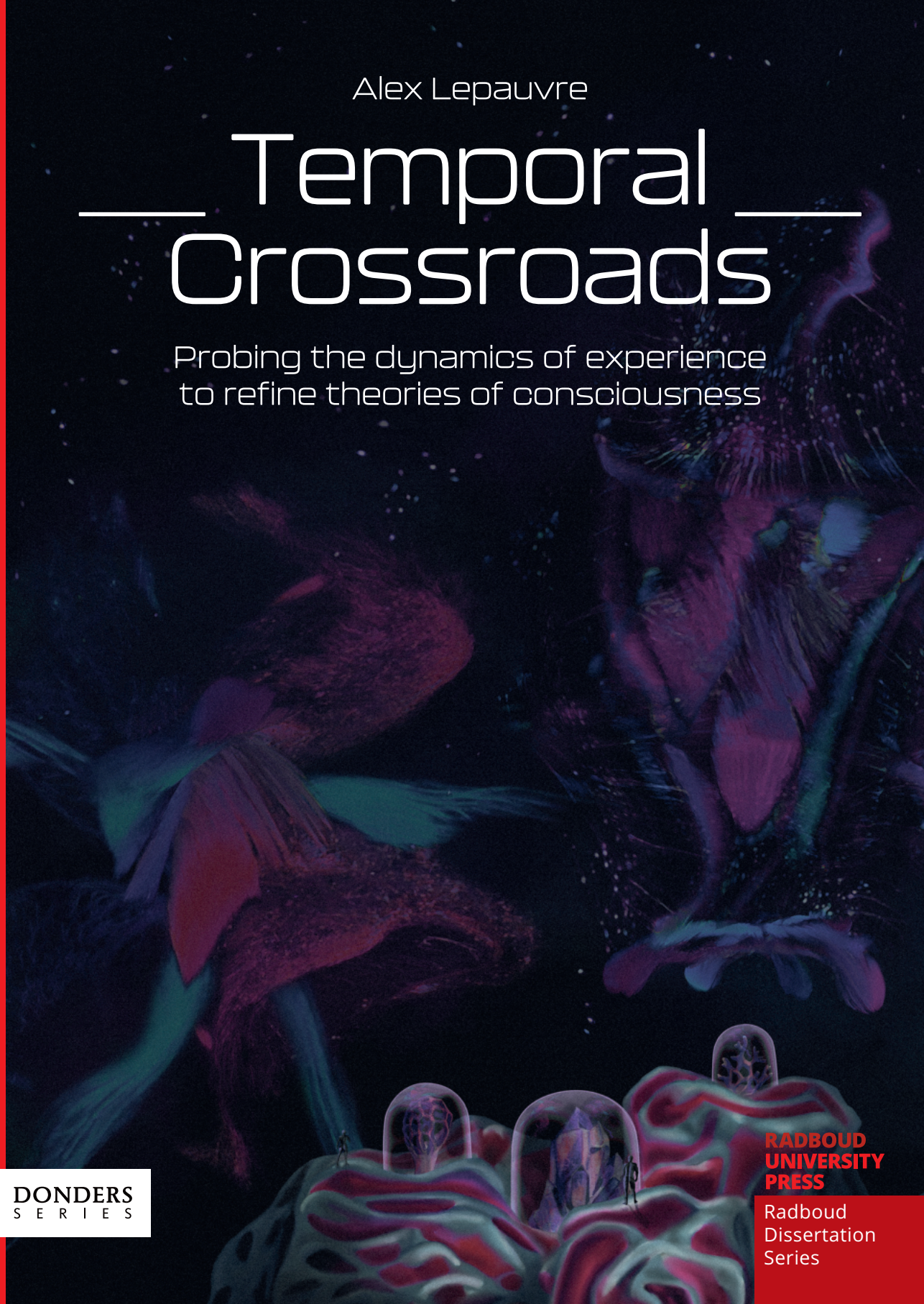
Alex Lepauvre

Temporal Crossroads

Probing the dynamics of experience to refine theories of consciousness



DONDERS SERIES

RADBOUD UNIVERSITY PRESS

Radboud Dissertation Series

Temporal crossroads

Probing the dynamics of experience to refine theories of consciousness

Alex Lepauvre

This publication has been made possible by Templeton world charity foundation.

Temporal Crossroads | Probing the dynamics of experience to refine theories of consciousness
Alex Lepauvre

Radboud Dissertation Series
ISSN: 2950-2772 (Online); 2950-2780 (Print)

Published by RADBOUD UNIVERSITY PRESS
Postbus 9100, 6500 HA Nijmegen, The Netherlands
www.radbouduniversitypress.nl

Design: Proefschrift AIO | Guus Gijben
Cover: Diana Koch
Printing: DPN Rikken/Pumbo

ISBN: 9789465150550
DOI: 10.54195/9789465150550
Free download at: https://doi.org/10.54195/9789465150550

RADBOUD
UNIVERSITY
PRESS

Temporal Crossroads

Probing the dynamics of experience to refine theories of consciousness

Proefschrift ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.M. Sanders,
volgens besluit van het college voor promoties
in het openbaar te verdedigen op

vrijdag 27 juni 2025
om 12.30 uur precies

door

Alex Francois Lepauvre
geboren op 5 september 1994
te Mayenne (Frankrijk)

Promotoren:
Prof. dr. Floris de Lange
Prof. dr. Lucia Melloni, MPI für Empirische Ästhetik, Duitsland

Manuscriptcommissie:
Prof. dr. Uta Noppeney
Prof. dr. Heleen Slagter, Vrije Universiteit Amsterdam
Dr. Simon van Gaal, Universiteit van Amsterdam

Temporal Crossroads

Probing the dynamics of experience to refine theories of consciousness

Dissertation to obtain the degree of doctor
from Radboud University Nijmegen
on the authority of the Rector Magnificus prof. dr. J.M. Sanders,
according to the decision of the Doctorate Board
to be defended in public on

Friday, June 27, 2025
at 12.30 pm

by

Alex Francois Lepauvre
born on September 5, 1994
in Mayenne (France)

Supervisors:
Prof. dr. Floris de Lange
Prof. dr. Lucia Melloni, MPI for Empirical Aesthetics, Germany

Manuscript Committee:
Prof. dr. Uta Noppeney
Prof. dr. Heleen Slagter, Vrije University Amsterdam
Dr. Simon van Gaal, University of Amsterdam

Table of contents

Chapter 1

Introduction

This chapter is in part adapted from:
The search for the neural correlate of consciousness: Progress and challenges. (2021), *Philosophy and the Mind Sciences* 2. doi: https://doi.org/10.33735/phimisci.2021.87
Alex Lepauvre, Lucia Melloni

When first encountering the images produced by scanning a human brain using magnetic resonance imaging (MRI), one might assume that these machines can "see" through tissues and bones to observe the brain lying underneath. A better understanding of how an MRI scanner works reveals that this is not quite true. An MRI scanner operates by generating a strong magnetic field to align hydrogen atoms in the tissue to be imaged and then disrupting this alignment with radio pulses [1]. As the atoms realign, they release energy, the amount of which depends on the concentration of hydrogen atoms present in different tissues. These variations are then processed by computer algorithms to reconstruct an image. This process appears far removed from the intuitive and immediate notion of "seeing"; the MRI is using tricks.

Now, consider what happens when we are presented with a picture: photons emitted by a light source meet the image; some are absorbed while others bounce off, depending on the frequency of their oscillations. Photons that bounce off the image travel to our eyes and are absorbed by photoreceptors in the retina, resulting in electrical signals. These signals are processed and integrated along the visual system, to construct a representation of the outside world [2]. Both in the case of the MRI and the visual system, certain physical properties of the outside world are measured by specialized sensors, and the information they gather is integrated to construct a representation of what was measured.

This analogy illustrates that while it is true MRI measurements are very indirect, the same can be said for the human visual system. Yet despite the similarities, a key difference remains between the scanner and a human. For all the complex processing involved in constructing an image from the MRI scanner sensors' measurements, the scanner arguably does not "see" the tissue being imaged in the way we see the reconstructed image. In our case, somewhere along the way, we become conscious of this reconstructed image. There is something it is like for us to "see" something [3]; there is nothing it is like for the MRI to reconstruct the images from its sensors' measurements. This fundamental difference constitutes the very basis of our existence. Without it, there would be no reality to speak of—the color blue would not exist, nor would the smell of coffee in the morning.

Something special must be happening in the human brain to give rise to conscious experience, and since the time of Hippocrates, mankind has sought to understand what consciousness is, what it does, and how physical systems such as the brain —but not others like the heart—can instantiate it [4,5]. David Chalmers famously referred to the latter aspect as "The Hard Problem" [6]: why does matter, such as the

brain, give rise to perceptions and emotions that have subjective, phenomenological qualities? Solving this problem is not only a matter of philosophical intrigue; it has profound medical, societal, and moral implications [7–11].

Studying consciousness scientifically poses a unique challenge, due to the inherent subjective nature of the phenomenon [8], which historically placed it outside the scope of empirical science. However, this changed when Crick and Koch proposed a narrow and tractable framework to attack consciousness, suggesting to focus on identifying the neural mechanisms present during conscious states and absent during unconscious ones [12,13]. By pinpointing such mechanisms, known as the neural correlates of consciousness (NCC) [14], scientists hope to reveal the common denominator of these processes, to eventually formulate theories to answer the hard questions of consciousness [6].

For the past thirty years, this bottom-up agenda has dominated the field, producing a detailed cartography of brain areas and refined spatio-temporal patterns of brain activity thought to be associated with conscious experience. The question is then, are we closer to finding the mechanism(s) responsible for subjective, phenomenal qualities than we were thirty years ago? Empirical findings have been vastly inconsistent, with different studies proposing different spatiotemporal neural activation patterns to constitute the NCCs [15–19]. These inconsistencies relate to the controversies regarding which experimental conditions truly allow to capture consciousness, leading to debates over which of these findings revealed the true NCCs—and the conclusions differ depending on underlying theoretical commitments. Consequently, the supposedly theory neutral bottom-up approach has resulted the proliferation theories of consciousness, each shaped by the specific biases inherent to the experimental paradigms used to test them [20]. These theories are pursued in parallel and seldom converge or challenge one another, dismissing each other's evidence on methodological grounds. This situation highlights that this agenda, while helpful, has significant limitations in its ability to provide a unified scientific explanation of consciousness.

In my research, I propose that it is time to move beyond this bottom-up approach and instead focus on rigorous testing of existing theories of consciousness. By concentrating on finding dissociations between conscious experience and the mechanisms proposed by different theories, we can overcome the limitations associated with the NCC debates. In this thesis, I will demonstrate that this approach leads to the refinement of theoretical models and to a better characterization of consciousness itself, uncovering novel empirical avenues to progress toward a unified scientific account of consciousness.

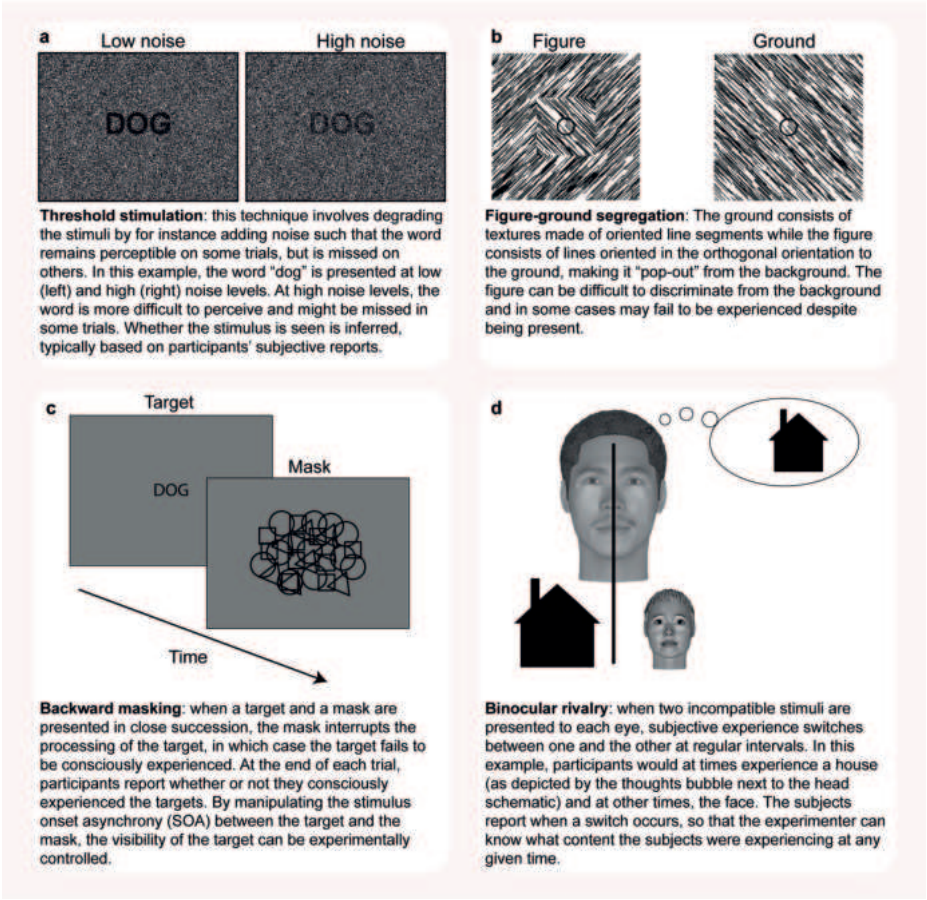
The contrastive method to identify the NCCs: a bottom-up agenda

The NCCs are formally defined as the minimal set of neural activities that are jointly sufficient to give rise to consciousness [14]. In any given conscious state—whether it is a general state such as wakefulness versus dreamless sleep, or the experience of specific content such as seeing a red ball or smelling coffee—a subset of the brain's ongoing neural activity is directly responsible for consciousness, while the rest is associated with unconscious processes. If we were to disrupt or prevent the neural activation responsible for the experience of the content, then the experience of that content would cease. In contrast, if neural activity associated with unconscious processes were disrupted without affecting the content-NCCs (as opposed to state-NCCs, which are linked to overall states of consciousness, such as being in dreamless sleep or fully awake and alert [14,18,19]), consciousness would remain unaffected. Identifying these mechanisms should reveal how they differ from those associated with unconscious processes and reveal what the neural underpinnings of consciousness are.

Isolating the NCC requires modulating the content of experience while keeping other parameters constant, such as the sensory input [12,13]. This is not an easy feat, as there is a tight association between what is presented to our sensorium and the content of our experience; we tend to see what is in front of our eyes. Over the years, psychophysicists have developed several methods to achieve a dissociation between sensory input and conscious experience, rendering the same stimuli either visible or invisible [21] (see box 1a-d). Under such conditions, as external factors are maintained constant, contrasting the neural activity between the “seen” and “unseen trials” removes any neural activation associated with unconscious sensory processing and should in principle yield the NCC (see Figure 1.1).

However, this raises another issue. If different conscious experiences can occur under matched sensory input, how can we know which content is experienced by a participant at a given moment? Subjective experiences are by definition private and cannot be measured from a third person’s perspective. Instead, experimenters must infer participants' experience indirectly, relying on overt or covert markers of conscious experience. The most common way to do so is to rely on participants' introspective reports of their subjective experiences. In other words, we can simply ask participants which stimulus they saw and which one they did not and compare brain activity based on these subjective reports.

In summary, isolating the content-NCCs requires relating subjective experience with the underlying neural activity, while keeping all other factors constant. Provided that the only difference between conditions is the perceptual states themselves, the isolated neural activity should be the one that is minimally sufficient for consciousness to arise, thus constituting a NCC (see Figure 1.1). This approach of comparing neural activity between conscious states is referred to as the contrastive method and according to some constitutes a gold standard in consciousness research [12,13,22].



Box 1.1. Example of experimental paradigms to manipulate conscious experience independently of sensory input

Box b is adapted from [23]

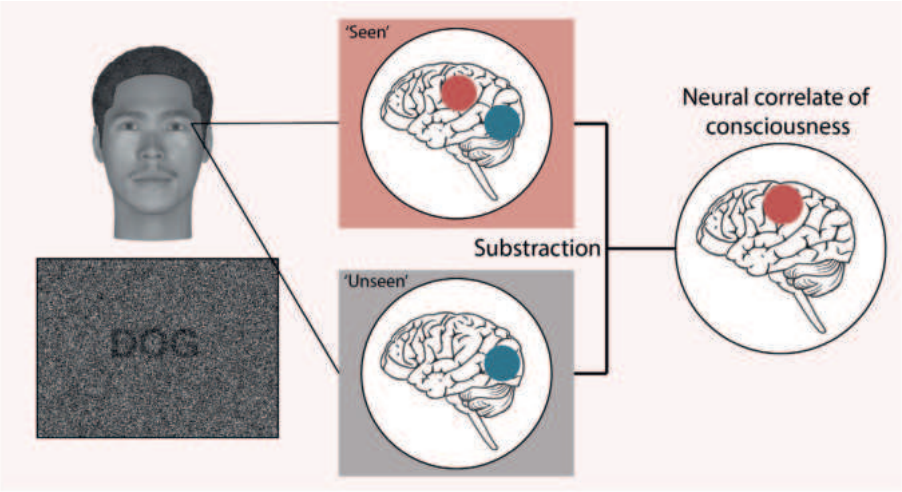


Figure 1.1. The contrastive method
By recording the neural activity of participants when presented with a sensory input that participants only sometimes see, the recorded activity can be sorted into "seen" and "unseen" groups. As the sensory input is the same across both conditions, sensory encoding should be present in both conditions (depicted in blue on the schematic brains), while some activity should be unique to the seen condition (depicted in red). When subtracting the neural activation between seen and unseen conditions, sensory encoding is removed and the neural activation that is unique to the conscious condition remains, i.e. the Neural Correlate of Consciousness (NCC).

Empirical attempts to isolate the NCCs

Equipped with experimental methodologies to contrast consciousness under matched sensory conditions, many studies have attempted to isolate the content NCC [17–20,24]. These efforts were fueled by both the development of new experimental methods as well as advances in human and animal brain recording and analysis technologies.

Unfortunately, no unique and consistent neural correlate of conscious vision was identified [17,18,20,25,26]. Some studies using various experimental paradigms (some described in box 1) and recording techniques such as electroencephalography (EEG), functional MRI (fMRI), and single neuron recordings in humans and animal models, showed that activation in sensory cortex start to differ between the seen and unseen conditions as early as 100ms after stimulus onset, suggesting that these activations constitute NCCs [27–34]. In contrast, other studies suggested that while activation in sensory cortices is necessary for consciousness to emerge, it is not sufficient; it cannot alone give rise to conscious experience. These studies found that cortical regions associated with visual processing were similarly activated by seen and unseen trials but that a fronto-parietal network was uniquely activated when participants reported seeing the stimulus later than 250 to 300ms from stimulus onset onward [35,35–41].

These discrepancies can be partly attributed to the methodological shortcomings of the contrastive method. The core assumption of the contrastive method is that when comparing between conditions in which the external input is controlled for and only perception varies, the resulting difference directly reflects neural activity involved in consciousness. However, this strategy is too simplistic. When comparing neural responses associated with two perceptual states, as the contrastive method does, two other families of internal processes co-occur with the ones directly reflecting consciousness: the NCC-precursors (NCC-pr) and the NCC-consequences (NCC-co) [42,43]. NCC-pr refers to processes that precede the NCC proper. They might enable a given stimulus to reach consciousness but are not conscious themselves. The NCC-co refer to processes that might follow the NCC proper. They result from consciousness but are not conscious either.

A typical example of NCC-pr is attention. While the role of attention in consciousness is still being debated [26,42,44,45], it is now widely agreed that they are indeed two different mechanisms. In classical contrastive paradigms, covert shifts of attention may explain why some stimuli are seen while others are not. For example, in masking paradigms, differential engagement of attention across trials might well explain why an otherwise identical physical stimulus is perceived in some trials but missed in others. The NCC-co are processes triggered by conscious experience but not responsible for it. Some theories of consciousness assign consciousness a function. Therefore, the fact that a critical stimulus was perceived entails that additional processes will follow it. Such processes include encoding in working memory [46] and/or episodic memory, reflecting about the perceived stimulus, and in the case of most experimental paradigms employing the contrastive method, reporting about it [47]. Therefore, the NCCs discovered across studies may have been inflated by these two types of processes.

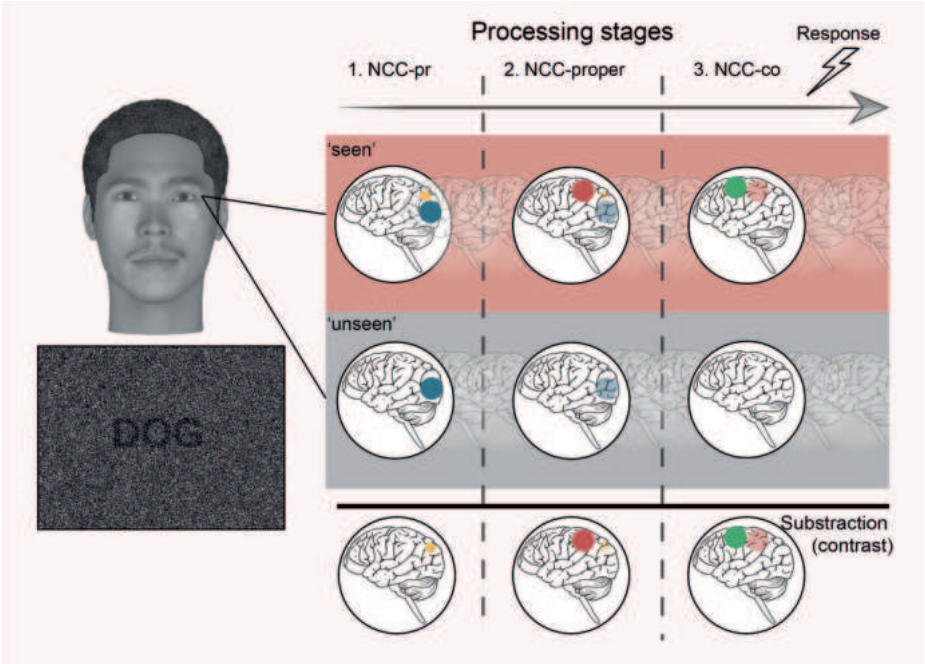


Figure 1.2 The three NCCs problem

The contrastive method was argued to be insufficient to isolate the so-called NCC-proper (2). The results of the subtraction between seen and unseen conditions are confounded by the occurrence of temporally flanked mechanisms, termed NCC-pr (1) and NCC-co (3) (see [48,49]). The former refers to the mechanisms enabling a stimulus to reach consciousness without directly accounting for consciousness itself (yellow in the figure), while the latter refers to mechanisms that are consequences of the conscious perception of a stimulus but need not be conscious themselves (red in the plot). Examples of NCC-pr are attention and prior expectations [48,50,51]. As most theories of consciousness assign a function to consciousness, typical NCC-co accompanies it: encoding in memory, decision making to name a few. Additionally, as many contrastive studies rely on reports to differentiate conscious versus non-conscious conditions, motor responses and planning, as well as self-monitoring, often confound the NCC-proper [47].

Following the acknowledgment of the 3-NCC problem, report and the associated task-relevance confounds have received the most attention. No-report studies were designed in which the content of consciousness is inferred without overt-report [47]. This can be achieved by relying on eye movements to classify trials as seen or unseen, or on-task instructions to manipulate the visibility of stimuli. Several no-report studies suggest that late activation over the fronto-parietal relates to report-related cognitive processes rather than consciousness, as such effects are only present when seen stimuli have to be reported, not when seen stimuli do not have to be reported [52–58]. These studies further show that mid-latency activation over posterior brain regions seems to track closely conscious perception, both when participants report and do not report the content of their experience.

While these studies seem to relegate late, frontal activation to post-perceptual processes only, several studies have highlighted that even though frontal activation is reduced in the absence of a report, it is not absent [54,59–63]. Thus, while some of the previously observed prefrontal activation might index confounding factors such as report, leading to an exacerbated prefrontal cortex activation, a selective portion might still correspond to the NCC proper and the question as to whether the front or the back of the brain constitutes the best candidate NCC remains open [64–67].

The limitations of the bottom-up approach

Despite significant advancements in the empirical investigation of consciousness over the past 30 years, the bottom-up approach has failed to identify a unified NCC. While no-report paradigms can help eliminate post-perceptual processes, not all confounds are related to report. For example, prior expectations also influence conscious perception of stimuli and modulate associated neural responses and might confound the search for the NCC [50,51,68–70]. In paradigms with passive viewing, such as in the experiment by Cohen et al. [52], encoding in working memory occurs in the seen but not in the unseen condition, also confounding the empirical findings. Thus, there are a number of confounds, some already known: report, prior expectations, working memory, episodic memory, task demands, attention, decision making, and neuromodulatory states. This list is not exhaustive, and as our understanding of the processes related to consciousness is improved, this list is likely to be expanded. Aru and colleagues [48] have proposed that isolating the NCC proper will require controlling for all these potential confounds—a task that could prove to be a century-long enterprise.

Though time-consuming, this perspective seems pragmatic: by carefully designing experiments accounting for all possible confounds, the NCC proper will eventually be identified. However, I argue that this approach overlooks a more fundamental limitation of the bottom-up agenda related to the measurement problem [71,72]. A perfect experiment (or series of experiments) to isolate the NCC would require two components: first, a valid measure of consciousness that accurately and reliably reflects a person's subjective experience; second, experimental designs that manipulates consciousness independently of all other cognitive processes [48,49].

However, we cannot achieve the first requirement, because consciousness can never be measured directly due to its subjective nature. Instead, we must rely on proxies–such as subjective reports, behavioral measurements, or physiological indicators–

that are assumed to track consciousness. Testing the validity of such proxies would require measuring them alongside the phenomenon they are supposed to capture to test how tightly they are associated. As subjective experience can never be directly measured, we can never establish with certainty that these proxies are tracking consciousness rather than an associated phenomenon [72].

In other words, we can never establish with certainty that the proxy we have chosen accurately measures participant's experience. Because we can never be sure that our proxy is valid, we also cannot be sure that our experimental manipulations are affecting consciousness itself rather than other cognitive processes that our measure might inadvertently be tracking. Even if we meticulously design experiments to control for known confounds, we have no definitive way to confirm that consciousness is manipulated independently of those confounds and that our results reflect the NCC proper. Researchers may consider different proxies as the most appropriate measures of consciousness, but without the ability to validate these proxies, there is no objective way to determine whose measure is most accurate [73–75]. Thus, the lack of convergence in empirical findings may not reflect the complexity of consciousness itself, but rather the fact that different paradigms and their associated measures are tracking non-overlapping concepts [73,76,77].

Based on the findings of experiments aiming to isolate the NCCs, theories of consciousness were formulated. However, the empirical basis on which each of these research programs were developed depends on unverified commitments regarding which experimental conditions are adequate to isolate the NCCs. Different programs may therefore have studied different phenomena labeled by each theory as "consciousness" [73]. This fragmentation creates a situation where findings from one research program may not be comparable to those from another, allowing theories to progress in parallel without ever converging or challenging each other and resulting in a proliferation of theories [78–81]. This view has recently been substantiated in a study by Yaron and colleagues [15] who further exposed a strong validation bias in the field. Their extensive literature review showed that the outcome of experiments is heavily dependent on the chosen set of experimental parameters and that the alignment of empirical results with theoretical framework was could be accurately predicted from the experimental parameters alone.

Ultimately, we cannot determine which research program, if any, is truly tracking consciousness as opposed to related phenomena. Because we cannot establish which paradigm and measures track consciousness, there is no convincing proponents of different theories that they are studying different phenomena. Designing

increasingly sophisticated experiments will not resolve the discord in the field, as researchers will likely continue exploring their own version of 'consciousness', disregarding the foundational assumptions of other programs, making convergence or meaningful challenges between research programs unlikely. Combined, these limitations indicate that the bottom-up approach, while providing valuable data on neural processes correlated with consciousness, is insufficient for developing a unified theory of consciousness.

Dissociating experience from its hypothesized mechanisms

Acknowledging the limitations of the traditional bottom-up approach, I have instead applied a top-down strategy in my research. An abundance of theories propose different and incompatible neural underpinnings of conscious experience [22,79,81,82]. Provided that consciousness is a unified phenomenon, these theories cannot all be true at the same time [78]. I therefore aimed to test existing theories of consciousness rigorously, to challenge and refine them.

Testing the necessity of proposed mechanisms to escape the validation bias

At first, adopting a top-down approach may not seem to resolve the issues that I have described. There are many different ways to operationalize consciousness and we cannot a priori know which is the most appropriate. If we want to test a given theory, we need to adopt an operationalization that is compatible with the theory being tested. Otherwise, evidence can be discarded by arguing that the experimental conditions did not allow to measure consciousness appropriately. However, this introduces a bias, as adopting the operationalization of a given theory can skew findings toward what the theory predicts [20]. This raises a critical question: How can we test theories of consciousness if the operationalization we have to adopt is biased in favor of the theory being tested?

Before answering this question, I need to introduce what testing a theory of consciousness entails and how it can be achieved. Falsifying a theory of consciousness (or at least its predictions) requires finding a dissociation between the content of consciousness inferred by a theory and the neural mechanisms proposed to instantiate it by that same theory [83].

One way to test a theory of consciousness is through the problematic contrastive method: if the mechanism proposed by a theory appears in both the seen and unseen conditions, then it cannot be sufficient to give rise to consciousness, thus

challenging the theory. There is however another way in which a theory can be falsified. As highlighted by Chalmers [14], a NCC need not to be necessary for conscious experience, as there may be several neural correlates of a conscious state. Theories of consciousness however aim to provide a mechanistic explanation of the target phenomenon they claim to explain (consciousness). If the proposed mechanisms is present without the target phenomenon, the proposed mechanism is not sufficient to give rise to that target phenomenon. If the target phenomenon occurs without the proposed mechanism, then that mechanism is not necessary. Either cases constitute a dissociation between the proposed mechanism and the target phenomenon, challenging the theory.

Accordingly, we can test a theory of consciousness by seeking conditions in which consciousness (as defined by the theory being tested) occurs without the mechanism(s) proposed to give rise to it, showing that the proposed mechanism(s) is not necessary for consciousness. I will refer to this as the necessity dissociation approach. Unlike the contrastive method, this approach does not require an unconscious condition, as it suffices to find one conscious condition where the proposed mechanism fails to occur. This allows theories to be tested under a broader range of conditions, reducing the validation bias by enabling to test theories regarding novel aspects of conscious experience and push the field forward [20].

Going beyond seen and unseen contrast by investigating the temporal dynamics of conscious experience

One aspect of consciousness that has remained underexplored due to the limitation of the contrastive method is the temporal dynamics of conscious experience. Most studies have focused on the entry of content into awareness [16] by relying on brief stimulus presentation, typically under 500ms [23,28,32,35,36,41,55,56,58,84], as it is difficult to render sustained stimuli unconscious. One exception is binocular rivalry (and binocular flash suppression [85]) where participants' experience typically oscillates between the percept of one or the other eyes with the interval between reported switches typically of the order of a couple of seconds [29,34,37,60,86]. Surprisingly, these studies focused primarily on identifying the brain regions that are selectively activated at the moment of the switch, without further investigation of the neural mechanisms associated with the persistence of contents in consciousness [29,30,61]. This approach leaves significant gaps in our understanding, particularly regarding how conscious experience unfolds and persists over time.

In my thesis, I address this gap by applying the necessity dissociation approach to investigate the temporal dynamics of conscious experience. Specifically, I investigated

whether the mechanisms proposed by theories of consciousness can account not only for the onset of experience but also for its persistence over longer durations. In the two empirical chapters of my thesis (**Chapter 2** and **Chapter 3**), I relied on a simple experimental paradigm in which highly visible stimuli of different categories were presented for three distinct durations (0.5, 1.0, and 1.5s).

As we will see, under these conditions, theories of consciousness can a priori infer the expected temporal dynamics of conscious experience and, in turn, predict the corresponding neural activations that should give rise to them. If the predicted neural activation is not observed, the theory is challenged, in line with the necessity dissociation approach. Crucially, this method allows us to test theories without relying on an unconscious condition, which would be difficult to achieve when manipulating stimuli duration.

Obtaining theories predictions and testing them

Given the abundance of theories of consciousness, it was not feasible to test them all in my research. Therefore, my efforts focused on two prominent theories in the field: the integrated information theory (IIT) [87–92] and the global neuronal workspace theory (GNWT) [25,93,94]; both of them are described in box 2. Both theories offer distinct and influential explanations of consciousness [20,79,81], and testing them is highly relevant for the field.

Global neuronal workspace theory (GNWT)

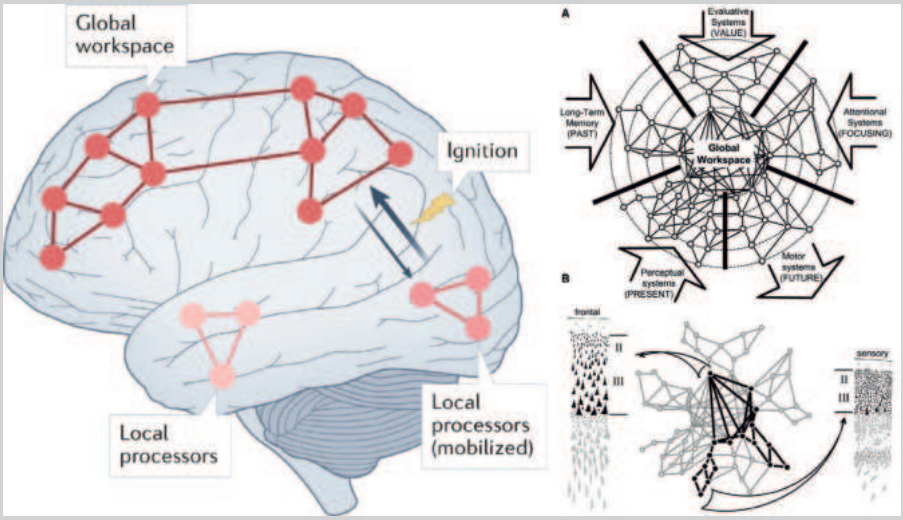


Figure adapted from [81,94]

The global neuronal workspace theory (GNWT) has received the most empirical attention over the past 30 years [20]. This theory specifies the neural implementation of the previously formulated global workspace theory, which was only defined in cognitive terms [12]. According to GNWT consciousness is the result of the broadcast of information through a fronto-parietal network of interconnected local processors to engage cognitive processes such as evaluative systems and working memory. The clearly stated explanatory target is access consciousness [25,93,94]. GNWT assumes a functional role of conscious experience: rendering information available to many cognitive systems enables flexible processing and behavior that would not be possible under automatic, unconscious processing. This theory is deeply rooted in empirical studies highlighting the differential activation of the prefrontal cortex between seen and unseen stimuli [35,36,41]. Within this framework, a marker of conscious experience is ignition (a non-linear, all or none increase) in the PFC. According to GNWT, the PFC plays a critical role in conscious perception, mediating conscious processing depending on a combination of signal intensity, attentional gain, and the current state of the workspace.

Integrated information theory (IIT)

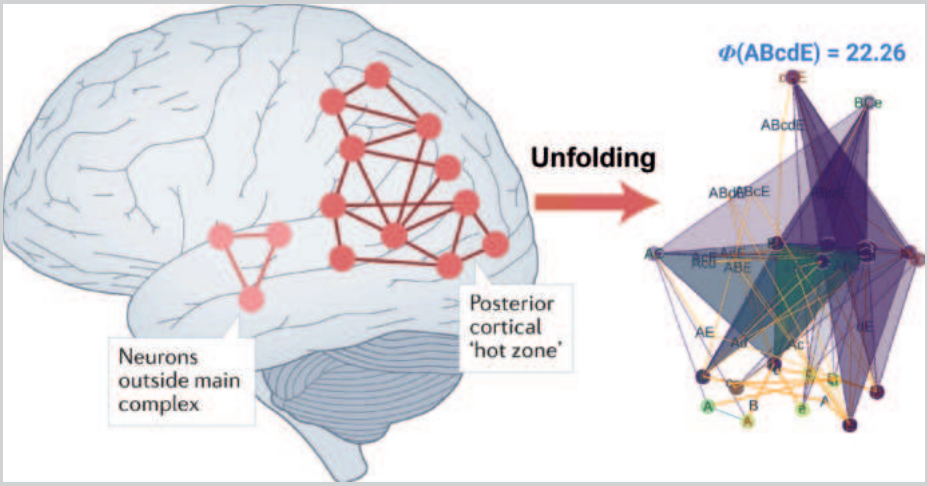


Figure adapted from [81,87]

In contrast to GNWT (and other theories of consciousness), the integrated information theory IIT does not start from the brain but instead from phenomenology, by defining five axioms that are considered to be true to any conscious experience [87–92]. IIT states that all experiences exist intrinsically, are structured, specific, unitary, and definite [92]. Based on these axioms, physical properties were derived in mathematical terms (postulates) that a system must satisfy to instantiate these axioms. From this approach, they conclude that consciousness is identical to the cause-effect structure specified by a partition of a system (called a complex in IIT) which has the maximum integrated cause-effect power [87]. In other words, for any system in which units have causal power over each other (such as the brain), partitions of that system have integrated cause-effect power (ϕ) if that subsystem has causal power over itself. The complex has borders: units are part of the system (or complex) if removing them decreases the amount of ϕ while other units are not part of it if removing them does not change ϕ. There can be many such complexes in the brain at any given time but the one with the highest ϕ is the conscious one. The experience itself *is* the unfolded cause-effect structure of this particular complex; its particular phenomenological properties are given by the form of that structure.

Due to its phenomenology-first approach, early empirical attempts at isolating the NCCs were not instrumental in the formulation of IIT. As such, these studies do not lend any direct support or challenge to the theory. Nonetheless, it is possible to derive predictions as to which brain regions are most propitious to allow for the highest levels of information integration in the brain. According to IIT, the neuro-

architecture of the so-called posterior hot zone spread across parietal, temporal, and occipital areas is most favorable for high ϕ due to its organization in a 'pyramid-of-grids'-like structure and the theory predicts that the NCCs should be found within this zone [87]. As such, IIT can accommodate the findings associating conscious experience with activation in sensory areas and as such is considered part of the 'back of the brain' camp in the ongoing debates.

Box 1.2: Theories of consciousness of interest for the thesis

For the necessity dissociation approach to work, two conditions must be met. First, theories of consciousness must state a priori their inference regarding the content of consciousness under the experimental conditions. This means that theories should explicitly define what they believe participants will consciously experience in those conditions, based on theoretical considerations alone (e.g., all attended stimuli should be experienced) or measurements (introspection, behavior, physiological measures). Second, the observation that is predicted by a given theory must accurately reflect the theory. Both conditions are difficult to meet when testing theories under novel conditions, as theories of consciousness are poorly defined beyond the confines of the restricted set of contrastive method parameters that they have adopted. Failure to meet either of these conditions will undermine the whole approach. If the criterion used to infer whether consciousness is present does not align with the theory, evidence can be discarded by arguing that the stimulus may not have been consciously experienced. If the predictions are not accurately related to the framework being tested, it can simply be argued that the predictions were wrong and that the theory is not challenged.

However, if both conditions are met and the predictions of a theory are falsified, the theory itself must be updated by either changing its mechanistic account for conscious experience to accommodate negative findings or by updating its assumptions as to which are the necessary conditions for a stimulus to be consciously experienced, which both constitutes a refinement of the theory. It is important to highlight that under this approach, a negative finding bears more significance than a positive one: validating a prediction does not necessarily imply that a theory as a whole is correct, but falsifying a prediction implies that there is something wrong with it (I will extend on this point in the discussion).

To ensure that both conditions are met, I have actively collaborated with key proponents of two previously described theories of consciousness, IIT and GNWT. In the work that I will present, theorists themselves were required to formulate their

predictions regarding experimental paradigms that were later conducted. Critically, these predictions were pre-registered before observing any data [95–97]. This procedure limits hindsight bias and makes explicit which results were truly predicted and which had to be accommodated after the fact.

Outline

In my thesis, I will present the results of studies that go beyond the traditional bottom-up approach by rigorously evaluating theories of consciousness, investigating whether the mechanisms they propose are necessary for conscious experience. I investigated two theories, IIT and GNWT, in the context of sustained visual stimuli presentation. The experiments were designed to bring about novel predictions of the theories regarding an aspect of experience that they ought to be able to explain. Their failure to do so would imply that they need further refinements.

In **Chapter 2,** I will present the results of a large-scale adversarial collaboration between IIT and GNWT. This procedure consists of resolving debates between disagreeing scholars by testing their theory in a joint empirical effort [98] and it has been argued to constitute a gold standard to settle scientific disputes [99,100]. In the case of consciousness research, this approach is particularly valuable, as it compels theories to agree on a common set of experimental conditions that all parties accept, preventing them from dismissing evidence due to disagreements over the measurement of consciousness. Furthermore, this process pushes theories to venture beyond the biased set of experimental conditions under which they are typically tested, leading to novel predictions that can be empirically tested. These novel predictions were tested on a large multi-modal dataset (intracranial EEG, fMRI, and magneto-encephalography (MEG)) and significant challenges to both theories were revealed.

In **Chapter 3,** I will present the results of a study investigating the temporal dynamics of conscious experience under similar presentation conditions as that of Chapter 2. In light of the evidence contradicting GNWT prediction in Chapter 3, the theory refined its inference regarding the temporal dynamics of conscious experience, leading to novel predictions regarding behavioral observations expected in such conditions. These predictions were tested and revealed that the temporal dynamics of conscious access may in fact become dissociated from sensory input, challenging initial assumptions regarding the temporal dynamics of conscious experience.

In **Chapter 4,** I will provide an independent analysis of the results presented in Chapter 2 and discuss the broader implications of the results of both studies from a vision neuroscience and consciousness research perspective. I will explore how the findings contribute to our understanding of the neural dynamics underlying visual perception of persistent stimuli and propose that investigating the temporal dynamics of conscious experience may provide a way forward to dissociate access from phenomenal consciousness. Beyond the empirical results, I will engage in a critical discussion about the value of theory testing in consciousness research. I will emphasize that we should refrain from adopting extreme and naive falsificationist views on theory testing and instead consider theory testing as a tool for theoretical self-improvement. I will finish by providing recommendations for future top-down efforts in consciousness research to push us closer to a comprehensive understanding of consciousness through iterative refinement of theories of consciousness.

1

Chapter 2

An adversarial collaboration to critically evalutate theories of consciousness

This chapter has been published as:
An adversarial collaboration to critically evaluate theories of consciousness, *BioRxiv*.
doi: https://doi.org/10.1101/2023.06.23.546249

Cogitate Consortium, Oscar Ferrante[1], Urszula Gorska-Klimowska[1], Simon Henin[1], Rony Hirschhorn[1], Aya Khalaf[1], Alex Lepauvre[1], Ling Liu[1], David Richter[1], Yamil Vidal[1], Niccolò Bonacchi, Tanya Brown, Praveen Sripad, Marcelo Armendariz, Katarina Bendtz, Tara Ghafari, Dorottya Hetenyi, Jay Jeschke, Csaba Kozma, David R Mazumder, Stephanie Montenegro, Alia Seedat, Abdelrahman Sharafeldin, Shujun Yang, Sylvain Baillet, David J Chalmers, Radoslaw M Cichy, Francis Fallon, Theofanis I Panagiotaropoulos, Hal Blumenfeld, Floris P de Lange, Sasha Devore, Ole Jensen, Gabriel Kreiman, Huan Luo, Melanie Boly, Stanislas Dehaene, Christof Koch, Giulio Tononi, Michael Pitts, Liad Mudrik, Lucia Melloni;

[1]Shared first authorship. My contribution to this project entails Conceptualization, Data Curation, Data Quality, Formal analysis, Investigation, Methodology, Project Administration, Software, Validation, Visualization, Writing of the original draft, review and editing as defined by the credit taxonomy (https://credit.niso.org/)

Specifically, I contributed to the development of the experimental design (piloting); contributed to the deployment of experiments across sites; contributed to the data architecture of the platform for data sharing across sites for analysis; coordinated the collection of iEEG data across three data collection sites; conducted the iEEG data analysis of the project (including curation, validation, preprocessing, onset responsiveness, category selectivity, decoding analysis, duration analysis, representation similarity analysis) and contributed to data analysis of other recording modalities (principally MEG)

Summary

Different theories explain how subjective experience arises from brain activity [79–81]. These theories have independently accrued evidence, but confirmation bias and dependence on design choices hamper progress in the field [20]. Here, we present an open science adversarial collaboration which directly juxtaposes Integrated Information Theory (IIT) [87,92] and Global Neuronal Workspace Theory (GNWT) [93,94,101–103], employing a theory-neutral consortium approach [78,97,104]. We investigate neural correlates of the content and duration of visual experience. The theory proponents and the consortium developed and preregistered the experimental design, divergent predictions, expected outcomes, and interpretation thereof [97]. 256 human subjects viewed suprathreshold stimuli for variable durations while neural activity was measured with functional magnetic resonance imaging, magnetoencephalography, and intracranial electroencephalography. We find information about conscious content in visual, ventro-temporal and inferior frontal cortex, with sustained responses in occipital and lateral temporal cortex reflecting stimulus duration, and content-specific synchronization between frontal and early visual areas. These results align with some predictions of IIT and GNWT, while substantially challenging key tenets of both theories. For IIT, a lack of sustained synchronization within posterior cortex contradicts the claim that network connectivity specifies consciousness. GNWT is challenged by the general lack of ignition at stimulus offset and limited representation of certain conscious dimensions in prefrontal cortex. These challenges extend to some first-order and higher-order theories of consciousness that share some of the predictions tested here [105–108]. Beyond challenging the theories, we present an alternative approach to advance cognitive neuroscience through a principled, theory-driven, collaborative effort. We highlight the challenges to change people's mind [109] and the need for a quantitative framework integrating evidence for systematic theory testing and building.

Main

Philosophers and scientists have sought to explain the subjective nature of consciousness (e.g., the feeling of pain or of seeing a colorful rainbow) and how it relates to physical processes in the brain [13,14]. This ongoing endeavor has led to a number of theories of consciousness that have evolved in parallel [20,79,81]. Those theories offer incompatible accounts of the neural basis of consciousness [79,81]. Empirical support for a given theory is often highly dependent upon methodological choices, pointing towards a confirmation bias when testing these theories [20]. Convergence

upon a broadly accepted neuroscientific theory of consciousness will have profound medical, societal, and ethical implications.

With this goal as a starting point, we make a concerted effort to test two theories of consciousness, among several widely discussed ones [81], through a large-scale, open science, adversarial collaboration [78,97,98,110,111] aimed at accelerating progress in consciousness research by building upon constructive disagreement. This collaboration brings together proponents of Integrated Information Theory (IIT) [87,92] and Global Neuronal Workspace Theory (GNWT) [25,101], in addition to theory neutral researchers. Together, we identified divergent predictions of the theories and jointly developed an experimental design to test them (Figure 2.1a). We preregistered foundational and novel predictions from the two theories, including pass/fail criteria for each prediction, as well as expected outcomes and their interpretation ex-ante [78,97]. We focus on GNWT and IIT, two theories of consciousness out of several others widely discussed e.g., Recurrent processing theory and Higher-order theories [79,81], since these theories feature prominently in the field of consciousness science as shown in a recent systematic review of the literature [20].

IIT and GNWT explain consciousness differently: IIT proposes that consciousness is the intrinsic ability of a neuronal network to influence itself, as determined by the amount of maximally irreducible integrated information (phi) supported by a network. According to proponents, theoretical and neuroanatomical considerations suggest that a complex of maximum phi likely resides primarily in the posterior cerebral cortex, in a temporo-parietal-occipital "hot zone" [18,87,92,112]. GNWT instead posits that consciousness arises from global broadcasting and late amplification (or "ignition") of information across interconnected networks of higher-order sensory, parietal, and especially prefrontal cortex (PFC) [25,101,102].

IIT and GNWT both have a mathematical or computational core (concerning integrated information and the global workspace respectively) and a proposed biological implementation (primarily in posterior cortex vs. in prefrontal cortex and associated areas respectively). It is difficult to test the mathematical or computational core of these theories directly, so in this project we instead test their proposed biological implementations. The two proposed biological implementations are competing and incompatible proposals, and testing them is empirically tractable with current methods, enabling scientific progress. In the case of GNWT, we focus especially on PFC rather than the associated areas in higher-order sensory and parietal cortex, because this is where GNWT and IIT pose the most incompatible and hence maximally diagnostic predictions, enabling differential testing of the

theories. One consequence of this biological focus is that theorists could respond to challenging data by retaining the mathematical/computational core of a theory and changing the proposed biological implementation. Another consequence is that some predictions (and the associated consequences from testing these predictions) may be shared by other theories of consciousness with a similar proposed biological implementation, such as higher-order theories [107,108] implemented in prefrontal cortex, and local recurrency theories [23,105] implemented in visual cortex. These are natural features of a project designed to test theoretical proposals about the neural mechanisms of consciousness. For an extensive explanation and rationale, refer to our preregistration document (https://osf.io/92tbg/).

We scrutinize three preregistered, peer-reviewed predictions of IIT and GNWT for how the brain enables conscious experience [97]: **Prediction #1** pertains to which cortical areas hold information about different aspects of conscious content. IIT predicts that conscious content is maximal in posterior brain areas, while GNWT predicts a necessary role for PFC. **Prediction #2** pertains to how conscious percepts are maintained over time [113–115]: IIT predicts that conscious content is actively maintained by neural activity in the posterior 'hot zone' (PHZ) throughout the duration of a conscious experience. GNWT predicts, instead, that an ignition in PFC at stimulus onset, and at offset, updates the workspace, with activity-silent maintenance of information in between [116]. **Prediction #3** pertains to interareal connectivity between cortical regions during conscious perception. IIT predicts short-range connectivity within posterior cortex, including lower-level sensory (V1/V2) and high-level category-selective areas (e.g., fusiform face area, lateral occipital cortex). In contrast, GNWT predicts long-range connectivity between high-level category-selective areas and PFC. The combination of predictions places a high bar for either theory to pass considering the highly powered and multimodal studies we conducted. Predictions received differential weighting with respect to challenging the theories based on the centrality to the theory and methodological limitations (Extended Table 2.1). In addition to testing specific predictions of the theories, we also used this rich dataset for an exploratory analysis aimed at delineating cortical areas potentially participating in consciousness after excluding confounding factors related to cognitive/task-related processes (*putative Neural Correlates Consciousness (NCC) analysis in the supplementary section*).

To empirically test these predictions, we investigated the content and temporal extent of conscious visual experiences that are phenomenologically multifaceted and rich, even for a single stimulus. For example, when viewing the Mona Lisa (Figure 2.1b), one experiences it as located in a portion of visual space, having a specific identity, a

specific orientation, and for as long as one looks at the painting. To approximate this multifaceted aspect of consciousness, we manipulated several attributes of conscious content. Specifically, we presented suprathreshold stimuli belonging to four different categories (faces, objects, letters, false fonts), with each category containing twenty individual identities presented in three different orientations (front, left, right view) for three different durations (0.5, 1.0, 1.5 s) (Figure 2.1c). Participants viewed the stimuli while searching for two infrequent targets, making some stimuli task relevant and others task irrelevant (Figure 2.1d; See supplementary video depicting the task). This paradigm offers several advantages: first, it provides robust conditions to test the theories' predictions as it focuses on clearly experienced conscious content, studied through high signal-to-noise, suprathreshold, fully attended single stimulus at fixation, making any challenges of the theories' predictions more significant, thereby aligning with Lakatos's sophisticated falsification approach [117]. Second, it minimizes task and report confounds, thereby isolating neural activity specifically related to consciousness. Third, it allowed us to test novel predictions about questions previously unaddressed by the theories, contributing to theory refinement and advancing the field more broadly. For example, it diverges from the usual testing grounds of these theories to explore new predictions about how experience is maintained over time, thereby yielding more informative results. Additionally, this adversarial collaboration has prompted more specific predictions for existing claims, particularly regarding specific regions of interest, enhancing the detail of these theories in the process.

All research was conducted by theory-neutral teams to minimize confirmatory bias. We evaluated the theories' predictions in 256 subjects who performed the same behavioral task in three different neuroimaging modalities: functional magnetic resonance imaging (fMRI, *N*=120), magnetoencephalography (MEG, *N*=102), and intracranial electroencephalography (iEEG, *N*=34). Given the limitations of current methods for measuring and recording human brain activity, such as varying strengths in spatial or temporal resolution, we intentionally employed a combination of techniques to mitigate these shortcomings. The integration of whole-brain, non-invasive fMRI and MEG with invasive iEEG recordings maximizes sensitivity, spatiotemporal resolution, and spatial coverage, thereby providing stringent and comprehensive tests of the theories in humans. This approach, combined with the use of large sample sizes, reduces the likelihood that negative results are due to methodological or sensitivity issues. The selection of methods was pre-approved by the adversaries before the study was conducted and results were known, ensuring the entire protocol was deemed suitable for assessing their theories. Furthermore, each data type was collected by two (or three) independent laboratories to ensure

generalization across populations, recording systems, and experimenters. Altogether, we aimed at fostering informativeness, reproducibility, and robustness of the results by (1) dissociating theory leaders from researchers involved in data acquisition/analysis to minimize biases and post hoc interpretation, (2) using a multimodal approach to test theories with enough and adequate temporal and spatial precision in humans, (3) acquiring data in a large sample of subjects to increase statistical power, (4) using standardized [118] and preregistered protocols [97] to evaluate theories under the same experimental framework and further minimize confirmatory bias [110], and finally (5) combining an analysis optimization phase with a final testing phase using independent parts of our dataset to corroborate the robustness of the results [119]. Consequently, we present a large-scale international effort to evaluate two widely discussed theories of consciousness under an integrated, rigorous and comprehensive adversarial collaboration framework, setting a precedent for theory testing and proving the concept for an alternative scientific model aimed at reducing bias and enhancing scientific rigor in the bio-medical sciences.

