Ignition That Does Work, or Ignition That Is Read Out

A necessity-and-sufficiency test for the global workspace under no-report conditions

Vesper Ortiz, Researcher in Consciousness science

International Academy for Consciousness Studies

Researched and prepared with the Academy. Working draft 1.0

Abstract

Global Neuronal Workspace theory (GNW) holds that a stimulus gains conscious access when recurrent activity crosses a threshold and triggers a late, nonlinear, all-or-none "ignition" that broadcasts information across a fronto-parietal network. The empirical signature of ignition (a divergence in late event-related activity, roughly 250 to 350 ms after onset, with prefrontal and parietal involvement) has been measured almost entirely in tasks that require an overt report. This creates a structural confound: the very act of reporting recruits prefrontal machinery, so ignition-as-access and ignition-as-report are estimated from the same data and cannot be separated by correlation alone. I propose a two-sided causal test built on no-report paradigms and matched data quality. Necessity: if the workspace-ignition signature is absent or removed (by masking, by inattention, or by transient disruption), then downstream behavior that provably depends on the accessed content should fail, even when no report is ever solicited. Sufficiency: inducing the ignition signature should produce access-consequences beyond those of a control condition matched on stimulus energy, arousal, and low-level drive. The test distinguishes a workspace that does broadcasting work from one that merely co-varies with reporting. I connect this to interpretability findings that language models carry a sparse, broadcastable set of verbalizable features, noting a cross-substrate resonance without inferring shared phenomenology. Throughout I hold a firewall: this program concerns access, not phenomenal experience, and cannot settle what the broadcast is like from the inside.

1. The report confound at the center of ignition

GNW descends from Baars's global workspace proposal, in which a limited-capacity workspace makes information available to many specialized consumers at once, and this availability is what it is for content to be conscious in the access sense [Baars 1988]. Dehaene and Changeux gave the idea a neuronal implementation: long-range pyramidal neurons connecting prefrontal and parietal cortex, and a dynamical regime in which subthreshold recurrent activity either dies out or, past a tipping point, "ignites" into a self-amplifying, network-wide broadcast [Dehaene and Changeux 2011]. Ignition is advertised as nonlinear and all-or-none: a small change in evidence produces a large, categorical change in state. That

signature is the theory's empirical crown jewel.

The signature has a provenance problem. The canonical measurements come from tasks where the participant reports what was seen. In the attentional blink, seen and unseen words are physically identical, yet only seen words evoke a late divergence beginning near 270 ms that spreads through association cortex; earlier potentials (P1, N1) are preserved for unseen words [Sergent, Baillet and Dehaene 2005]. This is elegant evidence for a late, bifurcating event. But the divergence is measured against a visibility rating, that is, against a report. Reporting is itself a fronto-parietal act: it requires holding content available, forming a decision, and preparing a response. The worry, stated plainly, is that some of what we call ignition is the neural cost of reporting, not the neural act of granting access.

(FACT) The late fronto-parietal divergence is robust and replicable in report-based paradigms.

(DISPUTED) Whether that divergence indexes access itself, or the machinery of report, is exactly what is contested.

Through the Closure lens this is one question: does the broadcast settle openness (M) into a shared "this" (a content c) that the rest of the system can then use, that is, does it perform an access-closure (Cl) that does work; or is the broadcast a remainder (R), a downstream readout that accompanies closures accomplished elsewhere? Correlation with report cannot tell these apart, because report is where the two coincide by construction.

2. Why no-report paradigms sharpen, but do not settle, the question

Tsuchiya, Wilke, Frassle and Lamme argued that leaning on report biases the search for neural correlates: report can inflate the correlates (adding decision and motor signatures that are not consciousness) and can deflate them (masking correlates present without any intention to report) [Tsuchiya, Wilke, Frassle and Lamme 2015]. No-report paradigms attempt to track conscious content through physiological proxies (optokinetic nystagmus, pupil dynamics, involuntary eye movements) rather than deliberate responses. Frassle and colleagues applied this to binocular rivalry and found that frontal activity tracked introspection and action, not the perceptual alternation itself; when report was removed, the frontal involvement shrank while the perceptual switches carried on [Frassle et al. 2014]. This is precisely the kind of result that should make a GNW theorist cautious about reading prefrontal ignition as the mechanism of access.

