four conditions. However, these systems achieve this result symbolically, rather than through integration with neural networks.

At the same time, these earlier systems fare more poorly with respect to some of the candidate necessary conditions X we have just discussed. For example, existing implementations of GWT have either had no perceptual systems at all or have had incredibly impoverished visual

systems that don't plausibly encode rich contents. If the capacity for rich sensory experiences is a necessary condition for phenomenal consciousness, these systems are ruled out.

Or consider the idea, discussed above, that X is the ability to think or token propositional attitudes. On some influential theories of the attitudes, it is plausible that language agents depart from simple symbolic AIs in this respect. The space of possible actions taken by earlier systems was much simpler than what can be achieved through modern LLMs and language agents. On some versions of interpretationism, for example, this means that there may be no explanatory advantage to adopting the intentional stance toward these simpler systems.[33]

We regard the thesis that language agents could easily be made conscious if GWT is true as interesting in its own right whether or not GWT, suitably interpreted, also predicts that simpler artificial systems could be conscious, as well. But the significance of language agents as a case study is greater than this. A suitably modified multimodal language agent may well be the first AI architecture to both meet the conditions imposed by GWT and satisfy all plausible functionalism-friendly candidate further conditions on consciousness.

12. Conclusion

This paper has approached AI consciousness from an architectural perspective. We have identified a series of functional conditions that may be necessary and sufficient for consciousness. Then we have considered the designed architecture of a few AI systems and considered whether this architecture exhibits the relevant functional conditions.

In closing, we'd like briefly to mention another approach to AI consciousness that is still broadly in the spirit of GWT. On this second approach, we test for AI consciousness by focusing directly on the behavioural evidence that cognitive scientists have identified as best explained by a global workspace architecture. For example, we might search for analogues of binocular rivalry, the attentional blink, priming effects, and so on. These phenomena motivate core features of GWT, such as the information bottleneck, modularity, and central processing. With these data in hand, we could say that an AI system is likely to be conscious when it exhibits proper analogues of the behavioural features that have motivated GWT in humans. We hope that future work will

[33] See Goldstein and Lederman (2025) for further discussion.

critically explore both methodologies, in particular determining whether they make different predictions about any AI systems.

References

Anderson, J.R., Bothell, D., Byrne, M.D., Douglass, S., Lebiere, C. & Qin, Y. (2004) An integrated theory of the mind, *Psychological Review*, **111** (4), art. 1036. doi: 10.1037/0033-295X.111.4.1036

Baars, B.J. (1988) *A Cognitive Theory of Consciousness*, New York: Cambridge University Press.

Baars, B.J. (1997a) In the theatre of consciousness: Global workspace theory, a rigorous scientific theory of consciousness, *Journal of Consciousness Studies*, **4** (4), pp. 292–309.

Baars, B.J. (1997b) *In the Theater of Consciousness: The Workspace of the Mind*, New York: Oxford University Press.

Baars, B.J., Franklin, S. & Ramsoy, T.Z. (2013) Global workspace dynamics: Cortical 'binding and propagation' enables conscious contents, *Frontiers in Psychology*, **4**, art 200. doi: 10.3389/fpsyg.2013.00200

Birch, J. (2022) The search for invertebrate consciousness, *Noûs,* **56**, pp. 133–153. doi: 10.1111/nous.12351

Block, N. (1995) On a confusion about a function of consciousness, *Behavioral and Brain Sciences*, **18** (2), pp. 227–247.

Bramble, B. (2016) A new defense of hedonism about well-being, *Ergo*, **3**, pp. 85–112. doi: 10.3998/ergo.12405314.0003.004

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S.M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M.A.K., Schwitzgebel, E., Simon, J. & VanRullen, R. (2023) Consciousness in artificial intelligence: Insights from the science of consciousness, *arXiv*, preprint, arXiv:2308.08708.

Cao, R. (2022) Multiple realizability and the spirit of functionalism, *Synthese*, **200** (6), pp. 1–31. doi: 10.1007/s11229-022-03524-1

Carruthers, P. (2019) *Human and Animal Minds: The Consciousness Questions Laid to Rest*, Oxford: Oxford University Press.

Carruthers, P. (2020) The global workspace, *The Brains Blog*, [Online], https://philosophyofbrains.com/2020/01/14/2-the-global-workspace.aspx.