We first established that our task manipulations were effective and comparable behaviorally across data modalities and experimental sites (see supplementary section 1-2 for the full set of results). Subjects' performance in the task was excellent, with high hit rates (M=96.84%, SD=4.19%), low false alarm rates (M=1.45%, SD=4.30%), and high fixation stability (mean accuracy <2°=89.62%, SD=10.61%; Figure 2.1e-g). Subjects' performance across laboratories within each data modality was similar (all p=1.000 after multiple comparison correction, BF<0.12). Epilepsy patients showed slightly lower behavioral performance compared to neurotypical subjects, yet, behavior was still comparatively high (hit rate 93.90%, SD=12.29; false alarm rate M=4.25%, SD=20.17). We confirmed that subjects were conscious of the stimuli both in the task relevant and irrelevant trials in a separate experiment which included a surprise memory test (see supplementary section 3).

As part of our testing framework, after excluding a limited number of subjects due to data quality checks, we conducted an initial optimization phase on 1/3 of the MEG (N=32) and fMRI (N=35) datasets to evaluate data quality across sites and to optimize analysis pipelines. Following the optimization phase, pipelines were preregistered (https://osf.io/92tbg/), and applied to the novel datasets containing twice as much data (MEG, N=65 and fMRI, N=73). In what follows we report results obtained on the novel, previously unexamined datasets (see methods for the strategy used for iEEG and text for numbers of subjects that entered in each analysis). Results from the optimization phase and preregistered replication phase were subsequently compared and deemed to be largely compatible, with some minor exceptions (see supplementary section 4).

2

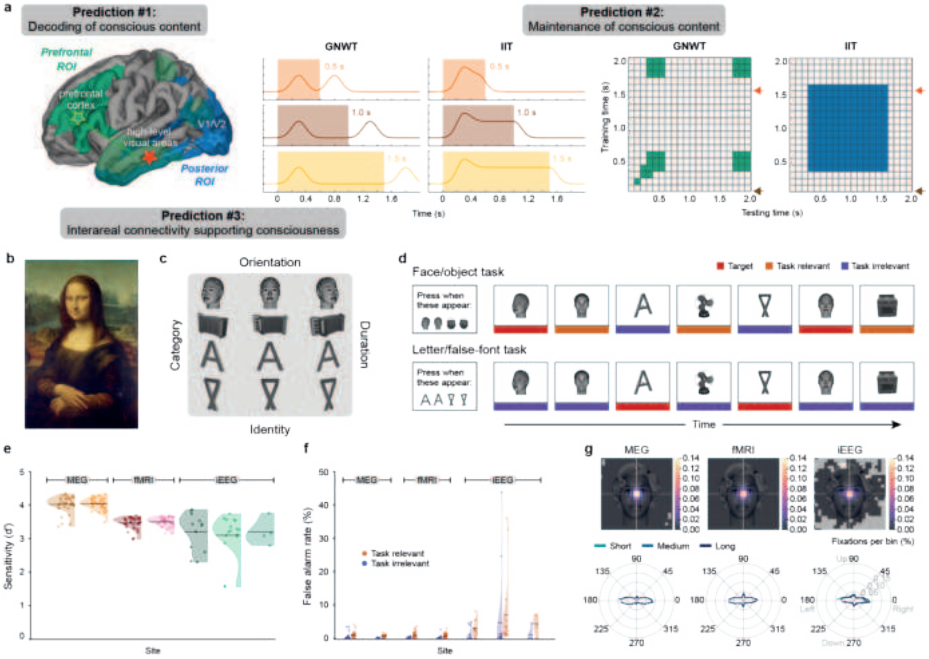

a
Prediction #1:
Decoding of conscious content
Prefrontal ROI
prefrontal cortex
high-level visual areas
V1/V2
Posterior ROI
Prediction #2:
Maintenance of conscious content
GNWT
IIT
0.5 s
1.0 s
1.5 s
Time (s)
Training time (s)
Testing time (s)
Prediction #3:
Interareal connectivity supporting consciousness
b
c
Orientation
Category
Duration
Identity
d
Face/object task
Target
Task relevant
Task irrelevant
Press when these appear:
Letter/false-font task
Time
e
MEG
fMRI
iEEG
Sensitivity (d')
Site
f
False alarm rate (%)
Task relevant
Task irrelevant
g
MEG
fMRI
iEEG
Short
Medium
Long
Fixations per bin (%)

Figure 2.1: GNWT and IIT predictions tested in an adversarial collaboration

a. *Three key contrasting predictions of Integrated Information Theory (IIT) and Global Neuronal Workspace Theory (GNWT) tested in an adversarial collaboration framework.* ***Prediction #1:*** *Decoding of conscious content, evaluating which cortical areas hold information about different aspects of conscious content. IIT predicts that conscious content is maximal in posterior brain areas, while GNWT predicts a necessary role for PFC.* ***Prediction #2:*** *Maintenance of conscious content over time, evaluating the temporal dynamics by which the temporal extent of the conscious content is instantiated. IIT predicts that conscious content is actively maintained in posterior cortex throughout the extent of a conscious experience; while GNWT predicts brief content-specific ignition in PFC ~0.3-0.5 s after stimulus onset and offset (when the workspace is updated), with content stored in a non-conscious silent state resembling activity-silent working memory in between. Waveforms and temporal generalization matrices depict the amplitude- and information-based temporal profiles predicted by the theories, respectively (left: colored rectangles indicate the three different stimulus durations, GNWT predicted waveforms pertain to PFC, IIT predictions to posterior cortex; right: brown arrows indicate stimulus onset, red arrows stimulus offset, green and blue colors reflect the predicted patterns of temporal generalization of conscious content according to each theory in PFC and posterior cortex for GNWT and IIT, respectively).* ***Prediction #3:*** *Interareal communication, evaluating the topological and temporal patterns of interareal connectivity subserving consciousness. The stars and arrows on the brain (left) depict the different predictions about the expected synchrony patterns (green: GNWT; blue: IIT).*

b. *Conscious experience is multifaceted in content. Looking at the image of Mona Lisa by Leonardo da Vinci underscores the fact that conscious experiences are rich: The painting is experienced as occupying a location in space, pertaining to a given category (i.e., a face and not an object, or any other category), specifying an identify (i.e., Mona Lisa and not any other face), and a particular orientation (i.e., leftward oriented and not rightward or any other orientation). Moreover, the conscious experience is maintained over time for as long as one appreciates the painting, endowing it with a temporal extent (i.e., it feels extended in time).*

c. *To experimentally capture the multifaceted aspect of phenomenological experience, we manipulated the content of consciousness by varying stimuli along four dimensions: category (faces, objects, letters and false fonts), identity (each category contained different exemplar), orientation (left, right, and front view), and duration (stimuli were presented for three durations i.e., 0.5 s, 1.0 s, and 1.5 s). Example stimuli used in the study are shown for reference.*

d. *Overview of the experimental paradigm: At any one point in time, no more than one high-contrast, stimulus was present at fixation. In each trial, subjects were asked to detect target stimuli: either a face and an object or a letter and a false font in any of the three different orientations. Thus, each trial contained three stimuli types: targets (depicted in red), task relevant stimuli (belonging to the same categories as the targets, depicted in orange-red), and task irrelevant stimuli (belonging to the two other categories, depicted in purple). The pictorial stimuli (faces/objects) were task relevant in half of the trial blocks, while the symbolic stimuli (letters/false fonts) were relevant in the other half of the blocks. For illustration purposes only, a color line was added to depict the different trial types. Blank intervals between stimuli are not depicted here.*

e. *Distribution of behavioral sensitivity scores (d') separate per data modality and acquisition site. Crossing lines depict average d' per site/modality. Dots depict individual participants d's. Colors depict data modality: MEG N=65 (orange), fMRI N=73 (red), and iEEG N=32 (green), while the hue depicts each site within a modality.*

f. *Distributions of false alarm (FA) rates per site and data modality, separated by task condition: Orange-red depicts task relevant stimuli. Purple depicts task irrelevant stimuli. Dots are individual participants FA rates. Other conventions as in f.*

g. *Top row: Average fixations heatmaps computed over a 0.5 s window after stimulus onset. Heatmaps are displayed per data modality, zoomed into the stimulus area. Bottom row: Average saccadic direction maps per data modality. The three stimulus durations are shown separately.*

Prediction #1: Decoding of conscious content

According to IIT, PFC is not necessary for consciousness. Consequently, proponents of IIT predict that decoding of conscious content should be maximal from the posterior cortex, and should not increase when PFC is added. According to GNWT, PFC is necessary for consciousness and consequently predicts that every content of consciousness should be decodable from the PFC. IIT's prediction of maximal decoding in the posterior cortex was regarded as a non-core test of the theory because, for IIT, what matters is not how much information can be decoded from the extrinsic perspective of an observer, but how much information is available to a neural substrate from its intrinsic, causal perspective. IIT and GNWT further specify that brain areas evidencing conscious content should do so irrespective of other cognitive processes, e.g., report. This implies that conscious content should be present irrespective of task manipulations [61,86]. To empirically test prediction #1, we measured multivariate decoding of stimulus category (pictorial: faces/objects and symbolic: letters/false fonts), and orientation (left/right/front facing). In each block, the subjects' task was to identify two stimuli belonging to either the pictorial or the symbolic group of stimulus categories, e.g., a specific face and a specific object (Figure 2.1d), making these two categories task relevant in that block. Hence, all categories were task relevant and task irrelevant in different blocks. Stimulus orientation was orthogonal to the task, and thus task irrelevant in all blocks.

Based on our preregistered predictions and pre-approved interpretations (Extended Table 2.1, and https://osf.io/92tbg/), the theories would be challenged if we observe decoding of one stimulus category pairing (e.g., faces/objects or letters/false fonts) *but* not decoding of orientation (or vice versa) in at least one of the four categories, in the relevant brain regions and time windows. Thus, the theories would pass this test if decoding is possible for both category and orientation, but would fail otherwise. Testing for decoding of both category *and* orientation constitutes a more stringent test of the theories as it requires two conditions to be satisfied, making it more likely for the test to fail [120], while also capturing a critical aspect of conscious content, i.e., its multidimensionality, or phenomenological richness (Figure2.1b). For decoding of category, we also sought to demonstrate that information is present in the relevant regions irrespective of the task by training a classifier in one task and evaluating whether it generalizes to the other task condition, i.e., cross-task generalization.

Here, we report the most robust results for decoding of category (faces/objects) and orientation (left/right/front views of faces). Qualitatively similar results were observed for decoding of letters/false fonts (Extended Data Figure 2.2a-d). Results for

orientation decoding were consistent across stimulus categories and data modalities in posterior cortex, yet mostly absent in PFC (see supplementary section 5.1.2).

In the iEEG data, we trained pattern classifiers on high gamma frequency band activity (70-150 Hz) at each time-point in the task irrelevant condition and tested across all time-points in the task relevant condition, for each stimulus duration, category, and across all electrodes within the theory-relevant ROIs (Figure 2.2a for a visualization of ROIs and methods section for a list of anatomical ROIs). In the posterior ROIs, face/object decoding showed significant cross-task generalization (>95% accuracy) for the approximate duration of the stimulus (Figure 2.2b, top row). In the PFC ROIs, significant cross-task face/object decoding accuracy (~70%) was also evident, but the temporal generalization of this decoding was restricted to ~0.2-0.4 s (Figure 2.2b, bottom row). Training on task relevant and testing on task irrelevant trials showed similar results (Extended Data Figure 2.2e; within-task decoding provided in Extended Data Figure 2.3). The sustained (posterior) and phasic (PFC) patterns of cross-task temporal generalization of decoding thus matched both IIT's and GNWT's predictions, respectively.

While electrode coverage across our sample of iEEG patients (N=29 for the decoding analyses) was exceptional in the relevant brain regions (Figure 2.2a, PFC ROIs $N_{electrodes}$=576, Posterior ROIs $N_{electrodes}$=583), we also evaluated these predictions in a larger population of healthy subjects (N=65) in MEG. Results from the cross-task decoding of stimulus categories using the MEG cortical time series (see methods section) combining all parcels within the theory-relevant ROIs were consistent with the iEEG observations. Cross-task generalization of face/object decoding was significant in both posterior and prefrontal ROIs (Figure 2.2c) within the theory-predicted time-windows. The extent of cross-temporal generalization of decoding in MEG was sustained in posterior ROIs. In PFC ROIs, decoding was brief for all three stimulus durations (see supplementary section 5.1.1.2).

A limitation of MEG is its spatial imprecision, which can impact source localization. We thus also tested the theories' predictions in a large sample of healthy subjects (N=73) exploiting the high spatial resolution of fMRI. Using a searchlight approach (see methods section), we found distributed and robust cross-task generalization (~75%) in striate and extrastriate, ventral temporal, and intraparietal cortex (Figure 2.2d; see Extended Data Table 2.4 for anatomical details). Generalization in prefrontal cortex had lower accuracy (~60%), and was spatially restricted to middle and inferior frontal cortex regions (Figure 2.2d). We obtained similar results with a decoding approach using theory-relevant ROIs defined in the Destrieux atlas

(see supplementary section 5.1.1.3). These results also closely matched a theory-relevant ROIs analysis in the iEEG data restricted to the time windows specified by the theories (Figure 2.2e). Hence, across recording modalities, we observed that face/object decoding was present both in the posterior and the prefrontal ROI, in line with IIT and GNWT predictions.

As the representation of conscious content is rich and multidimensional including features beyond category, we turned to decoding of stimulus orientation (which was always task irrelevant). Here, we found divergent results for the predictions of IIT and GNWT: decoding of face orientation (left/right/front views) was found in posterior ROIs but not in prefrontal ROIs, both in the iEEG theory-relevant ROIs decoding approach (Figure 2.2f, h; accuracy improved to ~95% with pseudotrial aggregation as shown in Extended Data Figure 2.5a) and in the fMRI searchlight approach (Figure 2.2g, ~45%). From the MEG cortical time series, decoding of face orientation was robust in posterior ROIs (~75% with pseudotrial aggregation), and reached above chance levels, albeit weakly (35%) in prefrontal ROIs (Figure 2.2i). Notably though, control analyses could not conclusively rule out that MEG decoding in the PFC ROIs stemmed from signal leakage from posterior regions (Extended Data Figure 2.5b). Decoding of orientation for the other stimulus categories (letters and false fonts but not for objects) was observed in posterior ROIs but not in the prefrontal ones across the three data modalities (see supplementary section 5.1.2).

Finally, we tested the preregistered prediction by IIT that prefrontal regions do not contribute further information beyond that specified by posterior areas (or may even degrade performance as it could introduce noise into the classifiers) [121]. The results of this test would challenge IIT if the inclusion of PFC was found to increase decoding accuracy, while a lack of an increase would be consistent with both theories as GNWT holds that workspace neurons in PFC broadcast information from posterior processors rather than adding information. We compared decoding performance from classifiers exclusively trained on posterior ROIs with classifiers trained on posterior and prefrontal ROIs together (Extended Data Figure 2.5c) (see methods section). The results across all critical data modalities for testing (iEEG, MEG) indicate that neither category nor orientation decoding improves, and in some cases decreases, when adding prefrontal ROIs to posterior ROIs (Extended Data Figure 2.5d-e). These results are robust to the selection of ROIs, as a control analysis using a broader definition of prefrontal cortex, yielded comparable results (see supplementary section 5.1.3).

Prediction#1 is a central prediction for GNWT, while it is subsidiary for IIT. Considering the primary preregistered tests, and their implications for both theories: for **prediction #1**, we found mixed results for GNWT. On the one hand, we found robust decoding of category in PFC across all three imaging modalities. However, for decoding of orientation, results differed across modalities: only for MEG did cortical activity show decoding of orientation for faces but not for any other stimulus category in PFC. Yet, possible signal leakage from posterior sources could not be conclusively ruled-out. Considering the negative decoding results for orientation from fMRI and iEEG, which provide higher spatial resolution than MEG, this overall pattern of results challenges one of GNWT's predictions. For IIT's predictions, decoding of conscious content (both category and orientation) was robust in posterior cortex, independent of the task manipulation, and consistent across data modalities (iEEG, MEG and fMRI). Also, decoding of category and orientation was found to be the same, or to decrease, when adding PFC to posterior regions.

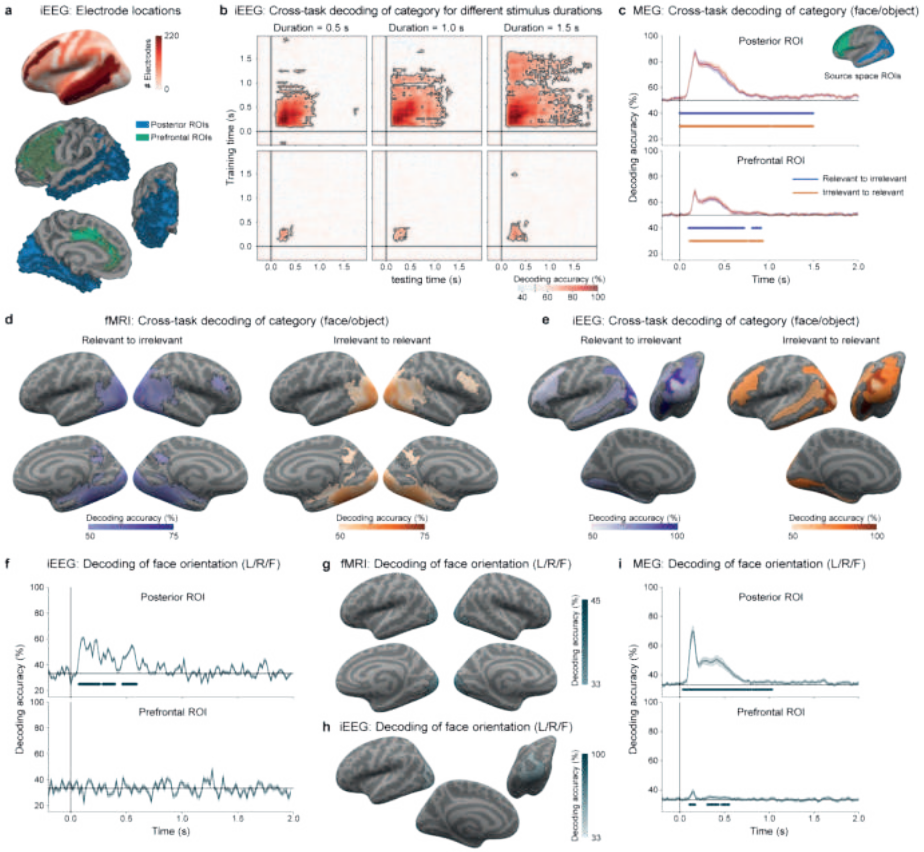

a iEEG: Electrode locations
Electrodes
220
0
Posterior ROIs
Prefrontal ROIs
b iEEG: Cross-task decoding of category for different stimulus durations
Duration = 0.5 s
Duration = 1.0 s
Duration = 1.5 s
Training time (s)
testing time (s)
Decoding accuracy (%)
c MEG: Cross-task decoding of category (face/object)
Posterior ROI
Source space ROIs
Prefrontal ROI
Relevant to irrelevant
Irrelevant to relevant
Decoding accuracy (%)
Time (s)
d fMRI: Cross-task decoding of category (face/object)
Relevant to irrelevant
Irrelevant to relevant
Decoding accuracy (%)
e iEEG: Cross-task decoding of category (face/object)
Relevant to irrelevant
Irrelevant to relevant
Decoding accuracy (%)
f iEEG: Decoding of face orientation (L/R/F)
Posterior ROI
Prefrontal ROI
Decoding accuracy (%)
Time (s)
g fMRI: Decoding of face orientation (L/R/F)
Decoding accuracy (%)
h iEEG: Decoding of face orientation (L/R/F)
Decoding accuracy (%)
i MEG: Decoding of face orientation (L/R/F)
Posterior ROI
Prefrontal ROI
Decoding accuracy (%)
Time (s)

Figure 2.2: Prediction #1-Decoding of conscious content

a. *Spatial coverage of intracranial electrodes across all patients included in the decoding analysis ($N_{subjects}$=29), displayed on a standard inflated cortical surface map (top), and within the regions of interest (ROIs) for the two theories (bottom): posterior (blue, $N_{electrodes}$=583), prefrontal (green, $N_{electrodes}$=576).*

b. *Cross-task temporal generalization of decoding of high gamma signal in iEEG in which pattern classifiers were trained to discriminate stimulus category (faces vs. objects) in the task irrelevant condition at each time-point and tested in the task relevant condition across all time-points. The three stimulus durations are plotted in columns (left: 0.5 s; center: 1.0 s; right: 1.5 s) and the two theory ROIs in rows (top: posterior ROIs; bottom: prefrontal ROIs). Significantly above-chance (50%) decoding is indicated by the outlined pink-red regions in the temporal generalization matrices. Contours indicate statistically significant decoding evaluated through a cluster-based permutation test.*

c. *Cross-task decoding of stimulus category (faces vs. objects) in MEG cortical time series (N=65) when classifiers were trained on relevant stimuli and tested on irrelevant stimuli (purple); or trained on irrelevant stimuli and tested on relevant stimuli (red). Decoding was done separately within the whole posterior ROIs (top) and prefrontal ROIs (bottom). The inset shows inflated cortical surfaces depicting the two ROIs used for theory testing (posterior: blue; prefrontal: green) in the decoding. These decoding results combine data across the three stimulus durations, and used pseudotrial aggregation. The purple and red lines underneath the decoding functions indicate time-periods showing significantly above-chance (50%) decoding as assessed by cluster-based permutation test. Error bars depict 95% CI estimated across subjects.*

d. *Cross-task decoding of stimulus category (faces vs. objects) in fMRI (N=73) using a searchlight approach, collapsed across the three stimulus durations. Left panel (purple): Pattern classifiers trained on relevant stimuli and tested on irrelevant stimuli. Right panel (orange-red): Pattern classifiers trained on irrelevant stimuli and tested on relevant stimuli. Regions showing significantly above-chance (50%) decoding, evaluated through a cluster-based permutation test, are indicated by the outlined colored regions on the inflated cortical surfaces (top: left/right lateral views; bottom: right/left medial views).*

e. *Cross-task decoding of stimulus category (faces vs. objects) in iEEG within the theory-specific ROIs, collapsed across stimulus duration. Decoding accuracies are indicated in purple for classifiers trained on relevant stimuli and tested on irrelevant stimuli, and in orange-red when trained on irrelevant stimuli and tested on relevant stimuli, and are displayed on inflated surface maps from a left lateral view (top left), posterior view (top right) and left medial view (bottom).*

f. *Decoding of stimulus orientation (left vs. right vs. front view faces) which was always task irrelevant, in single trial iEEG data, within posterior ROIs (top) and prefrontal ROIs (bottom), collapsed across the three stimulus durations. Lines under the decoding functions indicate time-points showing above chance (33%) decoding from a cluster-permutation test. Decoding using pseudotrial aggregation is shown in Extended Data Figure 2.5a. Error bars depict 95% CI estimated across cross-validation folds.*

g. *Decoding of orientation (left vs. right vs. front view faces) in fMRI using the searchlight approach. Regions with significantly above-chance (33%) decoding accuracies are indicated in outlined blue on the inflated cortical surface maps (top: left/right lateral views; bottom: right/left medial views).*

h. *Decoding of orientation (left vs. right vs. front view faces) in iEEG within the ROIs. Regions with electrodes showing above-chance (33%) accuracies are indicated in outlined blue on the inflated surfaces (top left: left lateral view; top right: posterior view; bottom: left medial view).*

i. *Decoding of orientation (left vs. right vs. front view faces) in MEG cortical time series within the ROIs (top: posterior; bottom: prefrontal). Time-points showing significantly above-chance (33%) decoding are indicated by lines below the decoding functions. Error bars depict 95% CI estimated across subjects.*

Prediction #2: Maintenance of conscious content over time

According to IIT, the state of the network that specifies the content of consciousness in posterior cortex is actively maintained for the duration of the conscious experience (manipulated here via different stimulus durations). In contrast, GNWT predicts brief content-specific ignition in PFC ~0.3-0.5s after stimulus onset, when the workspace is updated [97]. Then, activity decays to baseline, with information being maintained in an latent state, until another ignition marks the offset of the current percept and the onset of a new percept (in our paradigm, the fixation screen following stimulus offset). Thus, while the underlying brain response (the workspace update) is temporally discrete (i.e., an onset and an offset response), the conscious experience can be temporally continuous (lasting from one workspace update to the next).

Based on our preregistered predictions and interpretations (Extended Table 2.1, and https://osf.io/92tbg/), the theories would be challenged unless we observe the predicted temporal dynamics for maintenance of conscious content, i.e., sustained vs. phasic for IIT and GNWT (Figure 2.1a), respectively, for a minimum of one conscious feature (category, identity or orientation), in the relevant brain regions and time windows. Specifically, IIT would be challenged if we failed to observe sustained content-specific information *and* activation tracking stimulus duration in posterior cortex for the above-mentioned features. GNWT would be challenged if prefrontal phasic activation (at onset and offset) associated with the maintenance of conscious content over time was absent for those features. We tested those predictions by evaluating both the strength of activation as a function of stimulus duration, and the informational content of that activation in each of the theory-relevant ROIs. Here, both activation and information content were deemed central predictions for IIT, such that they jointly determine the overall interpretation of results. For GNWT, activation alone was considered essential for theory evaluation due to the challenges in precisely measuring the reinstatement of content specificity at the time of stimulus offset.

We focused on the task irrelevant condition as it is most diagnostic for neural activity related to consciousness, minimizing the contribution of other, potentially confounding, cognitive processes (see supplementary sections 6.1 and 6.2.9 for results on the task relevant condition). Due to the temporal nature of the predictions, they were tested on the two data modalities with millisecond temporal resolution, iEEG and MEG.

First, we tested the theories' predictions investigating neural activation as a function of stimulus duration. In the iEEG data, we used linear mixed models (LMMs, see methods section) to model the time course of neural activity in the high gamma (HG) frequency band (70-150 Hz), which correlates with spiking activity [122,123], per electrode and theory-relevant ROI as a function of the theories' predicted temporal models (Figure 2.1a. middle panel) and stimulus duration (LMMs, see methods section). To increase sensitivity and to accommodate the (category) selective responses expected in higher-order sensory areas, we included an interaction term with category.

Although we lacked control over the placement of electrodes, the sampling density of electrodes in both the posterior cortex and the prefrontal cortex (PFC) was consistently high and evenly distributed across ROIs pertinent to the theories. This enabled us to fairly and exhaustively test theories' predictions directly in the human brain. Across the 31 epilepsy patients in this analysis, 194 of 657 (29.5%) posterior ROI electrodes and 123 of 655 (18.7%) PFC ROI electrodes exhibited HG activity in response to the stimuli (see supplementary section 6.1.2).

In posterior cortex ROIs, the results of the LMMs revealed a total of 25 electrodes (out of 657) that exhibited sustained activity that tracked stimulus duration (Extended Data Table 2.6 for electrode localization and supplementary section 6.1.1 for results of the full model), in line with IIT's prediction (Figure 2.3a). A subset of 12 electrodes showed sustained duration tracking irrespective of stimulus category predominantly in early visual areas (Figure 2.3b for an example electrode in occipital pole). The remaining 13 electrodes showed category-selective tracking (mostly to face stimuli) localized to the ventral temporal cortex (Figure 2.3b for an example electrode in lateral fusiform gyrus). Overall, the proportion of electrodes showing category-specificity and duration tracking was rather small, e.g., only 15% (8/53) of face selective electrodes showed sustained duration tracking as predicted by IIT, pointing to a rather sparse underlying neural substrate. These responses mostly localized to the lateral fusiform gyrus. The remaining majority face selective electrodes exhibited transient activations at stimulus onset, localized across striate, extrastriate and ventral areas (see supplementary section 6.1.2).

In PFC ROIs, 99 and 24 electrodes showed non-selective and category-selective onset responses, respectively (Figure 2.3d). Yet, none of the 655 electrodes tested matched the temporal profile predicted by GNWT (i.e., onset and offset). This null result was not due to the analysis approach, as the LMM was indeed sensitive to picking up the pattern predicted by GNWT in 10 electrodes outside the predicted ROI, i.e., in striate/extrastriate cortex (Figure 2.3b). An exploratory analysis to decode stimulus duration

with unrestricted temporal profiles and time windows revealed a single electrode in the inferior frontal sulcus showing the GNWT-predicted pattern, yet earlier than expected (0.15 s post-onset and post-offset) (Figure 2.3d). The very same electrode exhibited a biphasic event-related potential with a positive deflection early on (0.15 s) and a negative deflection at a later latency (see supplementary section 6.1.1). Additional control analyses, including time-locking the analyses to stimulus offset to increase statistical sensitivity, corroborated the temporal profile predicted by IIT in posterior ROIs, and the absence of the temporal profile predicted by GNWT in PFC ROIs (see supplementary sections 6.2.1-6.2.3).

For MEG, we used LMMs to investigate the temporal patterns of gamma frequency band power within the posterior (15 parcels) and the PFC (11 parcels) ROIs. Although gamma frequency band activity was strong in posterior areas, none of the theory-based models provided a good fit to the data (see supplementary section 6.1.3.1). We also examined activity in the alpha band, recognizing its potential as a surrogate for neuronal spiking. This is based on its well-documented inverse relationship with neural spiking activity [124,125]. Results on alpha frequency in iEEG and MEG were inconclusive and did not provide strong support for either of the theories. In iEEG, none of the prefrontal electrodes showed the predicted combination of an onset and offset response, but instead this pattern was found in some posterior sites. In MEG, temporal profiles consistent with GNWT were found in most areas in posterior cortex and in the anterior cingulate cortex, but those results were highly dependent on parameter choices and contamination from posterior sites could not be ruled-out (see supplementary sections 6.1.1 and 6.1.3.2).

Together, the results from the temporal activation analysis are compatible with IIT's predictions of sustained activation within posterior cortex. In contrast, we found no evidence in iEEG for GNWT's prediction concerning late phasic ignition of PFC at both stimulus onset and offset. MEG evidence in the alpha band was inconclusive, and not supported by iEEG despite the ample coverage of PFC. These patterns of results accordingly challenges GNWT's predictions.

After analyzing the temporal profile of brain activity, we used cross-temporal Representational Similarity Analysis (RSA) both in the iEEG and MEG source data to test in which time windows the content of consciousness was represented (Figure 2.1a. middle panel). For IIT, a critical prediction is that conscious content should be maintained as long as the conscious experience lasts. GNWT instead predicts a phasic ignition of the workspace at stimulus onset with no active representation of the conscious content until another ignition marks the offset of

the percept. This prediction was tested, but it was classified as non-essential for the evaluation of GNWT. Within each of the theory-relevant ROIs, we performed cross-temporal RSA for each stimulus dimension (category, identity, orientation) and correlated them with the temporal models predicted by the theories (Figure 2.1a, right panel). Here, we report the results for face and object stimuli. Qualitatively similar results were observed for letters/false fonts (Extended Data Figure 2.7).

In iEEG, we calculated the correlation distance between the patterns of HG activity across 583 electrodes in posterior ($N_{subjects}$=28) and 576 electrodes in PFC ROIs ($N_{subjects}$=28), separately. Then, we applied principal component analysis (PCA) to visualize the similarity structure (see methods section). We investigated the 1.5 s duration trials only, because they enable the best contrast between the temporal profiles predicted by the theories.

In posterior cortex ROIs, the cross-temporal RSA revealed sustained face/object categorical representation, with larger correlation distances between categories (face/objects) than within category (face, object) (compare Figure 2.3e left with the predicted pattern in Figure 2.1a). The RSA matrix significantly correlated with the temporal model predicted by IIT, and outperformed the GNWT model (see supplementary section 6.3 for results of all contrasts).

In PFC ROIs, the cross-temporal RSA revealed transient face/object categorical representation at stimulus onset, but not at stimulus offset. In line with this observation, we did not find any significant correlation with the GNWT onset & offset model (compare Figure 2.3f left with the predicted pattern in Figure 2.1a). This was also the case for the task relevant condition, where face/object information was stronger, more stable and longer lasting. Further evidence for the absence of GNWT predicted patterns in PFC ROIs was found in three control analyses using (a) feature selection, which improved RSA in PFC; (b) modified time-windows to investigate the possibility of an earlier ignition at stimulus offset; and (c) a decoding analysis time-locking trials to stimulus offset to maximize sensitivity (see supplementary section 6.4). None of these control analyses changed the overall results. These results nicely align with two independent studies using comparable methods [126,127], attesting to their robustness.

It has been argued that because conscious experiences are specific, the representation of identity and orientation are more stringent tests of the neural substrate of conscious experience [128], than category. We thus also evaluated whether information about stimulus identity (and orientation) matched the theories' predictions.

In posterior ROIs, object identity information was sustained throughout the stimulus duration, with objects of the same identity showing smaller distances than different object identities (Figure 2.3e, middle right). The IIT model significantly correlated with the observed RSA matrix, and also better explained the data compared to the GNWT model. Comparable results were found for letter and false-font identity, but not for face identity (Extended Data Figure 2.7). For the PFC ROIs, identity information was absent for all categories, both at stimulus onset, offset, and generally throughout the time windows (for objects, see Figure 2.3f, middle right). Finally, we tested for the presence of orientation information. In posterior cortex ROIs, information about face orientation was weakly present at stimulus onset, yet was not sustained, decaying after 0.5 s (Figure 2.3e, right), contrary to IIT's predictions. In PFC ROIs, no information about face orientation was found (Figure 2.3f right). MEG time series were inconclusive, as none of the theories' predictions were borne out when testing information about category, identity, or orientation (see supplementary section 6.5).

Considering the primary preregistered tests, their respective weight and interpretations for both theories (Extended Table 2.1), for **prediction #2**, results were in line with IIT's prediction, as activation and representation of conscious content was sustained in posterior cortex, including representation of category and identity across multiple stimuli. Yet, sustained responses were rather rare in posterior cortex (found only in 3.8% of the electrodes in the iEEG data). Also, there was no sustained representation of orientation.

GNWT was challenged as we found no convincing evidence in iEEG or MEG for a late phasic ignition of PFC at stimulus offset, despite the presence of robust ignition at the onset of the stimuli. With regards to the information content, which was considered a non-critical prediction for GNW, the RSA analysis demonstrated category information in PFC, exclusively at stimulus onset and earlier than predicted; while information about stimulus identity and orientation was completely absent.

2

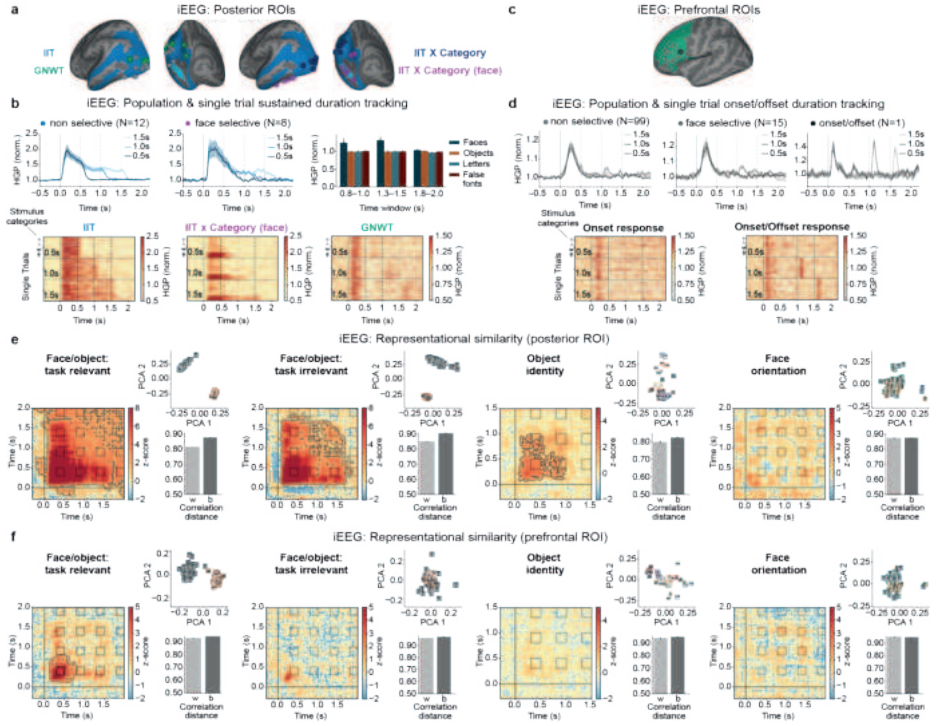

a
iEEG: Posterior ROIs
IIT
GNWT
IIT X Category
IIT X Category (face)
c
iEEG: Prefrontal ROIs
b
iEEG: Population & single trial sustained duration tracking
non selective (N=12)
face selective (N=8)
Faces
Objects
Letters
False fonts
Time window (s)
Stimulus categories
IIT
IIT x Category (face)
GNWT
Single Trials
HGP (norm.)
Time (s)
d
iEEG: Population & single trial onset/offset duration tracking
non selective (N=99)
face selective (N=15)
onset/offset (N=1)
Onset response
Onset/Offset response
e
iEEG: Representational similarity (posterior ROI)
Face/object: task relevant
Face/object: task irrelevant
Object identity
Face orientation
PCA 1
PCA 2
z-score
Correlation distance
f
iEEG: Representational similarity (prefrontal ROI)

Figure 2.3: Prediction #2-Maintenance of conscious content over time

a. *Intracranial electrodes in posterior ROIs, depicted in blue ($N_{subjects}$=31, $N_{electrodes}$=657) showing the sustained duration profile compatible with IIT's predictions, found for category-selective electrodes (N=13, dark blue), specifically for faces (N=8, purple), and non-category selective electrodes (N=12, light blue). Additionally, a small number of electrodes exhibited a biphasic duration profile (11 electrodes, green). Although this biphasic profile corresponds with the GNWT predictions, it was expected to appear in PFC, not in posterior regions. We present these findings to highlight the sensitivity of our analytical approach. However, this specific observation does not directly support GNWT, as the original prediction pertained exclusively to the PFC.*

b. *Top panels. Averaged waveforms in posterior ROIs for non-category selective (left) and face-selective (middle) sustained duration tracking electrodes, separately per stimulus duration, marked in shades of blue. Error bars depict standard error of the mean. (Right) Bar plot depicting mean high-gamma power averaged across all face-selective electrodes for each stimulus category separate per stimulus duration (faces: dark blue, objects: orange, letters: turquoise, false fonts: dark red). Bottom panels. Raster plots of example electrodes depicting non-category selective sustained duration tracking (left), face-selective sustained duration tracking (middle), and phasic onset and offset duration tracking responses predicted by GNWT for PFC ROIs (right). Rows depict single trials, sorted per stimulus duration (from top: 0.5, 1.0, 1.5 s), and then category (from top: false fonts, letters, objects, faces).*

c. *Electrodes in PFC ROIs, depicted in green ($N_{subjects}$=31, $N_{electrodes}$= 655) exhibiting phasic onset responses only (gray, N=114), 1 electrode (black) exhibiting a phasic onset and offset response but significantly earlier (0.15s) than the time window predicted by GNWT (>0.3s). None of the 655 electrodes showed phasic onset and offset response (with activity silence in between) at the time windows predicted by GNWT.*

d. *Top panels. Averaged waveforms in PFC ROIs for non-category selective (left) and face-selective (middle) onset only responsive electrodes, separately per stimulus duration, marked in shades of gray (as their pattern does not comply with any of the theory predictions). Error bars depict standard error of the mean. (Right) Averaged waveforms for the electrode showing an onset & offset response that occur earlier than the predicted time-window. Bottom panels: Raster plots for one example electrode exhibiting an onset response only (left), and the early onset and offset response (right). Y Axis labels as in b.*

e. *Cross-temporal representational dissimilarity matrices across all electrodes in posterior ROIs ($N_{subjects}$=28, $N_{electrodes}$=583) for category (left and middle-left), identity (middle-right) and orientation (right). Sustained representation of category was found irrespective of task (compare task relevant and task irrelevant RSA matrices). Principal component analysis revealed the stable separability across faces and objects, again irrespective of task. Bar plots show the within class dissimilarly (distances within the face and object category) and between class dissimilarity (faces vs. object distances). Larger between than within class separation was observed, consistent with the presence of category information. Sustained information about object identity was observed in posterior cortex, with larger between identity distances and within identity distances. Information about face orientation was weak and not sustained across the stimulus duration in posterior cortex.*

f. *Cross-temporal representational dissimilarity matrices across all electrodes in PFC ROIs, as in Figure 2.3e. Transient representation of category was found irrespective of task (compare task relevant and task irrelevant RSA matrices). Principal component analysis revealed the stable separability across faces and objects, again irrespective of task. Bar plots as in Figure 2.3e. Larger between than within class separation was observed, consistent with the presence of category information. There was no identity nor orientation information in PFC ROIs in the relevant time windows predicted by GNWT, or at any other time point.*

Prediction #3: Interareal communication

IIT predicts neural connectivity within the posterior cortex in the gamma band, i.e., between high-level and low-level sensory areas (V1/V2), throughout any conscious visual experience. In contrast, GNWT postulates a brief and late metastable state (>0.25 s) with information sharing between PFC and category-specific areas manifested in long-range synchronization in the gamma/beta band [129].

Based on our preregistered predictions and *a-priori* interpretations (Extended Table 2.1), the theories would be challenged if we fail to observe interareal connectivity between the cortical nodes specified by the theories in the relevant time windows. For IIT, this implies *sustained* content-specific synchronization between face/object selective areas and V1/V2; while for GNWT connectivity should be *phasic* (0.3-0.5 s) between the category selective areas and PFC. Due to the temporal nature of the predictions, iEEG and MEG provide the most informative test. We computed pairwise phase consistency (**PPC**) [130] between each category-selective time series (face- and object-selective nodes) and either the V1/V2 or the PFC time series in the intermediate (1.0 s) and long-stimulus-duration (1.5 s), task irrelevant trials (see supplementary section 7.1.2 for task relevant trials). We focused on gamma activity, which is held to closely reflect neuronal spiking activity [131]. Furthermore, within the framework of IIT, spiking activity is considered a constituent property of the physical substrate of consciousness [92].

For iEEG, we restricted analyses to electrodes showing face and object selectivity, using a different subset of electrodes to test connectivity with V1/V2 and PFC (see methods section, Figure 2.4a for ROIs and for examples of face and object selective electrodes). Due to the sparse coverage, the requirement to focus on 'activated' electrodes (see methods section) was relaxed. However, restricting the analysis to only activated electrodes does not change the pattern of results. We found increased category selective, e.g., faces>objects synchrony between category-selective and V1/V2 electrodes (Figure 2.4b, top row). However, these effects were early and short-lived (e.g., <0.75 s), observed only at low frequencies, i.e., 2-25Hz, and mostly explained by the synchronous activity elicited by the stimulus evoked response (Extended Data Figure 2.8). Thus, the findings did not match IIT predictions, as the activity was not found in the gamma frequency predicted by IIT, and was not sustained. No content-selective PPC was found between face- and object-selective electrodes and PFC electrodes in the relevant time window, in contrast to GNWT's prediction (Figure 2.4b, bottom row).

For MEG, we used Generalized Eigenvalue Decomposition (GED) [132] to extract face- and object-selective components from ventral temporal areas (Figure 2.4c) and then computed PPC. We found selective synchronization between face-selective areas and both V1/V2 and PFC. However, these effects were early and restricted to low frequencies (2-25 Hz), which was inconsistent with both IIT and GNWT (Figure 2.4d) and mostly explained by stimulus evoked responses (Extended Data Figure 2.8).

The results of the preregistered PPC metric for prediction #3, which was critical for both theories, thus supported neither of them. PPC was chosen based on the theories' mechanistic considerations, because it assesses oscillatory phase. However, phase estimation is challenging in neural signals due to noise. We thus relaxed the constraints and tested the theories exploring a connectivity metric sensitive to co-modulations of signal amplitude - dynamic functional connectivity (DFC; see methods section). We also removed the evoked responses given the observed impact in the PPC metric (Extended Data Figure 2.8 includes the evoked response).

In iEEG, we observed significant connectivity between object selective electrodes and V1/V2 (Figure 2.4e). Connectivity was evident in several frequency bands, most predominantly the gamma band. Yet, it was again brief, in contrast to IIT's predictions. Connectivity between face selective electrodes and V1/V2 was scarce. Significant connectivity was observed between PFC and both the face and the object-selective areas, in the frequency (gamma) and time range predicted by GNWT. For MEG, brief DFC in the alpha-beta frequency bands was found only between face-selective nodes and both PFC and V1/V2 (Figure 2.4f).

Together, the results of the exploratory DFC metric in iEEG were in line with GNWT's predictions, while challenging IIT's predictions, as connectivity with V1/V2 was not sustained. V1/V2 were however sparsely sampled with iEEG in our population, with only 12 electrodes localized to V1/V2 in contrast to 472 localized in PFC.

Finally, we then moved to fMRI, to evaluate connectivity across the entire cortex with homogeneous sampling. We computed generalized psychophysiological interaction (gPPI), defining Fusiform Face Area (FFA) and Lateral Occipital Complex (LOC) as seed regions per subject based on an anatomically constrained functional contrast (see methods section) and combining task relevant and irrelevant trials (the preregistered analysis performed separately on each condition can be found in the supplemental section 7.1.1. Here, conditions were pulled to increase statistical power. See supplementary section 12.). FFA showed content selective (face>object stimuli) connectivity with V1/V2, Inferior Frontal Gyrus (IFG) and Intraparietal Sulcus (IPS),

consistent with the predictions of both IIT and GNWT (Figure 2.4g). No selective increases in interareal connectivity between object selective nodes and PFC or V1/V2 was found in fMRI, also when separating task relevant and irrelevant trials (Extended Data Figure 2.8). To determine whether connectivity to PFC and V1/V2 might be driven by the task in gPPI, we explored the iEEG data separating trials by the task. We found task independent, selective DFC connectivity (face>objects) for face selective electrodes with both IFG and V1/V2 (Figure 2.4h).

The results for **prediction #3**, considering the preregistered hypotheses and their pre-approved interpretation, provided no evidence for IIT or GNWT. Neither the frequency band nor the temporal patterns of the PPC results were consistent with either theory. Yet, when exploring amplitude-based metrics of connectivity (DFC and gPPI), we did find support for GNWT predictions, as both in the iEEG and fMRI we observed connectivity with PFC, further matching the timing (~0.3 s) and spectral composition (gamma frequency) predicted by GNWT. For IIT, though connectivity with V1/V2 was present both in the iEEG and fMRI data, with the expected spectral signature (gamma frequency), it was not sustained throughout the duration of the stimulus, contrary to IIT's prediction.

2

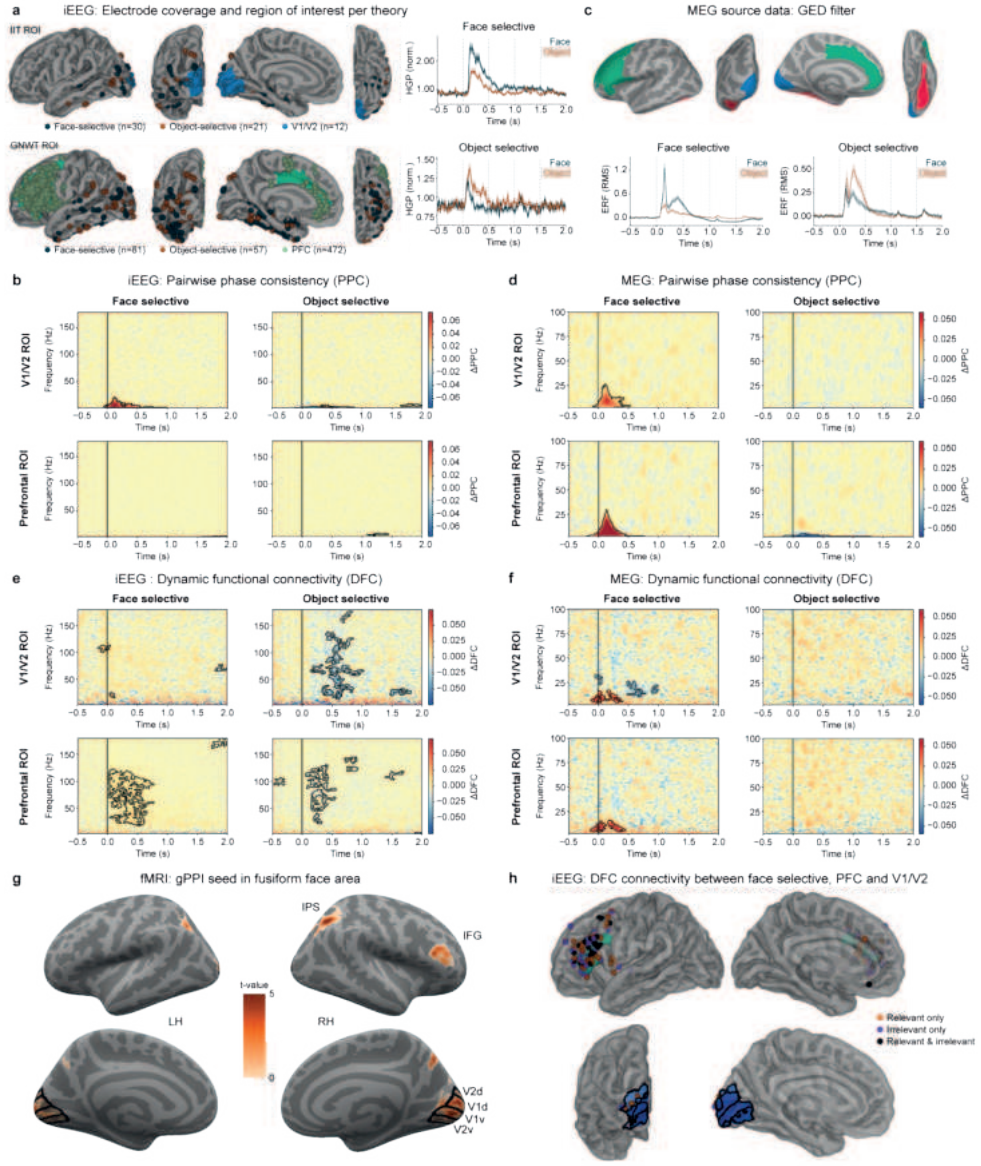

a
iEEG: Electrode coverage and region of interest per theory
IIT ROI
Face-selective (n=30)
Object-selective (n=21)
V1/V2 (n=12)
GNWT ROI
Face-selective (n=81)
Object-selective (n=57)
PFC (n=472)
Face selective
Object selective
HGP (norm.)
Time (s)
c
MEG source data: GED filter
ERF (RMS)
b
iEEG: Pairwise phase consistency (PPC)
d
MEG: Pairwise phase consistency (PPC)
V1/V2 ROI
Prefrontal ROI
Frequency (Hz)
ΔPPC
e
iEEG : Dynamic functional connectivity (DFC)
f
MEG: Dynamic functional connectivity (DFC)
ΔDFC
g
fMRI: gPPI seed in fusiform face area
IPS
IFG
LH
RH
t-value
V2d
V1d
V1v
V2v
h
iEEG: DFC connectivity between face selective, PFC and V1/V2
Relevant only
Irrelevant only
Relevant & irrelevant

Figure 2.4: Prediction #3-Interareal communication

a. *iEEG electrode coverage used to assess content-selective synchrony for IIT ROIs (top, $N_{subjects}$=4) & GNWT ROIs (bottom, $N_{subjects}$=21). Electrode coverage varied between ROIs as interareal connectivity was assessed between electrodes on a per-subject basis. In addition, two example category-selective electrodes are shown (right): one face-selective, and one object-selective. Error bars depict standard error of the mean.*

b. *iEEG Pairwise phase consistency (PPC) analysis of task irrelevant trials reveals significant content-selective synchrony (e.g. faces > objects for face-selective electrodes; objects > faces for object-selective electrodes) in V1/V2 ROIs (top row), but not in PFC ROIs (bottom row).*

c. *MEG cortical time series were extracted per participant from cortical parcels in V1/V2 (blue), PFC (green) and in a fusiform (red) ROIs. Category-selective signals were obtained by creating a category-selective GED filter (i.e., contrasting face/object trials against any other stimulus category trials) on the activity extracted from the fusiform ROI. Face- (bottom left) and object-selective (bottom right) responses averaged across participants are shown at the bottom. Error bars depict 95% CI.*

d. *MEG PPC analysis of task irrelevant trials (N=65) reveals significant category-selective synchrony below 25 Hz for the face-selective GED filter (i.e., faces > objects for face-selective electrodes) in both V1/V2 (top row) and PFC ROIs (bottom row) and for the object-selective synchrony (objects > faces for object-selective electrodes) in the PFC ROI only.*

e. *iEEG Dynamic functional connectivity (DFC) analysis of task irrelevant trials reveals significant content-selective synchrony only for object-selective electrodes in V1/V2 (e.g., top-right), but reveals significant content-selective synchrony for both categories in the PFC ROI (bottom row).*

f. *MEG DFC analysis of task irrelevant trials (N=65) reveals significant content-selective synchrony below 25 Hz for the face-selective GED filter in both V1/V2 (top left) and PFC (bottom left), but not for the object-selective GED filter.*

g. *fMRI gPPI (N=70) on task relevant and task irrelevant trials combined reveals significant content-selective connectivity when FFA is used as the analysis seed. A cluster-based permutation test was used to evaluate the statistical significance of the face > object contrast parameter estimates ($p < 0.05$). Various significant regions showing task related connectivity with the FFA seed were observed including V1/V2, right intraparietal sulcus (IPS), and right inferior frontal gyrus (IFG).*

h. *Analysis of face-selective DFC synchrony across tasks is shown at the single electrode level in PFC (top) & V1/V2 (bottom) ROIs. Electrodes showing significant synchrony in relevant (orange-red), irrelevant (purple), or both relevant & irrelevant (black) task conditions combined are shown (averaged over 70-120 Hz and 0-0.5 s time window). DFC synchrony was observed in both tasks, but restricted to IFG for the GNWT analysis and V2 regions for IIT analysis, consistent with fMRI gPPI analysis shown in panel g.*

Finally, as an additional goal, we used our rich fMRI dataset in a more exploratory manner to delineate the cortical areas presumably involved in (visual) consciousness (i.e., 'putative NCCs'), after ruling out cortical areas that are only responsive to other, accompanying (but confounding) cognitive processes [48]. This test, while being excessively broad and thus not critical for the theories, nonetheless carries implications for both theories, considering their distinct predictions regarding the NCC. IIT predicts that the cortical substrate of consciousness should include posterior areas while agreeing that certain PFC areas should be excluded due to task confounds. GNWT predicts an involvement of PFC even after ruling out task-based effects (see methods section for analysis strategy).