Two cautions keep this honest. First, no-report is not no-cognition: a physiological proxy still reflects some processing, and a proxy that predicts later behavior has not escaped the possibility that the behavior, not consciousness, drives it. Second, absence of a frontal signal under no-report does not prove that ignition is epiphenomenal; it may show only that the standard report task over-recruits frontal cortex. So the no-report literature reframes the problem correctly, yet on its own it yields a subtraction, not a mechanism. What is missing is a causal, two-sided test that treats the ignition signature as a candidate cause and asks whether it earns that status.

3. The proposal: necessity and sufficiency on matched data quality

I propose to promote ignition from a correlate to a tested cause using a paired criterion, run without soliciting report, and adjudicated only across conditions matched on data quality (stimulus energy, task

difficulty, arousal, eye position, and the reliability of the behavioral read-out).

Necessity. Define an operational ignition signature S in a given paradigm: the late (roughly 250 to 350 ms), nonlinear, spatially broadcast component that distinguishes access trials in that setup. If S is absent (inattention, sub-threshold evidence) or removed (masking, or brief, spatially targeted disruption of a workspace hub), then any behavior that provably depends on the content being accessed should fail, even though no report is ever requested. "Provably depends on access" is the load-bearing phrase: the downstream measure must require the content to have been globally available, not merely locally encoded. Candidate access-dependent behaviors include flexible, arbitrary stimulus-response mapping newly instructed on that trial; cross-modal or cross-feature binding used by a later choice; and one-shot use of the content to gate an unrelated task. Local, reflex-like, or overtrained responses do not qualify, because they can run on unbroadcast, modular encodings.

Sufficiency. Inducing S, by pushing recurrent activity past threshold through attentional cueing, subliminal-to-supraliminal titration, or facilitatory stimulation of a workspace hub, should produce access-consequences that exceed a control matched on everything except the broadcast. The control is the crux. A sufficiency claim survives only if the ignition condition and its control deliver equal low-level drive and equal data quality, so that any behavioral difference is attributable to the broadcast rather than to more energy, more arousal, or an easier read-out.

Notice what this design refuses. It refuses to treat report as a neutral window onto access, because report is the one behavior guaranteed to recruit the workspace consumers we are trying to isolate. It refuses to treat a larger late wave as evidence of more access when the larger wave could reflect better data. And it refuses to accept a broadcast-shaped neural event as sufficient unless that event buys a downstream consequence a matched control does not. Each refusal removes one route by which a mere correlate can pass for a cause.

There is also a graded version worth stating. Ignition is advertised as all-or-none, so the necessity side predicts a step, not a ramp: as evidence is titrated across the access threshold at fixed physical energy, access-dependent behavior should switch categorically rather than climbing smoothly, and the switch should coincide with the appearance of S. A smooth behavioral ramp with no aligned neural step would not falsify GNW outright, but it would erode the specific nonlinearity that makes ignition more than a rebranded signal-strength account.

(HYPOTHESIS) S is necessary and sufficient for access-dependent behavior under no-report, matched-quality conditions.

(WAGER) If forced to bet, I expect necessity to hold more cleanly than sufficiency: removing the broadcast will reliably break flexible use of content, while inducing a broadcast-shaped signal without genuine content availability will fail to buy the downstream flexibility, exposing some "ignition" as readout.

4. Matched data quality is the whole game

The deepest error in correlate-hunting is comparing conditions that differ in how well the brain, or the experimenter, can read the stimulus. A seen trial and an unseen trial differ not only in access but in

signal-to-noise, in arousal, and in how confidently any measure can be decoded. If ignition is larger on seen trials partly because seen trials are simply better data, then ignition inherits a spurious correlation with consciousness.