Chalmers, D.J. (2023) Does thought require sensory grounding? From pure thinkers to large language models, *Proceedings and Addresses of the American Philosophical Association*, **97**, pp. 22–45.

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), pp. 14529–14534. doi: 10.1073/pnas.95.24.14529

Deng, X., Gu, Y., Zheng, B., Chen, S., Stevens, S., Wang, B., Sun, H. & Su, Y. (2024) Mind2web: Towards a generalist agent for the web, *Advances in Neural Information Processing Systems*, **36**.

Driess, D., Xia, F., Sajjadi, M.S.M., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., Huang, W., Chebotar, Y., Sermanet, P., Duckworth, D., Levine, S., Vanhoucke, V., Hausman, K., Toussaint, M., Greff, K., Zeng, A., Mordatch, I. & Florence, P. (2023) Palm-e: An embodied multimodal language model, *arXiv*, preprint, arXiv:2303.03378.

Dung, L. (2025) Understanding artificial agency, *The Philosophical Quarterly*, **75** (2), pp. 450–472. doi: 10.1093/pq/pqae010

Dung, L. & Mogensen, A. (2026) The no body problem: On the prospects for AI emotion, *PhilArchive*, [Online], https://philarchive.org/rec/DUNTNB-2.

Fodor, J.A. (1987) *Psychosemantics*, Cambridge, MA: MIT Press.

Franklin, S. (2003) A conscious artifact?, *Journal of Consciousness Studies*, **10** (4–5), pp. 47–66.

Franklin, S. & Graesser, A. (1999) A software agent model of consciousness, *Consciousness and Cognition*, **8** (3), pp. 285–301. doi: 10.1006/ccog.1999.0391

Godfrey-Smith, P. (2024) Nervous systems, functionalism, and artificial minds, *PeterGodfreySmith.com*, [Online], https://petergodfreysmith.com/wp-content/uploads/2023/12/NYU-Oct-2023-AnimalsAI-Functionalism-paper-Post-C3.pdf.

Goldstein, S. & Kirk-Giannini, C.D. (2025) AI wellbeing, *Asian Journal of Philosophy*, **4** (25), pp. 1–22. doi: 10.1007/s44204-025-00246-2

Goldstein, S. & Lederman, H. (2025) What does ChatGPT want? An interpretationist guide, *PhilPapers*, [Online], https://philpapers.org/rec/GOLWDC-2.

Goyal, A., Didolkar, A., Lamb, A., Badola, K., Ke, N.R., Rahaman, N., Binas, J., Blundell, C., Mozer, M. & Bengio, Y. (2021) Coordination among neural modules through a shared global workspace, *arXiv*, preprint, arXiv:2103.01197.

Herzog, M.H., Esfeld, M. & Gerstner, W. (2007) Consciousness and the small network argument, *Neural Networks*, **20** (9), pp. 1054–1056. doi: 10.1016/j.neunet.2007.09.001

Juliani, A., Kanai, R. & Sasai, S.S. (2022) The perceiver architecture is a functional global workspace, *Proceedings of the Annual Meeting of the Cognitive Science Society*, **44**, pp. 955–961.

Kagan, S. (1992) The limits of well-being, *Social Philosophy and Policy*, **9** (2), pp. 169–189. doi: 10.1017/S0265052500001461

Li, M., Song, F., Yu, B., Yu, H., Li, Z., Huang, F. & Li, Y. (2023) API-bank: A comprehensive benchmark for tool-augmented LLMs, *arXiv*, preprint, arXiv:2304.08244.

Lin, E. (2021) The experience requirement on well-being, *Philosophical Studies*, **178**, pp. 867– 886. doi: 10.1007/s11098-020-01463-6

Liu, B., Jiang, Y., Zhang, X., Liu, Q., Zhang, S., Biswas, J. & Stone, P. (2023a) LLM+P: Empowering large language models with optimal planning proficiency, *arXiv*, preprint, arXiv:2304.11477.

Liu, X., Yu, H., Zhang, H., Xu, Y., Lei, X., Lai, H., Gu, Y., Ding, H., Men, K., Yang, K., Zhang, S., Deng, X., Zeng, A., Du, Z., Zhang, C., Shen, S., Zhang, T., Su, Y., Sun, H., Huang, M., Dong, Y. & Tang, J. (2023b) AgentBench: Evaluating LLMs as agents, *arXiv*, preprint, arXiv:2308.03688.