The full results of the pNCC analysis are described in the supplementary section 8.1; here we focus on the PFC given its relevance to the theories. In PFC, the observed pattern of candidate areas was more spatially restricted than anticipated by the rather extensive preregistered GNWT ROIs. Specifically, the MFG, IFG and orbital cortex might participate in consciousness, as predicted by GNWT. Furthermore, the scant activation patterns found in PFC compared to the widespread deactivations was surprising, and suggests a reconsideration of the strong focus on activations (relative to deactivations) when assessing this region's role in conscious perception.

General Discussion

This adversarial collaboration was aimed at overcoming researchers' confirmation biases, breaking theoretical echo chambers [20], identifying strengths and weaknesses of the theories [79,133] by forcing them to be explicit and committal about their respective empirical predictions, rigorously testing them on common methodological grounds [109,110], and providing the means for theorists to change their minds given conflicting results [109]. In doing so, this approach enables progress in the field by catalyzing our ability to evaluate and arbitrate between current theories of consciousness. Embracing this spirit, we opted for a discussion in three voices because even if we provide a stringent test that brought together incompatible theoretical views, different interpretations of the same evidence still remain due to how observers differentially weigh evidence. In what follows, the theory-neutral consortium presents the main challenges our study poses to the theories, based on the predictions, methods and analysis that were preregistered, and agreed upon with the adversaries prior to conducting the study and the disclosure of its results. Then, the adversaries offer their interpretation to the results and future directions. This process follows the guidelines for structuring adversarial collaborations [111].

Cogitate consortium

Figure 2.5 provides a detailed summary of the key results, including criteria for determining if the results support or contradict the theories being tested. This summary covers both the main findings and those less central to the theory evaluation. The consortium aimed to rigorously test these theories, adopting a Lakatos's sophisticated falsificationist approach to the philosophy of science[104,117]. As such, challenged predictions are considered more informative than predictions that are borne out by the data. Predictions and outcomes are weighted differentially across the three predictions and so are the methodologies deemed pertinent for the interpretation of the outcome (Extended Table 2.1).

For IIT, the **lack of sustained synchronization within posterior cortex** represents the most direct challenge, based on our preregistration. Across several analyses, with various degrees of sensitivity, we only observed transient synchronization between category selective and early visual areas. This is incompatible with IIT's claim that the state of the neural network, including its activity and connectivity, specifies the degree and content of consciousness [92]. Although this null result could stem from methodological limitations (e.g., limited iEEG sampling of V1/V2 areas), our multimodal and highly powered study provided the best conditions so far for the predicted patterns to be found. We urge IIT proponents to direct future efforts to evaluate this prediction and to determine its significance and the extent of this failure.

More broadly, although IIT passed the predefined criteria for the duration prediction (#2), there was **no evidence for a sustained representation of orientation**, despite being a property of the consciously perceived stimuli, which should have accordingly showed sustained representation [112]. This is an informative challenge for IIT, as orientation decoding was robust across all three data modalities, leaving open the question of whether and how information about orientation is maintained over time.

Finally, our pNCC analysis suggested that portions of PFC might be important for consciousness. While the most consistent activation and decodability of content was found in posterior cortex, IIT must explain the finding that the MFG and the IFG (for which we also found results in the decoding and synchrony analysis), were visually responsive and not ruled out as being task-related. This finding is particularly important to explain in the context of the current experiment where additional cognitive processing of the task irrelevant stimuli was minimized [65].

For GNWT, the most significant challenge based on our preregistered criteria pertains to its account for the maintenance of a conscious percept over time; and in particular, **the lack of ignition at stimulus offset**. In most of our main tests and control analyses across data modalities (for details, see supplementary sections 5-6), we failed to reveal an offset response in PFC (both in activation which was a critical test, and also in reinstatement of decoded content of any type, which was predefined as non-critical). This result is less likely to stem from sensitivity limitations, since offset responses were robustly found elsewhere (e.g., visual areas); and in PFC, strong onset responses were found to the very same stimuli. The lack of ignition at stimulus offset is especially surprising given the change of conscious experience at the onset of the blank fixation screen. This clear update to the content of consciousness should have been represented somehow by the global workspace [97]. Thus, as our results do not support GNWT's predictions regarding the maintenance of conscious experience, that aspect of consciousness remains unexplained within the GNWT framework.

Another key challenge for GNWT pertains to representing the contents of experience: though we found representation of category in PFC irrespective of the task, hereby demonstrating the sensitivity of our methods, **no representation of identity was found, and representation of orientation was only evident in MEG (without being able to exclude source leakage effects)**, although these dimensions were clearly a part of subjects' conscious experience of the stimuli. This raises the question of whether PFC is involved in broadcasting *all* conscious content as predicted by GNWT [25] or only a subset (e.g., abstract concepts and categories, rather than low-level details), in which case the role of PFC in consciousness might need to be redefined.

Finally, the highly spatially restricted decoding of conscious content in PFC, alongside the restricted activations and deactivations in PFC observed in the pNCC analysis, point to a "localized spark" rather than the "wide-spread ignition" predicted by the theory, further challenging it [93].

Prior to the current study, the predictions from IIT and GNWT had mostly been tested with one data modality at a time [18,25], leaving interpretational freedom for negative results, which can easily be attributed to the limitations of a given modality [134]. Here, the combination of techniques allowed us to cross-compensate for their respective limitations to thoroughly and systematically assess the theories' predictions. This methodological approach was mutually agreed upon by the theory leaders prior to data collection and results disclosure as the most powerful and conclusive approach, making both positive and negative findings more meaningful.

Although this study was designed around IIT and GNWT, the results may have implications for other theories of consciousness. For example, GNWT's prediction #1 about PFC is shared by some (but not all) higher-order theories of consciousness that also give a central role for PFC [107]. As a result, the challenges to this prediction challenge not only GNWT but also those higher-order theories. Predictions #2 and #3 about timing and connectivity are more distinctive to GNWT but could also be shared by other theories in principle. Likewise, IIT's non-core prediction #1 about posterior cortex is also shared by many other theories (e.g., recurrent processing theory [105]), and its prediction #2 about timing may be shared by some posterior theories of consciousness, such as the local recurrency theory [106]. Its prediction #3 about interareal connectivity is more distinctive to IIT (e.g., it is not shared by synchrony theory [135]), so the challenge here is more specific as well.

All this highlights that our adversarial collaboration is designed more to challenge theories than to confirm them. Both theories have some predictions confirmed, but these predictions are also consistent with other theories, so the successful predictions cannot serve as evidence for IIT or GNWT specifically. However, the disconfirmed predictions are certainly challenges to both theories (and to others, as discussed above). These challenges can be met by altering the theories or their proposed biological implementation, but such alteration typically comes at some cost to the theoretical framework, because the relevant features of the theory or the implementation were motivated by the framework. In this respect, our adversarial collaboration approach subscribes to the approach advocated by Lakatos [117], a sophisticated version of Popper's falsificationism [136], whereby scientific knowledge advances through a process of conjectures and refutations. When a theory makes an unsuccessful prediction, the challenged theory can survive by refining its details. But if unsuccessful predictions continue, the theory can be deemed a degenerate rather than a progressive research program [104]. This process is expected to be continued by the results of our second experiment (reported in a future manuscript), alongside those of a follow-up adversarial collaboration using a comparable experimental design in animal models (i.e., mice and non-human primates). With time, we hope that substantial evidence will be gathered, allowing the scientific community to form an informed judgment about both theories and possibly others (through the open data). This might be important, as some have proposed a theory-inspired approach to inferring consciousness in non-responsive populations such as unresponsive patients, infants, non-human animals and artificial systems [137–139].

Conceptually, our study focused on the mechanisms by which the content of the conscious experience of A differs from the experience of B (i.e., category, identity,

orientation and duration), which addresses how the link between brain activity and subjective phenomenology changes between distinct conscious experiences. As such, we departed from the mainstream contrastive method in which the presence of conscious experience is contrasted with its absence to study the neural differences between conscious and unconscious processing. Though widely used, the standard contrastive approach suffers from shortcomings which preclude it from directly revealing the processes related to consciousness, as it confounds consciousness with other cognitive processes such as decision-making, reporting, or the formation of episodic memory traces after a conscious experience [24,47,48]. Studying the content of consciousness more directly links phenomenology to brain activity and overcomes several of the limitations of the contrastive method. Yet, some might argue that in doing so, we are tracking mere stimulus processing rather than consciousness per se. Within the framework of this adversarial collaboration, our aim is to challenge and potentially falsify [117,136] IIT and GNW, by examining where their predictions differ, rather than to discover the neural correlates of consciousness. In this context, what might seem like a weakness — focusing on the presence of fully attended, consciously experienced stimuli to test the theories' primary positive predictions and their failures — is actually beneficial. This approach effectively tests if the neural mechanisms suggested by these theories are indeed necessary for consciousness, since if they are, they must be found in such clear-cut cases, where the stimuli are undoubtedly experienced, and the evoked signal is strong (so null results cannot stem from noisy or weak signals). Therefore, our method provides a rigorous and principled examination of both IIT and GNWT.

Our study, while comprehensive, is not without its limitations. First, despite our best efforts to minimize the contribution of task relevance by making some stimulus features relevant on some trials and irrelevant in others, we cannot rule out some residual task engagement with respect to category. However, this potential bias is addressed by our deliberate choice to make features like orientation and duration always irrelevant to the task. This approach strengthens the test for the theories we are examining, as any detected effects on these features cannot be attributed to selective attention driven by task requirements, no matter how minimal. Second, although we made our best efforts to capture the richness of experience by investigating multiple dimensions of conscious experience (i.e., category, orientation, identity and duration), we acknowledge that our efforts are still far from measuring consciousness in a way that truly captures its apparent phenomenal richness (e.g., an object's brightness and hue, its precise shape and location, the highly specific viewpoint from which an object is perceived, etc.). Future studies will be needed to address this further. Third, although our study offers superior spatial and temporal resolution

across the brain by integrating three distinct brain imaging techniques—fMRI, MEG, and iEEG—it falls short of incorporating single-unit recordings. Such recordings, typically reserved for a small subset of epilepsy patients and limited to certain brain areas like the Medial Temporal Lobe, are impractical for directly testing our theories. Studies in other animal models, including Neuropixels and causal manipulations, are underway as part of a different adversarial collaboration, and are expected to complement our findings. Despite the inherent challenges of using animal studies to probe consciousness (difficulty of measuring consciousness in non-human subjects and the limited spatial coverage of Neuropixels probes, and overtraining), we see these two adversarial collaborations as synergistic, providing a stronger test for the theories than either one alone.

Beyond the direct challenges to the theories, our study raises a number of important questions for theory testing and theory building, which apply broadly across most fields, e.g., how to weigh different theory predictions, and how to combine evidence across predictions, analyses and measures (in our case, fMRI, MEG and iEEG data). From the outset, we defined an independent set of predictions, setting criteria for failure to then weigh the results against these predictions. We opted for a lenient approach with respect to falsificationism, sufficing with some evidence for a prediction to pass (e.g., for decoding of category and orientation, we deemed a result in at least one of the tested features sufficient to rule out a failure, instead of requiring results to be seen across all tested categories and orientations). Yet, a formal framework that quantitatively integrates evidence by weighing and quantitatively integrating over passes and failures, accounting for the centrality of the predictions for the theory, measurement error, and consistency across samples and measurements is direly needed to enable systematic theory building in the era of accumulation of results [140].

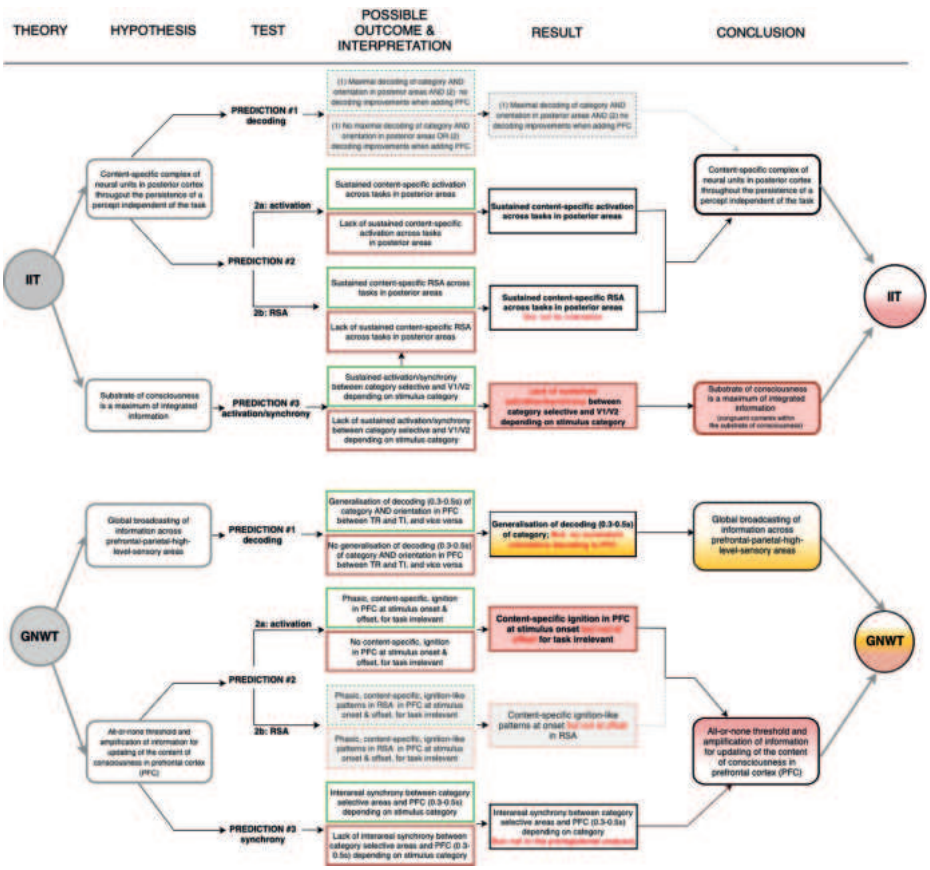

THEORY
HYPOTHESIS
TEST
POSSIBLE OUTCOME & INTERPRETATION
RESULT
CONCLUSION
IIT
Content-specific complex of neural units in posterior cortex throughout the persistence of a percept independent of the task
Substrate of consciousness is a maximum of integrated information
PREDICTION #1 decoding
2a: activation
PREDICTION #2
2b: RSA
PREDICTION #3 activation/synchrony
Sustained content-specific activation across tasks in posterior areas
Lack of sustained content-specific activation across tasks in posterior areas
Sustained content-specific RSA across tasks in posterior areas
Lack of sustained content-specific RSA across tasks in posterior areas
Sustained-activation/synchrony between category selective and V1/V2 depending on stimulus category
Lack of sustained activation/synchrony between category selective and V1/V2 depending on stimulus category
Sustained content-specific activation across tasks in posterior areas
Sustained content-specific RSA across tasks in posterior areas
between category selective and V1/V2 depending on stimulus category
Content-specific complex of neural units in posterior cortex throughout the persistence of a percept independent of the task
Substrate of consciousness is a maximum of integrated information
IIT
GNWT
Global broadcasting of information across prefrontal-parietal-high-level-sensory areas
All-or-none threshold and amplification of information for updating of the content of consciousness in prefrontal cortex (PFC)
PREDICTION #1 decoding
2a: activation
PREDICTION #2
2b: RSA
PREDICTION #3 synchrony
Generalisation of decoding (0.3-0.5s) of category AND orientation in PFC between TR and TI, and vice versa
No generalisation of decoding (0.3-0.5s) of category AND orientation in PFC between TR and TI, and vice versa
Phasic, content-specific, ignition in PFC at stimulus onset & offset, for task irrelevant
No content-specific, ignition in PFC at stimulus onset & offset, for task irrelevant
Interareal synchrony between category selective areas and PFC (0.3-0.5s) depending on stimulus category
Lack of interareal synchrony between category selective areas and PFC (0.3-0.5s) depending on stimulus category
Generalisation of decoding (0.3-0.5s) of category;
Content-specific ignition in PFC at stimulus onset for task irrelevant
Interareal synchrony between category selective areas and PFC (0.3-0.5s) depending on category
Global broadcasting of information across prefrontal-parietal-high-level-sensory areas
All-or-none threshold and amplification of information for updating of the content of consciousness in prefrontal cortex (PFC)
GNWT

Figure 2.5: An overview of theoretical predictions, experimental outcomes and interpretations

On the left, the original predictions made by the IIT (top) and GNWT (bottom) that were preregistered (see also [97]; Figure 2.1). The table describes the key hypotheses (second column, 'Key hypotheses') made by the theories (see also Figure 2.1a), and probed in three different tests analyses (third column, 'Test'; decoding (prediction #1; Figure 2.2), activation & RSA (prediction #2; Figure 2.3) and synchrony (prediction #3; Figure 2.4)). Next, we describe the possible outcomes of each of these analyses, and how they would inform the theoretical predictions (fourth column, 'Possible outcome and interpretation'). Outcomes that conform with the prediction are presented in a green frame (i.e., 'pass'), outcomes that contradict the prediction are presented in a red frame (i.e., 'fail'). Outcomes in a solid frame reflect critical predictions for the theories; dotted and grayed frames indicate non-critical predictions for the theories. Thus, the left side of the table presents the a-priori predictions, expected outcomes and their centrality for the theory evaluation, prior to conducting the experiment. The right side of the figure presents the actual findings of this experiment, integrating over the three modalities and multiple tests. We first summarize the key findings with respect to each prediction (fifth column; 'Result'). Here, white indicates results that are aligned with the theories' predictions, red indicates results that challenge them, the mixture of white/red indicates cases in which the combination of results yielded a mixture of a pass and a fail with the respective explanation for the failure. Yellow marks cases in which we considered that the results did not allow a strong interpretation. We integrate over these results to generate the final conclusion based on the key hypotheses, with the same color coding. For IIT, our conclusion includes a mixture of a passed prediction (of content-specific complex of neural units in posterior cortex, throughout the persistence of a percept, independent of the task) and a failure (of maximum of integrated information) and for GNWT, a mixture of a partly challenged prediction (of an all-or-none threshold and amplification of information updating the content of consciousness in PFC) and a partly supported one, given the inconclusive result for orientation (of global broadcasting of information in the PFC). See the main text for how these results might also challenge other theories of consciousness.

Integrated Information Theory: Melanie Boly, Christof Koch, Giulio Tononi

The results corroborate IIT's overall claim that posterior cortical areas are sufficient for consciousness, and neither the involvement of PFC nor global broadcasting are necessary. They support preregistered prediction #1, that decoding conscious contents is maximal from posterior regions but often unsuccessful from PFC, and prediction #2, that these regions are sustainedly activated while seeing a stimulus that persists in time. They do not support prediction #3 concerning sustained synchrony, although this negative finding is quite possibly the result of sparse electrode coverage (see supplementary section 9). Below we illustrate how these predictions were motivated by IIT.

Posterior regions are often considered mere 'information processors'; their activation, it is claimed, may be necessary but not sufficient for experiencing specific contents. For example, they may show activations during deep sleep or anesthesia and for unreported stimuli under contrastive, near-threshold paradigms [94]. This seems to warrant the need for additional ingredients, such as 'global broadcasting' [94] or 'higher-order monitoring' by PFC [103].

For IIT, however, posterior regions are sufficient for consciousness as long as they satisfy the requirements for maximal integrated information. Why this prediction? Unlike other approaches, IIT infers the essential, physical requirements for the substrate of consciousness from the essential properties of experience [87,92]. This leads to the claim that the quality and quantity of an experience are accounted for by the 'cause–effect structure' specified by a substrate with maximal integrated information, called the 'main complex' [87,92]. We conjectured that posterior cortical regions should provide an excellent substrate for the main complex owing to their dense local connections arranged topographically into a hierarchical, divergent–convergent 3D lattice [92], leading to prediction #1. Nevertheless, by IIT, posterior regions can only support consciousness if their physiology ensures high integrated information—which indeed breaks down [141] due to bistability when consciousness is lost in deep sleep and anesthesia [142–144].

Much of PFC, in contrast, seems to be organized not as a grid but as a patchwork of segregated columns [145], unfavorable for high integrated information. Even so, any PFC region organized in a grid-like way with dense interconnections with posterior regions may well be part of the main complex. As previously emphasized [39], "*…we bear no preconceived enmity to the prefrontal cortex. Indeed, searching for the NCC of specific aspects of experience…in certain anterior regions is an important task ahead*." For example,

parts of IFG might contribute to, say, an abstract/evaluative/actionable experiential aspect of faces, which could be consistent with some pNCC analysis results. However, IIT predicts that we would still experience faces (sans aspects contributed by PFC regions) if PFC were selectively inactivated.

For IIT, all quality is structure: all properties of an experience are accounted for by properties of the cause–effect structure specified by the main complex. Every conscious content (face, object, letter, blank screen) is thus a (sub)structure of integrated information (irreducible cause-effects and their overlaps [87]); it is neither a message that is encoded and broadcasted globally [12,94,146], nor a distributed activity pattern, nor a neural process. Indeed, IIT's research program aims to account for specific consciousness contents—why space feels extended, time feels flowing, and phenomenal objects feel like binding general concepts (invariants) with particular features—all exclusively in terms of their corresponding cause–effect structures [87,112]. As highlighted in the Introduction, when we see Mona Lisa, we see that it is a face, with her particular features, at a particular location on the canvas, and we see her for as long as we look at her. This is why we predicted (prediction #2) that the NCC in posterior cortex would last for the duration of the percept, notwithstanding the widespread evidence for neural adaptation and onset/offset neural responses (probably due to transient excitation/inhibition imbalance), and (prediction #3) that synchrony would occur (reflecting causal binding) between units in higher and lower areas, supporting respectively invariant concepts and particular features.

To conclude, moving beyond the contrastive paradigm between seen and unseen stimuli and beginning to account for how experience feels is one key reason why the experiments reported in this adversarial collaboration mark an important development. Another is that they inaugurate a powerful new way of making progress on a problem often considered beyond the reach of science. The group that carried out this endeavor did so in a way that was explicit, open, and truly collaborative—in short, in a way that is paradigmatically scientific.

Global Neuronal Workspace Theory: Stanislas Dehaene

This unprecedented data collection effort brings several new insights relevant to our theory. Most importantly, the results confirm that PFC exhibits a metastable bout of activity ("ignition") for about ~200 ms, in a content-specific manner, even for task irrelevant stimuli, irrespective of stimulus duration (Figures 2.2b, 2.3f, Supplementary Figure 2.23), and with a concomitant transient increase in long-distance dynamic functional connectivity with face- and object-selective posterior areas (Figure 2.4e-h). Those findings, unpredicted by IIT but predicted by GNWT, support previous findings

that PFC contains a detailed code for conscious visual contents [40,61,147–150]. They also counter previous conclusions that were, in our opinion, too hastily drawn on the basis of insufficient evidence [86]: with suitably sensitive experiments, content-specific PFC regions do show a transient ignition even for irrelevant stimuli. While agreeing with previous results [35,41,59,150,151], the convergence of iEEG, MEG and fMRI in the same task alleviates concerns associated with a possible mis-reconstruction of MEG sources. It also resolves a controversy related to the timing of conscious ignition, which was initially thought to be associated with the P300 ERP waveform [94], but can obviously arise earlier (~200 ms post-onset) [41,59]. GNWT would further predict that this latency should vary depending on the strength of both bottom-up accumulating evidence (e.g., contrast [152]) and top-down attention/distraction by other tasks [41,151,153].

While some results do challenge GNWT, they do not seem insurmountable given experimental limitations. First, note that there is a considerable asymmetry in the specificity of the theories' predictions. None of the massive mathematical backbone of IIT, such as the ϕ measure of awareness, was tested in the present experiment. Instead, what are presented as unique predictions of IIT (posterior visual activation throughout stimulus duration) are just what any physiologist familiar with the bottom-up response properties of those regions would predict, since visual neurons still respond selectively during inattention or general anesthesia [154–156]. Such posterior stimulus-specific, duration-dependent responses are equally predicted by GNWT, but attributed to non-conscious processing.

Unfortunately, here, it is impossible to decide which of the activations reflected conscious versus non-conscious processing, because the experimental design did not contrast conscious versus non-conscious conditions (fortunately, a second experiment by the Cogitate consortium will include such a contrast). The present experiment relied on the seemingly innocuous hypothesis that stimuli were "indubitably consciously experienced" for their entire duration. However, it is well known that perfectly *visible* stimuli, depending on attention orientation, may fail to be *seen* (attentional blink, inattentional blindness) [157,158] or may become conscious at a time decoupled from stimulus presentation (psychological refractory period, retro-cueing) [153,159–161]. Here, it seems likely that subjects briefly gained awareness of all the images (since they remembered them later), but then reoriented their conscious thoughts to other topics, without waiting for image offset – and this interpretation perfectly fits the ignition profile that was found in PFC. It would be surprising if participants' consciousness remained tied to each image for its full duration on every trial of this long experiment. It is also unclear whether participants were ever aware of stimulus orientation, which was always irrelevant. A new experiment,

using quantified introspection [153], will be needed to assess for how long participants maintained the visual image in consciousness.

For the same reason, the absence of decodable activation at stimulus offset, while challenging, may simply indicate that participants never consciously attended to that event, which was always uninformative and irrelevant. Making stimulus offset more attractive, for instance by turning it into an occlusion event where an object hides behind a screen, could yield different results.

For GNWT, the prefrontal code for a conscious mental object is thought to involve a vector code distributed over millions of neurons which, unlike in posterior regions, are not clustered but spatially intermingled [61,162]. Thus, we are not surprised that PFC responses are hard to decode from the macro- or mesoscopic signals measured by fMRI, MEG, or large intracranial electrodes that pool over tens of thousands of neurons. Therefore, the present positive results, indicating transient PFC ignition and decoding of faces and objects, seem to us more important than the null ones, especially as there is already much single-neuron evidence that PFC contains even more precise stimulus-specific neural codes [40,61,147,148].

Finally, while the theories concern the necessary regions for conscious experience, the present methods are purely correlational and do not evaluate causality. This limitation is not unique to the present work, but applies to any brain-imaging experiment. While applauding the present efforts, we therefore eagerly await the results of other adversarial collaborations using causal manipulations in animal models.

Conclusion (Cogitate consortium)

At this point, the reader might expect the consortium to draw a final conclusion regarding the two theories we have evaluated. Instead, we invite readers to form their own conclusions, considering the relative evidence we presented for each of the preregistered predictions, the scope of the evidence and the sophisticated techniques, the role of hindsight bias, and the many challenges in changing people's minds. Science is a social enterprise and evidence is interpreted based on prior beliefs and expectations. The reader is as much a part of this social enterprise as any of the authors from this consortium. We have aimed to present the evidence, and the adversaries' reactions, as straightforwardly and openly as possible. This aligns with our belief that science needs openness to collectively converge to true explanations of complex phenomena in nature, such as consciousness.

Methods

Preregistration and data availability

The full study protocol is available in the preregistration on the OSF webpage, including: (a) an exhaustive description of the experimental design, (b) the theories' predictions and agreed upon interpretations of the results, (c) iEEG, MEG, and fMRI data acquisition details; (d) preprocessing pipelines; and (e) data analysis procedures. All data and code will be shared upon publication. Below, the main methods are concisely summarized. Deviations from the preregistration are noted throughout the manuscript and summarized in Section 12 of the supplementary materials.

Ethics Statement

The experiment was approved by the institutional ethics committees of each of the data-collecting labs (see supplementary 10 for details). All volunteers and patients provided oral and written informed consent before participating in the study. All study procedures were carried out in accordance with the Declaration of Helsinki. Epilepsy patients were also informed that clinical care was not affected by participation in the study.

Participants

Healthy volunteers and patients with pharmaco-resistant focal epilepsy participated in this study. The datasets reported here consist of: (1) Behaviour, eye tracking and invasive electroencephalogram (iEEG) data collected at the Comprehensive Epilepsy Center at New York University (NYU) Langone Health, Brigham and Women's Hospital, Boston Children's Hospital (Harvard), and University of Wisconsin School of Medicine and Public Health (WU). (2) Behaviour, eye tracking, magnetoencephalographic (MEG) and electroencephalographic (EEG) data collected at the Centre for Human Brain Health (CHBH) of the University of Birmingham (UB), and at the Center for MRI Research of Peking University (PKU). (3) Behaviour, eye tracking and functional magnetic resonance (fMRI) data collected at Yale Magnetic Resonance Research Center (MRRC) and at the Donders Centre for Cognitive Neuroimaging (DCCN), of Radboud University Nijmegen. For both the MEG and fMRI datasets, a 1/3 of the data that passed quality tests (henceforth, *Optimization dataset*; see preregistration for details about quality test criteria) were used to optimize the analysis methods, which were subsequently added to the preregistration as an additional amendment. These preregistered analyses were then run on the remaining 2/3 of the data (henceforth, *Replication dataset*) and constitute the data reported in the main study. For comparison, results from the optimization phase are reported in the supplementary 4. This procedure was not used for the iEEG data due to the

serendipitous nature of the recording and electrode placement, the rarity of this type of data and the increased difficulty of data collection due to the COVID-19 pandemic.

For the iEEG arm of the project, a total of 34 patients were recruited. Two patients were excluded due to incomplete data. Demographic, medical and neuropsychological scores for each patient, when available, are reported in Supplementary Table 2.25. Three iEEG patients whose behavior fell slightly short of the predefined behavioral criteria (i.e. hits < 70%, FA > 30%) were nonetheless included given the difficulty to obtain additional iEEG data (see supplementary section 12). A total of 97 healthy subjects were included in the MEG sample (mean age 22.79 ± 3.59 years, 54 females, all right-handed), 32 of those datasets were included in the optimization phase (mean age 22.50 ± 3.43 years, 19 females, all right-handed), and 65 in the replication sample (mean age = 22.93 ± 3.66, 35 females, all right-handed). Five additional subjects were excluded from the MEG dataset: two due to failure to meet predefined behavioral criteria (i.e., hits < 80%, and/or FA > 20%), two due to excessive noise from sensors, and one due to incorrect sensor reconstruction. A total of 108 healthy participants were included in the fMRI sample (mean age 23.28 ± 3.46 years, 70 females, 105 right-handed), 35 of those datasets were included in the optimization sample (mean age 23.26 ± 3.64 years, 21 females, 34 right-handed), and 73 in the replication sample (mean age = 23.29 ± 3.37, 49 females, 71 right-handed). Twelve additional subjects were excluded from the fMRI dataset: eight due to motion artifacts, two due to insufficient coverage, and two due to incomplete data (with respect to these last two subjects, see supplementary section 12. Deviations from the preregistration document).

Experimental procedure

Experimental design

To test critical predictions of the theories, five experimental manipulations were included in the experimental design: (1) four stimulus category (faces, objects, letters and false fonts), (2) twenty stimulus identity (20 different exemplars per stimulus category), (3) three stimulus orientation (front, left and right view), (4) three stimulus duration (0.5 s, 1.0 s, 1.5 s), and (5) task relevance (relevant targets, relevant non-targets, irrelevant).

Stimulus category, stimulus identity and stimulus orientation served to test predictions about the representation of the content of consciousness in different brain areas by the theories. In addition, stimulus duration served to test predictions about the temporal dynamics of sustained conscious percepts and interareal synchronization between areas. Task relevance served to rule out the effect of task

demands, as opposed to conscious perception per se, on the observed effects [163]. This aspect of the experimental design was inspired by Farooqui & Manly [53].

Stimuli

Four stimulus categories were used: faces, objects, letters and false fonts. These stimuli naturally fell into two clearly distinct groups: pictures (faces and objects) and symbols (letters and false fonts). These natural couplings were aimed at creating a clear difference between task relevant and task irrelevant stimuli in each trial block (see Procedure). All stimuli covered a squared aperture at an average visual angle of 6˚ by 6˚. Face stimuli were created with FaceGen Modeler 3.1; letter and false fonts stimuli were generated with MAXON CINEMA 4D Studio (RC - R20) 20.059; object stimuli were taken from the Object Databank [164]. Stimuli were gray-scaled and equated for luminance and size. To facilitate face individuation, faces had different hairstyles and belonged to different ethnicities and genders. Equal proportion of male and female faces were presented. The orientation of the stimuli was manipulated, such that half of the stimuli from each category had a side view (30° and -30° horizontal viewing angle, left and right orientation) and the other half had a front view (0°).

Procedure

Subjects performed a non-speeded target detection task (see supplementary video). The experiment was divided into runs, with four blocks in each run (see Trial counts below). On a given block, subjects viewed a sequence of single, supra-threshold, foveally presented stimuli belonging to one of four stimulus categories and presented for one of three stimulus durations onto a fixation cross that was present throughout the experiment. Within each block, half of the stimuli were task relevant and half task irrelevant. To manipulate task relevance, at the beginning of each block subjects were instructed to detect the rare occurrences of two target stimulus identities, one from each relevant category (pictures: face/object or symbols: letter/false-font), irrespective of their orientation. This was specified by presenting the instruction "detect face A and object B" or "detect letter C and false-font D", accompanied by images for each target (See Figure 2.1d). Targets did not repeat across blocks. Each run contained two blocks of the Face/Object task and two blocks of the Letter/False-font task, with block order counterbalanced across runs.

Accordingly, each block contained three different trial types: i) *Targets*: the two stimuli being detected (e.g., the specific face and object identities); ii) *Task Relevant Stimuli*: all other stimuli from the task relevant categories (e.g., the non-target faces/objects); and iii) *Task Irrelevant Stimuli*: all stimuli from the two other categories

(e.g., letters/false fonts). An advantage of this design is that the three trial types enabled a differentiation of neural responses related to task goal, task relevance, and simply consciously seeing a stimulus.

Stimuli were presented for one of three durations (0.5 s, 1.0 s or 1.5 s), followed by a blank period of a variable duration to complete an overall trial length fixed at 2.0 s. For the MEG and iEEG version, random jitter was added at the end of each trial (mean inter-trial interval of 0.4 s jittered 0.2-2.0 s, truncated exponential distribution) to avoid periodic presentation of the stimuli. The mean trial length was 2.4 s. For the fMRI protocol, timing was adjusted as follows: the random jitter between trials was increased (mean inter-trial interval of 3 s, jittered 2.5-10 s, with truncated exponential distribution), with each trial lasting approximately 5.5 s. This modification helped avoid non-linearities in BOLD signal which may impact fMRI decoding [165]. Second, to increase detection efficacy for amplitude-based analyses, three additional baseline periods (blank screen) of 12 s each were included per run (total = 24). The identity of the stimuli was randomized with the constraint that they appeared equally across durations and tasks conditions.

Subjects were further instructed to maintain central fixation on a black circle with a white cross and another black circle in the middle throughout each trial (see Figure 2.1g).

Trial counts

The MEG study consisted of 10 runs containing 4 blocks each with 34-38 trials per block, 32 non-targets (8 per category) and 2-6 targets, for a total of 1,440 trials. The same design was used for iEEG, but with half the runs (5 runs total), resulting in a total of 720 trials. For fMRI, there were 8 runs containing 4 blocks each with 17-19 trials per block, 16 non-targets (4 per category) and 1-3 targets, for a total of 576 trials. Rest breaks between runs and blocks were included.

Data Acquisition

Behavioral data acquisition

The task was run on Matlab (PKU: R2018b; DCCN, UB and Yale: R2019b; Harvard: R2020b; NYU: R2020a, WU: 2021a) using Psychtoolbox v.3 [166]. The iEEG version of the task was run on a Dell Precision 5540 laptop, with a 15.6" Ultrasharp screen at NYU and Harvard and on a Dell D29M PC with an Acer 19.1" screen in WU. Participants responded using an 8-button response box (Millikey LH-8; response hand(s) varied based on the setting in the patient's room). The MEG version was run on a custom PC

at UB and a Dell XPS desktop PC on PKU. Stimuli were displayed on a screen placed in front of the subjects with a PROPixx DLP LED projector (VPixx Technologies Inc.). Subjects responded with both hands using two 5-button response boxes (NAtA or SINORAD). The fMRI version was run on an MSI laptop at Yale and a Dell Desktop PC at DCCN. In DCCN, stimuli were presented on an MRI compatible Cambridge Research Systems BOLD screen 32" IPS LCD monitor, and in Yale they were presented on a Psychology Software Tools Hyperion projection system to project stimuli on the mirror fixed to the head coil. Subjects responded with their right hand using a 2x2 Current Designs response box at Yale and a 1x4 Current Designs response box at DCCN.

Eye tracking data acquisition

For the iEEG setup, eye tracking and pupillometry data were collected using a EyeLink 1000 Plus in remote mode, sampled monocularly at 500 Hz (from the left eye at WU, and depending on the setup at Harvard), or on a Tobii-4C eye-tracker, sampled binocularly at 90 Hz (NYU). The MEG and fMRI labs used the MEG and fMRI compatible EyeLink 1000 Plus Eye-tracker system (SR Research Ltd., Ottawa, Canada) to collect data at 1000 Hz. For MEG, eye tracking data were acquired binocularly. For fMRI, data were acquired monocularly from either the left or the right eye, in DCCN and Yale, respectively. For all recordings, a nine-point calibration was performed (besides Harvard, where thirteen-point calibration was used) at the beginning of the experiment, and recalibrated as needed at the beginning of each block/run.

iEEG data acquisition

Brain activity was recorded with a combination of intracranially subdural platinum-iridium electrodes embedded in SILASTIC sheets (2.3 mm diameter contacts, Ad-Tech Medical Instrument and PMT Corporation) and/or depth stereo-electroencephalographic platinum-iridium electrodes (PMT Corporation; 0.8-mm diameter, 2.0-mm length cylinders; separated from adjacent contacts by 1.5 to 2.43 mm), or Behnke-Fried depth stereo-electroencephalographic platinum-iridium electrodes (Ad-Tech Medical, BF08R-SP21X-0C2, 1.28 mm in diameter, 1.57 mm in length, 3 to 5.5 mm spacing). Electrodes were arranged as grid arrays (either 8 × 8 with 10 mm center-to-center spacing, 8 x 16 contacts with 3 mm spacing, or hybrid macro/micro 8 x 8 contacts with 10 mm spacing and 64 integrated microcontacts with 5 mm spacing), linear strips (1 × 8/12 contacts), depth electrodes (1 × 8/12 contacts), or a combination thereof. Recordings from grid, strip and depth electrode arrays were done using a Natus Quantum amplifier (Pleasonton, CA) or a Neuralynx Atlas amplifier (Bozeman, MT). A total of 4057 electrodes (892 grids, 346 strips, 2819 depths) were implanted across 32 patients with drug-resistant

focal epilepsy undergoing clinically motivated invasive monitoring. 3512 electrodes (780 grids, 307 strips, 2425 depths) that were unaffected by epileptic activity, artifacts, or electrical noise were used in subsequent analyses. To determine the electrode localization for each patient, a post-operative computed tomography scan and a pre-operative T1 MRI were acquired and co-registered.

MEG data acquisition

MEG was acquired using a 306-sensor TRIUX MEGIN system, comprising 204 planar gradiometers and 102 magnetometers in a helmet-shaped array. The MEG gantry was positioned at 68 degrees for optimal coverage of frontal and posterior brain areas. Simultaneous EEG was recorded using an integrated EEG system and a 64-channel electrode cap (EEG data is not reported here, but is included in the shared dataset). During acquisition, MEG and EEG data were bandpass filtered (0.01 and 330 Hz) and sampled at 1000 Hz. The location of the head fiducials, the shape of the head, the positions of the 64 EEG electrodes and the head position indicator (HPI) coil locations relative to anatomical landmarks were collected with a 3-D digitizer system (Polhemus Isotrack). ECG was recorded with a set of bipolar electrodes placed on the subject's chest. Two sets of bipolar electrodes were placed around the eyes (two at the outer canthi of the right/left eyes and two above/below the center of the right eye) to record eye movements and blinks (EOG). Ground and reference electrodes were placed on the back of the neck and on the right cheek, respectively. Subjects' head position on the MEG system was measured at the beginning and end of each run, and also before and after each resting period, using four HPI coils placed on the EEG cap, next to the left and right mastoids and over left and right frontal areas.

Anatomical MRI data acquisition

For source localization of the MEG data with individual realistic head modeling, a high resolution T1-weighted (T1w) MRI volume (3T Siemens MRI Prisma scanner) was acquired per subject. Anatomical scans were acquired either with a 32-channel coil (TR/TE = 2000/2.03ms; TI = 880 ms; 8° flip angle; FOV = 256×256×208 mm; 208 slices; 1 mm isotropic voxels, UB) or a 64-channel coil (TR/TE = 2530/2.98ms; TI = 1100 ms; 7° flip angle; FOV = 224*256*192mm, 192 slice, 0.5*0.5*1mm voxels, PKU). The FreeSurfer standard template was used (fsaverage) for participants lacking an anatomical scan (N=5).

fMRI data acquisition

MRI data were acquired using a 32-channel head coil on a 3T Prisma scanner. A session included high-resolution anatomical T1w MPRAGE images (GRAPPA acceleration factor = 2, TR/TE = 2300/3.03 ms, 8° flip angle, 192 slices, 1 mm isotropic voxels), and a

whole-brain T2*-weighted multiband-4 sequence (TR/TE = 1500/39.6 ms, 75° flip angle, 68 slices, voxel size 2 mm isotropic, A/P phase encoding direction, FOV = 210 mm, BW = 2090 Hz/Px). A single band reference image was acquired before each run. To correct for susceptibility distortions, additional scans using the same T2*-weighted sequence, but with inverted phase encoding direction (inverted RO/PE polarity) were collected while the subject was resting at multiple points throughout the experiment.

Preprocessing and analysis details

For readability, we first detail the preprocessing protocols for each of the modalities (iEEG, MEG, and fMRI) separately. Then, we describe the different analyses, combining information across the modalities, while noting any differences between them.

iEEG preprocessing

Data were converted to BIDS [167] and preprocessed using MNE-Python version 0.24 [168], and custom-written functions in Python and Matlab. Preprocessing steps included downsampling to 512 Hz, detrending, bad channel rejection, line noise and harmonic removal, and re-referencing. Electrodes were re-referenced to a Laplacian scheme [169] while bipolar referencing was used for electrodes at the edge of a strip, grid or sEEG and the signal was localized at the midpoint (Euclidean distance) between the two electrodes. Electrodes with no direct neighbors were discarded. Seizure onset zone electrodes, those localized outside the brain, and/or containing no signal or high amplitude noise level were discarded. Line noise and harmonics were removed using a one pass, zero-phase non-causal band-stop FIR filter.

The high gamma power (HG, 70-150 Hz) was obtained by bandpass filtering the raw signal in 8 successive 10 Hz wide frequency bands, computing the envelope using a standard Hilbert transform, and normalizing it (dividing) by the mean power per frequency band across the entire recording. To produce a single HG envelope time-series, all frequency bands were averaged together [170]. Most analyses focused on the HG power as it closely correlated with neural spiking activity [171] and with the BOLD signal [122]. To obtain the Event Related Potentials (ERPs), the raw signal was low pass filtered at 30 Hz with a one pass, zero-phase non causal low pass FIR filter. Epochs were segmented between 1 s pre-stimulus until 2.5 s post-stimulus of interest.

Surface reconstruction and electrode localization

Electrode positions were determined based on a computed tomography scan coregistered with a pre-implant T1 weighted MRI. A three-dimensional reconstruction of each patient's brain was computed using FreeSurfer

(http://surfer.nmr.mgh.harvard.edu). For visualization, the individual subject's electrode positions were converted to Montreal Neurological Institute (MNI)152 space. As each theory specified a set of anatomical regions of interest (ROIs), after electrode localization, electrodes were labeled according to the Freesurfer based Destrieux atlas segmentation [172,173] and/or Wang atlas segmentation [174].

Identification of task responsive channels

To identify task responsive electrodes, we computed the Area Under the Curve (AUC) for the baseline (-0.3-0 s) and the stimulus-evoked period (0.05-0.35s) separately for the task relevant and irrelevant conditions, and compared them per electrode using a Wilcoxon sign-rank test, corrected for False Discovery Rate (FDR [175]). A Bayesian t-test [176] was used to quantify evidence for non-responsiveness.

Identification of category selective channels

To determine category selectivity for faces, objects, letters and false fonts on the HG, we followed the method of Kadipasaoglu and colleagues [177]. Per category, we computer a d' (AUC, 0.05 -0.4 s) comparing the activation between the category-of-interest (u_j) and each of the other categories (u_i), normalized by the standard deviation of each category:

$$d' = \frac{u_j - \frac{1}{N}\sum_i^N u_i}{\sqrt{\frac{1}{2}(\sigma_j^2 + \frac{1}{N}\sum_i^N \sigma_i^2)}}; \; i \neq j$$

A permutation test (10,000 permutations) was used to evaluate significance. d' was computed for the task relevant and irrelevant conditions, separately. An electrode was considered selective if it showed selectivity on both tasks.

Multivariate analysis electrodes combination

Due to the sparse and highly variable coverage of iEEG data, all collected electrodes were combined into a "super subject" multivariate analyses (RSA and decoding). To create a single trial matrix for the super subject, we equated the trial matrices of all our subjects by subsampling to the lowest number of trials in the relevant conditions. Subjects that did not complete the full experiment were discarded (N=3), resulting in a total of 29 subjects with 583 electrodes in posterior and 576 electrodes in prefrontal ROIs, respectively. In the case of analyses on stimuli identities, stimuli that were presented less than three times to any of the participants across intermediate and long trials in the task relevant and irrelevant trials were discarded. We then subsampled the trials for each identity to three trials per participant. The subsampling procedure

was repeated 100 times to avoid random fluctuation induced by the subsampling. The analysis was computed for each repetition and average across repetitions.

MEG preprocessing

The MEG data were converted to BIDS [178] using MNE-BIDS [179], and preprocessed following the FLUX Pipeline [180] in MNE-Python v0.24.0 [168]. Preprocessing steps included MEG sensor reconstruction using a semi-automatic detection algorithm and Signal-Space Separation (SSS) [181] to reduce environmental artifacts. FastICA [182] was used to detect and remove cardiac and ocular components from the data for each subject (M=2.90 components, SD=0.92). Prior to ICA, data were segmented, and segments containing muscle artifacts were removed. After preprocessing, data were epoched into a 3.5 s segment (1 s pre-stimulus to 2.5 s post-stimulus onset). Trials where gradiometers values exceeded 5000 fT/cm, magnetometers exceeded 5000 fT, and/or contained muscle artifacts were rejected from the MEG dataset. Finally, to be included in the analyses, participants should have a minimum of 30 clean trials per condition. No participants were excluded because of not meeting this criterion.

Source modeling

MEG source modeling was performed using the dynamic statistical parametric mapping (dSPM) method [183], based on depth-weighted minimum-norm estimates (MNE [184,185]), on epoched and baseline (-0.5 s to 0 s prior to stimulus onset) corrected data. To build a forward model, the MRI images were manually aligned to the digitized head shape. A single shell Boundary Elements Model (BEM) was constructed in MNE-Python based on the inner skull surface derived from FreeSurfer [172,173], to create a volumetric forward model (5 mm grid) covering the full brain volume. The lead field matrix was then calculated according to the head-position with respect to the MEG sensor array. A noise covariance matrix for the baseline and a covariance matrix for the active time window were calculated and the combined (i.e., sum) covariance matrix was used with the forward model to create a common spatial filter. Data were spatially pre-whitened using the covariance matrix from the baseline interval to combine gradiometer and magnetometer data [186].

fMRI Preprocessing

Source DICOM data were converted to BIDS using BIDScoin v3.6.3 [187]. This includes converting DICOM data to NIfTI using dcm2niix [188] and creating event files using custom Python codes. BIDS compliance of the resulting dataset was controlled using BIDS-Validator. Subsequently, MRI data quality control was performed using MRIQC [189] and custom scripts for data rejection. All (f)MRI data were preprocessed using

fMRIPrep 20.2.3 [190], based on Nipype 1.6.1[191]. For further details on the fMRIprep pipeline, see preregistration.

Analysis-specific functional preprocessing

Additional, analysis-specific, fMRI data preprocessing was performed using FSL 6.0.2 (FMRIB Software Library; Oxford, UK [192]), Statistical Parametric Mapping (SPM 12) software [193], and custom Python scripts after the above outlined general preprocessing. Functional data for univariate data analyses were spatially smoothed (Gaussian kernel with full-width at half-maximum of 5 mm), grand mean scaled, and temporal high-pass filtered (128 s). No spatial smoothing was applied for multivariate analyses.

Contrast of parameter estimates

We modeled BOLD signal responses to the experimental variables by fitting voxel-wise General Linear Model (GLM) to the data of each run using FSL FEAT. The following regressors were modeled in an event-related approach, with event duration corresponding to the stimulus duration (i.e., 0.5, 1.0, 1.5 s), and convolved with a double gamma hemodynamic response function: 12 regressors of interest (Targets, task relevant and task irrelevant stimuli per stimulus category i.e., faces, objects, letters, false fonts; and a regressors of no interest i.e., target screen display). We included the first-order temporal derivatives of the regressors of interest, and a set of nuisance regressors: 24 motion regressors (FSL's standard + extended set of motion parameters) plus a CSF and a WM tissue regressor.

Each of the 12 regressors of interest was contrasted against an implicit baseline (used in the putative NCC analysis). Additionally, we obtained contrast of parameter estimates for 'relevant faces vs. relevant objects', 'relevant letters vs. relevant false fonts', 'irrelevant faces vs. irrelevant objects', 'irrelevant letters vs. irrelevant false fonts' (used for the definition of decoding ROIs), 'relevant and irrelevant faces vs. relevant and irrelevant objects' and 'all stimuli vs. baseline' (used for the definition of seeds for the generalized psychophysiological interaction analysis).

Data were averaged across runs per subject using FSL's fixed effects analysis and subsequently averaged across participants using FSL's FLAME1 mixed effect analysis. Gaussian random-field cluster thresholding was used to correct for multiple comparisons, using the default settings of FSL, with a cluster formation threshold of one-sided $p < 0.001$ ($z \geq 3.1$,) and a cluster significance threshold of $p < 0.05$.

Anatomical Regions-of-interest (ROIs)

ROIs were defined a priori in consultation with the adversarial theories. They were determined per subject based on the Destrieux atlas [173] including both hemispheres, and then resampled to standard MNI space (see Supplementary Table 2.26). For the connectivity analysis, areas V1/V2 (combining dorsal and ventral) were defined based on the Wang cortical parcellation [174]. For details on the process of selecting the ROIs and the justification of the ROIs selection in the context of this study, see supplemental section 11.

Behavioral analyses

Log-linear corrected d'prime [194], false alarms (FA) and reaction times (RT) were computed per category and stimulus duration, separately (FAs were also calculated per task relevance, without duration), and per modality (iEEG, MEG, fMRI). These measures were compared with Linear/Logistic mixed models, where appropriate. For the former, we report ANOVA omnibus F tests, and for the latter, omnibus χ^2 test from an analysis of deviance. We approximated degrees of freedom using the Satterthwaite method [195]. Pairwise t-tests following significant interactions were Bonferroni corrected. To estimate Bayesian Information Criterion (BIC) differences between the original and null logistic models, we used the p-values and sample size ([196]; p_to_bf package in R).