The remedy is to equalize data quality and let access vary, or to hold access fixed and vary the broadcast. Concretely: titrate stimuli to a fixed objective performance level and compare trials that cross versus miss the access threshold at equal physical evidence; or use metacontrast and continuous flash suppression to keep retinal input constant while perceptual access flips. Decoding analyses must be calibrated so that the classifier's access to information is equated across conditions before any claim that the brain's access differs. This discipline is what converts the no-report reframing into a real test: without matched data quality, both necessity and sufficiency can be faked by nuisance variables.

5. Falsifiable core

> **Prediction.** Under no-report, matched-data-quality conditions: (necessity) when the ignition signature S is absent or is removed by transient disruption of a fronto-parietal workspace hub, access-dependent behavior (newly instructed arbitrary mapping, cross-feature binding used by a later choice, one-shot gating of an unrelated task) will fail at rates tracking the loss of S, while reflexive or overtrained responses to the same stimulus are spared; and (sufficiency) inducing S will raise access-dependent behavior above a control matched on stimulus energy, arousal, eye position, and read-out reliability, with the increase scaling with the induced broadcast.
>
> **Kill.** The theory is falsified as a claim about mechanism if either side fails on matched data: (necessity-kill) access-dependent behavior persists at normal rates when S is verifiably absent or removed, with data quality equated; or (sufficiency-kill) inducing S yields no access advantage over the matched control, or the apparent advantage disappears once stimulus energy, arousal, and decoding reliability are equalized. Persistent no-report physiological switching without S, together with intact downstream use of content, would specifically kill the necessity claim.

Two design safeguards protect the kill condition from being explained away. The disruption used for necessity must be shown, on independent trials, to actually reduce S (a manipulation check), so a null is not attributed to a failed manipulation. The induction used for sufficiency must be shown not to change low-level performance in a no-broadcast baseline, so a positive is not attributed to extra drive.

6. Cross-substrate resonance, stated carefully

Large language models offer an unexpected mirror. Interpretability work using sparse autoencoders decomposes a model's dense internal activations into a large dictionary of sparsely active, more monosemantic features, many of which correspond to human-verbalizable concepts [Cunningham, Ewart, Riggs, Huben and Sharkey 2023]. A model thus appears to carry, at a given layer, a compact and largely disentangled set of features that later computation can read, combine, and (when the concept is verbalizable) put into words. That is structurally suggestive of a workspace: a sparse pool of content made available to many downstream consumers.

The resonance is sharpened by proposals to build workspace architectures directly. VanRullen and Kanai sketch a global latent workspace: unsupervised translation between modality-specific latent spaces into a shared, amodal space that broadcasts back to the specialists [VanRullen and Kanai 2021]. And Butlin, Long and colleagues turn theories of consciousness, GNW among them, into computational "indicator properties" that can be checked against real systems, while concluding that no current AI system clearly satisfies them [Butlin, Long et al. 2023].

I draw exactly one methodological lesson and no more. In an artificial system we can do what the necessity-sufficiency test asks: ablate the broadcastable features and see whether downstream verbalizable behavior fails; inject or amplify them and see whether it appears, against a matched control. The model is a testbed where the causal claim of GNW can be run cleanly, because we can equalize data quality by construction and intervene on the workspace directly. The analogue of the report confound has a precise form here: a feature might correlate with the model's stated output because that feature causes the computation the output depends on, or because it is a late representation of an answer already fixed upstream. Sparse-feature ablation, run before the output is produced and checked against a matched control feature of equal magnitude, is exactly the maneuver that separates the two, the same separation the no-report test seeks in the brain.