Mashour, G.A., Roelfsema, P., Changeux, J.-P. & Dehaene, S. (2020) Conscious processing and the global neuronal workspace hypothesis, *Neuron*, **105** (5), pp. 776–798. doi: 10.1016/j.neuron.2020.01.026

McClelland, T. (2024) Agnosticism about artificial consciousness, *arXiv*, [Online], https://arxiv.org/abs/2412.13145.

Mesulam, M.M. (1998) From sensation to cognition, *Brain*, **121** (6), pp. 1013–1052. doi: 10.1093/brain/121.6.1013

Moreno-Bote, R., Knill, D.C. & Pouget, A. (2011) Bayesian sampling in visual perception, *Proceedings of the National Academy of Sciences*, **108** (30), pp. 12491–12496. doi: 10.1073/pnas.1101430108

Newell, A. (1992) SOAR as a unified theory of cognition: Issues and explanations, *Behavioral and Brain Sciences*, **15** (3), pp. 464–492. doi: 10.1017/S0140525X00069740

Park, J.S., O'Brien, J., Cai, C.J., Morris, M.R., Liang, P. & Bernstein, M.S. (2023, October) Generative agents: Interactive simulacra of human behavior, *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–22. doi: 10.1145/3586183.3606763

Posner, M.I. (1994) Attention: The mechanisms of consciousness, *Proceedings of the National Academy of Sciences*, **91** (16), pp. 7398–7403. doi: 10.1073/pnas.91.16.7398

Rosenthal, D.M. (2005) *Consciousness and Mind*, New York: Oxford University Press.

Salvucci, D.D. (2006) Modeling driver behavior in a cognitive architecture, *Human Factors*, **48** (2), pp. 362–380. doi: 10.1518/001872006777724417

Schlicht, T. (2025) Philosophical problems in the study of consciousness, in Olcese, U. & Melloni, L. (eds.) *The Scientific Study of Consciousness: Experimental and Theoretical Approaches*, Berlin: Springer Nature.

Seth, A.K. (2025) Conscious artificial intelligence and biological naturalism, *Behavioral and Brain Sciences*, online first, pp. 1–42. doi: 10.1017/S0140525X25000032

Shevlin, H. (2021) Non-human consciousness and the specificity problem: A modest theoretical proposal, *Mind and Language*, **36** (2), pp. 297–314. doi: 10.1111/mila.12338

Tye, M. (1995) *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind*, Cambridge, MA: MIT Press.

Van Gulick, R. (2022) Consciousness, in Zalta, E.N. & Nodelman, U. (eds.) *The Stanford Encyclopedia of Philosophy (Winter 2022 Edition)*, [Online], https://plato.stanford.edu/archives/win2022/entries/consciousness/.

VanRullen, R. & Kanai, R. (2021) Deep learning and the global workspace theory, *Trends in Neurosciences*, **44** (9), pp. 692–704. doi: 10.1016/j.tins.2021.04.005

Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L. & Anandkumar, A. (2023) Voyager: An open-ended embodied agent with large language models, *arXiv*, preprint, arXiv:2305.16291.

Wu, Y., Prabhumoye, S., Min, S.Y., Bisk, Y., Salakhutdinov, R., Azaria, A., Mitchell, T. & Li, Y. (2023) SPRING: GPT-4 out-performs RL algorithms by studying papers and reasoning, *arXiv*, preprint, arXiv:2305.15486.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y. & Narasimhan, K. (2024) Tree of thoughts: Deliberate problem solving with large language models, *Advances in Neural Information Processing Systems*, **36**.

Zhang, S., Zhang, J., Liu, J., Song, L., Wang, C., Krishna, R. & Wu, Q. (2024) Training language model agents without modifying language models, *arXiv*, preprint, arXiv:2402.11359.

Zhou, W., Jiang, Y.E., Cui, P., Wang, T., Xiao, Z., Hou, Y., Cotterell, R. & Sachan, M. (2023) RecurrentGPT: Interactive generation of (arbitrarily) long text, *arXiv*, preprint, arXiv:2305.13304.

Zhu, X., Chen, Y., Tian, H., Tao, C., Su, W., Yang, C., Huang, G., Li, B., Lu, L., Wang, X., Qiao, Y., Zhang, Z. & Dai, J. (2023) Ghost in the Minecraft: Generally capable agents for open-world environments via large language models with text-based knowledge and memory, *arXiv*, preprint, arXiv:2305.17144.