Eye-tracking analyses

For Eyelink, gaze and pupil data were segmented, and trials with missing data were excluded. Blinks were detected using the Hershman algorithm [197], and removed with 200 ms padding [198]. The Eyelink standard parser algorithm was used for saccade and fixation detection. Saccades were further corroborated using the Engbert & Kliegl [199] algorithm. Fixations were baseline corrected (-0.25 s to 0 s). Mean fixation distance, mean blink rate, mean saccade amplitude and mean pupil size were compared in a Linear Mixed Model (LMM) with category and task relevance as fixed effects and subject and item as random effects. Separate analyses were carried out on the first 0.5 s after stimulus onset including all trials; and on the 1.5 s trials including time window (0-0.5 s, 0.5-1.0 s, 1.0-1.5 s) as fixed effects. BIC was used to test the models against the null hypothesis models. For Tobii, gaze coordinate data was segmented, missing data were excluded, and coordinates were baseline corrected to depict heatmaps of patients' gaze. Notably, the coordinate data was not added to the LMMs due to its poorer quality with respect to the EyeLink data.

Decoding analysis

All decoding analyses were performed using a linear Support Vector Machine (SVM, scikit learn, https://scikit-learn.org/) classifier. Below we explain how this was done for each one of the predictions.

iEEG Decoding was done on the HG response, averaged over non-overlapping windows of 0.02 s separately for electrodes located in the GNWT and IIT ROIs. The top 200 electrodes (selectKbest [200]), as determined by F-test within a given set of electrodes from the theory ROIs, were used as features for the classifier. 200 features were selected to provide a balance between model optimization (e.g., feature selection) and subject representation (e.g., electrodes/features coming from multiple subjects). Statistical significance of decoding performance was assessed via permutation test, randomly permuting the sample labels and repeating the decoding analysis 1000 times, corrected for multiple comparisons using a cluster-based correction (cluster mass inference with cluster forming threshold at $p < 0.05$ [201,202]). Also, to assess the decoding accuracy within unique ROIs (e.g., S_temporal_sup of the Destrieux atlas), separate classifiers were trained using all electrodes in a given parcel. Each classifier was fitted using all electrodes in a parcel and time window (GNWT: 0.3-0.5 s, IIT: 0.3-1.5 s) as features, resulting in a single accuracy value per parcel. SelectKbest (200 features iEEG) feature selection and 5-fold cross-validation with 3 repetitions was used. To assess the statistical significance of the decoding accuracy within unique ROIs (so only one accuracy score is obtained per ROI), p-values obtained via permutation tests were corrected for multiple comparisons across all ROIs using FDR correction ($q \leq 0.05$ [175]).

MEG Decoding was done on bandpass filtered (1-40 Hz) and downsampled (100 Hz) data. The reconstructed source-level MEG data within a subset of the predefined anatomical ROIs (GNWT: 'G_and_S_cingul-Ant','G_and_S_cingul-Mid-Ant', 'G_and_S_cingul-Mid-Post', 'G_front_middle','S_front_inf', 'S_front_sup', IIT: 'G_cuneus', 'G_oc-temp_lat-fusifor', 'G_oc-temp_med-Lingual','Pole_occipital', 'S_calcarine','S_oc_sup_and_transversal', as they show high response to the stimulus on the optimization dataset) were extracted for further analysis (500 vertices and 800 vertices per hemisphere for each of the anatomical ROI defined by the theories). We applied temporal smoothing (0.05 s window, 0.01 sliding window), computed pseudotrials [203], normalized the data, and selected the top 30 features within a given ROI as features for the different classifiers. A group-level one-sample t-test per time point was performed on the decoding accuracy results, corrected for multiple comparisons using a cluster-based correction [201].

The overall decoding strategy for fMRI was similar to that used on the iEEG and MEG data, yet with some differences. A Multi-Variate Pattern Analysis (MVPA) approach was used on the pattern of BOLD activity over voxels. A non-spatially-smoothed parameter estimate map was obtained by fitting a GLM per event with that event as the regressor of interest and all the other remaining events as one regressor of no interest [204] as implemented in NiBetaSeries 0.6.0 package. The model also included the 24 nuisance regressors described in the fMRI preprocessing section.

Decoding was performed using a whole-brain approach and an ROI-based approach. The whole-brain analysis was performed using a searchlight approach with 4 mm radius. For ROI-based decoding, decoding ROIs were defined based on functional fMRI contrasts (see fMRI preprocessing section) and constrained with pre-defined anatomical ROIs (see Extended Data Table 2.2: Anatomical Regions-of-interest (ROIs)). One-sample permutation test was used to determine if decoding significantly exceeds chance level within each ROI. FDR was used to correct for multiple comparisons across ROIs. For whole-brain decoding, a cluster-based permutation test was used to evaluate the decoding statistical significance across subjects ($p < 0.05$). Additionally, stimulus vs. baseline searchlight decoding was performed using leave-one-run out cross validation and the resultant decoding accuracy maps were used as input for the multivariate putative NCC analysis (see below). To perform stimulus vs. baseline decoding, we subsampled the stimuli trials to a 2:1 ratio with respect to baseline. The SVM cost function was weighted by the number of trials from each class.

Decoding schemes for the different predictions

To test GNWT and IIT decoding predictions, stimulus category (faces vs. objects and letters vs. false fonts) was decoded separately for the task relevant and task irrelevant conditions (*within-task category decoding*) while orientation (front view vs. left view vs. right view) was decoded on the combined data from the two task conditions. In addition, *cross-task category decoding* from task relevant to task irrelevant condition and vice versa was performed to test generalization by training classifiers on one condition and testing on the other condition. Both within-task category and orientation decoding were performed in a leave-one-run-out cross validation scheme for fMRI and in an k-fold cross validation scheme for MEG and iEEG.

For *category decoding*, trials from each task condition (i.e., task relevant, irrelevant) were extracted for each category comparison of interest: 160 face/160 objects classification, 160 letters/160 false fonts classification within each task relevance condition for MEG, and half the trials for iEEG. For fMRI, there were 64 trials for

each category in each task relevance condition. For *orientation decoding*, task relevant and task irrelevant trials were collapsed within category to increase Signal-to-Noise Ratio (SNR), resulting in 160 Front, 80 Left, and 80 Right trials per category for MEG, and half these numbers for iEEG. For fMRI, there were 64 Front, and 32 Left and Right trials per category. Decoding was evaluated using accuracy measures, tested against 50% chance level for category decoding (binary classification) and against 33% chance level for orientation decoding (3-class classification). For orientation decoding, balanced accuracy was used due to the unbalanced number of trials for the different orientations. The SVM cost function was weighted by the number of trials per class to reduce bias to the class with the highest number.

For *within-task decoding* (e.g., classification of categories across time), a classifier at each time-point was trained and tested separately using a 5-fold cross-validation (with 3 separate repeats of cross-validation). For cross-task decoding (task relevant -> irrelevant & task irrelevant -> relevant), each SVM model was trained on one task (e.g., faces/objects in the task relevant condition) and tested on the second task (e.g., faces/objects in the task irrelevant one). As cross-decoding in iEEG data is performed across all pooled electrodes, an additional cross-validation step was performed on this modality data to provide a confidence metric (e.g., confidence intervals) using a 5-fold cross-validation with 3 repetitions (e.g. train on 80% of task 1, and test on held-out 20% of task 2).

Within-task temporal generalization was performed by training a classifier at each time-point (using selectKbest feature selection) and testing its performance across all time-points using the same set of selected features and 3 repetitions of 5-fold cross-validation. To generalize from one task to another across all time-points, cross-temporal generalization was used: a classifier was trained at each time-point in task 1 (e.g., task relevant) using selectKbest feature selection, and tested across all time-points in task 2 (e.g., task irrelevant) using the same set of selected features. Cross-validation was performed in the same fashion as in cross-decoding.

Additional decoding analyses were performed on all trials aligned to the stimulus onset (e.g. -0.2-2 s relative to stimulus onset), and stimulus offset (-0.5-0.5 s around stimulus offset). For the latter analysis, all trials from different durations were aligned to the stimulus offset.

To assess the specific IIT prediction that including prefrontal regions along with posterior regions to the decoding of categories will not significantly affect decoding accuracy, we performed two additional decoding analyses in which the decoding

performance of electrodes from the IIT region were compared with the decoding performance when electrodes from both the posterior + PFC ROIs are included. The PFC ROI included all PFC ROIs, except for inferior frontal sulcus, as it belongs to the IIT extended ROIs. Posterior ROI included all IIT ROIs shown in Extended Data Table 2.2. The first analysis compared the decoding accuracy for a model including all electrodes from posterior regions to a separate model in which electrodes (features) from posterior & PFC regions were combined (e.g., feature combination). In the second analysis, the decoding accuracy of the model including all electrodes from posterior regions was compared to a combined posterior + PFC model, in which two separate classifiers were trained and calibrated on posterior & PFC regions separately using isotonic calibration [205], and posterior probabilities from each classifier were combined using a softmax normalization [206]. Training and testing of the individual models followed all previously described cross-validation procedures and model comparison was performed using a variance-corrected paired t-test [207] and complemented with Bayesian analysis. Following Benavoli and colleagues [208], the prior distribution of the mean difference in decoding scores between two classifier models was modeled as a Normal-gamma distribution conjugate to a normal likelihood, and the posterior distribution was obtained as a normal distribution. This posterior distribution was utilized to calculate the probability of one classification model being better than, worse than, or equivalent to the other model. As this estimation approach is applied using resampled datasets (e.g., using 5-fold cross-validation), the performance of the model becomes dependent on the folds, and thus a variance corrected t-distribution was used [207].

We also tested this prediction on the fMRI data. To select features to be used for both analyses, the face vs. object contrast for each subject was masked by a predefined anatomical posterior ROIs as well as a PFC anatomical ROIs, defined the same way as described above. Within each of the two ROIs, the 150 voxels that are most selective to each of the to-be-decoded stimuli were defined as the decoding ROIs (300 voxels total) for each subject. The first analysis compared the decoding accuracies for a model that included 300 voxels from the posterior ROIs as features to another model that included 600 voxels (300 features from each ROI). In the second analysis, two separate models were constructed, calibrated, and combined as described above. For the two analyses, model comparison was performed using a group-level one-sample permutation test to determine if accuracies obtained by combining posterior and PFC ROIs are significantly higher than the accuracies obtained based on posterior ROIs only. FDR was used to correct for multiple comparisons.

Duration analysis

Neural responses were extracted from three windows of interest (WoI) (0.8-1.0 s, 1.3-1.5 s, 1.8 -2.0 s) and compared using LMM. Four theory agnostic models were fitted: a null model, a duration model (3 durations), a WoI model, and a duration and WoI model. Two theory model were fitted: the GNWT model predicts activation (ignition) following stimulus offset (0.3-0.5 s) independent of duration, with virtually no response in between. The IIT model predicts sustained activation for the duration of the stimulus returning to baseline after stimulus offset. Both theoretical models were complemented with an interaction term between category (faces, objects, letters and false fonts) and the theories' predictors, to account for regions showing selective responses to categories. Bayesian Integration Criterion (BIC) was used to define the winning model.

Models for iEEG were fitted per electrode on the predefined ROIs, using the HG (AUC), alpha (8-13 Hz, obtained through Morlet wavelets, f=8-13 Hz, in 1 Hz steps; f/2 cycles, AUC), and ERPs (peak to peak) as signal, separately for task relevant and irrelevant condition.

MEG models were fitted to source data on the predefined ROIs, using the gamma (60-90 Hz) and alpha band (8-13 Hz) as signal, separately for task relevant and irrelevant conditions. Time-frequency analyses were performed on source-data using Morlet wavelets (f=8-13 Hz, in 1 Hz steps; f/2 cycles; f=60-90 Hz, in 2 Hz steps, f/4 cycles), and were baseline corrected. Spectral activity was computed for each vertex, baseline corrected and then averaged across trials within each parcel included in the ROIs, yielding a unique time-course per ROI parcel. In addition, a single source time-course capturing the entire prefrontal ROI and the posterior ROI was computed by averaging the spectral activity within an ROI. Models were fitted on each parcel and ROI, as defined by the theories.

Representational Similarity Analysis (RSA)

To examine how the neural representations evolved over time in response to the different stimulus properties (i.e., category, orientation and identity representation), we performed cross-temporal RSA on source level MEG data and iEEG HG power within each of the theory-defined ROIs, using all trials (see supplementary section 12). Specifically, at each set of data points, we computed a Representational Dissimilarity Matrix (RDM) by calculating the correlation distance (1- Pearson's r, Fisher corrected) between all pairs of stimuli (the preregistration document described a different method which was however updated to optimize trial numbers, see supplementary section 12 for a justification). Next, to quantify the representational space occupied

by one class vs. another, we computed the average within-class distances vs. the average between-class distances. This analysis was performed in a cross-temporal manner, in which RDMs were computed between all stimuli at time point t1 and the corresponding set of stimuli at time points t1,2,…n.

Long trials (1.5 s) were used to investigate category and orientation representation. Since specific identities were repeated a limited number of times per duration, both intermediate (1.0) and long (1.5 secs) trials were combined and equated in duration by cropping the 1-1.5s time interval for long trials. This was done to allow for the analysis of at least three (3) presentations of the same identity.

To evaluate the theoretical predictions about when significant content representation should occur, we subsampled the observed cross-temporal representational matrices in four time windows (0.3-0.5, 0.8-1.0, 1.3-1.5, 1.8-2.0 s). The subsampled matrices were correlated to the model matrices predicted by GNWT and IIT (see Figure 2.1a, right panel) using Kendall's Tau correlation. If the correlation was significant (see below) for at least one of the predicted matrices, we computed the difference between the transformed correlation () to each theory; and compared this difference against a random distribution to obtain a p-value. If the correlation with the theory predicted pattern in the theory ROI was significantly higher than the other model, we considered the theory prediction to be fulfilled.

To generate a null distribution of cross-temporal RSA surrogate matrices, we repeated the procedure outlined above 1024 times, randomly shuffling the labels. Next, the observed RSA matrix was z-scored using the null distribution as:

$$z_{i,j} = \frac{obs_{i,j} - \mu_{surr_{i,j}}}{\sigma_{surr_{i,j}}}$$

Where is the observed within-vs.-between class difference at time points i and j, and,and are the mean and standard deviation of the surrogate representational similarity matrix at time points i and j, respectively. Cluster based permutation tests [209], z-score threshold of z = 1.5 for clustering, were used to evaluate significance. RSA surrogates were also used to assess the significance of the correlation between the observed matrices and the theories' predicted matrices. First, a null distribution of possible correlations was generated for each of the theories by correlating each of the surrogate matrices to each of the theory predicted matrices. Next, a p-value was obtained for each theory predicted matrix, by locating its observed correlation

within the null correlation distribution. The same procedure was used to assess the significance of the difference in correlation to IIT and GNWT matrices (e.g., each of the surrogate matrices was correlated to each of the theory predicted matrices and the difference between the two was computed). P-values were FDR corrected ($q \leq 0.05$) [175].

For iEEG, the HG power per electrode within the predefined anatomical ROI was averaged in 0.02s non-overlapping windows. Electrodes were used as features for the RDM. The data were vectorized across all electrodes within a ROI (e.g., samples x significant electrodes) to compute the RDMs. 576 and 583 electrodes entered this analysis for the prefrontal and posterior ROI, respectively. The resultant RDM was subjected to a principal component analysis and the first two dimensions were plotted against each other to produce a 2-dimensional projection of dissimilarity scores across all pairs for each of the 100 subsampling repetitions. The PCA components were aligned across repetitions using Procrustes alignment and averaged together for visualization purposes [210,211].

For MEG, the same analysis was run on the source reconstructed data within the predefined anatomical ROIs used for the Decoding analysis, bandpass filtered (1-40 Hz) and downsampled (100 Hz). For the category and orientation analysis, pseudo-trials and temporal moving-average methods were used to optimize the RSA analysis and improve the SNR. For identity, single trials were used. Vertices within the ROIs were used as features. The statistical testing differed from that conducted on the iEEG data, as it was performed at the subject level. Like the iEEG analysis, we first tested if the correlation between the data and the model predicted by each theory was greater than zero using the Kendall's tau measure, and then compared between the theories using the Mann-Whitney U rank test on two independent samples.

Functional Connectivity analysis

For both iEEG and MEG, pairwise phase consistency (PPC [130]) was computed between each category-selective time series (face- and object-selective) and either the V1/V2 or the PFC time series.

For iEEG, the PPC analysis included electrodes in V1/V2 visual areas, in PFC ROIs (see Extended Data Table 2.2), and face and object selective electrodes (see *Identification of task responsive channels*), as long as they were "active" during the task. As both theories predict different types of activation (e.g., ignition vs. sustained activation), channels were categorized as active if they showed an increase in HG power relative to baseline (-0.5 to -0.3 s, $p<0.05$, signed-rank test) evaluated across all trials (task relevant +

irrelevant, intermediate + long trials, combined across both categories), for the 0.3-0.5 s window (GNWT), or in all time windows 0.3-0.5 s, 0.5-0.8 s, and 1.3-1.5 s (IIT).

For MEG, the category-selective single-trial time courses used to define the ROIs for PPC analysis were extracted using the Generalized Eigenvalue Decomposition (GED) method [132]. Two GED spatial filters were built by contrasting either faces or objects against all other categories during the first 0.5 s after stimulus onset. Single-trial covariance matrices were computed separately for signal and reference for all vertices within the fusiform ROI identified from the FreeSurfer parcellation using the Desikan atlas [212], and the Euclidean distance between them was z-scored. Trials exceeding 3 z-scores were excluded. The reference covariance matrix was regularized to reduce overfitting and increase numerical stability. The GED was then performed on the two covariance matrices, resulting in N (= rank of the data) pairs of eigenvectors and eigenvalues. The eigenvector associated with the highest eigenvalue was selected as a GED spatial filter, which in turn was applied to the data to compute the single-trial GED component time series. A GED spatial filter was extracted also for the PFC ROI, on parcels from the Destrieux atlas [173], to identify the distributed pattern of sources that are responsive to visually-presented stimuli. Specifically, a spatial filter was built by contrasting source-level frontal slow-frequency activity (30-Hz low-pass filter) after stimulus onset (0 to 0.5 s) against baseline (-0.5 to 0 s). V1/V2 areas were identified using the Wang Atlas [174] and a singular values-decomposition approach. For the GED, the 1.0 and 1.5 s duration trials were used to minimize overlap with the transient evoked at stimulus onset.

PPC was computed for each MEG time series/iEEG electrode pairing, for all face-trials and object-trials separately. Analyses were performed on 1.0 and 1.5 duration trials, separately on task relevant and irrelevant trials and also combined to maximize statistical power. To compute synchrony, time-frequency analysis of the broadband MEG and LFP signal was performed using Morlet wavelets (f=2-30 Hz, in 1 Hz steps; 4 cycles; f=30-180 Hz for iEEG or f=30-100 Hz for MEG, in 2 Hz steps, f/4 cycles), and PPC was then computed by taking the difference in phase angle between MEG time series/iEEG electrode at each time, t, and frequency f, for a specific trial and computing PPC across all trials in a category (e.g., faces) as:

$$PPC(f,t) = \frac{2}{(N(N-1))} \sum_{j=1}^{N-1} \sum_{k=j+1}^{N} cos\left(\theta_j(f,t) - \theta_k(f,t)\right), j = \{1 \dots N\ trials\}$$

, for all frequencies f, and at all times t.

For iEEG, PPC for each category-selective site was then averaged across all its pairings (e.g., all PFC electrodes pairings or all V1/V2 pairings within that patient). The variability in electrode coverage across patients precluded a within-subjects analysis. Therefore, to achieve sufficient statistical power, we pooled all derived PPC values from one electrode pairing (e.g., face-selective to PFC) across all patients into one ROI specific analysis. A similar approach was used on the MEG parcels.

To quantify content-specific synchrony enhancement, the difference in PPC was computed between within-category and across-category trials (e.g., for face-selective sites, the change in PPC was computed between faces vs. objects trials) using a cluster-based permutation test [201]. This was done for both modalities.

As an exploratory analysis, we also investigated dynamic functional connectivity using the Gaussian-Copula Mutual Information (GCMI [213]) approach to evaluate the dependencies between time series. This power-based measure of connectivity was implemented using the conn_dfc method from the Frites Python package [214]. We used the same parameters as for the PPC analysis, with the following exceptions: For both MEG and iEEG, power was estimated through a multitaper-based method (using a frequency dependent dynamic sliding window: 2-30 Hz, T= 4 cycles; 30-100 Hz, T4/f using a 0.25-s sliding window. For iEEG the high frequency range was extended from 30-180 Hz, T=4/f cycles). DFC was performed per frequency band, 0.1 s sliding window, 0.02s steps.

For fMRI, connectivity was assessed through generalized Psycho-Physiological Interaction (gPPI) implemented in SPM [215]. The Fusiform Face Area (FFA) and Lateral occipital cortex (LOC) were defined as seed regions per subject based on an anatomically constrained functional contrast. Anatomically, FFA seeds were constrained to the "Inferior occipital gyrus (O3) and sulcus" and "Lateral occipito-temporal gyrus (fusiform gyrus, O4-T4)". LOC seeds were constrained to the "Middle occipital gyrus (O2, lateral occipital gyrus)" and the "Middle occipital sulcus and lunatus sulcus" (Destrieux ROIs 2 and 21 for FFA and ROIs 19 and 57 for LOC, see Anatomical Regions-of-interest (ROIs)).

Candidate seed voxels within the above-mentioned anatomical ROIs were defined as those with a z value > 1 in the contrast of parameter estimates of all stimuli vs. baseline. Three subjects with less than 300 candidate seed voxels were excluded from the analysis. This was done to ensure that the seed voxels were visually driven. Next, using an unthresholded contrast of parameter estimates between 'relevant and irrelevant faces' and 'relevant and irrelevant objects', the 300 voxels most responsive

to faces within the FFA anatomical ROIs were selected for the FFA seed, and the 300 voxels most responsive to objects within the LOC anatomical ROIs were selected for the LOC seed.

gPPI analysis was performed per subject and seed region separately, including an interaction term between the seed time series regressor (physiological term) and the task regressor (psychological term) at the subject-level GLM [215], separately for task relevant and irrelevant conditions, and also combining across tasks to increase statistical power. For combined conditions, the model design matrix for each subject included regressors for task relevant and task irrelevant faces, objects, letters, and false fonts collapsed across conditions (four regressors) as well as a regressor for targets (irrespective of their category), yielding five regressors in total. As for separated conditions, the model design matrix included regressors for task relevant and task irrelevant faces, objects, letters, and falsefonts (eight regressors) as well as a regressor for targets (irrespective of their category), yielding nine regressors in total. For each seed, group level analysis was performed using a cluster-based permutation test (preferred over the preregistered FDR correction. See supplementary section 12 for a justification of this change) to evaluate the statistical significance of face > object contrast parameter estimates across subjects ($p < 0.05$; see supplementary section 12).

Putative NCC analyses

A series of conjunction analyses were performed on the fMRI data to identify a) areas responsive to task goal, b) areas responsive to task relevance, and c) areas putatively involved in the neural correlate of consciousness. We note that the contrasts proposed below might overestimate the neural correlates of consciousness and that the fast event-related design adopted here might be suboptimal to detect activity changes in the salience network [216], i.e., potentially underestimating some regions that might be involved in conscious processing. We therefore have adopted a conservative approach that distinguishes between areas that might participate in consciousness vs. those that definitely do not.

The conjunction defining *areas responsive to task* goals was defined as [TaskRelTar > bsl] & [(TaskRelNonTar = bsl) & (TaskIrrel = bsl)]. This contrast captures areas that show an increase of BOLD signal for targets but not for other stimuli. The following conjunction identified *areas responsive to task relevance*: [(TaskRelTar > bsl) & (TaskRelNonTar ≠ bsl)] & [TaskIrrel = bsl]. This contrast identifies areas displaying differential activity for all task relevant stimuli, but are insensitive to non-task relevant ones. Finally, the following conjunction was used to identify the

putative NCC areas: [(TaskRelNonTar (stim id) > bsl) & (TaskIrrel (stim id) > bsl)] OR [(TaskRelNonTar (stim id) < bsl) & (TaskIrrel (stim id) < bsl)], critically detecting areas that responsive to any stimulus category irrespective of task, with consistent activation or deactivation. Thus, this analysis casts a wide net to identify areas that can potentially be the neural correlates of consciousness, while excluding areas that do not respond to task relevant/irrelevant stimuli (meaning that areas that respond both to the task and to the content of perception are still included).

To compute conjunctions, we first ran a GLM (see above) corrected for multiple comparisons (Gaussian random-field cluster-based inference). Equivalence to baseline was established using a JZS Bayes Factor test, with a Cauchy prior (r scale value of 0.707). Evidence maps were thresholded at BF01 > 3. The thresholded z maps and the Bayesian evidence maps on the group level were used for the conjunction analysis. For conjunctions including an 'unequal to', a 'logical and' operation was used between the directional z maps, after thresholded maps were binarized. For the putative NCC contrast, conjunctions were performed separately for activations and deactivations, using a 'logical and' operator for the task relevant and irrelevant z maps. The resulting maps were combined using a 'logical or' operation to discard areas showing effects of opposite direction for task relevant and task irrelevant stimuli. This analysis was also done at the subject level, masked using the anatomical ROIs, to account for inter-subject variability. For each ROI, the proportion of subjects with voxels included in the conjunction is reported. The multivariate version of the putative NCC analysis was done using the thresholded statistical maps obtained from the whole-brain searchlight decoding based on a subject-level stimulus vs. baseline decoding accuracy maps (for details regarding the decoding approach used, see *Decoding Analysis*).

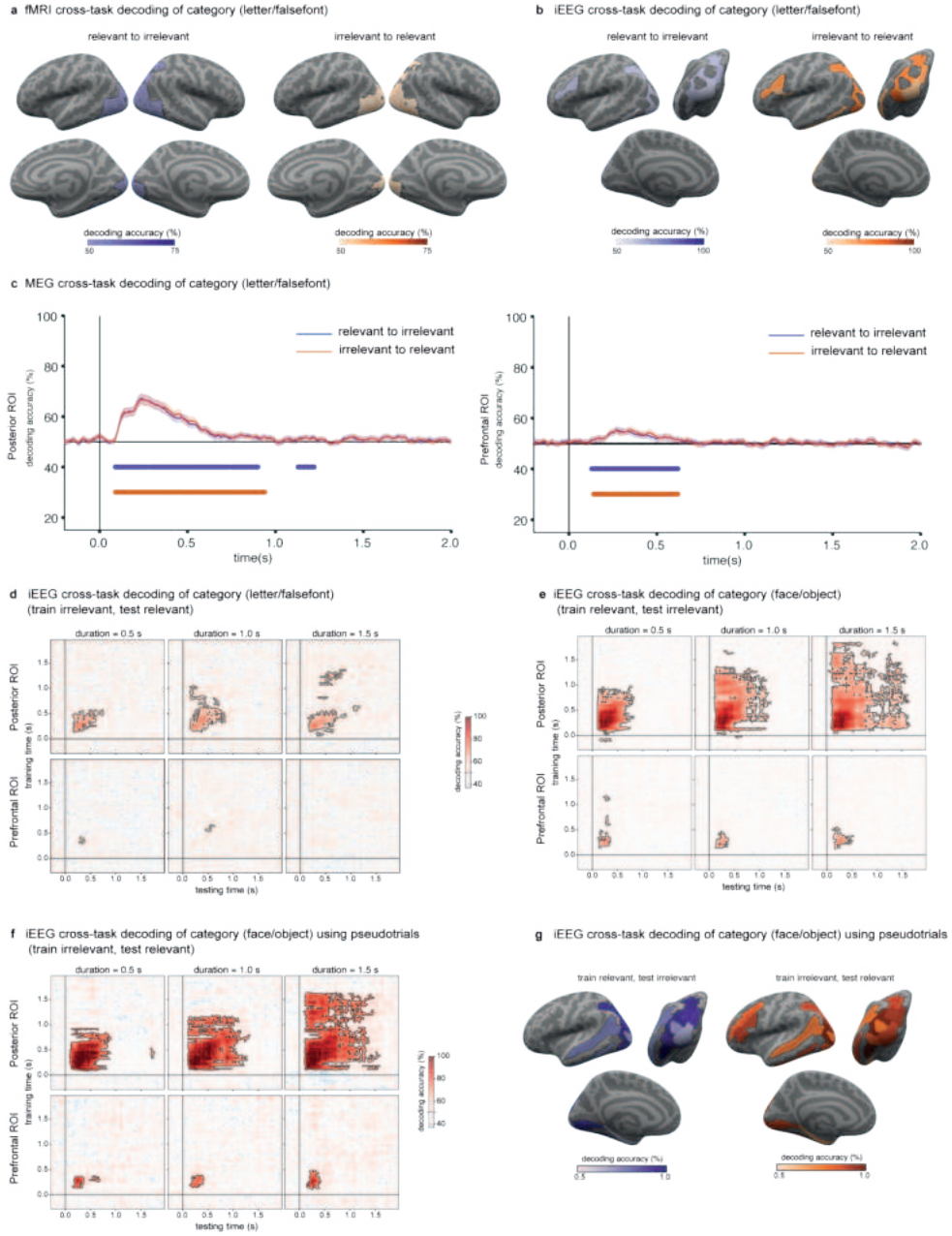
Extended Materials

GNWT predictions	IIT predictions
Prediction #1: Decoding analyses	
(A1) Cross-task generalization of decoding of ANY CATEGORY that showed decoding in the TR[1] condition in ANY PREFRONTAL areas from Task irrelevant (TI) to Task relevant (TR) OR from TR to TI, *during 300-500 ms post-stimulus onset*	*data taken from any time window***)**
(A2) Decoding of ORIENTATION for ANY category in ANY PREFRONTAL area, *during 300-500 ms post-stimulus onset*	(A2) Decoding of ORIENTATION, for ANY category in ANY POSTERIOR areas (*data taken from any time window)*
	(B1) NO Increase in decoding accuracy[6] for ANY CATEGORY that showed decoding in the TR condition when adding non-specialized frontal areas (only for task irrelevant, *data taken from any time window for posterior ROI and from 300-500 ms post-stimulus onset for frontal areas)*
	(B2) NO Increase in decoding accuracy[1] of ORIENTATION for ANY category that showed decoding when adding non-specialized frontal areas (only for task irrelevant), *data taken from any time window for posterior ROI and from 300-500 ms post-stimulus onset for frontal areas)*
A1&A2 should be TRUE for MEG OR iEEG	**B1&B2 should be TRUE for MEG OR iEEG**
Prediction #2: Activation and representational similarity analyses	
(A) Phasic ignition in ANY PREFRONTAL area at stimulus ONSET (300-500 post onset) AND OFFSET (300-500 post offset) in TI for ALL stimulus durations for at least ONE category in at least ONE measure of activation (ERP, High gamma, alpha)	**(A) Content-specific sustained activation (from 300 ms until the offset) in TI for ALL durations for at least ONE category in posterior cortical areas in at least ONE measure of activation (increased gamma, decreased alpha)**
(B) Phasic RSA during ONSET and OFFSET (300-500 ms post stimulus onset/offset) in TI for 1.0 and 1.5 durations for at least ONE content (category OR orientation OR identity) in any PREFRONTAL area	**(B) Sustained RSA (from 300 ms until the offset) in TI for 1.0 and 1.5 durations for at least ONE content (category OR orientation OR identity) in any POSTERIOR area (contingent on results of the blink control analysis)**

A should be TRUE for MEG OR iEEG	**A&B should be TRUE for MEG OR iEEG**
Prediction #3: Synchronization analyses	
(A) Stronger synchronization between PFC and FFA for faces vs. objects during the 300-500ms time window in ANY technique[2], AND the STIMULUS difference should be larger than the TASK difference	**(B) Stronger sustained (from 300 ms until the offset) synchronization between (activated) V1/V2 and FFA for faces vs. objects for ALL durations in MEG/iEEG, AND the difference in the pattern of synchronization should be more consistent with the STIMULUS than with the TASK**
(C) Stronger synchronization between PFC and LOC for objects vs. faces during the 300-500ms time window in ANY technique, AND the STIMULUS difference should be larger than the TASK difference	**(D) Stronger sustained (from 300 ms until the offset) synchronization between (activated) V1/V2 and LOC for objects vs. faces for ALL durations in MEG/iEEG, AND the difference in the pattern of synchronization should be more consistent with the STIMULUS than with the TASK**
A OR B	**A OR B**
Integration across predictions: **Prediction#1 (Decoding) AND** **Prediction#2 (Activation) AND** **Prediction#3 (Synchronization)**	**Integration across predictions:** **Prediction#2 (Activation) AND** **Prediction#2 (RSA) AND** **Prediction#3 (Synchronization)**

Extended Table 2.1: Key Predictions and Integration of Evidence across Planned analyses

Key predictions of each theory and plan for integrating outcomes across the different brain recording modalities and analyses. Each prediction (Bolded titles, light gray cells) is broken down to sub-predictions, which are then integrated together to provide the final conclusion per prediction (dark gray rows, appearing at the bottom for each prediction). Bolded predictions are the ones appearing on Figure 2.7 on the Preregistration, and are defined as the critical predictions for evaluating the theories. Numbered sub-predictions are the ones considered when integrating across sub-predictions to reach the final conclusion of each prediction (black rows). Finally, light red row denotes vertical integration across all predictions, to form the final conclusion for each theory based on its critical predictions.


a fMRI cross-task decoding of category (letter/falsefont)
relevant to irrelevant
irrelevant to relevant
decoding accuracy (%)
b iEEG cross-task decoding of category (letter/falsefont)
relevant to irrelevant
irrelevant to relevant
decoding accuracy (%)
c MEG cross-task decoding of category (letter/falsefont)
Posterior ROI
Prefrontal ROI
decoding accuracy (%)
relevant to irrelevant
irrelevant to relevant
time(s)
d iEEG cross-task decoding of category (letter/falsefont)
(train irrelevant, test relevant)
duration = 0.5 s
duration = 1.0 s
duration = 1.5 s
Posterior ROI
Prefrontal ROI
training time (s)
testing time (s)
decoding accuracy (%)
e iEEG cross-task decoding of category (face/object)
(train relevant, test irrelevant)
duration = 0.5 s
duration = 1.0 s
duration = 1.5 s
Posterior ROI
Prefrontal ROI
training time (s)
testing time (s)
f iEEG cross-task decoding of category (face/object) using pseudotrials
(train irrelevant, test relevant)
duration = 0.5 s
duration = 1.0 s
duration = 1.5 s
Posterior ROI
Prefrontal ROI
training time (s)
testing time (s)
decoding accuracy (%)
g iEEG cross-task decoding of category (face/object) using pseudotrials
train relevant, test irrelevant
train irrelevant, test relevant
decoding accuracy (%)
decoding accuracy (%)

Extended Data Figure 2.1: Prediction#1 Decoding of conscious content for letters, false fonts, faces and objects

a. *fMRI decoding accuracies (letters vs. false fonts) using a searchlight approach, collapsed across the three stimulus durations. Left: decoding for classifiers trained on task relevant and tested on task irrelevant stimuli (purple). Right: decoding for classifiers trained on task irrelevant and tested on task relevant stimuli (orange-red). Regions showing significantly above-chance (50%) decoding accuracies are indicated by the outlined colored regions on the inflated cortical surfaces (top: left/right lateral views; bottom: right/left medial views).*

b. *iEEG decoding accuracies (letters vs. false fonts) within the theory-relevant ROIs collapsed across stimulus duration. Left: decoding for classifiers trained on task relevant and tested on task irrelevant stimuli (purple). Right: decoding for classifiers trained on task irrelevant and tested on task relevant stimuli (orange-red). ROIs showing significantly above-chance (50%) decoding are displayed on inflated surface maps from a left lateral view (top left), posterior view (top right) and left medial view (bottom).*

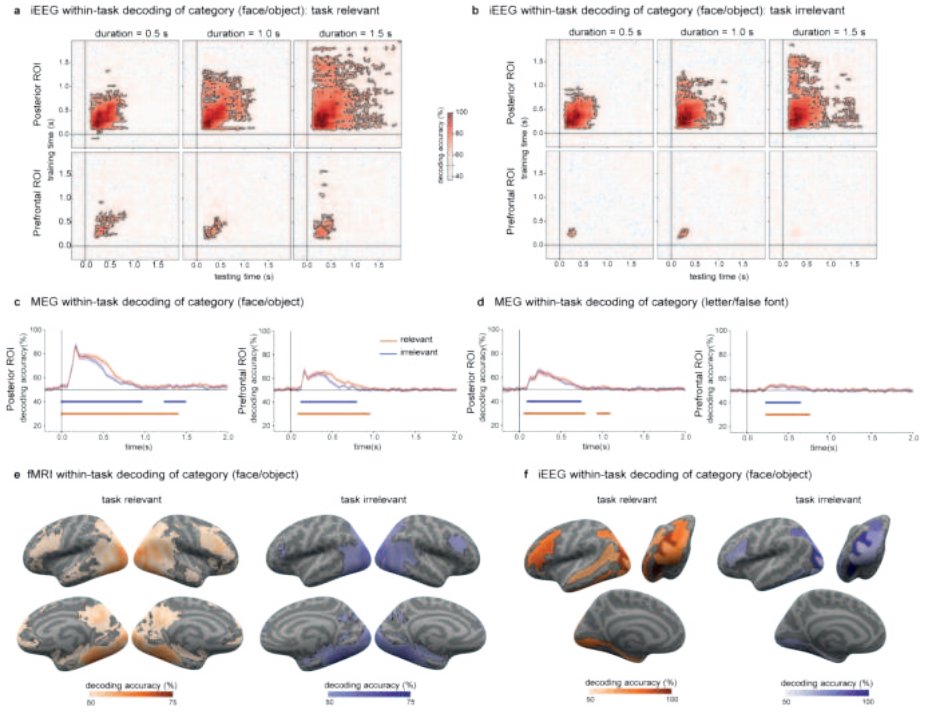
c. *MEG cross-task decoding of category for letter vs false font. (orange-red: train on test irrelevant, test on task relevant; purple: train on task relevant, test on task irrelevant). Left: results in posterior ROIs. Right: results in prefrontal ROIs. Error bars depict 95% CI estimated across subjects.*

d. *iEEG cross-task temporal generalization of category decoding (letters vs. false fonts) classifiers trained on task relevant stimuli and tested on task irrelevant stimuli. The three stimulus durations are plotted in columns (left: 0.5 s; center: 1.0 s; right: 1.5 s) and the two theory ROIs in rows (top: posterior ROIs; bottom: prefrontal ROIs). Significantly above-chance (50%) decoding is indicated by the outlined pink-red regions in the temporal generalization matrices.*

e. *iEEG cross-task temporal generalization of category decoding (faces vs. objects) in the opposite direction as in Figure 2.2b (classifiers trained on task relevant stimuli and tested on task irrelevant stimuli). Conventions as in c.*

f. *iEEG cross-task temporal generalization of category decoding (faces vs. objects), Classifiers are trained on task relevant and tested on task irrelevant stimuli. Pseudotrials are used to boost decoding accuracy. Conventions as in c.*

g. *iEEG decoding accuracies within the theory-relevant ROIs using pseudotrial aggregation to boost decoding accuracies, collapsed across stimulus duration. Conventions as in b.*



Extended Data Figure 2.2: Within-task temporal generalization of decoding of stimulus category (faces vs. objects).

a. *iEEG decoding accuracies for pattern classifiers trained and tested on task relevant stimuli. As in Figure 2.2b, the three stimulus durations are plotted in columns (left: 0.5 s; center: 1.0 s; right: 1.5 s) and the two theory ROIs in rows (top: posterior ROIs; bottom: prefrontal ROIs). Significantly above-chance (50%) decoding is indicated by the outlined pink-red regions in the temporal generalization matrices.*

b. *iEEG decoding accuracies for pattern classifiers trained and tested on task irrelevant stimuli. Same plotting conventions as in panel a.*

c. *MEG within task decoding of category for faces vs objects (red-task relevant; purple-task irrelevant). Left: results in posterior ROIs. Right: results in prefrontal ROIs.*

d. *MEG within task decoding of category for letters vs false fonts (red-task relevant; purple-task irrelevant). Left: results in posterior ROIs. Right: results in prefrontal ROIs. Error bars in c and d depict 95% CI estimated across subjects.*

e. *fMRI decoding using a searchlight approach, collapsed across the three stimulus durations. Left: decoding accuracies for pattern classifiers trained and tested on task relevant stimuli (orange-red). Right: decoding accuracies for pattern classifiers trained and tested on task irrelevant stimuli (purple). Regions showing significantly above-chance (50%) decoding accuracies are indicated by the outlined colored regions on the inflated cortical surfaces (top: left/right lateral views; bottom: right/left medial views).*

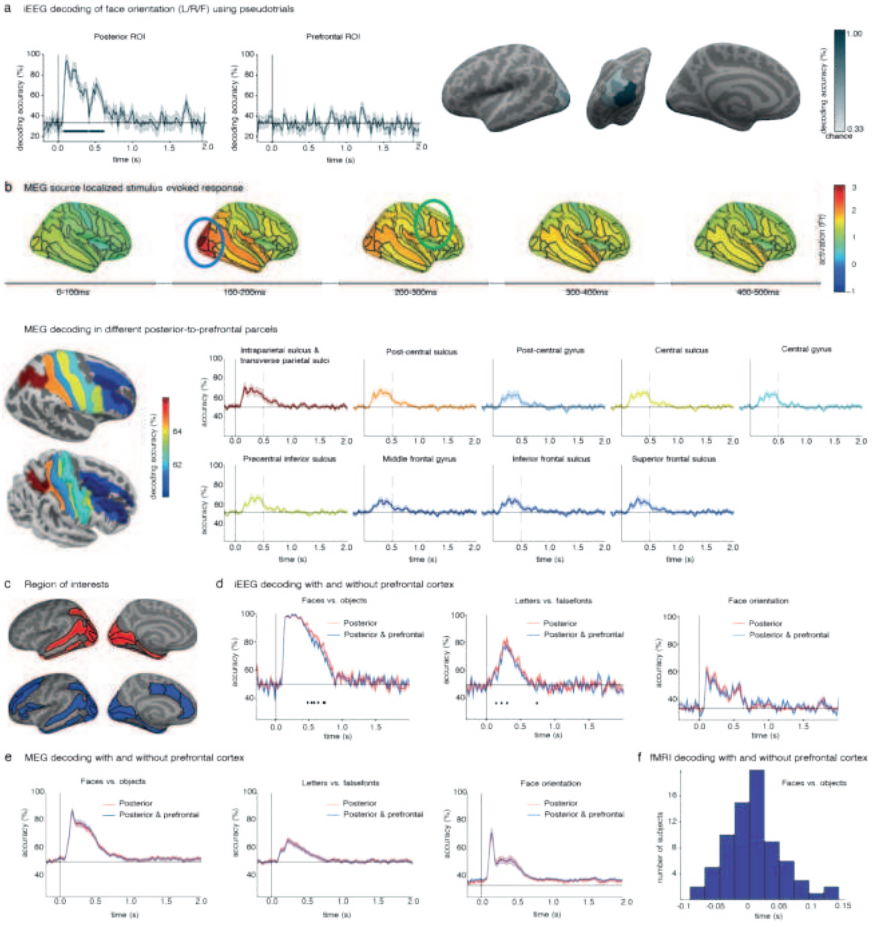
f. *iEEG decoding accuracies within the theory-relevant ROIs, collapsed across stimulus duration. Left: decoding for classifiers trained and tested on task relevant stimuli (orange-red). Right: decoding for classifiers trained and tested on task irrelevant stimuli (purple). ROIs showing significant above-chance (50%) decoding are displayed on inflated surface maps from a left lateral view (top left), posterior view (top right) and left medial view (bottom).*

Anatomical ROIs (Destrieux atlas)	Irrelevant-Relevant		Relevant-irrelevant		Irrelevant		Relevant	
	n voxels	% voxels	n voxels	% voxels	n voxels	% voxels	n voxels	% voxels
Posterior ROI								
G_and_S_occipital_inf	1868	93	1866	93	1868	93	1876	93
G_oc-temp_lat-fusifor	2549	98	2550	98	2542	98	2561	99
G_occipital_middle	1979	80	1952	79	1909	76	2096	85
S_oc_middle_and_Lunatus	1009	100	1008	100	1000	100	1010	100
G_cuneus	600	24	542	22	587	23	1233	49
G_occipital_sup	1351	69	1295	66	1299	66	1302	66
G_oc-temp_med-Lingual	1403	47	1374	46	1375	46	1499	50
G_oc-temp_med-Parahip	430	30	408	29	432	31	521	37
G_temporal_inf	686	47	692	47	756	52	859	59
Pole_occipital	1952	80	1934	80	1870	77	1968	81
Pole_temporal	0	0	0	0	0	0	15	2
S_calcarine	448	18	427	18	395	16	657	27
S_intrapariet_and_P_trans	261	7	287	8	799	21	1670	44
S_oc_sup_and_transversal	1163	82	1166	82	1225	87	1230	87
S_temporal_sup	1100	22	944	19	820	17	2264	46
PFC ROI								
G_and_S_cingul-Mid-Post	0	0	0	0	0	0	0	0
Lat_Fis-ant-Horizont	0	0	0	0	0	0	1250	23
Lat_Fis-ant-Vertical	6	1	1	0	3	1	36	8
G_and_S_cingul-Ant	0	0	0	0	5	0	278	8
G_and_S_cingul-Mid-Ant	0	0	0	0	0	0	200	1

Anatomical ROIs (Destrieux atlas)	Irrelevant-Relevant		Relevant-irrelevant		Irrelevant		Relevant	
	n voxels	% voxels	n voxels	% voxels	n voxels	% voxels	n voxels	% voxels
G_front_inf-Opercular	134	6	65	3	98	4	436	20
G_front_inf-Orbital	0	0	0	0	0	0	34	5
G_front_inf-Triangul	142	9	68	4	130	78	608	37
G_front_middle	50	1	15	0	154	3	1301	21
S_front_middle	0	0	4	0	29	1	86	4
S_front_sup	0	0	0	0	0	0	300	8
S_front_inf	164	8	89	4	184	9	1022	49

Extended Table 2.2: Decoding of faces vs. category in the theory-defined ROIs

The table presents the number of voxels in each theory-defined ROI that were detected in the searchlight decoding of category (faces vs. objects). The results are presented separately for cross-task decoding (i.e., when classifiers are trained on the task irrelevant trials and tested on task relevant ones, or vice versa), as well as for within task decoding (irrelevant and relevant conditions).


a iEEG decoding of face orientation (L/R/F) using pseudotrials
Posterior ROI
Prefrontal ROI
decoding accuracy (%)
time (s)
chance
b MEG source localized stimulus evoked response
0-100ms
100-200ms
200-300ms
300-400ms
400-500ms
activation (Pf)
MEG decoding in different posterior-to-prefrontal parcels
Intraparietal sulcus & transverse parietal sulci
Post-central sulcus
Post-central gyrus
Central sulcus
Central gyrus
Precentral inferior sulcus
Middle frontal gyrus
Inferior frontal sulcus
Superior frontal sulcus
c Region of interests
d iEEG decoding with and without prefrontal cortex
Faces vs. objects
Letters vs. falsefonts
Face orientation
Posterior
Posterior & prefrontal
e MEG decoding with and without prefrontal cortex
f fMRI decoding with and without prefrontal cortex
number of subjects

Extended Data Figure 2.3: Control analyses for the decoding prediction.

a. *Left panel: iEEG decoding results of orientation (left vs. right vs. front view faces) within the theory ROIs over time as in Figure 2.2, using pseudotrials akin to the MEG analysis. Right panel: Regions with electrodes showing above-chance (33%) accuracies are indicated in outlined blue on the inflated surfaces (left: left lateral view; middle: posterior view; right: left medial view). Error bars depict 95% CI.*

b. *Two analyses were performed to evaluate potential leakage in the MEG decoding results. These analyses were conducted on independent data from the optimization phase (N=32). Top panel: Stimulus-evoked response in face task relevant trials combined across three stimulus durations were investigated at different latencies and projected on the inflated surfaces. Blue and green ellipses denote posterior and prefrontal areas, respectively. Activity in posterior areas showed the highest peak ~0.1-0.2 s while prefrontal areas showed the highest peak in a later time window ~0.2-0.3 s. These differential peak timings serve as evidence against the leakage interpretation. Bottom panels: Face vs. object decoding performance in task relevant trials combining trials across the three durations was investigated separately within parcels in parietal and PFC to evaluate the possibility of a posterior to anterior decoding gradient. Left panel: Average face vs. object decoding accuracy in an early time window (0.25-0.5 s) projected on two differently inflated surfaces to better depict gyri and sulci in parietal and prefrontal areas. Right panel: Time-resolved decoding performance in parietal and frontal parcels. Decoding performance is highest in posterior areas and lowest in anterior areas, with fairly similar time courses, consistent with the possibility of leakage in decoding from posterior to anterior areas. This effect is better appreciated when considering the high decoding of faces vs. objects in motor related areas, with a gradient from postcentral to precentral sulcus. Error bars depict 95% CI estimated across participants.*

c. *Region of interest used in the decoding analysis including and excluding PFC areas.*

d. *Decoding analysis including or excluding prefrontal areas alongside posterior areas to evaluate changes in decoding performance. IIT predicts that including PFC to posterior areas should have either no effect or decreased decoding performance (Posterior + Prefrontal: blue; posterior only: red). iEEG decoding of faces vs. objects (left), letters vs. false fonts (middle) and face orientation (right). Lines underneath the decoding functions indicate time-periods showing significantly worse decoding accuracies when including PFC. Error bars depict 95% CI.*

e. *MEG decoding results, same order as iEEG. Error bars depict 95% CI estimated across participants.*

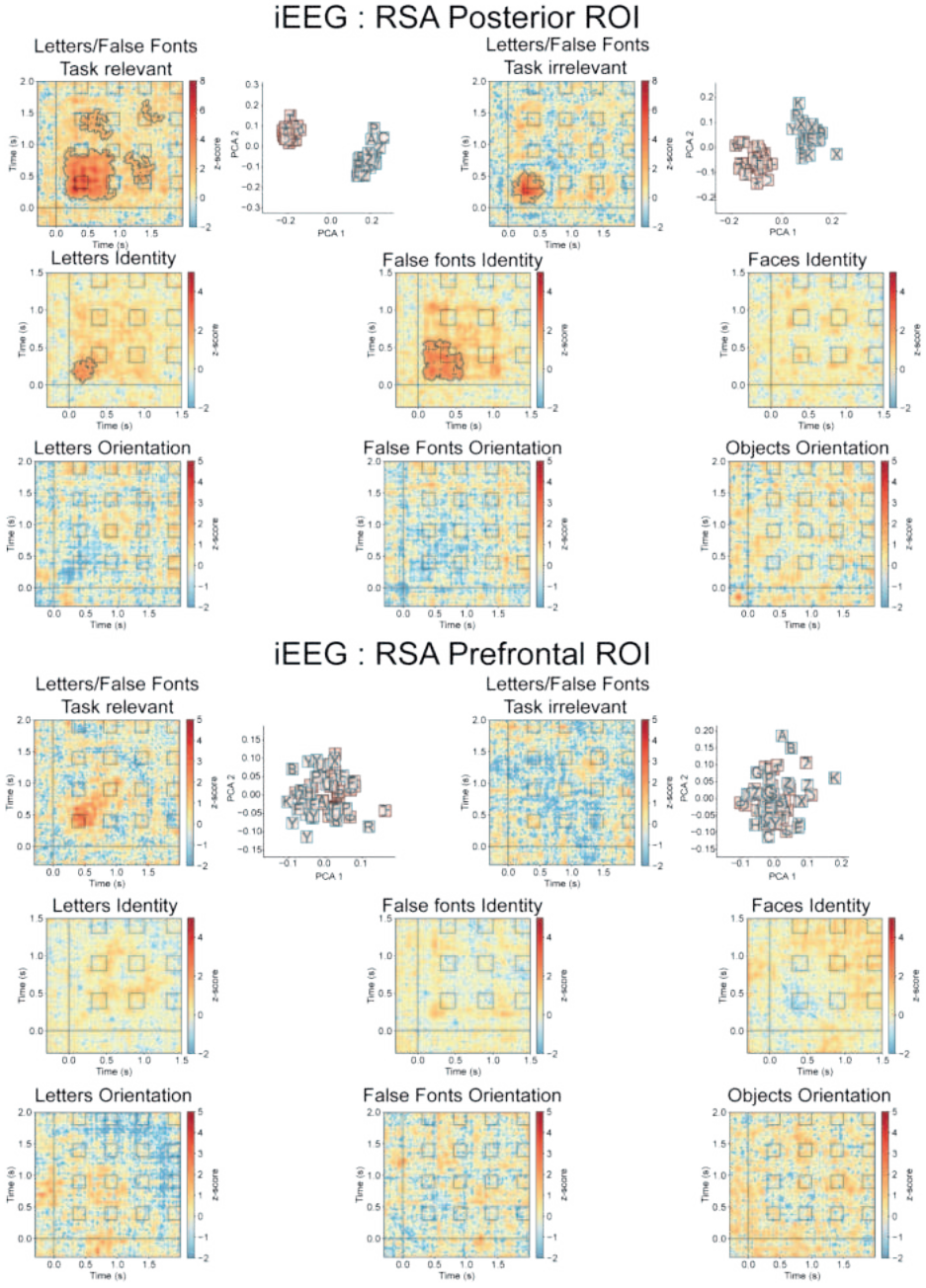
f. *fMRI decoding of faces vs. objects. Histogram shows the differences in classification including and excluding frontal areas. iEEG and MEG results consistently show similar (or worse) decoding performance when including prefrontal areas. fMRI accuracies of PFC + Posterior show slight increase of 1.2% on average compared to posterior accuracies, observed in 56% of the subjects. However, it is important to note that these increases are not considered robust due to several factors, including the small magnitude of the accuracy difference and the fact that this slight increase was observed only in the combined features analysis and not the combined models' analysis (see Methods). The negative outcomes observed in iEEG and MEG data support our interpretation of the fMRI results.*

Channel	x	y	z	Destrieux ROI
SE107-O2PH16	-0.03618	-0.08678	0.000733	S_oc_middle_and_Lunatus
SE120-T3bOT10	-0.05876	-0.06964	-0.02078	G_oc-temp_lat-fusifor
SE120-T3bOT9	-0.05712	-0.0689	-0.02016	G_oc-temp_lat-fusifor
SF102-LO1	-0.01976	-0.10359	0.001174	Pole_occipital
SF102-LO2	-0.02301	-0.09792	0.005426	Pole_occipital
SF103-PIT1	-0.04072	-0.06213	-0.02039	G_oc-temp_lat-fusifor
SF103-PIT2	-0.04156	-0.04393	-0.02499	G_oc-temp_lat-fusifor
SF104-LO1	-0.01396	-0.10275	0.008659	Pole_occipital
SF104-LO2	-0.01663	-0.10338	0.005258	Pole_occipital
SF109-IO3	0.006178	-0.07586	-0.00279	G_oc-temp_med-Lingual
SF109-IO4	0.005093	-0.07816	-0.0047	G_oc-temp_med-Lingual
SF113-RIT1	0.038119	-0.04974	-0.02225	G_oc-temp_lat-fusifor
SF113-RIT2	0.040545	-0.04845	-0.02346	G_oc-temp_lat-fusifor
SE107-O1b3	-0.01196	-0.06305	-0.00094	G_oc-temp_med-Lingual
SE107-O2PH14	-0.03383	-0.08203	6.93E-05	S_oc_middle_and_Lunatus
SE107-O2PH15	-0.0354	-0.08519	0.000512	S_oc_middle_and_Lunatus
SE108-O2b14	-0.0294	-0.09064	-0.00472	S_oc_middle_and_Lunatus
SE120-O2*5	-0.04225	-0.09646	-0.00451	G_and_S_occipital_inf
SE120-O2*6	-0.04354	-0.09769	-0.00357	G_and_S_occipital_inf
SE120-T3c6	-0.05264	-0.08681	0.025426	S_temporal_sup
SF104-LO3	-0.02255	-0.10253	0.000551	Pole_occipital
SF109-DL4	0.022039	-0.07051	0.008421	S_calcarine
SF109-DL5	0.02433	-0.07204	0.008081	S_calcarine
SF109-G45	0.04645	-0.08224	-0.00242	G_occipital_middle
SE108-O2b13	-0.02856	-0.08853	-0.00505	G_and_S_occipital_inf
SE110-O2*10	0.036288	-0.1042	-0.00079	G_and_S_occipital_inf
SE110-O2*7	0.031792	-0.09698	-0.00721	S_oc-temp_lat
SE110-O2*8	0.03359	-0.09987	-0.00464	G_and_S_occipital_inf
SE110-O2*9	0.035389	-0.10276	-0.00207	G_and_S_occipital_inf
SE120-O1b10	-0.02828	-0.11893	0.004408	Pole_occipital
SF102-LO3	-0.0356	-0.08904	-0.00424	G_occipital_middle
SF107-O1	0.024693	-0.10108	-0.00812	Pole_occipital
SF107-O2	0.027381	-0.09982	-0.00773	Pole_occipital
SF107-O3	0.042207	-0.08618	-0.00419	G_occipital_middle
SF113-RO1	0.034984	-0.08617	0.010333	G_occipital_middle
SF113-RO2	0.040244	-0.08034	0.011692	G_occipital_middle

Extended Table 2.3: Electrode locations found to be significant in the LMM analysis

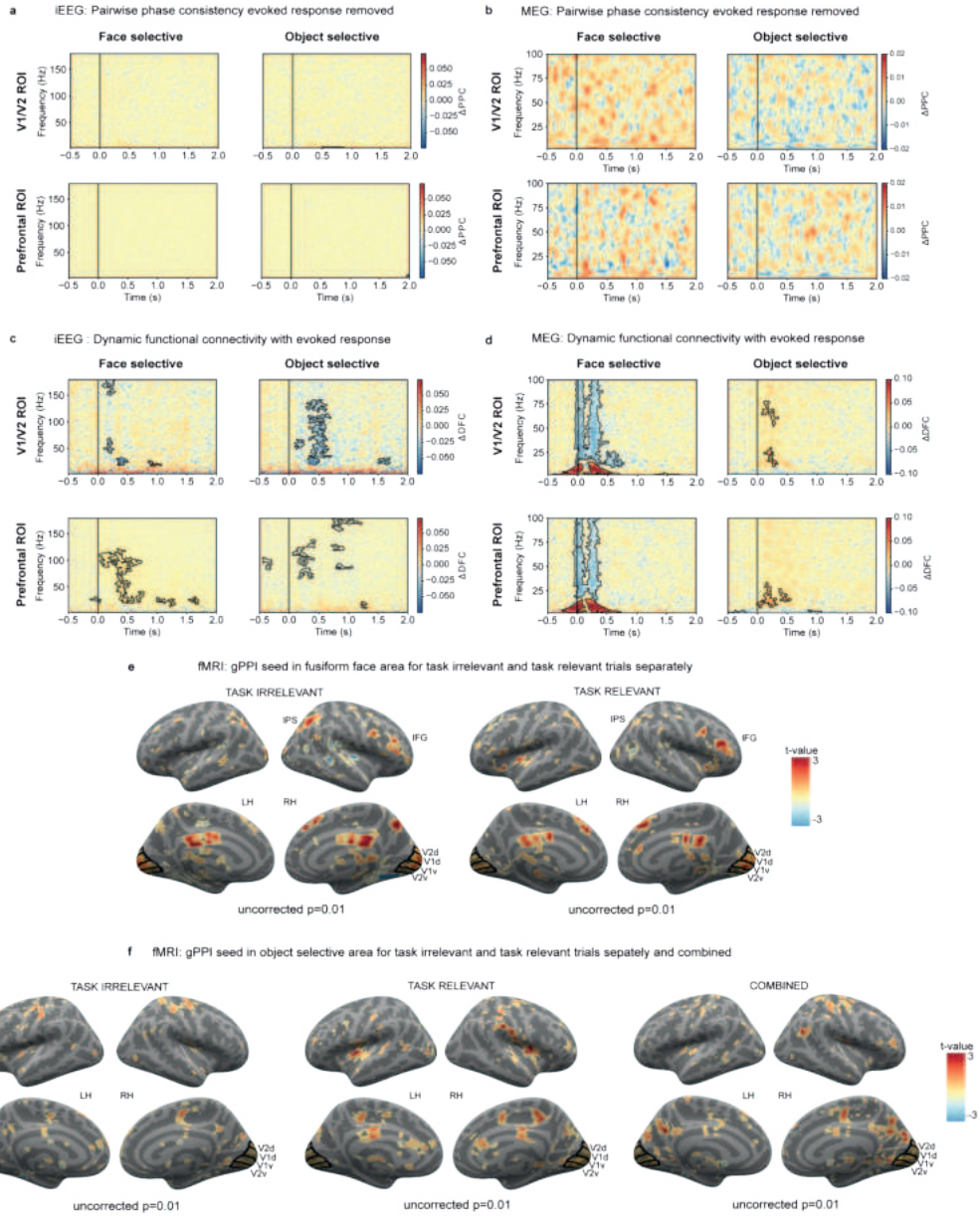
Electrodes location in MNI coordinates, as well as in the corresponding parcellations of the Destrieux Atlas, Wang Atlas and Desikan Atlas.