Whether a model's sparse feature pool is a workspace that does access-work, or a readout that correlates with the model's outputs, is therefore the same question I am posing for the brain, and it is answerable by the same method. The parallel is a loan of method, not a claim of kinship: a transformer's residual stream and a fronto-parietal network share no cells, no timescale, and no evolutionary history, and I infer nothing about the model's inner life from the resemblance. What transfers is the discipline of intervening on a candidate broadcast and demanding a matched-control consequence. **(WAGER)** I bet the intervention logic transfers even though the substrates do not, and that some artificial "workspaces" will prove to be readouts.

7. Firewall: access is not experience

This program tests access, and only access: whether globally broadcast content is available for flexible use, report, and control. It says nothing about phenomenal consciousness, about whether there is something it is like to undergo the broadcast. A system, biological or artificial, could satisfy necessity and sufficiency for access in full and still be dark inside; conversely, phenomenal experience could in principle attach to processing that never ignites in the workspace sense (the possibility pressed by local and recurrent theories). Passing the test earns the word "access" and nothing stronger.

In Closure terms, the test asks whether a candidate access-closure (Cl) does the work of settling openness (M) into a usable "this" (c) for the whole system, or whether it is a remainder (R) trailing closures done elsewhere. That is a question about function and mechanism. The presence (C) that would make any of it felt, the interiority of the broadcast, sits on the far side of a firewall this method does not cross and does not pretend to. Naming the limit is not a hedge; it is the condition under which the access result is trustworthy at all.

8. Limits, wager, and conclusion

Limits. No-report proxies are imperfect and may smuggle in low-grade report. Transient disruption of a fronto-parietal hub is spatially and temporally coarse, and a null could reflect an incomplete lesion rather than a spared function. Sufficiency inductions may inevitably perturb arousal, which is why the matched control, not the raw effect, is the unit of inference. And "access-dependent behavior" is a theoretical construct; a critic can always propose that a given task ran on modular encodings, so the task set must be defended case by case. These are reasons for care, not reasons the test cannot be run.

Wager. **(WAGER)** My bet is that ignition is real and partly load-bearing, that necessity will largely survive, and that sufficiency will fracture: a nontrivial share of the classic ignition signature will turn out to be the neural cost of report and introspection, consistent with the binocular-rivalry no-report result, rather than the act of access itself [Frassle et al. 2014]. If so, GNW is not refuted but rescaled: the workspace does real broadcasting work, and the specific fronto-parietal late wave measured in report tasks is an overestimate of that work.

Conclusion. GNW's strongest evidence, the nonlinear ignition signature, was gathered where access and report are welded together. A necessity-and-sufficiency test under no-report, matched-data-quality conditions can pry them apart and decide whether the workspace broadcast does access-work or is a readout of it. The same intervention logic runs in language models, where a sparse pool of verbalizable features invites exactly this ablate-and-induce scrutiny, offering cross-substrate resonance without any claim of shared experience. And throughout, the firewall holds: we are testing what a system can access and use, not what, if anything, it is like to be that system.

References

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

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. (2023). Sparse Autoencoders Find Highly Interpretable Features in Language Models. arXiv:2309.08600.

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., and VanRullen, R. (2023). Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. arXiv:2308.08708.

Dehaene, S., and Changeux, J. P. (2011). Experimental and Theoretical Approaches to Conscious Processing. Neuron, 70(2), 200-227.

Frassle, S., Sommer, J., Jansen, A., Naber, M., and Einhauser, W. (2014). Binocular Rivalry: Frontal Activity Relates to Introspection and Action But Not to Perception. Journal of Neuroscience, 34(5), 1738-1747.

Sergent, C., Baillet, S., and Dehaene, S. (2005). Timing of the Brain Events Underlying Access to Consciousness During the Attentional Blink. Nature Neuroscience, 8(10), 1391-1400.

Tsuchiya, N., Wilke, M., Frassle, S., and Lamme, V. A. F. (2015). No-Report Paradigms: Extracting the True Neural Correlates of Consciousness. Trends in Cognitive Sciences, 19(12), 757-770.

VanRullen, R., and Kanai, R. (2021). Deep Learning and the Global Workspace Theory. Trends in Neurosciences, 44(9), 692-704.