Wang ROI	Desikan ROI	Model
TO1	ctx-lh-lateraloccipital	IIT x Cate
Unknown	ctx-lh-fusiform	IIT x Cate
Unknown	ctx-lh-fusiform	IIT x Cate
V2d	ctx-lh-lateraloccipital	IIT x Cate
V2d	ctx-lh-lateraloccipital	IIT x Cate
Unknown	ctx-lh-fusiform	IIT x Cate
Unknown	ctx-lh-fusiform	IIT x Cate
V2d	ctx-lh-lateraloccipital	IIT x Cate
V2d	ctx-lh-lateraloccipital	IIT x Cate
V2v	ctx-rh-lingual	IIT x Cate
V2v	ctx-rh-lingual	IIT x Cate
Unknown	Cerebellum-Cortex	IIT x Cate
Unknown	Cerebellum-Cortex	IIT x Cate
Unknown	ctx-lh-lingual	GNW
LO2	ctx-lh-lateraloccipital	GNW
LO2	ctx-lh-lateraloccipital	GNW
LO1	ctx-lh-lateraloccipital	GNW
Unknown	ctx-lh-lateraloccipital	GNW
Unknown	ctx-lh-lateraloccipital	GNW
Unknown	ctx-lh-inferiorparietal	GNW
V2d	ctx-lh-lateraloccipital	GNW
Unknown	ctx-rh-pericalcarine	GNW
Unknown	ctx-rh-pericalcarine	GNW
Unknown	ctx-rh-lateraloccipital	GNW
Unknown	ctx-lh-lateraloccipital	IIT
Unknown	ctx-rh-lateraloccipital	IIT
Unknown	ctx-rh-lateraloccipital	IIT
Unknown	ctx-rh-lateraloccipital	IIT
Unknown	ctx-rh-lateraloccipital	IIT
V2d	ctx-lh-lateraloccipital	IIT
LO2	ctx-lh-lateraloccipital	IIT
Unknown	ctx-rh-lateraloccipital	IIT
Unknown	ctx-rh-lateraloccipital	IIT
Unknown	ctx-rh-lateraloccipital	IIT
V3B	ctx-rh-lateraloccipital	IIT
LO2	ctx-rh-inferiorparietal	IIT


iEEG : RSA Posterior ROI
Letters/False Fonts
Task relevant
Letters/False Fonts
Task irrelevant
Letters Identity
False fonts Identity
Faces Identity
Letters Orientation
False Fonts Orientation
Objects Orientation
iEEG : RSA Prefrontal ROI
Letters/False Fonts
Task relevant
Letters/False Fonts
Task irrelevant
Letters Identity
False fonts Identity
Faces Identity
Letters Orientation
False Fonts Orientation
Objects Orientation
Time (s)
PCA 1
PCA 2
z-score

Extended Data Figure 2.4: Maintenance of conscious content over time for stimulus categories, identity and orientation.

Cross temporal representational similarity matrices across all electrodes in posterior cortex for letters vs. false fonts (upper row), identity (middle) and orientation (bottom) for posterior (upper half) and PFC (lower half) ROI, respectively. Contours in the matrices represent statistical significance, established using cluster-based permutation tests (upper tail test at alpha=0.05). Clear separability between letters and false fonts in posterior cortex is illustrated using Principal Component Analysis at 0.3 s irrespective of the task (left – task relevant, right - task irrelevant). Separability was mostly sustained in the task relevant condition, but not from ~0.95 to 1.4 s. In the task irrelevant condition, however, separability was statistically significant for a brief period in the beginning. Identity information was statistically significant for letters and false fonts, but not faces. Identity information was not sustained for the entire stimulus duration (however, z-scores were elevated until 1 s, hinting at a limitation in statistical power). No statistically significant orientation information was evident for any of the categories. None of the contrasts yielded statistically significant results in the PFC ROI.


a
iEEG: Pairwise phase consistency evoked response removed
Face selective
Object selective
V1/V2 ROI
Prefrontal ROI
Frequency (Hz)
Time (s)
ΔPPC
b
MEG: Pairwise phase consistency evoked response removed
c
iEEG : Dynamic functional connectivity with evoked response
ΔDFC
d
MEG: Dynamic functional connectivity with evoked response
e
fMRI: gPPI seed in fusiform face area for task irrelevant and task relevant trials separately
TASK IRRELEVANT
TASK RELEVANT
IPS
IFG
LH
RH
V2d
V1d
V1v
V2v
t-value
uncorrected p=0.01
f
fMRI: gPPI seed in object selective area for task irrelevant and task relevant trials sepately and combined
COMBINED

Extended Data Figure 2.5: Control analysis for the interareal communication prediction

a. *iEEG Pairwise phase consistency (PPC) analysis of task irrelevant trials did not reveal any significant category-selective synchrony cluster neither in the posterior ROI nor in the PFC ROI after removing the evoked response. Colorbars represent the change in PPC (face-object trials) for each node (face-selective, object-selective). Positive values reflect stronger connectivity for faces. Negative values reflect stronger connectivity for objects.*

b. *MEG PPC analysis of task irrelevant trials did not reveal any significant category-selective synchrony cluster neither in the posterior ROI nor in the PFC ROI after removing the evoked response. The same conventions of Figure 2.8a are used here.*

c. *iEEG Dynamic functional connectivity (DFC) analysis of task irrelevant trials without removing the evoked response reveals significant content-selective connectivity between object-selective electrodes and V1/V2 electrodes (top-right), reflected as broadband (25-125 Hz) decrease in the change in DFC (e.g., faces < objects). Similar broadband content-selective changes in DFC (faces > objects) were observed for face-selective electrodes in PFC (bottom-left). Smaller, yet significant effects, were detected for connectivity between face-selective electrodes and V1/V2 electrodes (top-left) and for object-selective electrodes and PFC electrodes (bottom-right). Conventions as in Figure 2.8a.*

d. *MEG DFC analysis of task irrelevant trials without removing the evoked response reveal significant content-selective synchrony between the face-selective GED filter node and both V1/V2 (top-left) and PFC (bottom-left). This is reflected in an increase in low-frequency connectivity (<25 Hz) combined with a decrease in high-frequency connectivity (25-100 Hz). Smaller yet significant effects were detected for the object-selective GED filter (right). Conventions as in Figure 2.8a.*

e. *Generalized psychophysiological interactions (gPPI) task-related connectivity analysis of task irrelevant (left) and task relevant (right) conditions revealed weak clusters of content-selective connectivity when FFA is used as the analysis seed ($p < 0.01$, uncorrected). Common significant regions showing task related connectivity in task irrelevant, task relevant, and combined conditions (Figure 2.4) include V1/V2, right intraparietal sulcus (IPS), and right inferior frontal gyrus (IFG).*

f. *gPPI task-related connectivity analysis of task irrelevant (left), task relevant (middle), and combined conditions revealed weak clusters of content-selective connectivity when lateral occipital complex (LOC) is used as the analysis seed ($p < 0.01$, uncorrected). The results of the gPPI showed that there are no common significant regions showing task related connectivity in task irrelevant, task relevant, and combined conditions.*

Chapter 3

Investigating timing of conscious experience using a dual-task and quantified introspection

This chapter has been submitted to elife:
Investigating timing of conscious experience using a dual-task paradigm and quantified introspection

Alex Lepauvre[1], Micha Engeser[1], Stanislas Dehaene, Lucia Melloni

[1] Shared first authorship. My contribution to this project entails Conceptualization, Data Curation, Formal analysis, Methodology, Project Administration, Software, Supervision, Validation, Visualization, Writing of the original draft, review and editing as defined by the credit taxonomy (https://credit.niso.org/)

Specifically, I designed the experiment together with M.E. under the supervision of S.D. and L.M., supervised M.E. for collection and analysis of behavioral data, curated and validated the behavioral and eyetracking data, conducted some of the behavioral data and all eyetracking and iEEG data and wrote the manuscript together with M.E.

Abstract

What are the temporal dynamics of conscious perception? Intuitively, we believe we continuously experience the external world. Competing theories attribute conscious experience to different neural mechanisms, some emphasizing the prefrontal cortex (PFC). However, recent studies have shown that PFC activation does not reflect stimulus duration, challenging theories that assign the PFC a central role in consciousness, given the assumption that longer stimulus duration corresponds to prolonged conscious experience. This study tests that assumption by evaluating the impact of visual stimuli on response times in a dual-task psychological refractory period (PRP) paradigm and subjective timing through introspection. We found that: (1) visual stimuli gained conscious access for a fixed duration, regardless of actual duration; (2) this access occurred even for task-irrelevant stimuli, though it was extended for task-relevant ones; and (3) at stimulus offset, conscious processing weakened. Additionally, participants drastically underestimated delays caused by the PRP effect, further suggesting that conscious experience does not directly need to track stimulus dynamics. A reanalysis of Cogitate consortium data revealed that PFC decoding dynamics followed similar time courses to those predicted by PRP measures. We propose that the PRP effect is a reliable tool to track conscious access without the need for overt reports, offering new insights into the timing of consciousness.

Introduction

Do we always consciously perceive what is right in front of our eyes, at the moment when it occurs? Empirical evidence suggests that the answer is more complex than it might seem. For example, when we blink, visual stimulation is briefly interrupted by our eyelids, yet we are usually unaware of these interruptions unless we consciously focus on them [217]. Conversely, during the attentional blink or inattentional blindness, high-contrast stimuli appear, but participants report not seeing them because they are distracted by another task [41,55,84,158,218–221]. Conscious perception may also be severely delayed, as in the psychophysical refractory period [153,160,222,223], or it might be rescued by the presentation of a late attentional "retro-cue" [159]. In general, the timing and even the occurrence of conscious experience may not coincide with the timing of external sensory stimulation.

These considerations are essential in interpreting the results of recent studies investigating the neural dynamics associated with sustained stimulus presentation. Several studies have shown that when highly visible stimuli are presented for varying durations (0.3 to 1.5 seconds), sensory areas track stimulus duration, while the prefrontal cortex (PFC) only shows transient activation following stimulus onset, with no further activation correlating with stimulus duration [113,127,224]. Notably, one study found an activation increase following stimulus offset in fronto-parietal electrodes [126]. However, the design of this study differed from the others as participants were required to memorize the presented stimuli, and all stimuli were presented for the same duration (1.5 s). Accordingly, the offset activation may reflect memory-related processes or the predictability of stimulus disappearance in that specific study.

Superficially, the absence of a consistent neural signature in the PFC that tracks the persistence of perceptual content challenges theories like the Global Neuronal Workspace Theory (GNWT), which assumes that consciousness arises from the broadcast of information in a globally accessible workspace located in a fronto-parietal network [25]. According to GNWT, the maintenance of conscious access depends on a non-linear, ignition-like activity in a fronto-parietal network signalling new information entering the workspace at stimulus onset and offset of stimuli (if this offset is consciously detected) [96]. Critically, GNWT does not assume that the neural workspace remains active throughout a durable conscious experience, but at moments when conscious refreshes occur, thereby aligning to a reconstructive view of temporal experience [225]. During stable or predictable periods, no further refresh is needed, allowing the workspace to be occupied by other contents. This maintenance occurs in an activity-silent state, supported by short-term synaptic changes, with only occasional bursts of reactivation for conscious retrieval.

However, as noted earlier, evidence suggests that conscious access can be decoupled from stimulus timing, such that neither the onset nor the duration of this ignition have to be locked to stimulus properties. According to GNWT, the onset and duration of conscious experiences are dictated by the availability of the global neuronal workspace (GNW) rather than by external events. Furthermore, the persistence of a stimulus in consciousness does not necessarily mean that the activity is continuously sustained – it could simply be transiently activated to encode which stimulus was presented, together with a tag that encodes its onset and duration [225]. While an ignition needs to occur when contents enter into consciousness, typically 200-300 ms after stimulus onset, this activation (1) can be delayed when we are distracted by another task, and (2) may quickly return to baseline as the content is maintained in an activity-silent state. In other words, sustained activity is only optional: when watching a picture, we may effortfully continue to attend to it throughout its presentation duration, or we may simply encode its presence and timing, and let our thoughts wander elsewhere. In the latter case, when the stimulus disappears, if this disappearance is detected, the workspace would then have to refresh to change the temporal tag, leading to an ignition at stimulus offset. However, again, this would only occur if participants attend to stimulus duration and therefore to stimulus offset.

In summary, GNW theory does not predict persistence of GNW activation throughout a stimulus' duration, as this would unnecessarily tie up the GNW bottleneck. Instead, it predicts an ignition of the fronto-parietal network at the onset of the conscious experience and at its offset if and only if this experience is attended to and accordingly refreshed. Previous studies did not explicitly measure the persistence of conscious experience to avoid introducing task demands and attention to the temporal aspect of the stimuli [113,126,127,224]. The default assumption was that stimuli were consciously experienced throughout their duration because they were suprathreshold, presented in isolation and fixated upon. However, since conscious access is heavily dependent on attention, conscious perception may have been only transient after stimulus onset. If so, the transient PFC activation as well as the lack of offset ignition observed in previous studies would still align with GNWT's assertion that fronto-parietal ignition is necessary for consciousness. Thus, rather than challenging GNWT, the lack of association between PFC activation and stimulus duration might suggest that conscious experience was decoupled from visual presentation, exactly as GNWT predicts [224].

Here, using behavioural experiments, we test the hypothesis that the PFC activation observed by the Cogitate Consortium to visual stimuli of variable duration may have reflected the timing of participants' awareness. Using the very same stimuli,

our preregistered study aims to clarify this by directly measuring the duration of their conscious processing at stimulus onset and offset (https://osf.io/krjh7) – which according to GNWT is equivalent to the duration of their occupation of the Global Workspace. To do so, we relied on the psychological refractory period (PRP) effect [153]. This effect occurs when two tasks are presented in rapid succession, leading to a delay in reaction time (RT) to the second stimulus (T2) with decreased interval between the first and second task, reflecting a bottleneck in cognitive processing [223]. According to GNWT, this bottleneck is imposed by the GW, whereby the occupation of the workspace by a given conscious content prevents other contents from reaching consciousness [153,222,226] which implies that conscious access operates in a serial manner.

Evidence supporting this interpretation comes from time-resolved electrophysiological studies, which show that early sensory activation is unaffected by changes in stimulus onset asynchrony (SOA) between competing stimuli, while later processing stages are delayed when SOA is shorter [151,160,227–230]. Similarly, fMRI studies reveal that frontal and parietal regions exhibit delayed activity as SOA decreases [160,229,231]. These delays in PFC activation suggest that slower reaction times to the second task (RT2) occur because conscious access to the second stimulus is delayed due to the serial nature of conscious processing. This interpretation is supported by studies using quantified introspection, where participants provided subjective reports of when they became aware of stimuli. These studies show that while participants can accurately report their decision times, they are unaware of the large delays caused by the PRP effect [153,222], mistakenly believing they became conscious of the stimuli immediately. This introspective blindness supports the notion that the PRP effect reflects a delay in the conscious processing of the second stimulus, which in turn delays reaction time.

Building on the assumption that GW engagement imposes a cognitive bottleneck, we used the PRP effect as a time-resolved marker for conscious access to evaluate whether and when participants consciously experienced the onset, duration, and disappearance of a visual stimulus. We did this by measuring the response time to auditory stimuli presented at various SOAs relative to the onset and offset of visual stimuli. In line with Global Neuronal Workspace Theory (GNWT), we preregistered a series of hypotheses (https://osf.io/krjh7): (1) conscious access to an event should induce a PRP effect on a subsequent stimulus; (2) the duration of the PRP effect should be independent of the visual stimulus duration; and (3) there should be no PRP effect at the stimulus offset. Based on previous studies [113,126,127,224] showing a lack of prefrontal cortex (PFC) activation at stimulus offset, we hypothesized that participants might only experience the stimulus transiently and remain unaware of its disappearance, leading to a PRP effect at stimulus onset but not at offset.

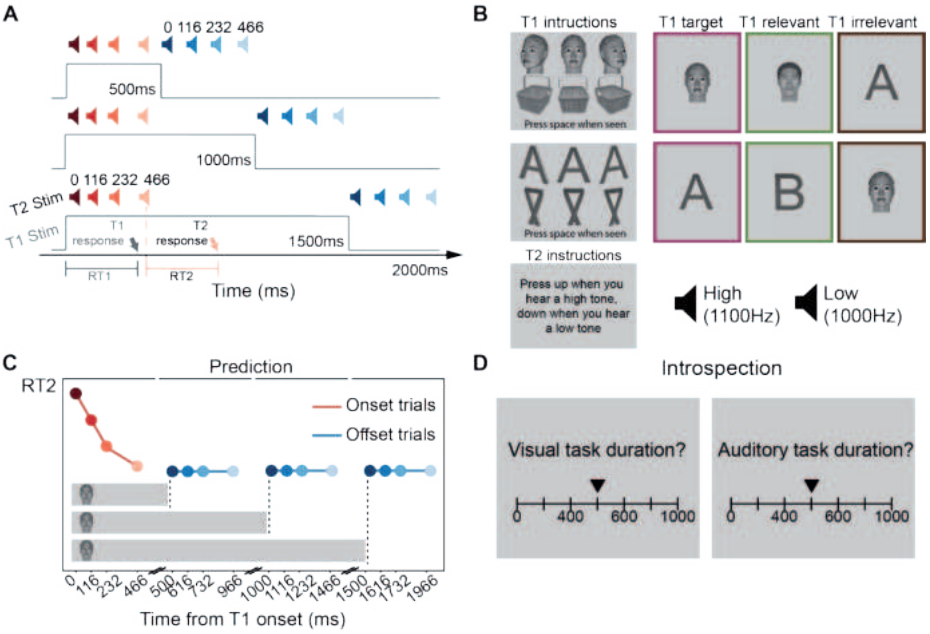


Figure 3.1: Overview of experimental design and predictions

a. *Timing of the task: Visual stimuli (T1) were always presented first, at one of three possible durations (500, 1000 and 1500ms) followed by a fixation cross. Auditory stimuli (T2) were presented at four different stimulus onset asynchronies (SOA: 0, 116, 232, 466ms) relative either to T1 stimulus onset (red) or to its offset (blue).*

b. *Tasks: For the visual T1 task, at the beginning of each block, two targets were presented (a face and an object or a letter and a symbol). Participants were asked to press a button whenever that target appeared. Stimuli of the same category as targets are labelled as task-relevant (green), stimuli of a different category as task-irrelevant (brown). For the auditory T2 task, participants were asked to discriminate high and low tone using two distinct buttons.*

c. *Predictions: RT to T2 stimuli (RT2) should decrease with increased SOA when T2 are presented following T1 onset (PRP effect, red) but not when following T1 offset (blue).*

d. *Introspective reports: In experiment 2, after each trial, introspection probes were also presented, asking participants to rate their decision time for T1 (left) and T2 (right).*

Our experiment also allowed us to investigate whether a PRP effect occurs even when the first stimulus is task-irrelevant and does not require an overt report. According to GNWT, this effect would support the serial nature of conscious processing. Since participants do not actively respond to the task-irrelevant stimulus, the collision between tasks would primarily affect central cognitive resources rather than motor preparation. In this condition, the Cogitate Consortium observed a small but significant PFC ignition, of fixed duration unrelated to stimulus duration. We argue that this reflected a brief moment of conscious experience of the task-irrelevant stimulus, and therefore predict that those stimuli should impose a similar short-lived, duration-independent PRP delay.

Finally, using quantified introspective reports of decision time, we aimed to replicate previously observed introspective blindness to the PRP effect even when no overt report is required for the T1 task. This allowed us to assess whether dual-task interference occurs without active task performance, reinforcing the GNWT hypothesis that the PRP effect represents a bottleneck of conscious processing. Additionally, the use of the PRP effect provided a means to probe the occupancy of the GW without drawing participants' attention to the temporal manipulation. This addresses critiques that PFC findings supporting GNWT may confound task demands and attention with consciousness itself.

Results

Experiment 1: visual events inducing a PRP effect

Our preregistered study builds on the experimental design and stimuli used by the Cogitate consortium, which demonstrated an absence of a neural offset response in the PFC using both invasive electrophysiology (iEEG) and source-localized magnetoencephalographic (MEG) signals. Our goal was to evaluate whether the appearance and disappearance of visual stimuli, presented for different durations, would lead to a PRP effect in the reaction time of a subsequent auditory stimulus. This allowed us to examine the relationship between conscious access to visual events and the PRP effect, particularly in relation to stimulus offset.

Our design consisted of a dual-task combining a visual go/no-go target detection task (Task 1, T1) and a pitch discrimination task (Task 2, T2). In T1, participants were shown visual stimuli (faces, objects, letters and false-fonts) for three durations (500, 1000, and 1500 ms, see fig. 3.1A). They were required to detect rare target stimuli (~11% of trials) from two categories (either a face and an object or a letter and a false-font) within each block. Non-target stimuli from the same category as the target were considered task-relevant (T1 relevant); while non-target stimuli from a different category were considered task-irrelevant (T1 irrelevant). In T2, high and low-pitch tones were presented at four different stimulus onset asynchrony (SOAs: 0, 116, 232, 466 ms) relative to either the onset or the offset of T1 visual stimulus. This setup allowed us to determine whether either event (onset or offset) triggered a PRP effect. Throughout the experiment, we collected motor responses and eye-tracking data to monitor participants' performance.

Following our preregistered protocols, we first confirmed that participants (N=21) performed well on both the visual (T1) and auditory tasks (T2). As expected, task

performance was high, with participants achieving 94.60% accuracy (SD = 3.00) in T1 and 94.64% accuracy (SD = 5.11) in T2 (see fig. 3.S1). No participant met our exclusion criterion (<80% hits or >20% false alarms in T1, <80% accuracy in T2). Eye-tracking data showed that participants consistently maintained fixation on the stimuli, spending 89% (± 15%) of their time within 6° of visual angle from its centre until 2.0 seconds after stimulus onset (fig. 3.S1).

We aimed to test the GNWT prediction about which events should trigger a PRP effect. According to GNWT, the PRP effect reflects the serial nature of conscious processing and indicates whether an event was consciously accessed. Since all T1 target stimuli required a button-press, we inferred that they were all consciously processed, predicting a PRP effect at T1 stimulus onset. This prediction aligns with previous findings of PFC ignition following the visual stimulus appearance [113,126,127,232].

We also predicted a PRP effect for non-target task-relevant stimuli, as the decision to withhold a button press still requires conscious processing. Furthermore, given that task demands likely maintain the workspace activated for a longer duration, we predicted a longer PRP effect for task-relevant stimuli compared to task-irrelevant ones.

Additionally, we predicted no PRP effect at stimulus offset, as this event is uninformative and irrelevant, meaning that subjects would not need to attend to it or consciously register the stimulus disappearance. This prediction addresses a challenge to GNWT posed by earlier studies that did not observe PFC ignition during stimulus offset. Superficially, the absence of ignition at stimulus offset appears to contradict the GNWT claim that PFC ignition is necessary for conscious access. However, if in fact subjects were not consciously aware of the stimulus offset, the lack of ignition would be entirely consistent with GNWT. These predictions were preregistered before the study (https://osf.io/krjh7).

To evaluate whether visual events (T1) affected the processing of auditory stimuli (T2), we analysed participants' reaction time to T2 (RT2) as a function of the interval between T1 and T2. We used a generalised linear mixed model (GLMM) to model participants' reaction times to T2 (RT2) as a function of the SOA, the relative timing of T2 (with respect to T1 onset and offset), and the relevance of T1 (task-relevant and task-irrelevant stimuli only). Due to the small number of T1 target trials (~11 %) and the requirement of an overt response in those cases; we excluded these trials from the main analysis and modelled them separately.

In T1 target trials, we observed a typical PRP effect: reaction times to T2 (RT2) decreased sharply as a function of the SOA of T2, locked to T1 onset ($\chi^2(3) = 362.47$, $p < .001$, see fig. 3.2a, red lines in left panel, table 3.1, and 3.S7). This interference was so pronounced that the effect lasted longer than the duration of T1. A similar PRP-like slowing effect was present for T2 locked to T1 offset. Yet, the PRP effect significantly decreased in magnitude with longer T1 durations (500 ms offset locked: $\chi^2(3) = 111.64$, $p < .001$, 1000 ms: $\chi^2(3) = 37.16$, $p < .001$, 1500 ms: $\chi^2(3) = 12.01$, $p = .007$, see table 3.S7 and fig. 3.2A). This suggests that for shorter T1 durations, the PRP effect at stimulus offset may be influenced by residual effects from T1 onset processing.

For non-target trials (T1 relevant and irrelevant), we also observed a typical PRP effect, with RTs to T2 being slower when T2 was presented closer to T1 onset (SOA main effect: $\chi^2(3) = 735.95$, $p < 0.001$, see table 3.S1 and fig. 3.2A). The magnitude of the PRP effect was strongly modulated by whether the auditory stimulus was locked to the appearance or disappearance of the T1 visual stimulus (SOA X onset/offset, $\chi^2(3) = 462.82$, $p < 0.001$). To further explore these differences, we modeled RT2 separately for trials locked to stimulus onset versus offset.

T1 task relevance influences central stage processing at stimulus onset

When the visual stimulus (T1) appeared, participants were required to decide on the appropriate behavioral response - whether to press a button or not. Therefore, we predicted that a PRP effect would be observed even in the no-go trials, regardless of whether the T1 stimulus was relevant or irrelevant to the task. First, we analyzed RT2 for T1 target trials during image onset, which revealed a strong effect of SOA (see fig. 3.2A, red lines, left panel). More importantly, and in line with our preregistered prediction, modeling RT2 for non-targets at image onset also showed a clear PRP effect (SOA main effect: $\chi^2(3) = 1109.31$, $p < 0.001$, see table 3.S2 and fig 3.2A, red lines, middle and right panels). This PRP effect was further supported by pupil dilation measurements. In line with earlier studies [233], pupil-evoked responses peaked later (90% peak) at shorter SOAs relative to T2 onset (SOA main effect: $\chi^2(3) = 15.78$, $p = 0.001$, see table 3.S8 and fig. 3.S2). Furthermore, RT2 was significantly correlated with pupil peak latency ($\beta = 0.24$, $p < 0.001$)

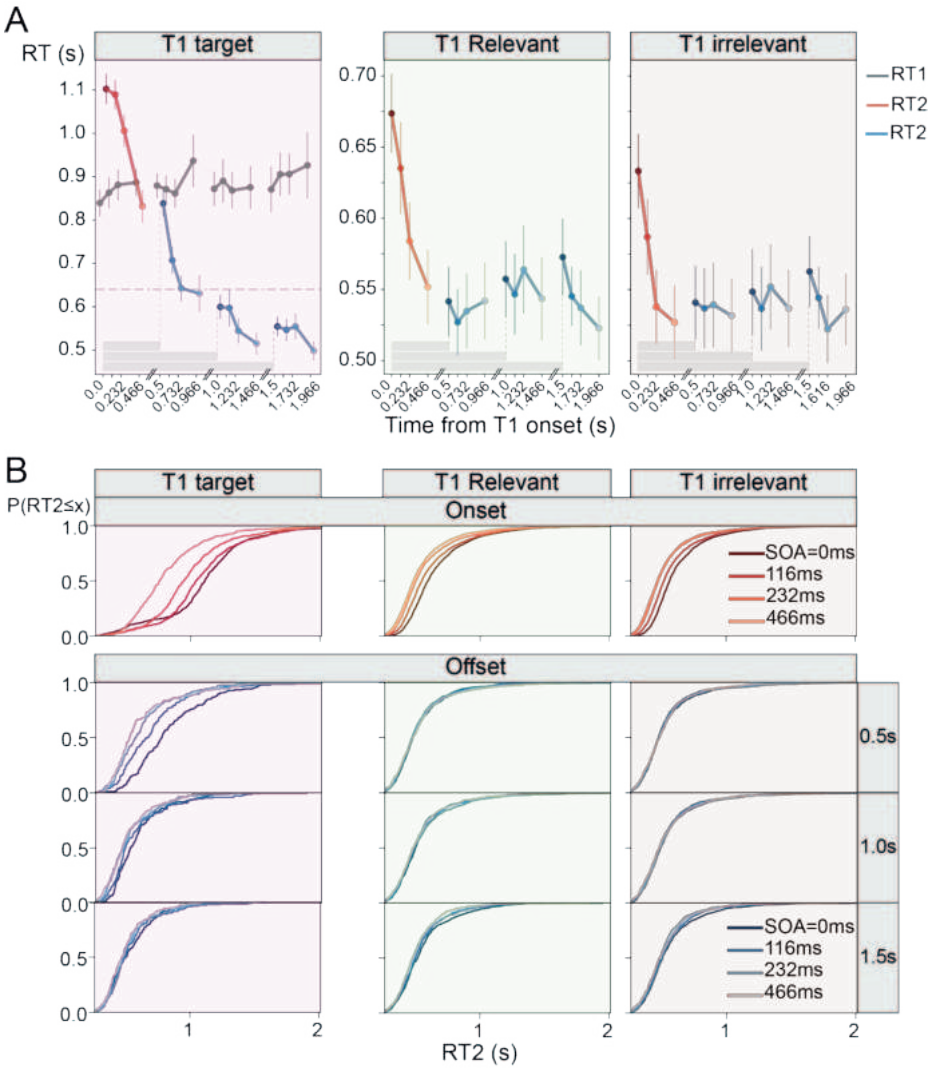


Figure 3.2 Experiment 1 Results: Modulation of the PRP effect by appearance and disappearance o T1 and T1 task-relevance

a. *Response time to the first (RT1, grey) and second task (RT2) as a function of SOA from T1 onset (x-axis). For RT2, trials in which T2 was presented following T1 onset are represented in red (collapsed across T1 duration), while trials in which T2 was presented following T1 offset are represented in blue. Grey boxes represent T1 stimuli durations. The horizontal dashed line in the upper left panel marks the mean RT to targets in the Cogitate study.*

b. *Empirical cumulative distribution function (ECDF) of RT2 separately for T1 target trials, T1 task-relevant, and T1 task-irrelevant trials. The upper row displays T1 onset-locked trials; the lower rows display T1 offset-locked trials. For offset trials, each T1 duration is shown separately (second row: 0.5, third row: 1.0, 4th row: 1.5 s).*

The PRP effect reflects the interference between two stimuli, where the processing of the first stimulus delays the processing of the second. For short SOAs, a delay in RT2 is expected in every trial due to T1 processing. A bifurcation of RT2, where only some trials are delayed while others are not, would contradict a proper PRP effect. To rule this out and to further validate our GLM results, we examined the empirical cumulative distribution for each SOA condition.

As shown in Figure 3.2B (upper row), consistent with a true PRP effect, the entire RT2 distribution shifts towards shorter values as SOA increased. This shift effect was confirmed by a Kolmogorov-Smirnov test, revealing a significant shift in the empirical cumulative distribution function (ECDF) between SOA 0 and 466 ms, for both T1-relevant and T1-irrelevant trials (T1 relevant: $D = 0.29$, $p < 0.001$; T1 irrelevant: $D = 0.29$, $p < 0.001$; see Table 3.S6).

We predicted that the PRP effect would be influenced by task relevance, with T1-relevant non-target stimuli requiring longer central-stage processing due to the difficulty of discriminating targets from similar stimuli. This extended processing should increase RT2. Indeed, we found a significant interaction between SOA and T1 relevance ($\chi^2(3) = 12.82$, $p = .005$; see Table 3.S2), indicating that central-stage processing persisted beyond T1, interfering with subsequent T2 stimuli. We also observed larger pupil sizes for T1-relevant trials, suggesting increased cognitive load (see Fig. 3.S2). These findings support our prediction that stimuli that are more relevant involve prolonged central processing.

Stimulus disappearance delays RT2

Unlike the appearance of the visual stimuli, the disappearance of the T1 stimulus was completely task-free but still represented a significant visual change. This raises the question: does a task-irrelevant event, if consciously perceived, affect subsequent stimulus processing? We observed that the disappearance of the visual stimulus delayed RT2 at short SOAs (main effect of SOA, $\chi^2(3) = 42.31$, $p < 0.001$; see Table 3.S3 and Fig. 3.2A, blue lines), superficially consistent with a PRP effect. However, this delay might simply reflect extended T1 processing during stimulus onset, mistakenly interpreted as an effect at stimulus offset. If that were the case, the lingering T1 processing should diminish with longer T1 durations, leading to a smaller PRP effect. Instead, we found the opposite: a stronger SOA effect with longer T1 duration (interaction between T1 duration and SOA; $\chi^2(6) = 35.44$, $p < 0.001$). Separate modelling of RT2 for each T1 duration revealed no SOA effect for short trials ($\chi^2(3) = 4.20$, $p = 0.240$) but a significant effect for longer T1 durations (T1 1.0s: $\chi^2(3) = 12.67$, $p = 0.005$; T1 1.5s: $\chi^2(3) = 69.644$, $p < 0.001$; see Table 3.S4). This suggests that

the offset of longer visual stimuli, even when task-irrelevant, does impact subsequent stimulus processing, supporting the idea that it is consciously perceived.

An alternative explanation for the effect of T1 disappearance on RT2 could be that participants respond faster at longer SOAs, rather than slower at shorter SOAs. This speed-up might result from the increased temporal predictability of the tone at large SOAs, similar to the foreperiod effect observed in hazard rate studies [234]. In our study, the offset may have acted as a cue, helping participants anticipate the tone and thus facilitating faster RT2 at longer SOAs.

To distinguish between interference and facilitation effects, we compared RT2 in offset trials for long T1 duration at SOA 0 ms (where the SOA effect is strongest) with RT2 in onset trials at SOA 466 ms (where interference with T1 is minimal). RT2 was significantly longer in offset trials, by 30 ms, compared to onset trials ($z = 6.22$, $p < 0.001$), indicating that the disappearance of the visual stimulus indeed delayed RT2.

Overall, our findings suggest that the disappearance of the visual stimulus did trigger a PRP effect, but this effect was contingent on the duration of T1. Even in long T1 trials, where the SOA effect is strongest, the effect size was significantly smaller than the PRP effect induced by the appearance of the visual stimulus (see Table 3.1). When comparing RT2 empirical cumulative distribution functions (ECDFs) between the SOA 0 and 466 ms conditions, significant differences emerged only in long T1 trials for both task-relevance conditions (T1 relevant: $D = 0.11$, $p < 0.001$; T1 irrelevant: $D = 0.07$, $p = 0.038$; see Table 3.S6). Additionally, intermediate T1 trials showed significant differences for the T1 relevant condition ($D = 0.08$, $p = 0.010$; see Table 3.S6), further supporting the reduced impact of stimulus offset compared to stimulus onset.

Visual inspection of the RT2 distribution in long T1 trials suggests that only a portion of trials were delayed by the disappearance (see Fig. 3.2b, bottom 3 rows). The difference in RT2 across durations emerged only when participants showed slower reaction times, indicating that in some trials, T2 processing was unaffected by T1 disappearance, while in others, RT2 was delayed.

These findings challenge the GNWT prediction that no PRP effect should occur at stimulus offset. However, the strong influence of task demands suggests that completely task-free events, like T1 disappearance, are processed only briefly and sporadically in the central stage. To further explore whether the PRP effect reflects a delay in conscious access to the T2 stimulus, we conducted a second study where participants provided introspective reports of their decision times.

Table 3.1: Experiment 1 magnitude and effect size of the PRP effect in RT2

Condition	PRP effect magnitude [s]	Effect size [Cohen's d]
Targets	0.60	1.40
Onset - T1 relevant	0.12	0.47
Onset - T1 irrelevant	0.11	0.44
Offset - T1 500 ms	0.01	0.02
Offset - T1 1000 ms	0.02	0.07
Offset - T1 1500 ms	0.04	0.21

Differences in RT2 between the shortest and longest SOA are reported. Magnitude refers to the difference in ms, while Cohen's d quantifies the effect size between these two SOAs. For non-targets, the PRP effect size was calculated separately for the visual stimulus onset (collapsed across the three visual stimulus durations) and offset (collapsed across task relevance). For targets, we focused on 1.5 sec visual stimulus and reported the difference between the shortest SOA locked to stimulus onset and the longest SOA to stimulus offset.

Experiment 2

In this experiment, we aimed to validate the PRP effect as a marker for conscious access, building on prior findings that both the appearance and disappearance of visual stimuli induced a PRP effect. Prior studies have shown that participants are introspectively blind to the delay in RT2 associated with SOA, suggesting that the PRP reflects the serial nature of conscious access.

To replicate this, 11 participants from the previous study were reinvited for a second experiment. The experimental design remained largely the same, with two modifications: introspective ratings were added, and one SOA (116 ms) was removed to reduce the overall duration of the study. At the end of each trial, participants provided introspective ratings of their decision time for both the visual (iT1) and auditory stimuli (iT2) on a continuous scale from 0 to 1000 ms (see Fig. 3.1A). They were instructed to focus on their decision time rather than their reaction time, as no overt response was required for T1 non-target trials.

In line with previous studies [153,222], we predicted that iT1 and iT2 would correlate with RT1 and RT2, indicating that participants can reliably introspect on their decision time. We also expected participants to report longer iT1 for T1-relevant compared to T1-irrelevant trials, due to the increased difficulty of target discrimination. However, we predicted that iT2 would remain unaffected by SOA, despite the strong PRP effect, suggesting that the PRP reflects a delay in participants gaining awareness of T2.

Additionally, following the central stage interference model, we anticipated that a significant portion of RT2 variance would be explained by iT1. Since the T1 and T2 tasks were identical to those in the first experiment, this allowed us to test the replicability of our previous findings (all predictions were preregistered, https://osf.io/krjh7).

Participants performed both tasks with high accuracy (T1: 95.86% accuracy, SD = 1.93; T2: 93.68% accuracy, SD = 5.86; see Fig. 3.S3). One participant was excluded for not meeting the preregistered accuracy criteria (<80% in the auditory task). Eye-tracking data showed that participants consistently maintained fixation, spending 94% (± 4%) of the time within 6° of visual angle from the stimuli until 2.0 seconds after stimulus onset (see Fig. 3.S3). These results indicate that participants were engaged and attentive throughout the experiment.

RT2 Experiment 1 replication

In the T1 onset trials, we successfully replicated the significant effect of SOA ($\chi^2(2) = 368.03$, $p < 0.001$; see Fig. 3.3A, red lines), but we did not replicate the interaction between T1 relevance and SOA ($\chi^2(2) = 1.99$, $p = 0.370$; see Table 3.S11). This lack of interaction might be due to the reduced power in our second study, as it involved only half the number of participants. Pupil size was larger in T1-relevant compared to T1-irrelevant trials (see Fig. 3.S4a), and the pupil peak latency mirrored the PRP effect observed in RT2 (SOA main effect: $\chi^2(3) = 9.24$, $p = 0.002$; see Table 3.S27), replicating the findings from Study 1. Additionally, RT was again correlated with pupil peak latency ($\beta = 0.40$, $p = 0.002$).

In T1 offset trials, both the SOA effect ($\chi^2(2) = 54.64$, $p < 0.001$; see Table 3.S12) and the interaction between T1 duration and SOA ($\chi^2(4) = 19.44$, $p = 0.001$) were replicated. SOA had a significant effect on RT2 for all T1 durations (500 ms: $\chi^2(2) = 6.46$, $p = 0.039$; 1000 ms: $\chi^2(2) = 42.82$, $p < 0.001$; 1500 ms: $\chi^2(2) = 30.73$, $p < 0.001$; see Table 3.S13). In contrast to Study 1, where this effect was only observed at 1000 and 1500 ms durations, here it was significant across all durations. As in Study 1, the SOA effect was much stronger in onset trials compared to offset trials (see Table 3.2).

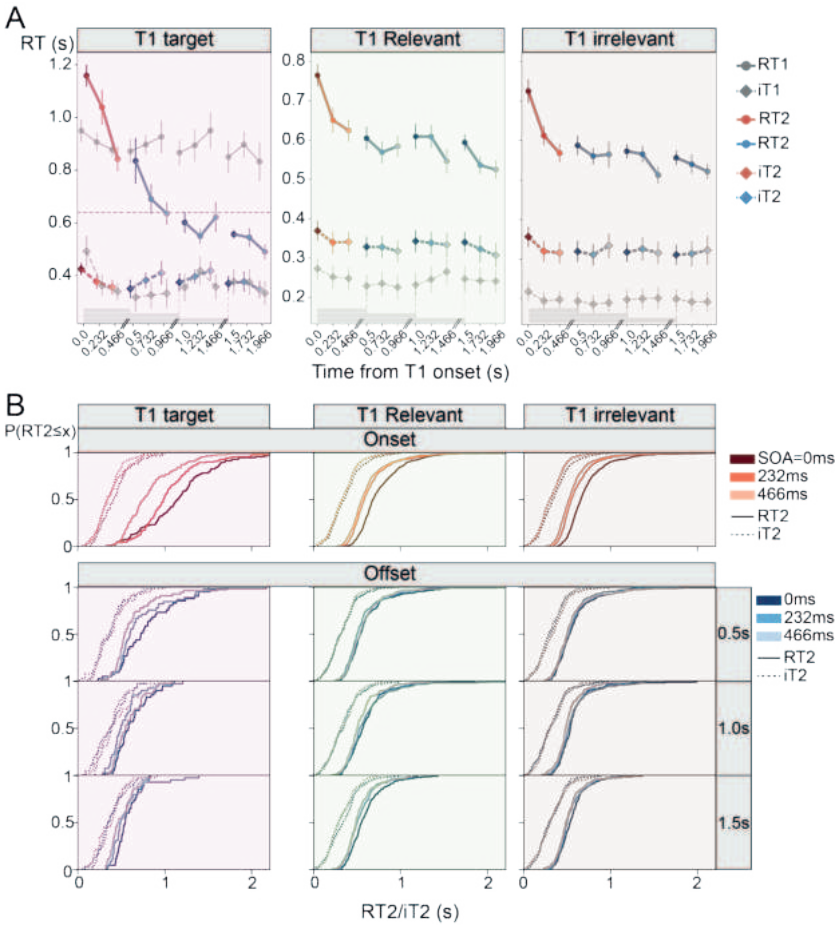


Figure 3.3: Experiment 2 results: Replication of study 1 and test of introspective awareness of the PRP effect

A. *Objective (solid lines) and introspective (dashed lines) response time to the first (RT1/iT1, grey) and second task (RT2/iT2) as a function of SOA from T1 onset (x-axis). For iT2 and RT2, trials in which T2 was presented following T1 onset are represented in red (collapsed across T1 duration), while trials in which T2 was presented following T1 offset are represented in blue. Grey boxes represent T1 stimuli durations. The horizontal dashed line in the upper left panel mark the mean RT to targets in the Cogitate study.*

B. *Empirical cumulative distribution function (ECDF) of RT2 and iT2 (solid and dashed lines respectively) separately for T1 target trials, T1 task-relevant, and T1 task-irrelevant trials. The upper row shows T1 onset-locked trials; the lower rows display T1 offset-locked trials. For offset trials, each T1 duration is shown separately (second row: 0.5, third row: 1.0, 4th row: 1.5 s).*

Introspective awareness of the PRP effect

First, we validated participants' introspective duration judgments using a calibration task in which they estimated the length of a tone presented for a random duration (20 to 1000 ms) Consistent with previous studies [153,222], estimated tone duration was strongly correlated with actual duration ($\beta = 0.91$, $p < 0.001$). Similarly, in the main experiment, participants' introspective decision times for T2 (iT2) were well correlated with their reaction times (RT2) (iT2-RT2: $r = 0.48$, $p < 0.001$).

For T1, objective reaction times (RT1) were only available for a few target trials (~11% of the trials), but despite the limited data, we observed a strong correlation between iT1 and RT1 (iT1-RT1: $r = 0.48$, $p < 0.001$). Consistent with previous studies, participants underestimated their objective RT2 ($\beta = 0.60$, $p < 0.001$; see Fig. 3.3A), likely reflecting the delay between perceptual processing and conscious awareness (Marti et al., 2010). Participants were instructed to report their decision time rather than their motor response time, as no overt response was required for T1 non-target trials. This underestimation of RT2 can be partly attributed to the exclusion of the response execution stages and the delayed onset of conscious experience.

Contrary to our preregistered prediction, we observed a significant interaction between SOA and stimulus onset/offset ($\chi^2(2) = 15.26$, $p < 0.001$; see Table 3.S17). When analyzing onset and offset separately, this effect was significant only for onset trials (Onset: $\chi^2(2) = 48.33$, $p < 0.001$; Offset: $\chi^2(2) = 3.83$, $p = 0.148$; see Tables S18-19 and Fig. 3.3). These findings suggest that participants were partially aware of the PRP effect at stimulus onset, but only when the effect was strong. However, as shown in Fig. 3.3A, participants significantly underestimated the slope of RT2 in their introspective reports. The effect size of SOA on iT2 was much smaller than on RT2 (see Table 3.2).

To quantify this underestimation, we modeled iT2 as a function of the mean RT2 (representing the SOA-related slope) and the trial-by-trial deviation from mean RT2 (representing fluctuations in RT2). Variance partitioning revealed that the RT2 slope explained only 1.6% of the variance in iT2, while 21.5% was explained by trial-by-trial RT2 variance. This suggests that participants significantly underestimated the effect of SOA on their RT2, indicating partial introspective blindness to the PRP effect.

The effect of SOA on iT2 in onset trials appears primarily driven by a significant delay at SOA 0 ms. Comparing iT2 across SOAs in onset trials revealed delays only at short SOAs (0-232 ms: $z = 6.10$, $p < 0.001$; 0-466 ms: $z = 5.95$, $p < 0.001$), but no difference between 232 ms and 466 ms ($z = -0.121$, $p = 1.000$) was observed. In contrast, RT2

progressively decreased as a function of SOA (0-232 ms: $z = 14.28$, $p < 0.001$; 0-466 ms: $z = 18.72$, $p < 0.001$; 232-466 ms: $z = 4.67$, $p < 0.001$). This dissociation suggests that the SOA effect on iT2 likely reflects participants' awareness of the simultaneous presentation of auditory and visual stimuli at SOA 0 ms.

Interestingly, iT1 was also affected by SOA in onset trials ($\chi^2(2) = 30.02$, $p < 0.001$; see Table 3.S21), despite the fact that RT1 is typically unaffected by SOA in classical PRP studies (Pashler, 1994). Although we lacked objective RT1 data for comparison, the iT1 effect was driven by longer iT1 at SOA 0 ms (0-232 ms: $z = 4.95$, $p < 0.001$; 0-466 ms: $z = 4.53$, $p < 0.001$; 232-466 ms: $z = -0.39$, $p = 1.000$), further suggesting that simultaneous stimulus presentation at SOA 0 ms influenced introspective judgements.

Despite participants' relative blindness to the large SOA effect on their objective RT, they were aware of the smaller difference associated with T1 task relevance. As predicted, task relevance significantly affected iT1 in both onset ($\chi^2(1) = 189.78$, $p < 0.001$) and offset trials ($\chi^2(2) = 7.25$, $p = 0.027$; see Table 3.S21). Participants were aware of the increased difficulty in classifying T1 non-target task-relevant stimuli. Surprisingly, T1 task relevance also influenced iT2 in onset trials (onset: $\chi^2(1) = 17.98$, $p < 0.001$; offset: $\chi^2(1) = 4.78$, $p = 0.092$; see Table 3.S18), which contradicts the serial access hypothesis that predicts consistent introspection regardless of T1 processing duration. These results suggest that participants were indeed aware of some delay in RT2 caused by the interference regime from T1 processing.

Table 3.2: Experiment 2 magnitude and effect size of the PRP effect

	RT2		iT1		iT2	
Condition	**Magnitude [s]**	**Cohen's d**	**Magnitude [s]**	**Cohen's d**	**Magnitude [s]**	**Cohen's d**
Targets	0.69	1.41	0.14	0.55	0.07	0.39
Onset - T1 relevant	0.14	0.51	0.02	0.14	0.03	0.16
Onset - T1 irrelevant	0.15	0.61	0.02	0.14	0.04	0.24
Offset - T1 500 ms	0.02	0.08	0.00	0.00	-0.01	-0.03
Offset - T1 1000 ms	0.06	0.30	-0.02	-0.16	0.00	0.02
Offset - T1 1500 ms	0.05	0.31	0.01	0.04	0.01	0.07

Differences in RT2, iT1, and iT2 between the shortest and longest SOA. Magnitude indicates the difference in ms, while Cohen's d quantifies the effect size between these two SOAs. For non-targets, the PRP effect size was calculated separately for the onset of the visual stimulus (collapsed across visual stimulus durations) and its offset (collapsed across task-relevance). For targets, the shortest SOA of the onset was compared to the longest SOA on the offset on the longest stimulus duration.

Introspective report reliability can be assessed in the absence of overt report

In our experiment, T1 did not require an overt response on non-target trials, so iT1 could not be directly validated against an objective RT. According to the conscious access bottleneck model, participants' introspection of RT should reflect the combined duration of the central and motor stages. Since participants were instructed to report only their decision time, excluding the motor stage (which was absent for T1 non-targets), iT1 should approximate the central stage occupation.

The central stage interference model explains the RT2 delay at short SOA by the central stage being occupied by the T1 stimulus. A direct prediction of this model is that iT1 should predict trial-by-trial variability in RT2 during the interference period. Therefore, we expected the correlation between iT1 and RT2 to be strongest at short SOA and to weaken as SOA increased.

This prediction was only partially validated. In onset trials, we observed a significant correlation between iT1 and RT2 ($\chi^2(1) = 614.01$, $p < 0.001$; see Table 3.S24), but there was no significant interaction with SOA ($\chi^2(2) = 1.34$, $p = 0.512$), suggesting that the correlation did not decrease as SOA increased. We had predicted that this correlation would decrease once the interference period ended - when T1 central stage processing was complete. One possible explanation is that T1 processing lasted longer than 0.466s in this study, which could explain the lack of interaction between iT1 and SOA in onset trials.

If T1 processing did exceed 0.466s, we would expect an interaction between iT1 and SOA in offset trials when T1 duration exceeded central stage processing. Consistent with this, we found a significant three-way interaction between iT1, SOA, and T1 duration ($\chi^2(2) = 13.13$, $p = 0.011$; see Table 3.S25) in offset trials. To explore this, we modeled RT2 as a function of SOA and iT1 separately for each T1 duration in offset-locked trials. Surprisingly, the interaction between iT1 and SOA was significant only for intermediate (1000 ms) trials ($\chi^2(2) = 11.37$, $p = 0.003$), but not for short (500 ms: $\chi^2(2) = 2.60$, $p = 0.27$) or long trials (1500 ms: $\chi^2(2) = 0.45$, $p = 0.798$; see Table 3.S26).

These results suggest that T1 processing may have ended between 1.0 and 1.466s. The absence of an interaction in short T1 trials might indicate that the interference period was still ongoing by 0.966s (the latest tone onset in offset-locked short trials), while the lack of interaction in long T1 trials suggests that the interference period had ended by 1.5s (the earliest tone onset in offset-locked long trials). However, it is unlikely that T1 processing lasted beyond 1s, as RT1 in target T1 trials averaged below 1s. These findings may instead reflect a conflation of the effects induced by the

T1 stimulus disappearance and T1 processing, which could not be separated in this study, presenting a potential confound in this analysis.

Overall, the introspective task results show that participants were largely unaware of the PRP effect, despite being sensitive to much smaller differences in their RT. This supports the idea that participants can only consciously access the T2 stimulus after completing their decision-making about T1. To further understand this cognitive bottleneck, we next explored the neural data to identify potential neural correlates of this cognitive bottleneck.

Neural Substrate of the cognitive bottleneck

Our novel application of the PRP paradigm provided detailed timing information on processing stages in our visual task, as indicated by its indirect effect on RT2, even though most T1 trials did not require an overt response. We leveraged findings from the Cogitate study, which used a similar T1 task and included high-resolution electrophysiological recordings from epilepsy patients (N=34), to identify potential brain regions involved in central stage processing [232].

In our dual-task paradigm, we found that non-target, task-relevant stimuli engaged central processing stages for a longer duration than non-target, task-irrelevant stimuli. This suggests that the stronger PRP effect observed for task-relevant stimuli is due to the increased difficulty in discriminating targets from similar non-targets, rather than simply being an artifact of the dual-task setup.

The original Cogitate study found decoding of T1 stimulus categories (faces/objects and letters/false-fonts) in brain regions such as the middle and inferior frontal gyri, under both task-relevant and task-irrelevant conditions. However, that study did not compare the duration of decoding between these conditions. Our findings, which showed a longer PRP effect for task-relevant conditions, suggest that central processing stages are 'occupied' for a longer period in these conditions. This extended processing likely correlates with longer decoding times for the stimulus category in the task relevant trials.

To identify brain regions involved in this extended processing, we re-analyzed the Cogitate data, focusing on differences in decoding duration between task-relevant and irrelevant conditions. We first confirmed that task relevance in the Cogitate study involved higher cognitive load by examining pupil dilation as a proxy [235], which was larger in task-relevant trials, consistent with our dual-task findings and other PRP studies (see Fig. 3.S6).

Next, to assess the 'occupation' of brain regions by specific perceptual content, we focused on decoding faces vs. objects. We selected this contrast because it had the highest decoding accuracy in the Cogitate study, maximizing our ability to detect differences between task relevance conditions. We performed time-resolved decoding of the high-gamma band across channels within specific brain regions, comparing task-relevant and irrelevant trials. Using a cluster-based permutation test, we identified brain regions with sustained differences in decoding accuracy between conditions (see methods).

Decoding of faces and objects occurred in the occipital, temporal, parietal, and prefrontal cortices, in both task-relevant and irrelevant conditions (Fig. 3.S4), with peak decoding accuracy similar across tasks. Using a stringent statistical threshold ($p < 0.01$) when comparing the decoding AUC between task-relevant and irrelevant trials, decoding was sustained for longer in task-relevant conditions specifically in the middle and inferior frontal gyri, inferior frontal sulci, and the fusiform gyrus (Fig. 3.4). Interestingly, extended decoding occurred earlier in frontal regions than in the fusiform gyrus. When applying a conventional threshold ($p < 0.05$), additional regions showed significant differences between task-relevant and irrelevant conditions, though most appeared in late time windows, beyond stimulus presentation (see Fig. 3.S8). Notably, at this threshold, the occipital pole exhibited more protruded representation of stimulus category for task-relevant stimuli following stimulus onset, although at a larger latency than what is observed in frontal regions. These findings suggest that the middle and inferior frontal gyri process sensory inputs for varying durations depending on the task context, highlighting their role in distinguishing targets from non-targets and in central-stage processing.

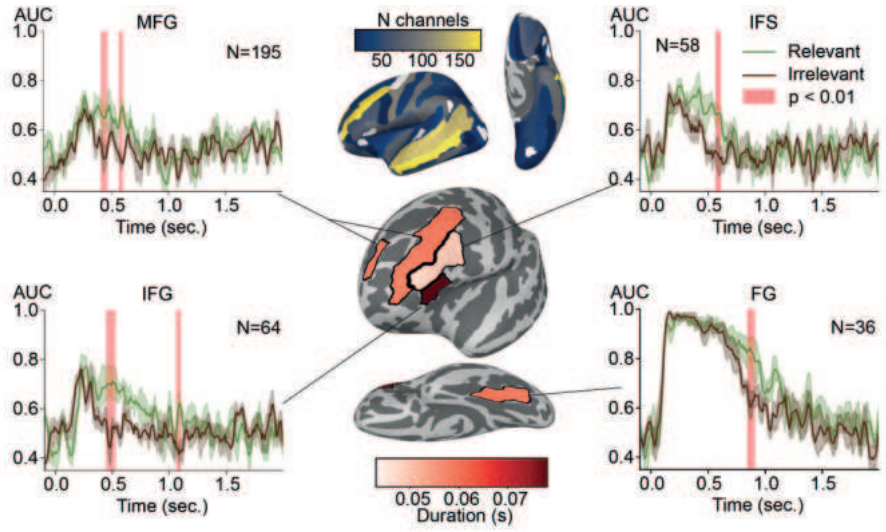


Figure 3.4: Test of hypothesis that Global Workspace access exists for task-irrelevant stimuli, but lasts longer for task-relevant stimuli

Middle panel: Destrieux parcels showing a significant difference between the task-relevant and irrelevant conditions. The colors indicate the duration of higher decoding in the task-relevant condition.
*Outer panels: Time-resolved ROC-AUC values in the high gamma band for face/object classification in the task-relevant (green) and task-irrelevant (brown) trials in regions of interest (ROI) showing significantly higher decoding in the task-relevant compared to the irrelevant condition (40ms uniform kernel smoothing, shading around curves: 95% confidence intervals, red shadings; p < 0.01, cluster-based permutation test). "N" indicates the number of channels. Abbreviations: MFG (middle frontal gyrus), IFS (inferior frontal sulcus), IFG (inferior frontal gyrus), FG (fusiform gyrus).*Upper middle panel: channel counts for each parcel of the Destrieux atlas (white represents parcels with fewer than 10 channels, which were excluded from the analysis).

Discussion

In the present study, we investigated the temporal dynamics of conscious experience by examining the Psychological Refractory Period (PRP) effect in response to visual stimuli. Previous studies have found transient activations only at stimulus onset in the PFC, but not at stimulus offset [113,126,127,224]. While these results diverge from the original GNWT prediction that conscious refreshes of the workspace occur at both stimulus onset and offset, the lack of an offset response could be explained by participants not consciously attending to the disappearance of the stimuli. This hypothesis was not tested in earlier studies, as they did not assess subjects' awareness of stimulus offset. Here, we employed the PRP effect and quantified introspection to directly evaluate the timing of access consciousness with respect to both stimulus onset and offset.

Our findings revealed a robust PRP effect at stimulus onset, supporting the GNWT prediction of conscious processing, even for task-irrelevant stimuli. While task relevance extended the duration of this processing, the effect remained transient, independent of stimulus duration, indicating that the workspace did not remain occupied for the entire duration of stimulus display. A small PRP effect was observed at stimulus offset, but it was not consistent, likely reflecting brief, occasional processing of that event. These results suggest that conscious processing can be transient and decoupled from visual presentation dynamics, aligning with previous findings of no sustained PFC activation or offset ignition [113,126,127,232].

Moreover, our findings reinforces the idea that the PRP represents a bottleneck in conscious processing, even for task-free events like stimulus disappearance, positioning the PRP effect as a powerful tool for studying conscious access without relying on overt reports, minimizing report-related confounds.

Transient conscious processing of visual stimuli

In our study, participants performed a go/no-go task with visual stimuli, and we observed a PRP effect in no-go trials (where no behavioural response is required), consistent with previous studies [236–238]. This finding indicates that motor planning is not the sole source of the interference between T1 and T2. As expected, T1 relevance modulated the magnitude of the PRP effect, with higher task relevance leading to longer processing of the T1 stimulus. This effect aligns with the central stage interference model and previous studies showing that manipulations affecting central stage processing duration have an additive effect on response times at short SOAs [239–242].

Our re-analysis of the Cogitate iEEG data [224] revealed that task relevance extended processing duration in a few brain regions, particularly in the PFC, supporting the hypothesis that the PRP effect is influenced by task relevance. These findings are consistent with previous PRP studies that link central-stage processing to the PFC [160,229,231] and further highlight its role in conscious processing, as suggested by the GNWT [25].

Importantly, while the relevance of T1 stimulus influenced the magnitude of the PRP effect, the duration of T1 stimuli did not. This supports the GNWT prediction in the Cogitate study [224] that conscious experience is transient and not sustained for the entire stimulus duration when participants are not required to attend to it. The duration-invariant transient activation in the PFC observed in previous studies [113,126,127,224] may thus reflect the dynamics of conscious experience, further

suggesting that conscious processing is decoupled from the duration of visual stimulus presentation, as initially proposed by GNWT [224].

Conscious processing of visual stimuli disappearance

A key aspect of our design involved presenting T2 stimuli at varying SOAs relative to T1 disappearance (offset trials). Based on the absence of fronto-parietal ignition at visual stimulus offset in previous studies [113,126,127,224], we hypothesized that participants would not consciously process T1 disappearance, predicting no PRP effect. Contrary to our prediction, we found a significant decrease in RT2 with increasing SOA following T1 disappearance. Although this effect was much smaller than the PRP effect observed around T1 onset, it calls for an explanation.

One possible interpretation is that the shorter RT2 at longer SOAs does not reflect a delay in RT2 at short SOAs, as a genuine PRP effect would predict, but rather a speeding up of RT2 as SOA increases, due to the increased probability of T2 appearance. Studies using a foreperiod design, where a target occurs at varying latencies from a cue, have shown that RT decreases with increased SOA, as the likelihood of the target appearing next increases [234,243]. In our task, T1 offset might have acted as a cue, increasing preparedness and speeding RT2 at longer SOAs.

However, one observation contradicts this explanation: RT2 at short SOAs in offset trials with long T1 duration was slower than RT2 at large SOAs in onset trials. This observation suggests that T1 disappearance did delay the auditory stimulus processing, consistent with a PRP effect. It may imply that, contrary to our hypothesis, participants did consciously experience the disappearance of the visual stimulus in at least some trials. Interestingly, the impact on T2 processing caused by T1 disappearance depended on the duration of T1, with no effect observed when T1 lasted only 0.5 seconds. This may be due to differences in attention driven by how predictable T1's disappearance was based on its duration. We used three discrete durations for the T1 stimuli (0.5, 1, and 1.5 s), which implies that as time elapsed, T1 offset became increasingly likely. Under the predictive processing framework, predictable events benefit from strong endogenous attention [244,245] and previous studies have shown that increased expectations enhance neural representations [246] and facilitate conscious perception [51,69]. Therefore, the increased predictability of longer T1 stimuli may have made their disappearance more likely to be consciously experienced compared to shorter T1 stimuli.

If participants sometimes consciously experienced the offset of T1 stimuli, why did previous studies fail to detect fronto-parietal ignition for this event? One explanation

is that the auditory task in our study made T1 disappearance more relevant, possibly serving as a cue for T2 and drawing more attention. This, combined with the predictability of the offset in longer T1 trials, could have increased the likelihood of the event being consciously experienced compared to previous studies. Alternatively, the relatively small PRP effect observed at T1 offset suggests that conscious processing of stimulus disappearance was either brief and/or sporadic across trials. As a result, PFC ignition may have occurred in too few trials to be detectable in neural recordings averaged across all trials.

Introspective awareness of the PRP Effect

Previous studies suggest that participants are introspectively blind to the PRP effect [153,222] -- even though objective RT differences are in the hundreds of milliseconds, they are only aware of the duration of the central stage of conscious decision making, not of the delays due to stimuli waiting in a preconscious buffer. This has been interpreted as evidence in favor of the serial nature of conscious access, implying that participants only become aware of the second stimulus after completing the central stage processing of the first.

Our results differ, as we observed that a small significant decrease in both iT1 and iT2 as SOA increased. Note that this effect was driven primarily by the shortest SOA (0 ms). Participants reported longer decision times when T1 and T2 were presented simultaneously, with no significant differences in iT2 at later SOAs (232 and 466 ms). This suggests that the impact of SOA on iT1 and iT2 is limited to simultaneous presentation, an effect that may stem from increased central competition at short SOAs. A previous study showed that without specific instructions, when visual and auditory stimuli were presented simultaneously, participants suffer from an additional central-stage slowing down due to task setting factors, i.e. the difficulty of deciding which stimulus to respond to [247]. Here, we explicitly instructed subjects to always respond to the visual stimulus first. That instruction could have been harder to maintain when a competing auditory stimulus was simultaneously presented. This effect may be compounded by the fact that auditory stimuli are typically processed faster than visual stimuli, especially in our task where visual stimuli were more complex and varied than the auditory stimuli, thereby increasing the competition for central resources.

Alternatively, the discrepancy may be related to task order. In several previous studies [153,222], T1 was auditory and T2 was visual, whereas in our study, T1 was visual and T2 was auditory. Several studies suggest that introspective blindness to the PRP effect occurs when T1 is auditory rather than visual [248–250]. Bryce and Bratzke [250]

suggest that auditory T2 allows participants to better introspect RT delays due to the sequence of sounds associated with the stimuli themselves but also the sounds elicited by button presses. In our task, non-target trials only involved the tone and the T2 response "click," which might have helped participants become aware of some delays in RT2 at short SOA. While our experiment cannot rule out this account, our findings, like others [249–251], indicate that despite a small effect of SOA on introspective time, participants still drastically underestimate the PRP effect, suggesting they have limited awareness of it.

PRP effect and conscious processing

Studies on the related phenomenon of the "attentional blink" (AB) have shown that merely consciously experiencing the first stimulus (T1) is sufficient to induce an attentional blink of a secondary stimulus. This finding indicates that the AB effect reflects a bottleneck in conscious processing [84] equivalent to the processing bottleneck in the PRP effect [151,226]. When a second stimulus (T2) is presented while the first is still being processed, it is stored in a decaying sensory buffer. If T1 conscious processing finishes before T2 decays, T2 will be processed at a delay; otherwise, it will be missed (i.e. not consciously experienced) [242]. In our studies, we observed that the PRP effect could be induced by events that are not associated with any tasks, such as the disappearance of the visual stimulus. This finding aligns with previous research showing that the PRP effect can occur when no task is associated with the T1 stimulus [226]. This further supports the interpretation that the PRP effect reflects the serial nature of conscious processing.

Some may argue that the PRP effect does not reflect the serial nature of conscious experience but instead, the cognitive processes associated with accessing the content of our conscious experience [252]. Under this view, our results might indicate that participants briefly access the content of their consciousness following the onset of the stimulus, while their conscious experience may continue beyond this initial access. If this is the case, the period during which the stimulus remains on the screen after the content has been accessed could constitute a state of phenomenal consciousness devoid of cognitive confounds. The distinction between phenomenal and access consciousness remains highly debated, however, as does the issue of whether consciousness can occur in the absence of cognitive processes [253]. The hypothesis that participants may have remained phenomenally conscious following conscious access, even in parallel to performing a secondary task, is untestable in the current study.

On a more positive note, the use of the PRP effect as a marker for conscious access constitutes a useful methodology for probing conscious experience in a time-resolve manner. In recent years, the reliance on reports to infer whether a stimulus was consciously experienced has been criticized, as the act of reporting itself introduces cognitive confounds related to the need to report [18,47–49]. Consequently, several no-report paradigms have emerged to investigate the neural correlates of consciousness by relying on alternative markers of conscious experience such as eye movements or delayed reports [54,55,86,254]. Such paradigms have however also received criticism, due to the difficulty of establishing whether content is experienced in the absence of a trial-by-trial report [74,75]. The PRP effect might offer a solution to these controversies. If serial processing is unique to conscious access, the PRP effect may constitute a reliable marker for conscious access which can be applied on a trial-by-trial basis and in a fully time-resolved fashion, without inducing additional cognitive confounds regarding the stimulus of interest. As such, the novel application of the PRP as a marker for conscious access holds significant potential in advancing our understanding of the neural underpinning of consciousness.

Methods

The experimental procedure, selection criteria and main hypothesis were pre-registered and can be accessed in OSF (https://osf.io/krjh7). Below we provide a summary of the experimental protocol. Further details are contained in the preregistration.

Study 1

Participants

Twenty-one adults (13 females, 25.18 ± 3.97 years old) with no hearing impairment (self-reported) and normal or corrected-to-normal vision participated in exchange for €14/hour. Experimental procedures were approved by the Ethics Council of the Max Planck Society and conducted in accordance with the Declaration of Helsinki. Participants provided written informed consent before the study. All subjects were included in the analysis, as none met the pre-registered exclusion criteria of low mean performance in the T1 visual task (<80% hits or >20% false alarms) or the T2 auditory task (<80% accuracy).

Dual task

Stimuli and procedure

To evaluate whether the appearance and disappearance of a visual stimulus perturb the processing of a subsequent auditory stimulus, subjects participated in a dual-task paradigm (visual and auditory).

Task 1 (T1): We replicated the design from the [224]. Grayscale images from four categories (faces, objects, letters, and false-fonts, referred to as T1 categories) were presented individually for three different durations (500, 1000, and 1500 ms, T1 duration), followed by a blank screen. Each trial lasted 2 seconds, with an added random jitter (mean of 1.0 seconds, range 0.7-2.0 seconds) to avoid periodic presentation. Half of the stimuli were displayed in side-view (+/-30° rotation), and half in front-view (T1 orientation). To manipulate task demands, participants were instructed to detect the rare occurrence of specific faces and objects or specific letters and false-fonts, depending on the experimental block, regardless of their orientation and duration. Within each block, stimuli were categorised into three task relevance conditions (T1 relevance): T1 target (stimuli to detect), T1 relevant (non-target stimuli of the same category as targets but different identity), and T1 irrelevant (non-target stimuli of a different category than targets). The task relevance manipulation was orthogonal to the stimulus category: in half the blocks, targets were faces and objects; in the other half, targets were letters and false-fonts.

Task 2 (T2): This consisted of a pitch discrimination task with high (1100 Hz) and low (1000 Hz) pitch tones (82 ms duration). These tones (T2 pitch) were presented at four stimulus onset asynchronies (SOA: 0, 116, 232, or 466ms) relative to the onset or offset of the T1 visual stimulus. Thus, offset T2 trials were presented at 12 different latencies from T1 onset, depending on T1 duration (500, 1000 and 1500 ms). Participants were instructed to respond as quickly and accurately as possible, making a decision on the T1 stimulus (go/no-go) before responding to the T2 task.

A total of 2,160 trials were presented, divided into 60 experimental blocks. At the beginning of each block, a target screen was shown displaying 2 target stimuli (a face and an object or a letter and a false-font) in three different orientations (fig. 3.1A, target screen). Each block consisted of 34-38 trials, including 2-6 targets. The remaining 32 trials comprised 16 T1 relevant and 16 T1 irrelevant trials (eight per category). The trial order was randomised and balanced with respect to task relevance, visual stimulus duration, orientation, category, SOAs, onset/offset, and auditory stimulus pitch frequency. Each unique combination of visual stimulus duration, SOA, and

onset/offset was presented 10 times for target trials and 40 times for non-target task-relevant and task-irrelevant trials.

Before the experiment, participants performed practice blocks: first only T2, then only T1, and finally both tasks together. Each practice block consisted of 40 trials. The entire experiment lasted roughly three hours, with participants completing it in a single session, including a mandatory break of at least 10 minutes halfway through. At the end of the experiment, participants filled a questionnaire asking participants if they noticed the various experimental manipulations and whether they experienced difficulty performing the task (see supplementary).

The experiment was programmed and controlled using Psychtoolbox-3 extensions [255], running in MATLAB [256] on Windows 10 Enterprise (64-bit). Visual stimuli were displayed on a BenQ XL2420Z 24-inch monitor at a 60Hz refresh rate, covering a 6° x 6° visual angle. Tones were played on dual speakers (Neumann, KH 120 A) at a constant volume across participants (74 dB). Responses were collected using a Cedrus RB-844 response box (mean latency of 5.13 ms ± 0.7 ms, measured for 100 responses). Participants used the index finger of one hand to respond to T1 and the index and thumb of the other hand to respond to T2, with hand attribution counterbalanced across subjects. Reaction times to both T1 (RT1) and T2 (RT2) stimuli were recorded.

Throughout the experiment, pupil and gaze data were continuously acquired using a high-speed, video-based eye tracker (EyeLink 1000 Plus, SR Research), sampled binocularly at 500 Hz. Participants' heads were stabilised using a chin rest (70 cm from the display) to ensure a stable head position. A 13 points calibration was performed at the beginning of the study, after the break or whenever participants displaced their heads from the chin rest.

Trial exclusion

Following preregistered criteria (https://osf.io/krjh7), trials were excluded if: no responses or incorrect responses to T2 were logged, if reaction times (RT) to T2 were shorter than 100 ms, if a false alarm was recorded to T1 and/or if responses to T2 preceded those to T1.

Reaction time predictions and analysis

Analyses were performed in R [257] using the lme4 extension [258]. All predictions and analyses described below were pre-registered (https://osf.io/krjh7), except if stated otherwise. T1 Target trials were analysed and treated separately from non-target

T1 trials to prevent contamination of over motor responses, which would affect comparison between task relevant and task irrelevant conditions.

We modelled reaction times to the auditory task (RT2) using a gamma distribution with an identity link function [259] in a generalised linear mixed model (GLMM). RT2 was modelled as a function of SOA, onset/offset (whether the auditory stimuli was locked to the onset or offset of the T1 stimuli) and T1 relevance as fixed effects (including interaction terms). We modelled inter-individual and inter-duration differences in slope and intercept for each fixed effects and their interaction as random effects, resulting in the following model:

$$RT2 \sim SOA \times \frac{Onset}{offset} \times T1\ relevance + (SOA \times Onset/offset \times T1\ relevance \mid Subject) + (SOA \times Onset/offset \times T1\ relevance \mid Duration)$$

(1)

We hypothesised that the workspace was occupied only by the appearance of the visual stimulus, not by its disappearance (offset). Accordingly, we predicted a main effect of SOA in model (1), with an interaction between SOA and onset/offset factors, capturing the lack of PRP effect in offset trials.

To further test the prediction that the disappearance of the visual stimulus did not occupy the workspace, we modelled RT2 in offset trials separately, as a function of SOA, duration and T1 task relevance:

$$RT2_{offset} \sim SOA \times T1\ relevance \times T1\ Duration + (SOA \times T1\ Task\ relevance \times T1\ Duration \mid Subject)$$

(2)

We predicted that no effect of SOA on RT2 in offset trials. To account for T1 duration, which influences the latency of T2 presentation, we included it as a fixed effect in the model. In short T1 trials, offset T2 stimuli were presented at 500 ms after T1 onset. If the workspace remained occupied by T1 until 500 ms, delay in RT2 at short SOAs might reflect lingering of T1 processing rather than T2 disappearance. Therefore, we predicted that an interaction between SOA and T1 duration might be observed. If the effect of SOA reflects a lingering of T1 processing, the effect should decrease with

increased T1 duration. We tested this by performing post-hoc pairwise comparisons of estimated marginal means using the "emmeans" package with Bonferroni correction. Observing a significant interaction, we further modelled RT2 of offset trials separately for each T1 duration (exploratory).

Finally, we hypothesised that in onset trials, the workspace should be occupied by relevant T1 trials compared to the irrelevant ones, as task-relevant trials required more extensive processing to decide whether a response was needed. Thus, we predicted a larger SOA effect for task-relevant T1 trials. To test this, we modelled RT2 in onset trials as a function of SOA and task relevance:

$$RT2_{onset} \sim SOA \times T1\ relevance\ +\ (SOA \times T1\ Task\ relevance \mid Subject) \quad \textbf{(3)}$$

We expected a significant main effect of T1 task relevance and/or a significant interaction between SOA and T1 relevance. For all models, p-values were obtained by likelihood-ratio chi square (χ^2) tests of the full model against the model without the respective effect.

In addition to the modelling approach, we computed the PRP effect size using the Cohen's d:

$$Cohen's\ d\ =\ \frac{(M2 - M1)}{Pooled\ SD} \quad \textbf{(4)}$$

where *M1* and *M2* are the mean RT2 for the short (0ms) and longest SOAs (466ms) respectively, *Pooled SD* is the standard deviation of the combined sample. Cohen's d was computed separately for onset and offset trials. For onset trials, effect size was computed across T1 durations separately for each T1 task relevance condition while for offset trials, effect sizes were computed across T1 task relevance conditions separately for each T1 duration. Moreover, for target trials, the interference regime seemed to persist beyond the visual stimulus duration; consequently, the shortest SOA of the onset was compared to the longest SOA and longest duration, time-locked to offset.

Eye-tracking analysis

Preprocessing

The eye-tracking data were analysed using python v3.12 [260] and the MNE toolbox v1.6.1 [261]. Blinks' periods were identified using the algorithm described in [197]. This method detects the onset and offset of blinks based on stereotypical pupillometry patterns associated with the occlusion of the pupil by the eyelid preceding and following a blink. For blinked segments lasting 1.5s or less, pupil size and gaze position missing samples were reconstructed using linear interpolation (mne.preprocessing.eyetracking.interpolate_blinks) with padding of 0.02 around blink event (remaining segments were discarded from further analysis). Interpolated data were epoched from -0.3 to 2.7s around the visual stimuli onsets. Epochs data were baseline corrected (divisive baseline). The same exclusion criterion as described for the reaction time data were applied. In addition, we removed trials in which the z-scored mean baseline amplitude (-0.2 to 0.0s) was superior to 2, as recommended by Mathôt and Vilotijevic [262] and trials in which participants spent less than 50% fixating within a 2° of visual angle from the centre of the screen. Two participants were removed from subsequent analysis as the total number of excluded trials exceeded 50%.

Task relevance and cognitive load

To test for an increase in cognitive load associated with the task relevance manipulation, we compared the pupil size between task relevant and irrelevant trials using a cluster-based permutation test [201]. The comparison was performed separately for trials where the tone was presented relative to the onset and offset, across all stimuli durations. Importantly, because the relevance manipulation pertained to T1, the data were aligned to the onset of the T1 stimulus, both in trials where the tones were presented relative to the onset and offset. We predicted a larger pupil size for task relevant compared to task irrelevant trials in both onset and offset locked trials (exploratory).

PRP effect in the pupil response

To investigate the manifestation of a PRP effect in the pupil size, we computed the evoked pupil response of each subject by averaging the pupil size across trials separately for each SOA, T1 duration and onset/offset trials. We then extracted the latency () of the evoked pupil response 90% peak [233] relative to the auditory stimuli onsets. The extracted latencies were modelled using a linear mixed model:

$$\tau \sim SOA \times Onset/offset \times Task\ relevance \times duration + (1 \mid Subject) \quad \textbf{(5)}$$

The random slopes for each factor had to be removed due to convergence issues. The latencies were then modelled separately for the onset and offset locked trials. The expected outcome was the same as described in the behavioural data.

We pre-registered an attempt to deconvolve the pupil response into latent components associated with the fast-paced events in each trial, following the method proposed by [263]. However, this analysis was unsuccessful since the events in our design occurred too closely in time, preventing the algorithm from accurately attributing the latent components to specific events. Consequently, we failed to resolve the deconvolved components, and thus the results are not reported.

Experiment 2

Participants

11 participants (6 females, aged 24.18 ± 2.17 years old) from the previous cohort were reinvited to participate for a compensation of €14/h. Experimental procedures were approved by the Ethics Council of the Max Planck Society and followed the guidelines from the declaration of Helsinki. Participants provided written informed consent before the study.

Stimuli and procedure

Stimuli, experimental procedure and apparatus were comparable to those used in study 1 with two exceptions: 1) three (0, 116, and 466ms) as opposed to four SOAs were used to decrease the overall duration of the experiment, 2) An introspective task was included in addition to a visual target detection task and a pitch discrimination task. Specifically, at the end of each trial, participants were prompted to report their introspective evaluation of their decision time to T1 and T2 i.e., estimate the time between the appearance of the stimulus and their decision to react rather than when a button press was executed as most trials did not require a response to T1. On each trial, shortened instructions were displayed to remind the participants of the task: (1) "Visual task duration?", and (2) "Auditory task duration?". Participants used a dial (Griffin PowerMate USB) to control a cursor on a linear scale (0-1s, fig. 3.1) presented in the middle of the screen. Subjects operated the dial with one hand, and used the other hand to respond to the visual T1 (ring finger), and auditory T2 (index finger for high pitch, and thumb for low pitch). The hand assignment was counterbalanced across participants.

Participants first conducted a duration estimation task whose purpose was to determine the fidelity of their duration estimation judgments. They were presented

binaurally with a single tone (800Hz) of variable duration (0.2 - 1s, 10ms interval), at the same loudness as during the main task, while fixating on a grey screen with a central fixation cross. At the end of each trial, subjects reported the duration of the tone on an analog scale using a response dial. Participants received visual feedback displaying their estimated duration against the true duration of the tone on the screen. They were also informed if their estimates were accurate (<20ms estimation error), too short or too long by displaying 'Well done!', 'Your estimate was too short!', 'Your estimate was too long!' respectively. 100 stimuli of different durations were presented in random order.

A total of 972 were presented to avoid exhaustion due to addition of the introspective task. 24 trials were presented per unique combination of T1 duration, orientation, category, task relevance, SOA and onset/offset for the T1 relevant/irrelevant condition and 6 times for T1 target condition. Subjects performed 24 blocks. Each block consisted of 38-44 trials (2-6 targets, 18 T1 relevant, 18 T1 irrelevant trials). Participants were reminded of the T1 target identities midway in the block, to avoid forgetting due to the increased block length. The experiment was divided into 2 sessions of 12 blocks each. A session lasted approximately 2h. A questionnaire was administered at the end of the second session asking participants whether they noticed the various experimental manipulations and whether they experienced difficulty performing the task (see supplementary).

Trial exclusion

Same exclusion criteria as in study 1 were used. One subject was excluded due to low T2 accuracy (<80%), resulting in a total sample of 10 subjects.

Reaction times (RT) and introspective Time (iT) analysis and predictions

The same modelling procedure as in the first study was applied to investigate the effect of SOA, task relevance and onset/offset on RT2, with the same expected outcome. These investigations constitute a replication of the first study, albeit with a lower sample size (N=10 instead of N=21). We therefore predicted (preregistered) that the analyses of RT2 in the second study should confirm the results of the first study. Two additional variables were measured in this experiment: the introspective decision time to the visual and auditory stimuli (iT1 and iT2, respectively). According to the GNWT, participants should only be able to introspect about the duration for which a given content was processed in the central stage. Accordingly, while iT should to some extent be sensitive to variation in RT, the delay in RT2 at short SOA should not be reflected in iT, as this delay reflects a delay in conscious processing of T2.

To test this, we derived additional predictions of the iT1 and iT2 patterns based on our experimental design, which were tested using R [257] and the lme4 extension [258]. All predictions and analyses described below were pre-registered, except if stated otherwise (https://osf.io/krjh7). As for previous models, both iT and RT data were modelled using a gamma distribution with an identity link function.

We investigated the correlation between iT and objective RT for each task separately. We z-scored RT and iT measures within subjects and then applied Pearson correlation coefficient (r) between iT1-RT1 and iT2-RT2, respectively. For iT1-RT1, the analysis was restricted to T1 target trials as only those required a response. In line with previous studies [153,222], we predicted that participants should accurately introspect on the time they required to reach a decision and therefore, iT should be strongly correlated with RT both for T1 and T2. However, if introspection is limited to the central stage, participants should underestimate RT due to the omission of sensory and motor stages in the introspective ratings, as shown by [153]. To assess whether participants underestimated their objective reaction time, we modelled RT2 as a function of iT2 (following centering of iT2 by the population mean) and investigated the significance of the intercept.

In addition, iT2 were modelled in the same fashion as RT2 to investigate the effects of our experimental manipulations on introspective rating of decision time:

$$\begin{aligned} &iT2 \sim SOA \times Onset/offset \times T1\ relevance \\ &+\ (SOA \times Onset/offset \times T1\ relevance \mid Subject) \\ &+ (SOA \times Onset/offset \times T1\ relevance \mid Duration) \end{aligned}$$

(6)

As the delay in RT2 is thought to reflect a delay in conscious processing of T2 in consciousness and because participants can only introspect about conscious processes, we predicted that there will not be any effect of SOA on iT2. As the full model revealed a main effect of SOA on iT2, we investigated the effect of SOA separately for onset and offset trials separately:

$$\begin{aligned} &iT2_{onset \mid offset} \sim SOA \times Duration \times T1\ relevance \\ &+\ (SOA \times duration \times T1\ relevance \mid Subject) \end{aligned}$$

(7)

iT1 was modelled in the same way. Typically, the PRP effect is characterised by a delay in RT2, while T1 processing remains constant across SOAs [223]. In our experiment, participants did not provide an overt response to T1 in most of the trials. Nonetheless, participants should be able to report the time it took them to decide not to reply once they gain awareness of a new stimulus. We predicted that iT1 should not be impacted by SOA. Importantly, we further hypothesised that our manipulation of T1 task relevance should impact duration of conscious processing: a stimulus of the same category as the target must be processed for longer than a stimulus of a different condition. We therefore predicted that participants should be able to perceive a difference in their decision time between task relevance conditions and we expected a main effect of this factor.

Furthermore, according to GNWT, the delay in T2 processing at short latencies should be explained by a transient occupation of the global workspace by T1. On the other hand, introspective report of decision time for each content arguably constitutes a quantification of the workspace occupation by a given content. If that is the case, RT2 should be tightly correlated with iT1 at short SOAs. To test this hypothesis, we added iT1 as a predictor in the RT2 model described in Eq 1. We predicted that there should be an interaction between iT1 and SOA, reflecting a stronger effect of iT1 at short SOAs, decreasing with increased SOA. We further modelled RT2 as a function of iT1 separately for onset and offset trials.

Cogitate iEEG data

While we could not directly test the role of the PFC in the cognitive bottleneck associated with the PRP effect, we conducted a reanalysis of the Cogitate [224] intracranial encephalography (iEEG) and eye tracking dataset. The similarity of their task with our T1 task allowed us to leverage this data to probe the neural underpinnings of our behavioural findings.

Participants

32 (18 females, aged 31.17 ± 13.45 years old) patients with pharmaco-resistant epilepsy who were monitored for epilepsy seizure localization were included in the analysis. Participants provided informed consent for their participation in the Cogitate study. iEEG data from a total of 4057 electrodes were collected across patients (1238 surface, 2819 depths). Three subjects were excluded from the analysis as they did not complete the entire study, resulting in a total of 29 subjects and 3613 (1070 surface) electrodes.

Stimuli and procedure

The experimental design of the original study was the same as our T1 task: grayscale images of 4 different categories (faces, objects, letters and false-fonts) presented for 3 durations (500, 1000 and 1500 ms) in 3 different orientations (half in centre orientation, quarter left, quarter right, ±30°). A blank screen was presented in between each stimulus, such that each trial lasted for 2s with a jittered inter-trial interval of 0.4s on average (truncated exponential distribution between 0.2 and 2.0s) to avoid periodic stimulus onset. Participants had to detect infrequent target stimuli (~11%).

Specific targets (a face and an object, or a letter and false-font) were presented at the beginning of each block. A block included 32 non-target trials (8 per category) and 2-6 target trials. Non-target trials consisted of 16 task-relevant (same category as targets) and 16 task-irrelevant stimuli (different category from the targets). A total of 720 trials were presented divided across 20 blocks. Trials were balanced across category and task relevance conditions (80 trials per combination of task relevance and category).

Eye-tracking data

The eye-tracking data were collected using an Eyelink 1000+ or a Tobii 4C eye tracker. Due to the lower sampling rate (90Hz) and signal quality of the Tobii 4C, as well as technical issues leading to the lack of pupil data in some of the subjects whose data were collected using the Tobii, we only analysed the data collected using the Eyelink. Out of the 29 subjects, eye-tracking data were collected from 14 subjects using the Eyelink. Out of those, data from 2 subjects could not be recorded due to technical issues during recordings in the clinic. Additionally, data from 2 participants were rejected as the validation of the calibration was not performed and one due to loss of tracking during the recording, resulting in a total of 9 subjects. The same preprocessing pipeline was applied as for the PRP study. Finally, we investigated the difference in pupil dilation between the task relevant and irrelevant trials in a time-resolved fashion using a cluster based permutation test as described in section task relevance and cognitive load.

iEEG preprocessing

We used the same preprocessing pipeline as described in [224]. The scripts can be retrieved from https://github.com/Cogitate-consortium/iEEG-data-release. First, data were downsampled to 512Hz and detrended. Channels marked by the epileptologist as epileptic onset zones and channels showing no signal or high level of noise (characterised by visual inspection) were discarded from further analysis. The

remaining 3156 channels (981 surface) were notch filtered at 60Hz (and harmonics) using a one pass, zero phase non-causal band-stop FIR filter to remove line noise. Channels were then re-referenced using a Laplacian scheme, subtracting the average activation of the two nearest channels on both sides from each channel within the same implant [169,264]. Contacts located at the edge of shafts, strips and grids were re-referenced using a bipolar scheme (subtracting the average of one neighbour only). The high gamma (HG) signal was then calculated as follows: the signal was bandpass filtered in 10 Hz frequency bins from 70 to 150 Hz (70-80 to 140-150Hz). For each frequency bin, the absolute of the Hilbert transform was computed to obtain the instantaneous amplitude and normalised by dividing each time point by the average across the entire recording to account for the 1/f power spectrum profile. The normalised envelopes were averaged across frequency bins to produce a single HG envelope time series. The signal was segmented in epochs from -1 to 2.5s from stimulus onset. Due to the variety in electrode coverage across participants, all channels were combined into a "super-subject". To that end, we first ensured that the trial matrices were equated across subjects and then combined all collected channels in a single subject.

Category decoding analysis

We hypothesised that task-relevant tasks require deeper central stage processing to decide on the appropriate behavioural response, resulting in a protruded PRP effect. Previous studies suggest that central stage processing occurs in the PFC. Importantly, the difference in central stage processing duration between task relevance conditions is not exclusive to dual tasks; it should also be observed when the T1 task is performed in isolation, as is the case in the Cogitate study. To explore this possibility, we used time-resolved multivariate patterns decoding of face/object category as a proxy for the timespan for which a given brain region processes perceptual information.

Specifically, we use a support vector machine (SVM) classifier to decode faces from objects separately for task relevant and irrelevant trials. SVW was performed per cortical label of the Destrieux atlas [173]. The following cortical labels contained less than 10 electrodes each and where omitted from the analysis: G_and_S_paracentral, G_cingul-Post-ventral, G_cuneus, Lat_Fis-ant-Horizont, Lat_Fis-ant-Vertical, S_cingul-Marginalis, S_collat_transv_post, S_interm_prim-Jensen, S_oc_middle_and_Lunatus, S_oc_sup_and_transversal, S_pericallosal, S_precentral-sup-part, S_suborbital, S_temporal_transverse, G and_S_frontomargin, G_subcallosal (see fig 3.4, middle panel). We limited our analysis to faces/objects comparisons as these were found to show the highest decoding accuracy in the Cogitate across all regions investigated [224]. After extracting trials in which either a face or an object were

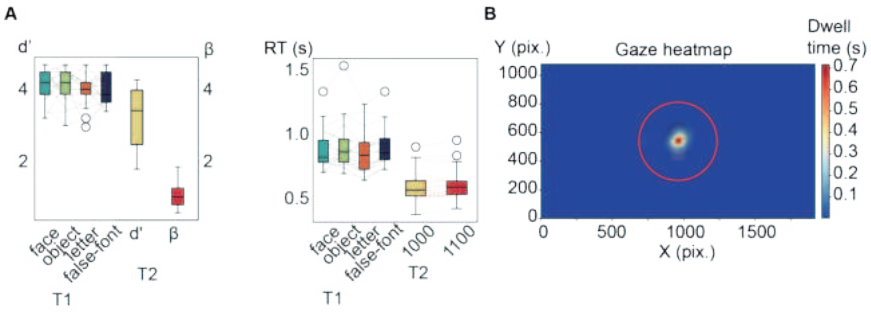
presented (80 trials each), we first averaged the HG signal in 0.01s non-overlapping window to smooth the data. We then used a time-resolved support vector machine (SVM) classifier to decode faces from objects with 5 fold cross-validation. The decoding accuracy was averaged across folds. We repeated this procedure (pseudo trials computations and classification) 5 times to avoid any bias from random splits in the cross-fold validation. The average decoding accuracy in each condition was obtained by averaging across folds and iterations.

Statistical significance of the difference in accuracy between task-relevant and irrelevant trials was obtained using a permutation test by shuffling category labels 10,000 times and repeating the decoding analysis. We corrected for multiple comparisons using cluster-based correction (cluster mass inference with cluster forming threshold as $p < 0.05$, [127,201,202]

Supplementary

Study 1

Behavioural results:



Supplementary figure 3.S1: Study 1 behavioural performance and fixation

A. *Participants behavioural performance and reaction time to each task in the first experiment. The left panel shows participants' behavioural performance in the first and second task. For T1, the d' to each stimulus category is displayed. For T2, the d' depicts the sensitivity in discrimination between high and low pitch sound and the β depicts whether participants' responses are biassed towards one or the other response (β=1 indicates no bias). The right panel depicts the reaction time to the first and second task, for each of the visual stimulus categories and tones respectively.*

B. *Fixation heatmap across all participants and experimental conditions. The x and y axis represent the gaze position on the screen in pixel units and the colour represents the dwell time, i.e. the amount of time (in seconds) spent at a particular location during the trial (from -0.2 to 2.7 s from stimulus onset). The red circle has a diameter of 6° of visual angle and the semi-transparent stimuli represent the stimuli in the dimension they were displayed on the screen.*

Supplementary table 3.S1: Experiment 1 full model results

	Chisq	**Df**	**Pr(>Chisq)**
SOA	735.95	3	**< 0.001*****
Onset/Offset	652.77	1	**< 0.001*****
T1 Task relevance	125.99	1	**< 0.001*****
SOA:Onset/Offset	462.82	3	**< 0.001*****
SOA:T1 Task relevance	10.00	3	**0.02***
Onset/Offset:T1 Task relevance	83.86	1	**< 0.001*****
SOA:Onset/Offset:T1 Task relevance	4.11	3	0.250

RT2 ~ SOA×Onset/Offset×T1 Task relevance + (SOA×Onset/offset×T1 Task relevance | Subject)+ (SOA×Onset/offset×T1 Task relevance | Duration)

Supplementary table 3.S2 Experiment 1 onset model results

	Chisq	Df	Pr(>Chisq)
SOA	1109.31	3	**< 0.001*****
T1 Task relevance	192.37	1	**< 0.001*****
SOA:T1 Task relevance	12.82	3	**0.005****

RT2* ~ SOA×T1 Task relevance+ (SOA×T1 Task relevance | Subject)+ (SOA×T1 Task relevance | Duration)
* Data restricted to onset-locked trials

Supplementary table 3.S3: Experiment 1 offset model results

	Chisq	Df	Pr(>Chisq)
SOA	42.31	3	**< 0.001*****
T1 Duration	10.93	2	**0.004****
T1 Task relevance	4.34	1	**0.037***
SOA:T1 Duration	35.44	6	**< 0.001*****
SOA:T1 Task relevance	1.06	3	0.786
Duration:T1 Task relevance	1.62	2	0.444
SOA:T1 Duration:T1 Task relevance	4.49	6	0.611

RT2* ~ SOA×Duration× T1 Task relevance+ (SOA×Duration×T1 Task relevance | Subject)
* Data restricted to offset-locked trials

Supplementary table 3.S4: Experiment 1 offset model, separately for each T1 durations

T1 Duration		Chisq	Df	Pr(>Chisq)
500 ms	SOA	4.20	3	0.240
	T1 Task relevance	0.19	1	0.659
	SOA:T1 Task relevance	1.32	3	0.724
1000 ms	SOA	12.67	3	**0.005****
	T1 Task relevance	4.95	1	**0.026***
	SOA:T1 Task relevance	0.50	3	0.919
1500 ms	SOA	69.44	3	**< 0.001*****
	T1 Task relevance	0.67	1	0.412
	SOA:T1 Task relevance	4.01	3	0.260

RT2* ~ SOA×T1 Task relevance+ (SOA×T1 Task relevance | Subject)
* Data restricted to offset-locked trials of corresponding T1 duration.

Supplementary table 3.S5: Pairwise comparison of RT2 between SOA 0 and 466 on offset trials separately for each T1 duration

T1 Duration	Difference (s)	z.ratio	Pr(>Chisq)
500 ms	0.009	1.75	0.484
1000 ms	0.015	2.73	**0.038***
1500 ms	0.036	6.40	**< 0.001*****

Supplementary table 3.S6: Comparison of RT2 empirical cumulative distribution of SOA 0ms against 0.466ms

Onset/offset	T1 relevance	Duration (s)	D	p
onset	T1 relevant	all	0.29	**< 0.001*****
	T1 irrelevant	all	0.29	**< 0.001*****
offset	T1 relevant	500 ms	0.05	0.130
		1000 ms	0.08	**0.010***
		1500 ms	0.11	**< 0.001*****
	T1 irrelevant	500 ms	0.04	0.319
		1000 ms	0.05	0.117
		1500 ms	0.07	**0.038***

Kilmogorov Smirnoff test separately for onset/offset T2 lock and T1 relevant/irrelevant trials. In the case of the offset trials, the test was conducted separately on each T1 durations.

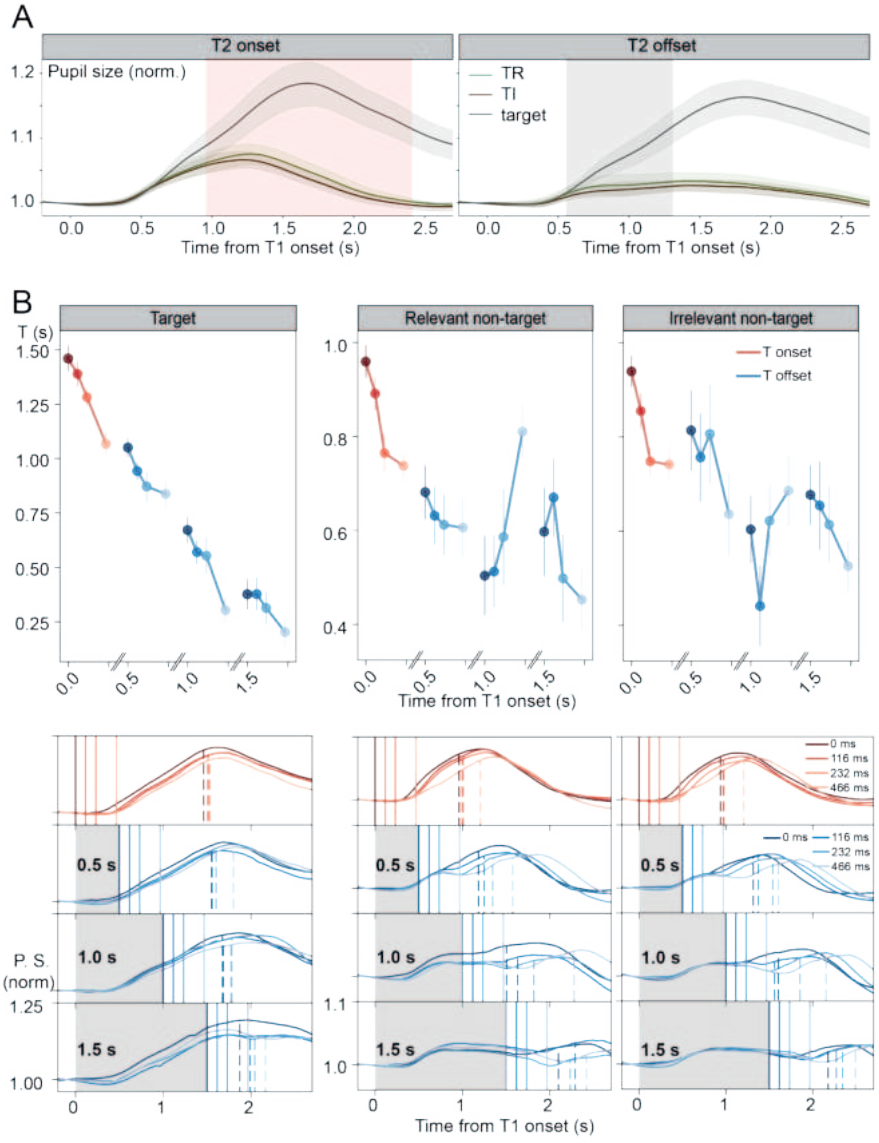
Supplementary table 3.S7: Results of experiment 1 target only models

		Duration (ms)	Chisq	Df	Pr(>Chisq)
SOA	Onset	All	362.47	3	**< 0.001*****
	Offset	500	111.64	3	**< 0.001*****
		1000	37.16	3	**< 0.001*****
		1500	12.01	3	**0.007****

RT2* ~ SOA×T1 Task relevance+ (SOA×T1 Task relevance | Subject)

*RT2 was modelled as a function of SOA separately for onset and offset trials. In the case of the offset trials, RT2 was modelled as a function of SOA separately for each T1 duration.

Eyetracker results:



Supplementary figure 3.S2: Pupil peak amplitude latency as a function of T1 stimulus appearance and disappearance.

a.. *Average pupil size (y-axis) in T1 relevant (green), irrelevant (brown) and target (grey) conditions as a function of time (x-axis) relative to the onset of T1 stimuli, separately for onset (left) and offset trials (right). Shaded areas around the curve represent 95% confidence intervals computed across subjects. Vertical box shading represent segments in which the pupil size is significantly larger in T1 relevant compared to irrelevant trials determined using a cluster based permutation test (red $\alpha < 0.05$, grey $\alpha < 0.1$)*

b. *Average pupil 90% peak latency as a function of SOA (x-axis) in auditory task time-locked to T1 onset (red) and offset (blue), separately for T1 target trials, and T1 non-target task relevant and task irrelevant trials. Upper, leftward panel displays peak latency for targets only (Go trials), red lines indicate peak latency per SOA (0, 116, 232, 466ms) locked to T1 onset.*

Below, average pupil size (y-axis) as a function of time separately for each SOA, onset/offset (red and blue respectively) and T1 duration conditions (each row). The vertical dashed lines represent the average 90% peak latency. The columns correspond to the T1 relevance conditions (left: T1 target, middle: T1 relevant, left: T1 irrelevant). The first rows display the results in onset locked trials and the 3 bottom row depict the pupil response in offset trials separately for each T1 duration, as indicated by the numbers in the margins.

Supplementary table 3.S8: Pupil peak latency onset

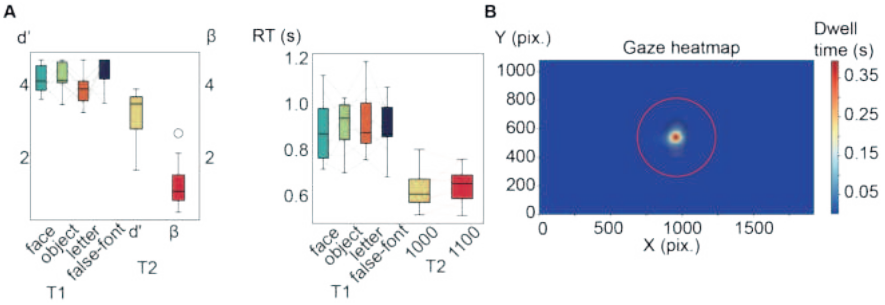
	Chisq	**Df**	**Pr(>Chisq)**
SOA	15.78	3	**0.001****
T1 Task relevance	1.70	1	0.192
SOA:T1 Task relevance	0.48	3	0.922

Supplementary table 3.S9: Pupil peak latency offset

	Chisq	**Df**	**Pr(>Chisq)**
SOA	0.11	3	0.990
T1 Duration	2.03	2	0.362
T1 Task relevance	1.75	1	0.186
SOA:T1 Duration	4.30	6	0.636
SOA:T1 Task relevance	0.71	3	0.871
Duration:T1 Task relevance	0.52	2	0.771
SOA:T1 Duration:T1 Task relevance	0.40	6	0.999

Study 2

Behavioral results:



Supplementary figure 3.S3: Study 2 behavioural performance and fixation

a. *Participants behavioural performance and reaction time to each task in the first experiment. The left panel shows participants' behavioural performance in the first and second task. For T1, the d' to each stimulus category is displayed. For T2, the d' depicts the sensitivity in discrimination between high and low pitch sound and the β depicts whether participants' responses are biassed towards one or the other response (β=1 indicates no bias). The right panel depicts the reaction time to the first and second task, for each of the visual stimulus categories and tones respectively.*

b. *Fixation heatmap across all participants and experimental conditions. The x and y axis represent the gaze position on the screen in pixel units and the colour represents the dwell time, i.e. the amount of time (in seconds) spent at a particular location during the trial (from -0.2 to 2.7 s from stimulus onset). The red circle has a diameter of 6° of visual angle and the semi-transparent stimuli represent the stimuli in the dimension they were displayed on the screen.*

Supplementary table 3.S10: Experiment 2 full model results

	Chisq	Df	Pr(>Chisq)
SOA	308.14	2	**< 0.001*****
Onset/Offset	413.53	1	**< 0.001*****
T1 Task relevance	50.83	1	**< 0.001*****
SOA:Onset/Offset	110.51	2	**< 0.001*****
SOA:T1 Task relevance	1.91	2	0.384
Onset/Offset:T1 Task relevance	10.85	1	**0.001*****
SOA:Onset/Offset:T1 Task relevance	1.14	2	0.566

RT2 ~ SOA×Onset/offset×T1 Task relevance+ (SOA×Onset/offset×T1 Task relevance | Subject)+ (SOA×Onset/offset×T1 Task relevance | Duration

Supplementary table 3.S11: Experiment 2 onset model results

	Chisq	Df	Pr(>Chisq)
SOA	368.03	2	**< 0.001*****
T1 Task relevance	51.32	1	**< 0.001*****
SOA:T1 Task relevance	1.99	2	0.370

RT2* ~ SOA×T1 Task relevance+ (SOA×T1 Task relevance | Subject)+ (SOA× T1 Task relevance | Duration)
* Data restricted to onset-locked trials

Supplementary table 3.S12: Experiment 2 offset model results

	Chisq	Df	Pr(>Chisq)
SOA	54.64	2	**< 0.001*****
T1 Duration	23.38	2	**< 0.001*****
T1 Task relevance	11.27	1	**0.001****
SOA:T1 Duration	19.44	4	**0.001****
SOA:T1 Task relevance	1.54	2	0.462
Duration:T1 Task relevance	1.48	2	0.476
SOA:T1 Duration:T1 Task relevance	3.63	4	0.458

RT2* ~ SOA×Duration× T1 Task relevance+ (SOA×Duration×T1 Task relevance | Subject)
* Data restricted to offset-locked trials

Supplementary table 3.S13: Experiment 2 offset model separately for each T1 duration

T1 Duration		Chisq	Df	Pr(>Chisq)
500 ms	SOA	6.46	2	**0.039***
	T1 Task relevance	2.22	1	0.136
	SOA:T1 Task relevance	0.99	2	0.610
1000 ms	SOA	42.82	3	**< 0.001*****
	T1 Task relevance	8.81	1	**0.003****
	SOA:T1 Task relevance	0.33	3	0.848
1500 ms	SOA	30.73	3	**< 0.001*****
	T1 Task relevance	2.05	1	0.152
	SOA:T1 Task relevance	5.29	3	0.071

RT2* ~ SOA×T1 Task relevance+ (SOA×T1 Task relevance | Subject)
* Data restricted to offset-locked trials of corresponding T1 duration.

Supplementary table 3.S14: Pairwise comparison of RT2 between SOA 0 and 466ms on offset trials separately for each T1 duration

T1 Duration	Difference (s)	z.ratio	Pr(>Chisq)
500 ms	0.026	2.25	0.0732
1000 ms	0.064	5.67	**< 0.001*****
1500 ms	0.053	4.80	**< 0.001*****

Supplementary table 3.S15: Experiment 2 comparison of RT2 cummulative distribution of SOA 0 ms against 0.466 s

Onset/offset	T1 relevance	Duration (s)	D	p
onset	T1 relevant	all	0.33	**< 0.001*****
	T1 irrelevant	all	0.37	**< 0.001*****
offset	T1 relevant	500 ms	0.06	0.420
		1000 ms	0.19	**0.001****
		1500 ms	0.21	**< 0.001*****
	T1 irrelevant	500 ms	0.10	0.103
		1000 ms	0.23	**< 0.001*****
		1500 ms	0.16	**0.003****

Kilmogorov Smirnoff test separately for onset/offset T2 lock and T1 relevant/irrelevant trials. In the case of the offset trials, the test was conducted separately on each T1 durations.

Supplementary table 3.S16: Experiment 2 target only model results.

		Duration (ms)	Chisq	Df	Pr(>Chisq)
SOA	Onset	All	62.59	3	**< 0.001*****
	Offset	500	5.07	3	0.080
		1000	4.22	3	0.121
		1500	7.13	3	0.028*

RT2* ~ SOA×T1 Task relevance+ (SOA×T1 Task relevance | Subject).
RT2 was modelled as a function of SOA separately for onset and offset trials. In the case of the offset trials, RT2 was modelled as a function of SOA separately for each T1 duration.

Supplementary table 3.S17: iT2 full model results

	Chisq	Df	Pr(>Chisq)
SOA	40.31	2	**< 0.001*****
Onset/Offset	27.49	1	**< 0.001*****
T1 Task relevance	21.51	1	**< 0.001*****
SOA:Onset/Offset	15.26	2	**< 0.001****
SOA:T1 Task relevance	0.77	2	0.680
Onset/Offset:T1 Task relevance	2.92	1	0.087
SOA:Onset/Offset:T1 Task relevance	4.88	2	0.087

iT2 ~ SOA×Onset/offset×T1 Task relevance+ (SOA×Onset/offset×T1 Task relevance | Subject)+ (SOA×Onset/offset×T1 Task relevance | Duration)

Supplementary table 3.S18: iT2 onset model results

	Chisq	Df	Pr(>Chisq)
SOA	48.33	2	**< 0.001*****
T1 Task relevance	17.98	1	**< 0.001*****
SOA:T1 Task relevance	0.99	2	0.609

iT2* ~ SOA×T1 Task relevance + (SOA×T1 Task relevance | Subject)
*Data restricted to onset-locked trials

Supplementary table 3.S19: iT2 offset model results

	Chisq	Df	Pr(>Chisq)
SOA	3.83	2	0.148
T1 Duration	4.76	2	**0.029***
T1 Task relevance	4.78	1	0.092
SOA:T1 Duration	4.79	4	0.091
SOA:T1 Task relevance	2.63	2	0.622
Duration:T1 Task relevance	0.68	2	0.712
SOA:T1 Duration:T1 Task relevance	1.73	4	0.785

iT2* ~ SOA×Duration× T1 Task relevance + (SOA×Duration×T1 Task relevance | Subject)
* Data restricted to offset-locked trials

Supplementary table 3.S20: iT1 full model results

	Chisq	Df	Pr(>Chisq)
SOA	15.62	2	**< 0.001*****
Onset/Offset	28.39	1	**< 0.001*****
T1 Task relevance	315.77	1	**< 0.001*****
SOA:Onset/Offset	17.04	2	**< 0.001*****
SOA:T1 Task relevance	0.30	2	0.862
Onset/Offset:T1 Task relevance	3.05	1	0.081
SOA:Onset/Offset:T1 Task relevance	0.70	2	0.704

iT1 ~ SOA×Onset/offset×T1 Task relevance + (SOA×Onset/offset×T1 Task relevance | Subject)+ (SOA×Onset/offset×T1 Task relevance | Duration)

Supplementary table 3.S21: iT1 onset model results

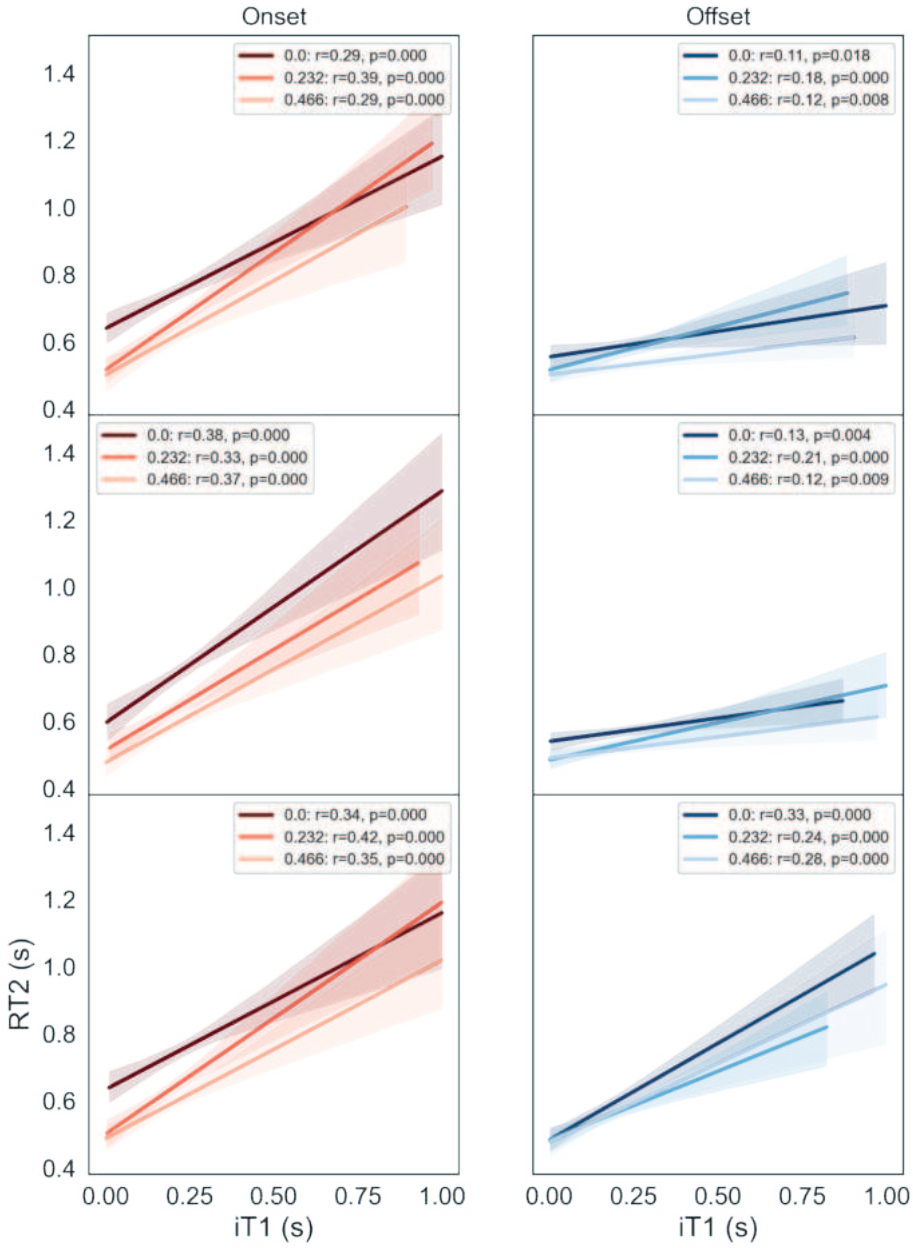
	Chisq	Df	Pr(>Chisq)
SOA	30.02	2	**< 0.001*****
T1 Task relevance	189.78	1	**< 0.001*****
SOA:T1 Task relevance	0.15	2	0.929

iT1* ~ SOA×T1 Task relevance+ (SOA×T1 Task relevance | Subject)
* Data restricted to onset-locked trials.

Supplementary table 3.S22: iT1 onset model results

	Chisq	Df	Pr(>Chisq)
SOA	0.62	2	0.734
T1 Duration	128.73	2	**<0.001*****
T1 Task relevance	7.25	1	**0.027***
SOA:T1 Duration	0.89	4	0.641
SOA:T1 Task relevance	4.43	2	0.352
Duration:T1 Task relevance	0.49	2	0.781
SOA:T1 Duration:T1 Task relevance	2.81	4	0.590

iT1* ~ SOA×Duration× T1 Task relevance+ (SOA×Duration×T1 Task relevance | Subject)
* Data restricted to offset-locked trials.



Supplementary figure 3.S4: Correlation between RT2 and iT1 for each duration and onset offset trials

Reaction time to the auditory stimulus (RT2, y-axis) as a function iT1 (x-axis) separately for each SOA condition (dark shades: SOA of 0s, light shade: SOA of 0.466s), T1 duration (top: 500, middle: 1000, bottom: 1500 ms) and T1 onset locked trials and offset locked trials (left column and red lines: onset locked, right column and blue lines: offset locked)

Supplementary table 3.S23:

	Chisq	Df	Pr(>Chisq)
SOA	306.76	2	**< 0.001*****
Onset/Offset	362.51	1	**< 0.001*****
T1 Task relevance	1.02	1	0.312
iT1	835.57	1	**< 0.001*****
SOA:Onset/Offset	99.47	2	**< 0.001*****
SOA:T1 Task relevance	3.19	2	0.203
Onset/Offset:T1 Task relevance	4.99	1	**0.026***
SOA:iT1	3.01	2	0.222
Onset/Offset:iT1	12.00	1	**< 0.001*****
Task relevance:iT1	0.01	1	0.937
SOA:Onset/Offset:Task relevance	1.08	2	0.582
SOA:Onset/Offset:iT1	0.64	2	0.726
SOA:T1 Task relevance:iT1	2.08	2	0.353
Onset/Offset:T1 Task relevance:iT1	0.54	1	0.463
SOA:Onset/Offset:T1 Task relevance:iT1	6.70	2	**0.035***

RT2 ~ SOA×Onset/offset×T1 Task relevance ×iT1 + (SOA×Onset/offset×T1 Task relevance ×iT1 | Subject)+ (SOA×Onset/offset×T1 Task relevance ×iT1 | Duration)

Supplementary table 3.S24: iT1-RT2 model results in onset trials

	Chisq	Df	Pr(>Chisq)
SOA	363.48	2	**< 0.001*****
T1 Task relevance	3.94	1	**0.047***
iT1	614.01	1	**< 0.001*****
SOA:T1 Task relevance	1.82	2	0.403
SOA:iT1	1.34	2	0.512
T1 Task relevance:iT1	0.00	1	0.975
SOA:T1 Task relevance:iT1	9.26	2	**0.010***

RT2* ~ SOA×T1 Task relevance ×iT1+ (SOA×T1 Task relevance×iT1 | Subject)
* Data restricted to onset-locked trials

Supplementary table 3.S25: iT1-RT2 model results in offset trials

	Chisq	Df	Pr(>Chisq)
SOA	56.81	2	**< 0.001*****
T1 Duration	19.55	1	**< 0.001*****
iT1	259.31	1	**< 0.001*****
SOA:T1 Duration	20.11	2	**< 0.001*****
SOA:iT1	1.91	2	0.386
T1 Duration:iT1	18.28	1	**< 0.001*****
SOA:T1 Duration:iT1	13.13	2	**0.011***

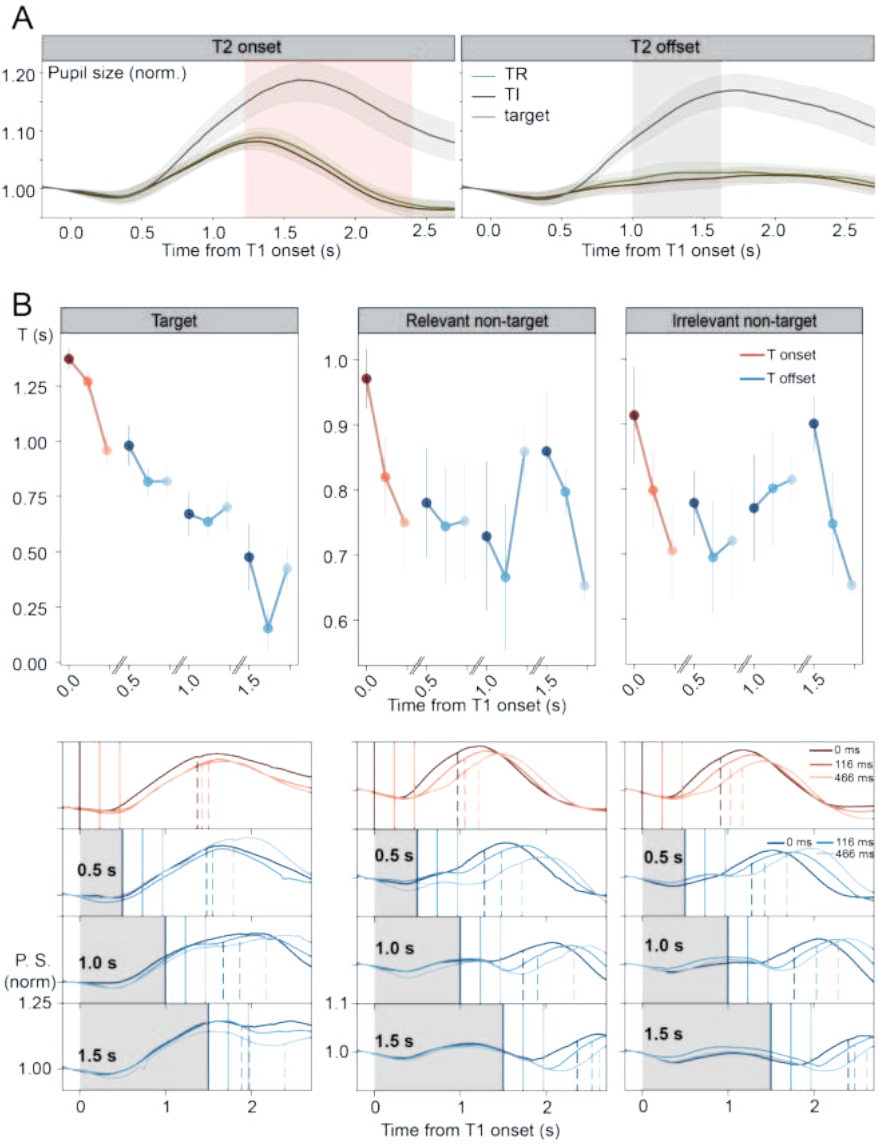
RT2* ~ SOA× Duration×iT1+ (SOA×Duration×iT1 | Subject
* Data restricted to offset-locked trials

Supplementary table 3.S26: iT1-RT2 offset model separately for each of the T1 durations

T1 Duration (ms)		Chisq	Df	Pr(>Chisq)
500	SOA	4.31	2	0.116
	iT1	133.74	1	**< 0.001*****
	SOA:iT1	2.60	1	0.272
1000	SOA	48.75	2	**< 0.001*****
	iT1	82.97	1	**< 0.001*****
	SOA:iT1	11.37	2	**0.003****
1500	SOA	30.92	2	**< 0.001*****
	iT1	41.14	1	**< 0.001*****
	SOA:iT1	0.45	2	0.798

RT2* ~ SOA×iT1 (SOA×iT1 | Subject)
* Data restricted to offset-locked trials of corresponding T1 duration.

Eyetracker results



Supplementary figure 3.S5: Pupil peak amplitude latency as a function of T1 stimulus appearance and disappearance

a. *Average pupil size (y-axis) in T1 relevant (green), irrelevant (brown) and target (grey) conditions as a function of time (x-axis) relative to the onset of T1 stimuli, separately for onset (left) and offset trials (right). Shaded areas around the curve represent 95% confidence intervals computed across subjects. Vertical box shading represent segments in which the pupil size is significantly larger in T1 relevant compared to irrelevant trials determined using a cluster based permutation test (red $\alpha < 0.05$, grey $\alpha < 0.1$)*

b. *Average pupil 90% peak latency as a function of SOA (x-axis) in auditory task time-locked to T1 onset (red) and offset (blue), separately for T1 target trials, and T1 non-target task relevant and task irrelevant trials. Upper, leftward panel displays peak latency for targets only (Go trials), red lines indicate peak latency per SOA (0, 232, 466ms) locked to T1 onset.*

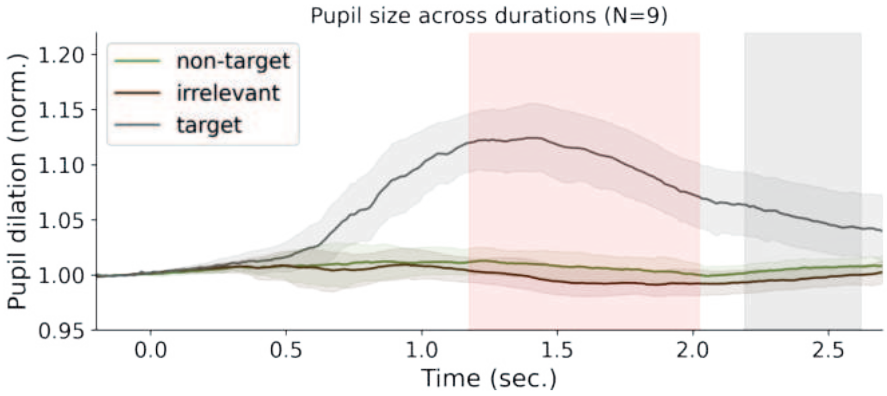
Below, average pupil size (y-axis) as a function of time separately for each SOA, onset/offset (red and blue respectively) and T1 duration conditions (each row). The vertical dashed lines represent the average 90% peak latency. The columns correspond to the T1 relevance conditions (left: T1 target, middle: T1 relevant, left: T1 irrelevant). The first rows display the results in onset locked trials and the 3 bottom row depict the pupil response in offset trials separately for each T1 duration, as indicated by the numbers in the margins.

Supplementary table 3.S27: Experiment 2 Pupil peak latency onset

	Chisq	**Df**	**Pr(>Chisq)**
SOA	9.24	3	**0.002**
T1 Task relevance	0.94	1	0.331
SOA:T1 Task relevance	0.09	3	0.764

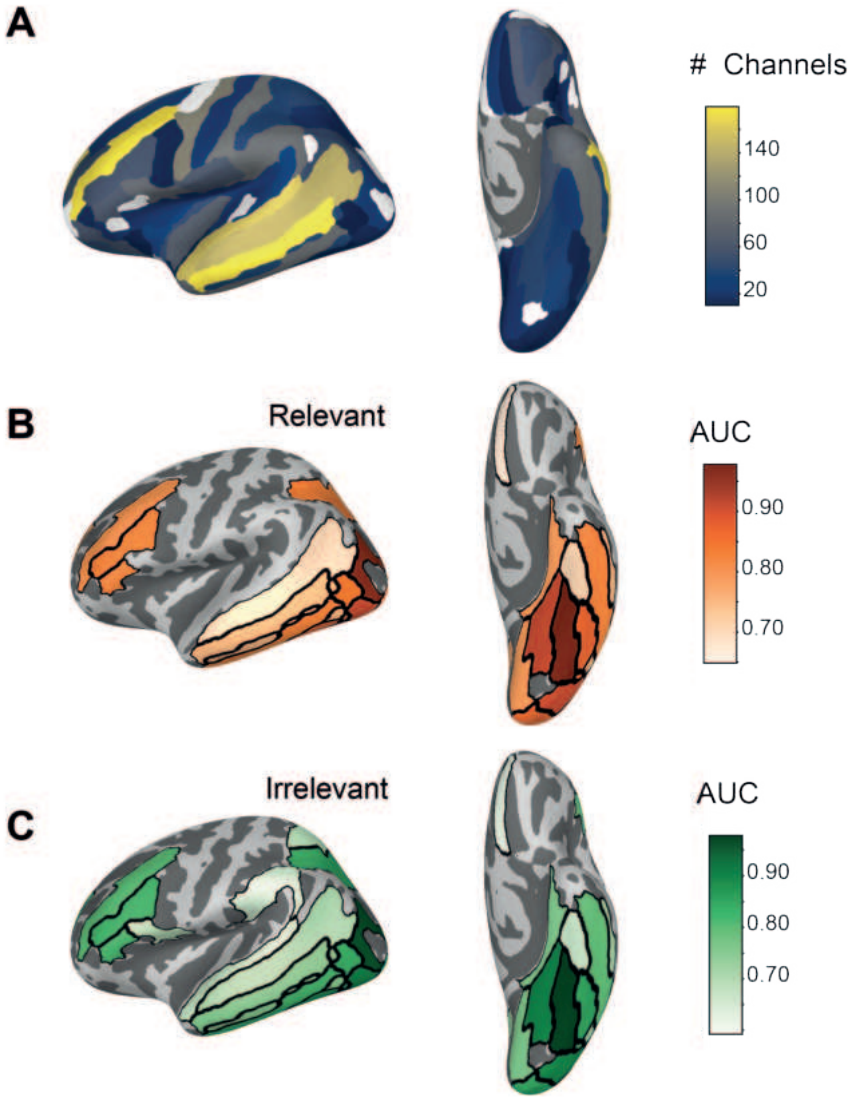
Supplementary table 3.S28: Experiment 2 pupil peak latency offset

	Chisq	**Df**	**Pr(>Chisq)**
SOA	0.33	3	0.565
T1 Duration	0.13	2	0.937
T1 Task relevance	0.01	1	0.932
SOA:T1 Duration	2.20	6	0.333
SOA:T1 Task relevance	0.10	3	0.751
Duration:T1 Task relevance	0.02	2	0.989
SOA:T1 Duration:T1 Task relevance	0.02	6	0.991



Supplementary figure 3.S6: Pupil dilation comparison between task relevant and irrelevant condition in the Cogitate data set

Average pupil size (y-axis) in relevant (green), irrelevant (brown) and target (grey) conditions as a function of time (x-axis) relative to the onset of T1 stimuli, separately for onset (left) and offset trials (right). Shaded areas around the curve represent 95% confidence intervals computed across subjects. Vertical box shading represent segments in which the pupil size is significantly larger in T1 relevant compared to irrelevant trials determined using a cluster based permutation test (red $\alpha < 0.05$*)*

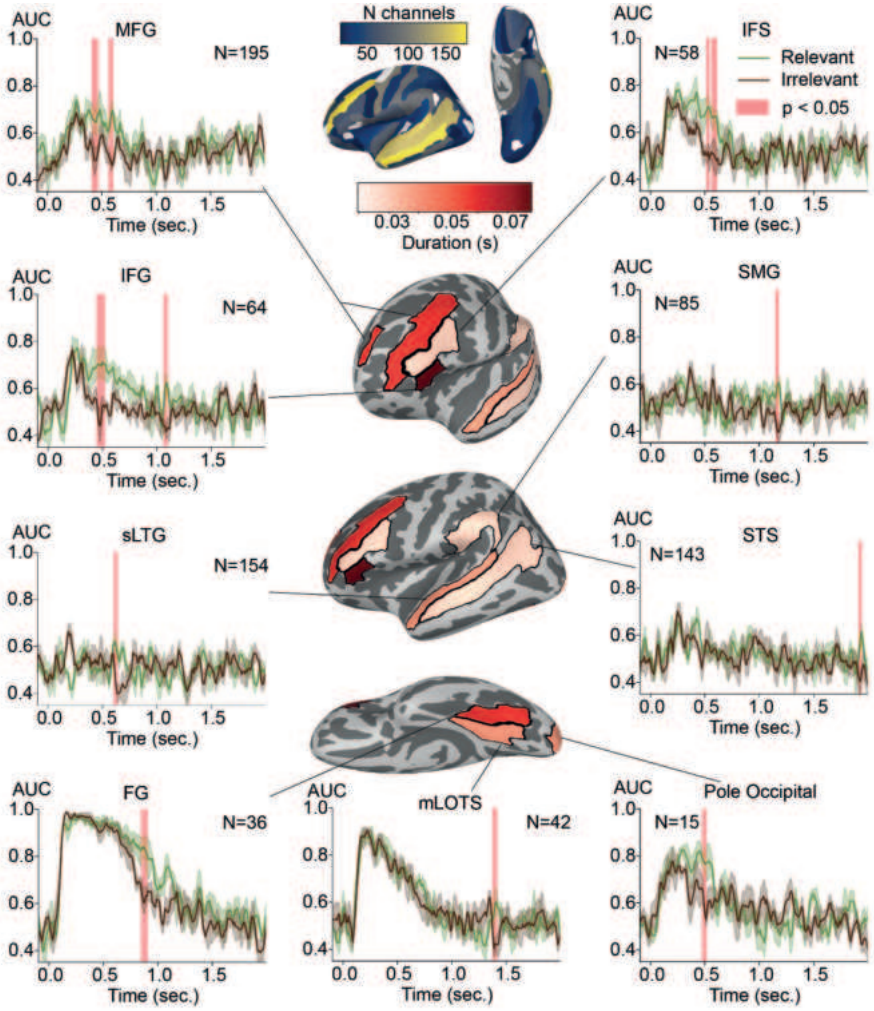


Supplementary figure 3.S7: Density of electrode coverage and decoding results

a. *brain surface coloured according to the number of electrodes found in each region (Destrieux Atlas) across subjects (N=29). Areas marked in white correspond to those excluded from the analysis as they contained less than 10 electrodes*

b. *Maximal decoding AUC (faces vs. objects) in the task relevant trial masked by significance decoding*

c. *Maximal decoding AUC (faces vs. objects) in the task irrelevant condition masked by significance decoding*



Supplementary figure 3.S8: Time-Resolved Face/Object Decoding using invasive electrophysiological data from the Cogitate Study

Using the Cogitate *iEEG data to compare face vs. objects decoding AUC in task-relevant (green) trials compared to task-irrelevant (brown) trials with cluster-based permutation test. Shading indicates the 95% confidence interval across cross-validation folds. Red shading represents significant clusters (p < 0.01, cluster-based permutation test). "N" indicates the number of channels per region. Temporal smoothing of the decoding time series using a uniform 40 ms kernel was applied for plotting purposes only. The upper middle panel depicts the coverage in the Cogitate iEEG sample, with each cortical parcel from the Destrieux atlas color-coded by the number of electrodes present. White areas indicate regions with fewer than 10 channels, which were omitted from the analysis. The middle panel shows a brain surface map highlighting four cortical parcels from the Destrieux atlas where decoding accuracy for task-relevant trials was significantly higher than for task-irrelevant trials. Colors indicate the duration of the higher decoding in the task-relevant condition. Abbreviations: MFG (middle frontal gyrus), IFS (inferior frontal sulcus), IFG (inferior frontal gyrus), FG (fusiform gyrus), sLTG (superior lateral temporal gyrus), mLOTS (medial and lingual occipital temporal sulcus), STS (superior temporal sulcus), SMG (inferior temporal supramarginal gyrus).*

Chapter 4

Discussion

In the introduction, I proposed to accelerate progress in consciousness research by identifying cases where a given content is experienced, but the mechanisms proposed by a theory to instantiate it are not observed. This approach constitutes a shift from the traditional search for the NCCs and requires less restrictive experimental conditions, as the need to control for unconscious processing is alleviated. Theories of consciousness can accordingly be tested across a wider set of experimental conditions previously untested, forcing them to formulate novel predictions to put their explanatory power to the test.

To that end, I have relied on one aspect of consciousness that has so far received little attention: the temporal dynamics of conscious experiences. We experience particular contents for particular durations, and if a theory truly explains consciousness, it must be able to account for this aspect (among all others). I relied on a rather simple experimental paradigm in which visual stimuli were presented for three distinct durations while collaborating with proponents of the theories to ensure that these conditions matched their criteria for consciousness and that the theories were truly based on their predictions rather than accommodating the results a posteriori.

In this discussion, I begin by providing an independent analysis of the results presented in Chapter 2, situating them within the broader context of vision science and conscious research. While this study was a collaborative effort involving many researchers, I will provide intellectual insights going beyond the collective interpretation presented in Chapter 2. I will discuss how our results have advanced our understanding of the neural mechanisms associated with sustained visual presentation. Building upon this, I will explore the broader implications of both Chapter 2 and 3 in conjunction, illustrating how they open new avenues for investigating the dissociation between access and phenomenal consciousness. Subsequently, I further elaborate on the fundamental goals of theory testing in consciousness research. In line with the Lakatosian view of scientific progress, I argue that theory testing should not aim to eliminate current theories, but should instead be viewed as a process of refining and improving them toward the goal of a unified theory of consciousness. I will specifically highlight the value of adversarial collaboration in generating novel predictions and formalizing theories more effectively than testing them in isolation, while acknowledging the difficulty of obtaining opposing predictions from competing theories in the field of consciousness research. Finally, I will propose several concrete steps to improve the theory-testing process, informed by the challenges I encountered and the insights I gained throughout my research. These recommendations entail general guidelines but also highlight additional scientific efforts I have undertaken to address obstacles in the

field. By sharing these insights and outlining practical measures, I aim to contribute to the advancement of consciousness research, helping to foster a more effective, integrative, and iterative approach to testing theories of consciousness.

Temporal dynamics of conscious experience and the underlying neural activity

In the previous two chapters, I presented the results of experiments in which highly visible stimuli were presented for three different durations. In the first study, we recorded neural data using three recording modalities (iEEG, MEG, and fMRI) to investigate the neural dynamics associated with such stimulus durations to test the predictions of IIT and GNWT regarding the neural dynamics to be observed under such stimulation conditions. In the posterior region of interest defined by IIT (encompassing the occipital and ventral temporal cortices, which I will refer to as the posterior ROI), sustained activation and content representation were observed, matching the duration for which the stimulus was on the screen. On the other hand, in the prefrontal region defined by GNWT (which I will refer to as the PFC), only transient responses and content representation were observed following stimulus onset, with virtually no further coupling with stimulus durations. Our results therefore align with IIT but challenge GNWT predictions.

IIT's prediction through the lens of vision neuroscience

In the discussion section of the second chapter, Prof. Dehaene argues that the prediction of IIT regarding sustained activation is trivial (the meaning of which I will elaborate on later in the discussion), as 'any physiologist familiar with the bottom-up response properties of those regions' would also have made the same prediction. There is a sense in which this is true. The posterior ROI defined by IIT contains the occipital and the ventral temporal cortices, which play a prominent role in visual information processing. Specifically, these combined regions largely overlap with the ventral stream, which is widely accepted as being functionally specialized to recognize shapes and objects [265–268]. As such, it is indeed fully expected that the category and identity of visual stimuli should be decodable from these brain regions, and had that not been the case, we indeed would have had bigger fish to fry [269]. In fact, when developing our analysis pipelines, finding the strongest activation and content representation in these regions acted as a sanity check regarding the sensitivity of our methods.

However, the prediction is less trivial when considering its temporal aspect. IIT does not only predict that there should be activation and visual content representation in these regions but also that these patterns should be sustained and stable for as long as a stimulus is presented on the screen. While the literature clearly shows that the ventral stream is involved in object recognition, most studies characterizing the functional specificity of cortical regions along this pathway relied on short, transient stimulus presentation [203,265–268]. Some studies have manipulated stimulus duration but only within a short range (i.e. below 500 ms) to characterize the speed with which object recognition can be performed in the visual system [265,270]. Therefore, it is unclear from this line of evidence whether the category-selective activation observed along the ventral stream is a transient process solely involved in the detection of a visual stimulus of a particular category, or if it instead plays a role in real-time monitoring of the presence of a particular percept, which would involve sustained activation.

In addition, visual adaptation has been shown to occur at various levels of the visual processing hierarchy. Neural responses following the appearance of a visual stimulus decrease rapidly following stimulus onset [271–273]. Similarly, when the same visual stimulus is presented in rapid succession, the amplitude of the neural responses induced by the stimulus appearance decreases, an effect known as repetition suppression [273–276]. A recent iEEG study (from which our experimental design was inspired) has shown that only a minority of visually responsive electrodes (21 out of 292) were sensitive to stimulus duration [113]. Furthermore, they showed that the proportion of duration-sensitive electrodes strongly decreases along the visual hierarchy and is minimal in inferior-temporal category-selective sites (50% of the visual responsive electrodes show duration tracking in the early visual cortex, only 1.5% in the inferior temporal cortex). Our study broadly replicated these findings, as only 25 out of 194 (13%) visually responsive electrodes showed an association with stimulus duration across the posterior ROI. Furthermore, only 8 out of 53 (15%) face-selective electrodes showed significant duration tracking and were located mostly in the ventral temporal cortex (see Fig. 4.1).

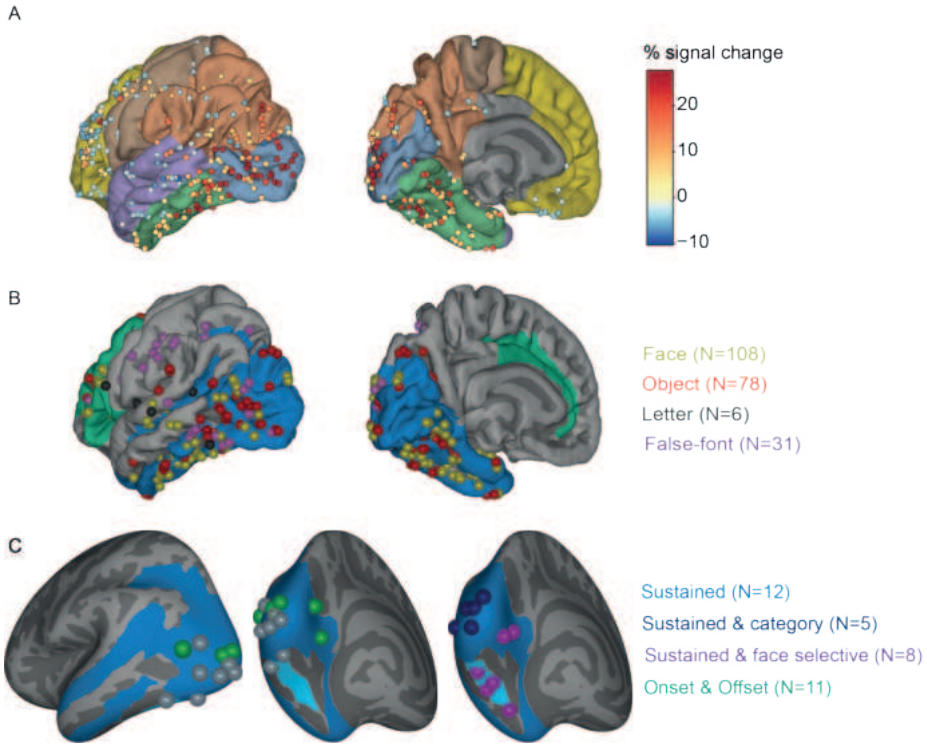


Figure 4.1: Channels responsivity, selectivity and duration tracking (adapted from Chapter 2 supplementary material)

a. *Localization of all channels showing a significant change in activation following stimulus onset (50-350 ms) color coded by the percentage of signal change displayed on the fsaverage pial surface. The colors on the brain surface correspond to broad anatomical regions (blue: occipital, green: ventral temporal, purple: temporal, orange: parietal, brown: motor, yellow: frontal cortices).*

b. *Localization of all category-selective channels (significantly higher activation in one condition compared to all others from 50-350 ms from stimulus onset). The color of each contact represents the category the contact is selective for (yellow: faces, red: objects, grey: letters, purple: false-fonts) and the color on the brain surface represents the regions of interest defined by the theories (blue: IIT, green: GNWT).*

c. *Localization of channels showing an association with stimulus duration (only depicting posterior ROI results, the light blue region on the brain is the fusiform gyrus). Electrodes are color-coded based on the response type that they display. Light blue electrodes show sustained activation for the duration of stimulus presentation regardless of category, dark blue show sustained activation that is further modulated by stimulus duration, purple electrodes show sustained activation only to faces stimuli and green electrodes show transient activation following both the onset and the offset of the visual stimulus.*

Sustained category-selective activation in the posterior ROI is far from ubiquitous, and I would therefore contend that IIT's prediction was not trivial. It can be argued that some of the sustained activation observed in posterior electrodes may reflect the fact that visual input was not truly stable, despite participants being instructed to fixate in studies presenting visual stimuli for longer periods [113]. Even when participants fixate, the position of the stimuli is not exactly fixed on the retina, as small eye movements remain (micro-saccades, tremors and drifts [277]). Early visual areas have been shown to be sensitive to these fast-paced and spatially constrained changes in the visual input [278–280]. It is however unlikely to be the case for higher-order areas in the ventral stream. As we progress along the posterior-anterior axis, several studies revealed that the spatial and temporal extent of the receptive fields increases [272,275,281,282]. In addition, several studies have revealed that higher-order visual area activation is invariant to changes in the visual input associated with saccades and blinks [114,217]. Furthermore, recent studies have highlighted that adaptation occurs at different time scales along the visual hierarchy, reflecting the difference in the stability of the input drive of different regions [271]. Accordingly, sustained activation observed in lower-level areas may reflect the constant change in visual input associated with minute eye movements. In contrast, as the input to higher-order areas is more stable, responses might become more adapted in most recording sites, accounting for the smaller proportion of channels showing sustained activation.

Nonetheless, sustained activation was observed at all levels of the visual hierarchy in Chapter 2 as well as in the study by Gerber and colleagues [113], including higher-order areas. As discussed by Gerber and colleagues, these results might indicate two different functional specializations in visual information processing. Some units' functions might be to detect and recognize novel objects transiently, while other units are specialized in the monitoring of visual information in real-time, keeping tabs on the external world [113]. Our results further confirm that sustained activation is scarcer than transient activation, and grows scarcer going up the hierarchy of the ventral stream.

Importantly, IIT not only predicted that such sustained category-selective responses should be observed, but it also predicted that the content of experience should be represented in a sustained and stable fashion. In light of the prominence of transient category-selective responses over sustained ones (which was already documented by Gerber and colleagues before the submission of the Cogitate pre-registration [96]), this prediction is also not trivial as it entails that the transient category-selective activations do not contribute to the representations underlying our perception, despite their prominence in magnitude. Furthermore, if the sustained activation

observed at lower levels of the visual hierarchy reflects fast-paced and local changes in the retinal input, these should not support sustained and stable perceptual representations. Under these circumstances, a prediction very similar to that of GNWT made regarding content representation in the PFC could have been imaginable. To spare energy resources, perceptual contents may be represented transiently in higher-order visual areas following the onset and offset of the visual stimuli.

In addition to ours, two separate studies investigated the predictions regarding the temporal dynamics of content representation [126,127]. The results of these studies align with ours, stable and sustained representations of visual information were observed in the occipital and ventral temporal regions. Importantly, not only category-level information was represented in a sustained fashion along the ventral stream, but also stimulus identity information. These results further indicate that perceptual content is not associated with activation magnitude, but instead with the multivariate activation patterns observed across neuronal populations [126].

Cogitate results through the lens of subjective experience

Beyond arguments regarding the triviality of IIT's predictions from a vision science point of view, another argument can be made from the perspective of subjective experience. All recent studies investigating the neural dynamics associated with sustained stimulus presentation (including ours) share the initial assumption that under such conditions, participants' subjective experience is linked in some way to stimulus duration. In the introduction from Broday Dvir and colleagues [126], the authors state the following question: "Simply put, if the magnitude of neuronal activity determines perceptual awareness, how does perception remain stable despite this massive reduction?". Similarly, Vishne and colleagues [127] open their paper by stating that "In essence, every perception has non-zero duration". In our study, this assumption was initially shared by both IIT and GNWT, which is reflected in their predictions.

An important clarification must be made regarding this shared assumption: it is not about the experience of the duration of the stimulus (i.e., time perception), which refers to how long a stimulus is felt to last. Instead, it was assumed that participants experience the persistence of stimuli on the screen, one way or another. According to IIT, at each time point where the stimulus is present on the screen, participants have the experience of the stimulus present on the screen and accordingly, when stimuli are presented for longer, the stimulus presented on the screen remains experienced for longer. This also seems to be the initial assumption of the studies mentioned above [113,126,127]. GNWT's initial assumption was that participants experience the onset

and the offset of the visual stimulus, and as they experience the offset of the visual stimuli, the brain concludes that the visual stimulus persisted since the onset, and it is in that conclusion that participants are experiencing the persistence of the stimulus [96,97,225]. Accordingly, participants experienced that longer stimuli persisted for longer on the screen. Regardless of whether the experience of persistence occurs in real-time or in a post hoc fashion (as GNWT suggests), the underlying neural activation should be modulated by stimulus duration in some way.

In our study, we focused primarily on the regions defined by the theories and observed that PFC activation was invariant concerning stimulus duration (neither in single-channel activation nor in multivariate activation patterns) and that only the posterior ROI showed such an association. Furthermore, Vishne and colleagues observed that only occipital and ventral-temporal cortices are modulated by stimulus duration, while PFC and parietal cortices' representations were invariant to stimulus duration [127]. Interestingly, Broday Dvir and colleagues observed an increase in fronto-parietal electrodes activation following both the onset and the offset of the visual stimulus [126]. However, they did not observe any multivariate content representations across these electrodes at any time points, which they attribute to the poor coverage of these regions in their data set. It must also be noted that in contrast to the study presented in Chapter 2 and the study of Vishne and colleagues [127], participants were required to memorize the presented stimuli to perform a memory task following the experiment and that all stimuli were presented for the same duration[126]. The ignition observed in fronto-parietal electrodes in their study might reflect memory-encoding processes, or expectation related processes, as highlighted in Chapter 3.

Under the assumption that the experience of persistence was influenced by stimulus duration in these studies, the brain regions consciousness involved in consciousness should reflect this temporal aspect of experience. In that sense, it matters little that IIT's prediction is trivial. Together, these studies highlight that the region showing the strongest association with stimulus duration is the ventral visual stream. If the experience was indeed affected by how long the stimulus was presented, this makes the ventral visual stream a better candidate for the neural substrate of visual conscious experience, no matter how unoriginal that may be.

A triple dissociation between access consciousness, phenomenal consciousness, and unconscious processing in sight?

Based on the sustained activation and representation observed over sensory regions and on the transient activation and representations observed in the prefrontal cortex, [127] conclude that 'to the extent conscious experience is continuous, it may rely

on sensory representations, and to the extent experience is discrete, it may rely on prefrontal representations'. However, I would argue that this conclusion mistakenly hinges on the distinction between the continuous vs. discrete nature of conscious experience [283–285]. I believe it is mistaken because it implies that if consciousness is continuous, it is necessarily linked to sensory input, and if it is discrete, it is necessarily dissociated from sensory input. This does not need to be the case. Conscious experience could be continuous and dissociated from sensory input, in which case participants would have continuous experience but no experience of the persistence of the stimulus on the screen. Alternatively, participants' experience could be discrete and still experience the persistence of the stimulus on the screen. In other words, a more accurate conclusion is that to the extent that participants' experience reflected the temporal dynamics of stimulus presentation, it cannot arise from the prefrontal cortex, and to the extent that participants' experience was dissociated from the temporal dynamics of stimulus presentation, it cannot arise from sensory regions.

As Prof. Deheaene suggests in the discussion of Chapter 2, the lack of PFC ignition at stimulus offset may indicate that participants' experience was indeed dissociated from the temporal dynamics of stimulus presentation and that they may not in fact experience the persistence of the stimuli on the screen. The result of the study in Chapter 3 provides evidence supporting this view. Using the PRP as a time-resolved marker for conscious access, we observed that while this effect was present at stimulus offset, it was much weaker compared to the onset of the visual stimuli and not systematic.

These results indicate that the offset is not systematically consciously accessed or that when they do so, they do so only very briefly. This is perhaps not too surprising: as the duration of the stimuli was not relevant to the task, participants may not have dedicated cognitive resources to 'do something with it' in most trials, because they did not have to. These results enable us to reconcile the lack of PFC activation in Chapter 2 with the GNWT. If participants access the offset only very briefly and only on some trials, PFC ignition may have been too brief or sporadic (across trials) to be detected in neural recordings.

These findings support the view that participants only accessed the visual stimuli transiently rather than in a sustained fashion. However, this brings us to one of the largest and longest lasting debate the field of consciousness, whether access and phenomenal consciousness are distinct or distinguishable [252,253]. Phenomenal consciousness refers to the subjective quality of an experience (seeing blue vs.

seeing red) while the latter refers to the capacity to 'do something' with that content (i.e. engaging cognitive functions such as evaluating it, thinking about it, and reporting it). Some authors argue that access has in fact little to do with consciousness and that studying it amounts to studying cognitive confounds [64]. Others argue that there is no distinction between the two as it is by accessing representation that we become aware of them [286]. Still, others argue that even if there is such a thing as inaccessible conscious experiences, such cases can never be investigated scientifically as such contents are by definition remain private and unmeasurable from a third-person perspective [45,253].

Provided that only accessed contents are experienced, our results would indicate that indeed, participants' experience of the stimuli was transient and decoupled from stimulus duration. Alternatively, participants might have accessed conscious representations only long enough to classify the stimuli and select a response. Afterward, they may have continued to experience the stimulus without actively accessing it, which would constitute the highly sought-after condition of conscious experience in the absence of cognitive confounds [48,49,64].

Arbitrating between these two views brings us to the very reason why the debate between access and phenomenal consciousness has endured for so long: how can we know if participants' experience is limited to what is accessed? It has been proposed that this problem is dealt with by no-report paradigms. If participants are not required to report about a stimulus, they would passively experience the stimulus without accessing it, because they have no reason to do so [47]. Accordingly, differences in neural activation between both conditions should reveal the neural correlates of phenomenal experience, in other words, the true NCCs [48,49]. However, even when participants are not instructed to report the content of their experience, they might still spontaneously engage in post-perceptual cognitive processes related to their experience, or as Ned Block would put it, 'You cannot stop […] monkeys thinking' [64]. Therefore, it is at present unclear whether studies relying on the no-report method have investigated phenomenal consciousness in the absence of access or access consciousness in the absence of a report [52,55,56,59–61,86].

Progress in the debate between access and phenomenal consciousness would greatly benefit from knowing whether unreported stimuli were accessed or not. This is tricky, as how can we know whether it was accessed if participants do not report it? Ned Block argues [64] that this can be ensured by carefully designing 'no-cognition' experimental paradigms, in which post-perceptual processes do not occur. However, it is quite challenging to do so, as it requires a priori knowledge on how to limit post-perceptual processes.

I would argue that the work presented in Chapter 3 constitutes a more practical solution to this problem. Under the assumption that conscious access is a serial process, if a particular stimulus is accessed, it will interfere with the processing of a subsequent stimulus, even if no report is required of the stimulus being accessed. Therefore, the PRP effect constitutes a time-resolved marker that can be used to infer whether a given event was consciously accessed, even in the absence of a report. By combining this approach with no-report paradigms, it becomes possible to know if unreported events were consciously accessed. In other words, the PRP method enables finessing the need to a priori define conditions in which post-perceptual processes will be limited [64] and instead directly test whether that is th e case in existing paradigms.

One problem remains: how can we distinguish unaccessed yet consciously experienced stimuli from unconsciously experienced ones? Unaccessed contents are by definition unreportable, so we cannot rely on subjective reports of participants to differentiate them from unconscious contents. As solution to this last issue, I appeal to the neural stance proposed by Lamme [287]. By combining no-report paradigms with the PRP method, we can determine whether the stimuli in the condition that is considered conscious (through any adequate means) were accessed. If they were not accessed but the neural response differs from the unconscious condition, this would show a three-way dissociation between conscious access, phenomenal consciousness, and unconscious processing. The existence of such conditions remains speculative, but at least we now have a method to start searching.

Perhaps evidence for such conditions is not as far as one might think. In a recent study, ambiguous face stimuli were presented for 500 and 1000 ms, and participants only became aware of the faces if informed beforehand (one-shot learning) [288]. Comparing the ERP components from the EEG recordings, the aware group displayed a sustained visual awareness negativity component (VAN) throughout the entire duration of the stimulus, unlike the unaware group. Clearly, there is something different between the conscious and unconscious condition. However, it is possible that participants consciously accessed the stimuli for as long as they were presented on the screen, accounting for the sustained difference between the seen and unseen conditions. This is of course unlikely given the results presented in Chapter 3, suggesting that under sustained visual presentation, conscious access is transient. Nonetheless, this has not been explicitly tested in this study. If a follow-up study were to demonstrate that a PRP effect is absent or only transient when participants are informed of the presence of the stimuli, these results would constitute a strong piece of evidence that participants' experience is indeed sustained for the entire duration of the stimulus but only sometimes and transiently access unreported contents.

The bearing of evidence on theories

As I have outlined in my introduction, consciousness research faces a major challenge: the proliferation of multiple and often incompatible theories evolving in parallel [20,79,81,82]. Not all these theories can be true at the same time and therefore, achieving a unified scientific theory of consciousness requires rigorous empirical testing of existing theories to begin the process of pruning the theoretical jungle [78]. To achieve this, I have proposed an alternative to the traditional contrastive method, by trying to identify necessity dissociations, cases in which consciousness occurs in the absence of the mechanisms proposed to instantiate it by the theory. Both in Chapters 2 and 3, some of the predictions made by the theories failed. This begs the question: do these failed predictions entail that we have unequivocally established the existence of such dissociation between consciousness and the mechanisms proposed by the theories? And should we, based on these results, consider the theories falsified and abandon them?

The Lakatosian view on scientific progress and theory testing as theory refinement

Answering yes to these questions would be committing to a Popperian view of falsificationism [136]: a failed prediction entails that the theory is wrong and any ad hoc explanations provided by the theorists would constitute desperate attempts to save their failed theories. In line the with the collective interpretation presented in Chapter 2, I do not believe that this is an adequate conclusion, for this would be throwing the baby with the bathwater. It would imply that the experiments I have presented functioned as an *experimentum crucis*–a decisive test that can definitely falsify a theory [289]. This is hardly tenable in our case and more generally in young developing fields as complex as the neuroscience of consciousness [100,104,290].

One reason to refrain from such a reading is the problem of underdetermination, as articulated by the Duhem-Quine thesis [291]. Underdetermination implies that empirical evidence alone is insufficient to establish the validity of a single scientific theory, as the same empirical observation can be predicted by many different theories rather than a single one. Furthermore, predictions are never tested in isolation; instead, we are testing a whole network of assumptions and hypotheses [292]. A failed prediction indicates that something within this network is wrong, but it does not specify which assumption or hypothesis is at fault. Consequently, a confirmed prediction in a single study does not entail that a theory is confirmed and a failed prediction does not entail that the theory is wrong [140].

This may seem rather obvious. In both chapters, when testing predictions, both theories relied on several auxiliary assumptions: that the signal of interest was strong enough to be detected in our data (or that it would be comparable in size to studies, on which we based our power calculations), that our analytical tools were sufficiently sensitive to detect those signals. These are of course the usual suspects among a large amount of auxiliary assumptions. Accordingly, when predictions fail, we can never establish with certitude that the prediction under test is itself at fault, or if for example, a more trivial issue is to blame. For example, in Chapter 2, IIT's proponent argues that the lack of synchrony between low and high-level visual areas might be due to the limitations of the spatial resolutions available in our recordings, rather than the absence of synchrony between these brain regions. Similarly, in Chapter 3, I suggested that offset ignition may have been present in Chapter 2, but in too few trials to be detected from the neural recordings.

In the case of consciousness research, the issue of underdetermination is particularly severe. I have explained in the introduction that consciousness cannot be measured directly and must therefore it must be inferred [71,72,83], in our case based on theoretical considerations and knowledge amassed in previous studies (in the case of Chapter 3). Accordingly, a failed prediction might entail that the inference about consciousness was misled, rather than the predicted observation [83]. This kind of argumentation can be seen across both chapters. In the second chapter, Prof. Dehaene argues that the lack of offset ignition (and content representation) in the PFC might indicate that the initial assumption of experience of persistence may have been misled, and had it been otherwise, the predicted neural activation would have been observed. As a result, it was inferred in Chapter 3 that participants never experienced the disappearance of the stimuli (based on the results of Chapter 2). Based on the observation of a PRP effect at the offset of stimuli, I myself proposed that participants might after all have experienced the disappearance of the stimulus on some occasions.

From the Duhem-Quine thesis, it follows that the ad-hoc justifications provided by theorists and scientists to account for failed predictions may very well be justified. It is indeed quite reasonable to suggest that predictions' failures can result from noise, flawed inferences, or many other factors that do not directly implicate the core theory. Thus, we can never, based on failed predictions, unequivocally conclude that the mechanisms proposed by the theories are not necessary and/or sufficient for conscious experience. However, this leads to a broader concern: if theories can always accommodate failed predictions, does this mean that they can never be empirically overthrown? And if so, is there any realistic empirical path towards a unified theory of consciousness?

The answer lies in how we understand the purpose of theory testing. The fact that a prediction can be accounted for by adjusting auxiliary assumptions does not entail that a failed prediction is inconsequential, as the very act of refinement of auxiliary assumptions constitutes scientific progress. As discussed in Chapter 2, Imre Lakatos formalized this approach in his concept of research program [117,293]. Rather than viewing theories as static entities that are falsified by a single failed prediction (as the Popperian view entails), Lakatos described theories as research programs, composed of a "hard core" which are the fundamental claims of the theory and auxiliary assumptions.

In this framework, failed predictions are not without consequences. When a prediction fails, the theory's protective belt of auxiliary assumptions is adjusted, allowing the core to remain intact. The core and this adjusted belt constitute a novel version of the theory, in a long line of versions of the theories in the research program history. In this view, research programs are never falsified, they are either progressive or degenerative. For a research program to be considered progressive, it must predict novel facts that were not predicted by the previous versions of the theory and some of these novel predictions must be corroborated empirically [294]. If it fails to do so, it is considered degenerative.

As we and several others have argued, this framework is well-suited to appreciate the value of theory testing in consciousness research [83,104,140,290]. This iterative refinement of theories is evident in Chapters 1 and 2, particularly in how GNWT evolved in response to empirical challenges. In Chapter 2, we tested a version of GNWT,, whose hard core is that consciousness is the result of information broadcast in the workspace, and inferred that participants experience the persistence of stimuli they consciously perceive (along with many other assumptions). Based on the observation that PFC activation was not modulated by stimulus duration, GNWT revised this inference, predicting that participants do not experience the disappearance of the stimulus unless task-relevant. Chapter 3 tested this updated version of the theory,, finding only partial validation of this novel prediction, as conscious access of stimulus disappearance occurred in a subset of trials. Now GNWT is further refined to, predicting that task relevance and predictive processes interact to influence conscious experience, paving the way for future studies.

Accordingly, theories can remain progressive despite their predictions being falsified, by adapting and making novel predictions, spurring novel scientific discoveries. Accordingly, engaging in testing theories of consciousness can be seen as a way toward theories' self-improvement [290], refining our understanding of consciousness at the same time [294], making it a worthwhile endeavor.

On the usefulness and limitations of Adversarial Collaboration in consciousness research

While we cannot definitively establish whether the current version of a theory is wrong, we can establish whether it is better or worse than a competing theory. As Lakatos observed, often in the history of science, theories are abandoned because another, better theory becomes available [117,293]. It is with that goal in mind that we relied on the framework of Adversarial Collaboration in Chapter 2.

The framework of Adversarial Collaboration aims to settle theoretical disputes and has recently been discussed as constituting a gold standard for theory testing [99,110]. As can be seen in Chapter 2, this approach requires theorists from opposing camps to band together and design an experiment they agree a priori is appropriate to arbitrate between their contradicting claims [98,111]. This approach is particularly useful in the case of consciousness research. As we have seen, theories are tested in parallel with operationalization parameters reflecting the particular commitment of the theory being tested. As a result, the evidence fitting a given research program is discarded by another based on disagreement with the experimental parameters, stalling progress [78,140]. This is prevented by the very nature of ACs: an experimental design is agreed upon a priori based on which theories' predictions will be tested. Accordingly, the results should enable us to establish which (current version) of the theories being tested better explains conscious experience [78].

Involving the advocates of each theory in the theories comparison process is crucial in the field of consciousness research, where theories are often under-defined and theoretical work is typically focused on explaining existing effects rather than advancing novel predictions under untested conditions [25,65,87,94,108,121,287,295]. When theories are compelled to step beyond their comfort zones, novel predictions are formulated and these predictions must be faithful representations of the theories themselves, which is ensured by involving proponents of each theory. This framework does not only yield novel predictions, but it also forces theories to become more precise to be able to attribute evidence to one or the other theory. For instance, in Chapter 2, theories have to provide precise definitions of the regions of the brain they consider to play a critical role in conscious experience–details that were not explicitly stated before. I believe that my work attests to the value of this approach: the results have led to the refinement of the theories and the generation of novel predictions, engaging in a progressive research agenda that may not have emerged had the theories not been challenged under novel experimental conditions.

However, there is an important limitation of the study presented in Chapter 2. Despite the theories being tested on common experimental grounds and some of their predictions being falsified, we cannot decide which of IIT or GNWT (in their current formulations) provides a better account of empirical data. That is because, despite our best efforts to bring the theories on common grounds, the predictions they committed to were not formally adversarial.

This can be seen in the pre-registered predictions of the project [96]. The first prediction relates to the decoding of conscious content. GNWT predicts that it can be decoded from the PFC, while IIT predicts it can be decoded from the posterior ROI. These predictions are not contradictory as they concern different data; they are therefore not mutually exclusive. As Prof. Dehaene mentions in the discussion of Chapter 2, GNWT would also predict that the content being experienced should be decodable from sensory regions. Furthermore, IIT does not deny that decoding in the PFC could be observed. Instead, IIT predicts that there should not be additional information regarding the content being experienced in the PFC beyond the information found in the posterior hot zone. Importantly, this prediction from IIT was not explicitly contested nor endorsed by GNWT. In Chapter 2, we observed decoding in both prefrontal and posterior ROIs, and we observed that adding prefrontal features to the posterior classifier did not lead to an increase in decoding accuracy. In other words, both theories saw all their predictions validated, and it is accordingly not possible to determine which theory accounts the best for the observed data.

Similarly, GNWT did not predict that **no** sustained activation should be observed in the posterior ROI and IIT did not predict that **no** onset and offset ignition should be observed in the PFC. The lack of offset-ignition in the PFC does challenge GNWT but it has no implications for IIT. GNWT did not predict that the connectivity between FFA and V1/V2 should **not** be larger when a face is presented, and IIT did not predict that connectivity between FFA and PFC should **not** be larger when a face is presented. Accordingly, the failure of IIT prediction challenges only IIT and has no bearing on GNWT.

This limitation does not change the implications that failed predictions have for the theories, but does entail that we cannot answer the question of whether one theory is better than the other.

There are several possible reasons why the theories did not commit to directly opposing predictions. A first possibility is that while the theories in their current form might be capable of making contradicting statements, the experimental

conditions may have not been adequate to bring about such contradicting statements. I acknowledge that our experimental choices may have limited the potential for direct opposition between the theories, and future studies could explore alternative paradigms. Alternatively, this limitation may not lie in the experimental design but rather reflect the lack of specification of theories in their current formulations. As theories have been mostly tested with a restricted set of experimental conditions, they may be reluctant to make predictions in experimental conditions falling outside of their current purview. A last possibility is that GNWT and IIT may have non-overlapping explanatory targets, that is they attempt to explain different concepts loosely understood to be covered by the concept of consciousness [76,77], in which case they will never be able to formulate truly contradicting predictions.

I do not believe it is possible to determine a posterioir which of these reasons is the cause of the lack of contradicting predictions in Chapter 2. However, our results highlight the importance of considering such factors in future adversarial collaborations. If theories are reluctant to commit to contradicting predictions in a given experimental paradigm, it is worth exploring whether alternative paradigms in which the theories being tested would be willing to commit to contradicting predictions. Importantly, if this proves to be unfeasible, this may reflect their current lack of specificity or their misaligned explanatory targets. I would argue that in both cases, adversarial collaboration projects remain far from vain, as they encourage the refinement and formalization of theories in ways that isolated testing cannot. In turn, such projects may produce the necessary refinements, making theories capable of formulating contradicting predictions in future studies. Alternatively, these refinements might help clarify the misalignment in explanatory targets of theories so far thought to address related phenomena. Both outcomes would contribute to refining the theoretical landscape. Therefore, researchers should not be discouraged from engaging in adversarial collaborations, even when obtaining directly competing predictions is challenging, as these collaborations drive both theoretical and experimental progress.

Improving the efficacy of theory testing efforts

I believe that my research demonstrates the significant progress that can be made by testing predictions from theories of consciousness. By attempting to identify dissociations between the content of conscious experience and the mechanisms proposed by various theories, I have uncovered key gaps in their explanations and in our understanding of conscious experience. These efforts, in line with the Lakatosian

view of scientific progress, have not only led to the refinement of these existing theories but have also deepened our understanding of consciousness itself.

While the Lakatosian view entails that that theories cannot be directly falsified by a single experiment, attempts at testing theories of consciousness should still aim to be as decisive as possible. Experiments should be designed such that when a prediction fails, the options for revising assumptions are limited to meaningful ones. In my research, I have developed and applied strategies to constrain, which assumptions need to be revised when a prediction fails. My work has also granted me hindsight wisdom, revealing additional strategies that can further enhance the efficacy of theory testing. By sharing both the methods I implemented and the lessons I gained, I offer practical guidelines to make future attempts at testing theories of consciousness more effective and integrative, fostering iterative progress toward a unified scientific theory of conscious experience.

Constraining auxiliary assumptions

A first insight is that while some assumptions inevitably need to be made when testing predictions, certain assumptions can—and should—be verified. A critical assumption, often taken for granted, is that the experimental conditions under which data are collected are appropriate and consistent. This is by no means guaranteed. In the large-scale, multi-lab study presented in Chapter 2, we observed that differences in hardware and software configurations, as well as variability in testing practices, could lead to significant discrepancies in experimental setup performances. Addressing this variability in the Cogitate project required the development of testing protocols to ensure that experiments were functioning consistently across all sites.

Based on this observation, I conducted an additional study (not included in this thesis) to investigate the extent to which it is representative of the community. We surveyed the field and revealed that while most researchers conduct some form of setup testing, the specific aspects tested varied greatly across researchers [118]. Furthermore, I demonstrated through simulations that even slight deviation in setup performance could have drastic impact on statistical results. In this state of affairs, failed predictions in theories could easily be attributed to trivial technical issues. To address this issue, the testing protocol developed in the Cogitate project was extended to a standardized framework applicable to any event-based studies in cognitive neuroscience. This framework includes a shorthand report, enabling researchers to document the performance of their setups comprehensively.

By applying this protocol, researchers can rule out trivial technical issues as potential explanations for failed predictions, directly reducing the underdetermination problem. This is particularly important in multi-lab studies, where cross-site consistency is critical and facilitate replication efforts by enabling researchers to match experimental performance across different studies. By systematically addressing these technical challenges, variability in results across studies can be reduced; helping to ensure that theory testing is based on robust, high-quality data.

A second, related insight deals with the inference problem. As discussed earlier, consciousness research relies on proxies to infer the content of consciousness, which introduces a layer of uncertainty. When a theory faces a failed prediction, it can often dismiss the evidence by arguing that the inference about consciousness was flawed. I believe that this specific type of post hoc justification can be mitigated by providing a priori a logical derivation of the inference from the theory itself and declaring the level of confidence the theory places on the inference.

In the experiment presented in Chapter 2, GNWT did not explicitly justify the rationale behind the inference that participants experienced the persistence of stimuli on the screen (nor did IIT, but I will stick to GNWT to illustrate this recommendation). As a result, it became easy to discard this inference when no offset ignition was observed. Had GNWT provided a theoretical motivation for this inference (based on attention and the lack of competing stimuli for example) and declared a high confidence in it, rejecting the inference would have incurred a large cost to the theory. Furthermore, by exposing, the logical derivation behind the inference would have made explicit the background assumptions supporting the theory's prediction, leading to a more precise and constructive update of the theory by identifying which part of the rational is flawed. In other words, declaring a priori the rationale and degree of confidence in the inference can help limit the degrees of freedom a theory has in terms of ad-hoc accommodation, leading to more precise revision of the theories.

Adherence to the open science ethos

Another critical recommendation to strengthen theory testing is adhering to the open science ethos [104,110]. Pre-registration is critical to establish the progression of research programs. Under the Lakatosian view, there is nothing wrong per se with theories to provide ad-hoc explanations when their predictions fail. There is however a clear demarcation between predictions (made based on theoretical considerations alone) and accommodation, which is a post hoc explanation of why a prediction fails [290]. Across both chapters, theories predictions were pre-registered, enabling a clear tracking of the evolution of theories, highlighting how accommodation leads to novel

predictions, which can then be empirically tested. This, in turns, makes it possible to assess whether a theory remains progressive or eventually becomes degenerative. In addition to pre-registering the predictions, the rationale and degree of confidence placed in the inference I have proposed above should also be pre-registered. By pre-registering both the predictions alongside the rationale behind the inference, the degrees of freedom available for post-hoc justifications ensures more precise revisions of theories based on empirical evidence.

In addition to pre-registration, data sharing is another essential component of the open science framework. As I have highlighted, theory testing is an inherently iterative and dynamic process; the findings of a single study are insufficient to conclusively confirm or refute a theory. For a theory to remain progressive when its predictions are refuted, it must generate novel predictions that are empirically tested. Importantly, as highlighted by Negro [290], the novelty of a prediction is understood in terms of 'use-novelty'—the prediction of a fact is considered novel if it was not part of the set of empirical observations used to motivate the theory up until that point. Therefore, testing these novel predictions does not always require new experiments; they can often be validated using existing observations or by reanalyzing existing datasets.

Furthermore, the validation of a theory's prediction in a single experiment demonstrates its validity only in this particular instance. Importantly, if a theory of consciousness is correct, the mechanisms it proposes to instantiate conscious experience should be observed across all experimental conditions under which consciousness occurs. In this sense theory testing can be seen as a generalization problem [296] and testing theory's predictions across diverse datasets is necessary to establish the robustness and the scope of a theory. Such an endeavor requires the aggregation of empirical data to test the robustness of theories' predictions across the wealth of empirical data amassed in empirical efforts to investigate the neural underpinnings of conscious experience.

However, in the current state of the literature, most data aggregation has occurred at the level of reviews written primarily from the perspective of the proponents of the theories [25,65,87,94,108,121,287,295]. However, a comprehensive meta-analysis to establish the generalizability of theories' predictions would require the sharing of the data collected across the wealth of empirical studies conducted over the past three decades. However, sharing data is not sufficient as data that are not properly documented or structured are of limited use. Instead, data should be shared following the FAIR principle—they should be Findable, Accessible, Interoperable, and Reusable [297].

Recognizing this necessity, I have dedicated significant effort to ensure that the data collected across both chapters are openly accessible and thoroughly documented. Specifically, I have authored a paper (submitted to *Scientific Data*) describing the extensive iEEG data set collected in Chapter 2 and developed a comprehensive Python package to facilitate data access and analysis for replications and reuse of the data for alternative purposes [298]. The data are structured and documented following the Brain Imaging Data Structure (BIDS) principle [167]. We extended the metadata beyond the BIDS specification to include detailed, machine-readable information about the experimental setup, data collection procedures, clinical information about participants, and more.

These efforts not only make the data more usable for replication and reanalysis for other purposes, they also constitute a stepping-stone to building a centralized database that can support meta-analysis and holistic theory testing. Such analyses can help resolve inconsistencies in the literature, facilitates the generalization and scope delineation of theories of consciousness, moving the field toward a more unified and empirically grounded theoretical framework.

Bayesian evidence accumulation: a comprehensive framework to test theories of consciousness

A final important insight is that the reliance on frequentist statistics suffers from key limitations when it comes to navigating complex empirical situations that arose in the studies I have presented, and more broadly, in consciousness neuroscience [100]. While the Lakatosian view provides a useful philosophical framework, it remains vague as to what constitutes a challenge to a theory requiring an update of the protective belt and how to determine when a research program becomes degenerative [299]. Across both Chapters, each theory made several predictions and a binary outcome (pass or fail) was determined based on a criterion of significance in the classical frequentist statistic tradition. According to this criterion, some predictions were supported by the data, others were not. Moreover, these predictions were tested across different modalities (iEEG, fMRI, MEG, behavior, and eye-tracking), and in some cases, predictions were only validated in some modalities but not others.

In this state of affairs, it becomes difficult to determine how much a theory should be revised in light of empirical evidence. Not all predictions are equally relevant to the theories; some are more critical than others [140]. For instance, it could be argued that for IIT, the prediction of decoding in the posterior cortex may be less central than that of sustained representation for the duration of the stimulus presentation. Similarly, one might argue that the lack of offset-ignition poses a stronger challenge than the

absence of content representation following stimulus offset for GNWT. Furthermore, a given theory may a priori place more trust in one recording modality compared to another, depending on the spatio-temporal resolution of the method in question. These important nuances are difficult to address using frequentist statistics, as all predictions are tested on equal footing and in isolation. While we attempted to provide a comprehensive reading of our experimental results in Chapter 2, taking into consideration the relevance of each prediction to the theory, this approach lacks formalism.

In contrast, Bayesian inference offers a formal framework for updating beliefs based on novel empirical observations, making it well-suited for evaluating and refining theories of consciousness, as recently proposed by Corcoran and colleagues [100]. While Bayesian methods have traditionally been associated with confirmation rather than falsification [300], it transcends the traditional dichotomy of confirmation versus falsification by providing a probabilistic quantification of evidence for a particular model. Under this framework, a generative model is defined and theories predictions are operationalized by specifying prior distributions reflecting expectations about the observed effects [100]. For a single prediction, evidence can be estimated through variational approximations by fitting the model to the observed data, thereby assessing how well the prior distribution aligns with the observed data. For a single prediction, the log of evidence of each theory can compared to determine which one has the most empirical support. Critically, evidence for each model can also be summed across multiple predictions, recording modalities, and experiments to establish which theory receives more overall support from the data.

Building on these advantages, the Bayesian framework offers several critical improvements over frequentist statistics for theory testing in consciousness neuroscience. Bayesian inference overcomes the limitation of testing predictions in isolation by allowing for the accumulation of evidence across predictions and recording modalities. In adversarial collaboration situations where each theory makes different predictions on separate modalities, this enables to determine which theory has the most empirical support in a principled way. It is important to note that the evidence of a given theory is only meaningful when compared to another that of another model making predictions regarding the same data. Accordingly, it does not account for the aforementioned limitations of the study presented in Chapter 2 where the theories predictions addressed different aspects of the data.

However, the straightforward evidence accumulation scheme possible under the Bayesian framework readily enables meta-analyses to assess the generalizability of

theories of consciousness across many studies. When combined with the advocated data sharing practices, this framework allows for the accumulation of evidence for competing theories across many data sets, thereby establishing which theory accounts for the most openly accessible data. This provides a tractable and low-cost agenda towards more empirically grounded theoretical frameworks.

In addition, the flexibility of operationalizing theory predictions through prior distributions enables the integration of several recommendation outlined earlier. For instance, if experimental performance are documented to have lower performance, this can be reflected by selecting less constrained priors, indicating lower confidence in detecting effects. Similarly, the degree of confidence a theory places in the inference regarding the content of consciousness in a particular experiment can be incorporated in the analysis, to reflect the associated confidence a theory has that the predicted effect will be observed.

Importantly, under the Bayesian framework, setting a less constrained prior has meaningful consequences: it implies that the amount of evidence gains if its prediction is correct is lower compared to if it had set a more constrained, confident prior. This mechanism ensures that precise predictions are rewarded over vague ones, encouraging them to put forth riskier predictions by rewarding them when they do so [100,290].

Moreover, the Bayesian framework aligns well with the open science practices of pre-registration. By pre-registering analysis plans, models, and priors, researchers enhance the transparency of theory testing and ensure that the specified priors were not inadvertently contaminated by knowledge of the data. This approach is embraced by a novel adversarial collaboration aimed at comparing IIT, predictive processing theory and neuro-representationalism [301], and this project holds the promise of providing an unprecedented quantification of empirical support for the theories involved.

In conclusion, adopting Bayesian evidence accumulation schemes constitutes a promising direction for a more formal effective theory testing in consciousness research. This approach aligns with the Lakatosian view of scientific progress, providing a formal method to determine the progressiveness of a research project [83,290]. Furthermore, this approach integrates our earlier recommendations regarding experimental rigor, explicit theoretical inferences, pre-registration and holistic theories testing. By embracing Bayesian inference, we can encourage a more rigorous, integrated and productive scientific environment, ultimately propelling the field towards more unified and empirically grounded theoretical frameworks.

Conclusion

In this thesis, I proposed a novel approach to advancing consciousness research by identifying instances where conscious experience occurs in the absence of the mechanisms proposed by existing theories. This method shifts away from the traditional search for neural correlates of consciousness (NCCs) and the limitations of the contrastive method, allowing for a broader range of experimental conditions and alleviating the need to control for unconscious processing. By focusing on the temporal dynamics of conscious experience—specifically, how we experience the persistence of particular contents—I aimed to test the predictions of Integrated Information Theory (IIT) and the Global Neuronal Workspace Theory (GNWT) under new experimental paradigms.

In collaboration with proponents of both theories, I conducted experiments presenting highly visible visual stimuli for three distinct durations. The results showed sustained activation and content representation in the posterior regions, as predicted by IIT. In contrast, only transient responses were observed in the PFC, challenging GNWT's initial predictions. However, my subsequent study suggested that participants might not experience the full duration of stimuli; instead, they may only access representations transiently, just long enough to reach a decision.

These results have significant implications for both vision science and consciousness research. They refine our understanding of the neural mechanisms associated with sustained visual presentation and provide insights into the temporal dynamics of conscious experience in this context. Additionally, they offer a new avenue for investigating the dissociation between access consciousness and phenomenal consciousness, providing a way forward for this longstanding debate.

Importantly, this work demonstrates that scientific progress in consciousness research is best achieved through the iterative refinement of existing theories rather than their outright rejection upon encountering contradictory evidence. Adopting a Lakatosian view of scientific progress, I observed that even when core assumptions of theories remain robust against empirical falsification, the process of testing and refining these theories leads to deeper insights and more comprehensive explanations.

While the adversarial collaboration framework used in this research facilitated rigorous testing and promoted constructive dialogue between competing theories, it also revealed limitations—particularly the difficulty for theories of consciousness to

commit to opposing predictions. This made it challenging to definitively establish which theory better accounts for the empirical data. Future research should aim to design experiments that elicit directly competing predictions and consider employing Bayesian inference methods to navigate complex empirical situations more effectively.

In conclusion, this thesis underscores the importance of an iterative, integrative approach to testing theories of consciousness. By focusing on refining and improving existing theories and embracing collaborative efforts, we can enhance the explanatory power of these theories and move closer to a unified understanding of consciousness. Adhering to open science practices, such as pre-registration and data sharing, will further strengthen the rigor and transparency of future research in this intricate field.

Appendices

References

1. Wager, T. D. & Lindquist, M. A. *Principles of fMRI*. (Leanpub, 2015).
2. Goldstein, E. & Brockmole, J. *Sensation and Perception*. (Cengage Learning EMEA, Australia, Boston, MA, 2016).
3. Nagel, T. 11. What Is It Like to Be a Bat? in *11. What Is It Like to Be a Bat?* 159–168 (Harvard University Press, 2013). doi:10.4159/harvard.9780674594623.c15.
4. Kos, H. of & Adams, F. *On the Sacred Disease*. (Dalcassian Publishing Company, 2023).
5. LeDoux, J. E., Michel, M. & Lau, H. A little history goes a long way toward understanding why we study consciousness the way we do today. *Proc. Natl. Acad. Sci.* **117**, 6976–6984 (2020).
6. Chalmers, D. Facing up to the problem of consciousness. *J. Conscious. Stud.* **2**, 200–19 (1995).
7. Casali, A. G. *et al.* A Theoretically Based Index of Consciousness Independent of Sensory Processing and Behavior. *Sci. Transl. Med.* **5**, 198ra105-198ra105 (2013).
8. Metzinger, T. Introduction: Consciousness Research at the End of the Twentieth Century. in *Neural Correlates of Consciousness: Empirical and Conceptual Questions* (ed. Metzinger, T.) (MIT Press, 2000).
9. Michel, M. *et al.* Opportunities and challenges for a maturing science of consciousness. *Nat. Hum. Behav.* **3**, 104–107 (2019).
10. Owen, A. M. *et al.* Detecting Awareness in the Vegetative State. *Science* **313**, 1402–1402 (2006).
11. Xu, C. *et al.* Statistical learning in patients in the minimally conscious state. *Cereb. Cortex* **33**, 2507–2516 (2023).
12. Baars, B. J. *A Cognitive Theory of Consciousness*. (Cambridge University Press, Cambridge, 1989).
13. Crick, F. & Koch, C. Toward a Neurobiological Theory of Consciousness. *Semin. Neurosci.* **2**, 263–275 (1990).
14. Chalmers, D. J. What is a Neural Correlate of Consciousness? in *Neural Correlates of Consciousness: Empirical and Conceptual Questions* (ed. Metzinger, T.) 17–39 (MIT Press, 2000).
15. Yaron, I., Melloni, L., Pitts, M. & Mudrik, L. *The Consciousness Theories Studies (ConTraSt) Database: Analyzing and Comparing Empirical Studies of Consciousness Theories*. http://biorxiv.org/lookup/doi/10.1101/2021.06.10.447863 (2021) doi:10.1101/2021.06.10.447863.
16. Förster, J., Koivisto, M. & Revonsuo, A. ERP and MEG correlates of visual consciousness: The second decade. *Conscious. Cogn.* **80**, 102917 (2020).
17. Koivisto, M. & Revonsuo, A. Event-related brain potential correlates of visual awareness. *Neurosci. Biobehav. Rev.* **34**, 922–934 (2010).
18. Koch, C., Massimini, M., Boly, M. & Tononi, G. Neural correlates of consciousness: progress and problems. *Nat. Rev. Neurosci.* **17**, 307–321 (2016).
19. Rees, G., Kreiman, G. & Koch, C. Neural correlates of consciousness in humans. *Nat. Rev. Neurosci.* **3**, 261–270 (2002).
20. Yaron, I., Melloni, L., Pitts, M. & Mudrik, L. The ConTraSt database for analysing and comparing empirical studies of consciousness theories. *Nat. Hum. Behav.* **6**, 593–604 (2022).
21. Kim, C.-Y. & Blake, R. Psychophysical magic: rendering the visible 'invisible'. *Trends Cogn. Sci.* **9**, 381–388 (2005).
22. Doerig, A., Schurger, A. & Herzog, M. H. Hard criteria for empirical theories of consciousness. *Cogn. Neurosci.* **12**, 41–62 (2021).
23. Lamme, V. A. F. & Roelfsema, P. R. The distinct modes of vision offered by feedforward and recurrent processing. *Trends Neurosci.* **23**, 571–579 (2000).

24. Lepauvre, A. & Melloni, L. The search for the neural correlate of consciousness: Progress and challenges. *Philos. Mind Sci.* **2**, (2021).
25. Mashour, G. A., Roelfsema, P., Changeux, J.-P. & Dehaene, S. Conscious Processing and the Global Neuronal Workspace Hypothesis. *Neuron* **105**, 776–798 (2020).
26. Pitts, M. A., Lutsyshyna, L. A. & Hillyard, S. A. The relationship between attention and consciousness: an expanded taxonomy and implications for 'no-report' paradigms. *Philos. Trans. R. Soc. B Biol. Sci.* **373**, 20170348 (2018).
27. Koivisto, M. & Revonsuo, A. An ERP study of change detection, change blindness, and visual awareness. *Psychophysiology* **40**, 423–429 (2003).
28. Lamme, V. A. F., Zipser, K. & Spekreijse, H. Masking Interrupts Figure-Ground Signals in V1. *J. Cogn. Neurosci.* **14**, 1044–1053 (2002).
29. Leopold, D. A. & Logothetis, N. K. Activity changes in early visual cortex reflect monkeys' percepts during binocular rivalry. *Nature* **379**, 549–553 (1996).
30. Logothetis, N. K. & Schall, J. D. Neuronal Correlates of Subjective Visual Perception. *Science* **245**, 761–763 (1989).
31. Logothetis, N. K. Single units and conscious vision. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **353**, 1801–1818 (1998).
32. Pins, D. & Ffytche, D. The neural correlates of conscious vision. *Cereb. Cortex N. Y. N 1991* **13**, 461–474 (2003).
33. Roelfsema, P. R., Lamme, V. A. F., Spekreijse, H. & Bosch, H. Figure—Ground Segregation in a Recurrent Network Architecture. *J. Cogn. Neurosci.* **14**, 525–537 (2002).
34. Tong, F., Nakayama, K., Vaughan, J. T. & Kanwisher, N. Binocular Rivalry and Visual Awareness in Human Extrastriate Cortex. *Neuron* **21**, 753–759 (1998).
35. Cul, A. D., Baillet, S. & Dehaene, S. Brain Dynamics Underlying the Nonlinear Threshold for Access to Consciousness. *PLOS Biol.* **5**, e260 (2007).
36. Dehaene, S. *et al.* Cerebral mechanisms of word masking and unconscious repetition priming. *Nat. Neurosci.* **4**, 752–758 (2001).
37. Lumer, E. D., Friston, K. J. & Rees, G. Neural Correlates of Perceptual Rivalry in the Human Brain. *Science* **280**, 1930–1934 (1998).
38. Lumer, E. D. & Rees, G. Covariation of activity in visual and prefrontal cortex associated with subjective visual perception. *Proc. Natl. Acad. Sci.* **96**, 1669–1673 (1999).
39. Marois, R., Yi, D.-J. & Chun, M. M. The Neural Fate of Consciously Perceived and Missed Events in the Attentional Blink. *Neuron* **41**, 465–472 (2004).
40. Panagiotaropoulos, T. I., Deco, G., Kapoor, V. & Logothetis, N. K. Neuronal Discharges and Gamma Oscillations Explicitly Reflect Visual Consciousness in the Lateral Prefrontal Cortex. *Neuron* **74**, 924–935 (2012).
41. Sergent, C., Baillet, S. & Dehaene, S. Timing of the brain events underlying access to consciousness during the attentional blink. *Nat. Neurosci.* **8**, 1391–1400 (2005).
42. Koch, C. & Tsuchiya, N. Attention and consciousness: two distinct brain processes. *Trends Cogn. Sci.* **11**, 16–22 (2007).
43. Naccache, L., Blandin, E. & Dehaene, S. Unconscious Masked Priming Depends on Temporal Attention. *Psychol. Sci.* **13**, 416–424 (2002).

44. Koivisto, M. & Revonsuo, A. Electrophysiological correlates of visual consciousness and selective attention. *NeuroReport* **18**, 753 (2007).
45. Dehaene, S., Changeux, J.-P., Naccache, L., Sackur, J. & Sergent, C. Conscious, preconscious, and subliminal processing: a testable taxonomy. *Trends Cogn. Sci.* **10**, 204–211 (2006).
46. Soto, D. & Silvanto, J. Reappraising the relationship between working memory and conscious awareness. *Trends Cogn. Sci.* **18**, 520–525 (2014).
47. Tsuchiya, N., Wilke, M., Frässle, S. & Lamme, V. A. F. No-Report Paradigms: Extracting the True Neural Correlates of Consciousness. *Trends Cogn. Sci.* **19**, 757–770 (2015).
48. Aru, J., Bachmann, T., Singer, W. & Melloni, L. Distilling the neural correlates of consciousness. *Neurosci. Biobehav. Rev.* **36**, 737–746 (2012).
49. de Graaf, T. A., Hsieh, P.-J. & Sack, A. T. The 'correlates' in neural correlates of consciousness. *Neurosci. Biobehav. Rev.* **36**, 191–197 (2012).
50. Aru, J. *et al.* Local Category-Specific Gamma Band Responses in the Visual Cortex Do Not Reflect Conscious Perception. *J. Neurosci.* **32**, 14909–14914 (2012).
51. Melloni, L., Schwiedrzik, C. M., Müller, N., Rodriguez, E. & Singer, W. Expectations Change the Signatures and Timing of Electrophysiological Correlates of Perceptual Awareness. *J. Neurosci.* **31**, 1386–1396 (2011).
52. Cohen, M. A., Ortego, K., Kyroudis, A. & Pitts, M. Distinguishing the Neural Correlates of Perceptual Awareness and Postperceptual Processing. *J. Neurosci.* **40**, 4925–4935 (2020).
53. Farooqui, A. A. & Manly, T. When attended and conscious perception deactivates fronto-parietal regions. *Cortex* **107**, 166–179 (2018).
54. Kronemer, S. I. *et al.* Human visual consciousness involves large scale cortical and subcortical networks independent of task report and eye movement activity. *Nat. Commun.* **13**, 7342 (2022).
55. Pitts, M. A., Martínez, A. & Hillyard, S. A. Visual Processing of Contour Patterns under Conditions of Inattentional Blindness. *J. Cogn. Neurosci.* **24**, 287–303 (2012).
56. Pitts, M. A., Padwal, J., Fennelly, D., Martínez, A. & Hillyard, S. A. Gamma band activity and the P3 reflect post-perceptual processes, not visual awareness. *NeuroImage* **101**, 337–350 (2014).
57. Schelonka, K., Graulty, C., Canseco-Gonzalez, E. & Pitts, M. A. ERP signatures of conscious and unconscious word and letter perception in an inattentional blindness paradigm. *Conscious. Cogn.* **54**, 56–71 (2017).
58. Shafto, J. P. & Pitts, M. A. Neural Signatures of Conscious Face Perception in an Inattentional Blindness Paradigm. *J. Neurosci.* **35**, 10940–10948 (2015).
59. Dellert, T. *et al.* Dissociating the Neural Correlates of Consciousness and Task Relevance in Face Perception Using Simultaneous EEG-fMRI. *J. Neurosci.* **41**, 7864–7875 (2021).
60. Dwarakanath, A. *et al.* Bistability of prefrontal states gates access to consciousness. *Neuron* **111**, 1666-1683.e4 (2023).
61. Kapoor, V. *et al.* Decoding internally generated transitions of conscious contents in the prefrontal cortex without subjective reports. *Nat. Commun.* **13**, 1535 (2022).
62. Noy, N. *et al.* Ignition's glow: Ultra-fast spread of global cortical activity accompanying local "ignitions" in visual cortex during conscious visual perception. *Conscious. Cogn.* **35**, 206–224 (2015).
63. Sergent, C. *et al.* Bifurcation in brain dynamics reveals a signature of conscious processing independent of report. *Nat. Commun.* **12**, 1149 (2021).
64. Block, N. What Is Wrong with the No-Report Paradigm and How to Fix It. *Trends Cogn. Sci.* **23**, 1003–1013 (2019).

65. Boly, M. *et al.* Are the Neural Correlates of Consciousness in the Front or in the Back of the Cerebral Cortex? Clinical and Neuroimaging Evidence. *J. Neurosci.* **37**, 9603–9613 (2017).

66. Michel, M. & Morales, J. Minority reports: Consciousness and the prefrontal cortex. *Mind Lang.* **35**, 493–513 (2020).

67. Odegaard, B., Knight, R. T. & Lau, H. Should a Few Null Findings Falsify Prefrontal Theories of Conscious Perception? *J. Neurosci.* **37**, 9593–9602 (2017).

68. Mayer, A., Schwiedrzik, C. M., Wibral, M., Singer, W. & Melloni, L. Expecting to See a Letter: Alpha Oscillations as Carriers of Top-Down Sensory Predictions. *Cereb. Cortex* **26**, 3146–3160 (2016).

69. Meijs, E. L., Slagter, H. A., Lange, F. P. de & Gaal, S. van. Dynamic Interactions between Top–Down Expectations and Conscious Awareness. *J. Neurosci.* **38**, 2318–2327 (2018).

70. Pinto, Y., van Gaal, S., de Lange, F. P., Lamme, V. A. F. & Seth, A. K. Expectations accelerate entry of visual stimuli into awareness. *J. Vis.* **15**, 13 (2015).

71. Irvine, E. Measures of Consciousness. *Philos. Compass* **8**, 285–297 (2013).

72. Browning, H. & Veit, W. The Measurement Problem of Consciousness. *Philos. Top.* **48**, 85–108 (2020).

73. Irvine, E. Explaining What? *Topoi* **36**, 95–106 (2017).

74. Overgaard, M. The challenge of measuring consciousness. in *Behavioral Methods in Consciousness Research* (ed. Overgaard, M.) 7–20 (Oxford University Press, 2015). doi:10.1093/acprof:oso/9780199688890.003.0002.

75. Overgaard, M. & Fazekas, P. Can No-Report Paradigms Extract True Correlates of Consciousness? *Trends Cogn. Sci.* **20**, 241–242 (2016).

76. Vilas, M. G., Auksztulewicz, R. & Melloni, L. Active Inference as a Computational Framework for Consciousness. *Rev. Philos. Psychol.* **13**, 859–878 (2022).

77. Wiese, W. Toward a Mature Science of Consciousness. *Front. Psychol.* **9**, (2018).

78. Melloni, L., Mudrik, L., Pitts, M. & Koch, C. Making the hard problem of consciousness easier. *Science* **372**, 911–912 (2021).

79. Signorelli, C. M., Szczotka, J. & Prentner, R. Explanatory profiles of models of consciousness - towards a systematic classification. *Neurosci. Conscious.* **2021**, niab021 (2021).

80. Kuhn, R. L. A landscape of consciousness: Toward a taxonomy of explanations and implications. *Prog. Biophys. Mol. Biol.* **190**, 28–169 (2024).

81. Seth, A. K. & Bayne, T. Theories of consciousness. *Nat. Rev. Neurosci.* **23**, 439–452 (2022).

82. Sattin, D. *et al.* Theoretical Models of Consciousness: A Scoping Review. *Brain Sci.* **11**, 535 (2021).

83. Kleiner, J. & Hoel, E. Falsification and consciousness. *Neurosci. Conscious.* **2021**, niab001 (2021).

84. Nieuwenstein, M., Van der Burg, E., Theeuwes, J., Wyble, B. & Potter, M. Temporal constraints on conscious vision: On the ubiquitous nature of the attentional blink. *J. Vis.* **9**, 18 (2009).

85. Tsuchiya, N. & Koch, C. Continuous flash suppression reduces negative afterimages. *Nat. Neurosci.* **8**, 1096–1101 (2005).

86. Frässle, S., Sommer, J., Jansen, A., Naber, M. & Einhäuser, W. Binocular Rivalry: Frontal Activity Relates to Introspection and Action But Not to Perception. *J. Neurosci.* **34**, 1738–1747 (2014).

87. Albantakis, L. *et al.* Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms. *PLOS Comput. Biol.* **19**, e1011465 (2023).

88. Oizumi, M., Albantakis, L. & Tononi, G. From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0. *PLOS Comput. Biol.* **10**, e1003588 (2014).

89. Tononi, G. An information integration theory of consciousness. *BMC Neurosci.* **5**, 42 (2004).

90. Tononi, G. Consciousness as Integrated Information: a Provisional Manifesto. *Biol. Bull.* **215**, 216–242 (2008).

91. Tononi, G. The Integrated Information Theory of Consciousness: An Updated Account. *Arch. Ital. Biol.* **150**, 56–90 (2012).

92. Tononi, G., Boly, M., Massimini, M. & Koch, C. Integrated information theory: from consciousness to its physical substrate. *Nat. Rev. Neurosci.* **17**, 450–461 (2016).

93. Dehaene, S. Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition* **79**, 1–37 (2001).

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

95. Engeser, M., Lepauvre, A., Dehaene, S. & Melloni, L. Reconstructed time. *Open Sci. Framew.* (2023) doi:10.17605/OSF.IO/KRJH7.

96. Melloni, L. *et al.* Adversarial Collaboration to test GNW and IIT. (2019) doi:10.17605/OSF.IO/MBCFY.

97. Melloni, L. *et al.* An adversarial collaboration protocol for testing contrasting predictions of global neuronal workspace and integrated information theory. *PLOS ONE* **18**, e0268577 (2023).

98. Kahneman, D. Experiences of collaborative research. *Am. Psychol.* **58**, 723–730 (2003).

99. Clark, C. J. & Tetlock, P. E. Adversarial Collaboration: The Next Science Reform. in *Ideological and Political Bias in Psychology: Nature, Scope, and Solutions* (eds. Frisby, C. L., Redding, R. E., O'Donohue, W. T. & Lilienfeld, S. O.) 905–927 (Springer International Publishing, Cham, 2023). doi:10.1007/978-3-031-29148-7_32.

100. Corcoran, A. W., Hohwy, J. & Friston, K. J. Accelerating scientific progress through Bayesian adversarial collaboration. *Neuron* **111**, 3505–3516 (2023).

101. Dehaene, S. *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts*. (Penguin books, New York (N. Y.), 2014).

102. Dehaene, S., Kerszberg, M. & Changeux, J.-P. A neuronal model of a global workspace in effortful cognitive tasks. *Proc. Natl. Acad. Sci.* **95**, 14529–14534 (1998).

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

104. Melloni, L. On keeping our adversaries close, preventing collateral damage, and changing our minds. Comment on Clark et al. *J. Appl. Res. Mem. Cogn.* **11**, 45–49 (2022).

105. Lamme, V. Predictive Coding Is Unconscious, so that Consciousness Happens Now. in *Open MIND* (eds. Metzinger, T. & Windt, J. M.) (Open MIND. Frankfurt am Main: MIND Group, 2015). doi:10.15502/9783958571105.

106. Malach, R. Local neuronal relational structures underlying the contents of human conscious experience. *Neurosci. Conscious.* **2021**, niab028 (2021).

107. Brown, R. The HOROR theory of phenomenal consciousness. *Philos. Stud.* **172**, 1783–1794 (2015).

108. Brown, R., Lau, H. & LeDoux, J. E. Understanding the Higher-Order Approach to Consciousness. *Trends Cogn. Sci.* **23**, 754–768 (2019).

109. Kahneman, D. *Adversarial Collaboration: An EDGE Lecture*. (2022).

110. Clark, C. J., Costello, T., Mitchell, G. & Tetlock, P. E. Keep your enemies close: Adversarial collaborations will improve behavioral science. *J. Appl. Res. Mem. Cogn.* **11**, 1–18 (2022).

111. Mellers, B., Hertwig, R. & Kahneman, D. Do Frequency Representations Eliminate Conjunction Effects? An Exercise in Adversarial Collaboration. *Psychol. Sci.* **12**, 269–275 (2001).

112. Haun, A. & Tononi, G. Why Does Space Feel the Way it Does? Towards a Principled Account of Spatial Experience. *Entropy* **21**, 1160 (2019).

113. Gerber, E. M., Golan, T., Knight, R. T. & Deouell, L. Y. Cortical representation of persistent visual stimuli. *NeuroImage* **161**, 67–79 (2017).

114. Podvalny, E. *et al.* Invariant Temporal Dynamics Underlie Perceptual Stability in Human Visual Cortex. *Curr. Biol.* **27**, 155–165 (2017).

115. Stigliani, A., Jeska, B. & Grill-Spector, K. Encoding model of temporal processing in human visual cortex. *Proc. Natl. Acad. Sci.* **114**, (2017).

116. Stokes, M. G. 'Activity-silent' working memory in prefrontal cortex: a dynamic coding framework. *Trends Cogn. Sci.* **19**, 394–405 (2015).

117. Lakatos, I. Falsification and the Methodology of Scientific Research Programmes. in *Philosophy, Science, and History* (Routledge, 1976).

118. Lepauvre, A., Hirschhorn, R., Bendtz, K., Mudrik, L. & Melloni, L. A standardized framework to test event-based experiments. *Behav. Res. Methods* (2024) doi:10.3758/s13428-024-02508-y.

119. Tversky, A. & Kahneman, D. Belief in the law of small numbers. *Psychol. Bull.* **76**, 105–110 (1971).

120. Tversky, A. & Kahneman, D. Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychol. Rev.* **90**, 293–315 (1983).

121. Koch, C., Massimini, M., Boly, M. & Tononi, G. Posterior and anterior cortex — where is the difference that makes the difference? *Nat. Rev. Neurosci.* **17**, 666–666 (2016).

122. Nir, Y. *et al.* Coupling between Neuronal Firing Rate, Gamma LFP, and BOLD fMRI Is Related to Interneuronal Correlations. *Curr. Biol.* **17**, 1275–1285 (2007).

123. Ray, S., Crone, N. E., Niebur, E., Franaszczuk, P. J. & Hsiao, S. S. Neural Correlates of High-Gamma Oscillations (60–200 Hz) in Macaque Local Field Potentials and Their Potential Implications in Electrocorticography. *J. Neurosci.* **28**, 11526–11536 (2008).

124. Haegens, S., Nácher, V., Luna, R., Romo, R. & Jensen, O. α-Oscillations in the monkey sensorimotor network influence discrimination performance by rhythmical inhibition of neuronal spiking. *Proc. Natl. Acad. Sci.* **108**, 19377–19382 (2011).

125. Iemi, L. *et al.* Ongoing neural oscillations influence behavior and sensory representations by suppressing neuronal excitability. *NeuroImage* **247**, 118746 (2022).

126. Broday-Dvir, R., Norman, Y., Harel, M., Mehta, A. D. & Malach, R. Perceptual stability reflected in neuronal pattern similarities in human visual cortex. *Cell Rep.* **42**, (2023).

127. Vishne, G., Gerber, E. M., Knight, R. T. & Deouell, L. Y. Distinct ventral stream and prefrontal cortex representational dynamics during sustained conscious visual perception. *Cell Rep.* **42**, (2023).

128. Jackendoff, R. *Consciousness and the Computational Mind*. xvi, 356 (The MIT Press, Cambridge, MA, US, 1987).

129. Gaillard, R. *et al.* Converging Intracranial Markers of Conscious Access. *PLoS Biol.* **7**, e1000061 (2009).

130. Vinck, M., Van Wingerden, M., Womelsdorf, T., Fries, P. & Pennartz, C. M. A. The pairwise phase consistency: A bias-free measure of rhythmic neuronal synchronization. *NeuroImage* **51**, 112–122 (2010).

131. Cardin, J. A. *et al.* Driving fast-spiking cells induces gamma rhythm and controls sensory responses. *Nature* **459**, 663–667 (2009).

132. Cohen, M. X. A tutorial on generalized eigendecomposition for denoising, contrast enhancement, and dimension reduction in multichannel electrophysiology. *NeuroImage* **247**, 118809 (2022).

133. Northoff, G. & Lamme, V. Neural signs and mechanisms of consciousness: Is there a potential convergence of theories of consciousness in sight? *Neurosci. Biobehav. Rev.* **118**, 568–587 (2020).

134. *The Cognitive Neurosciences*. (MIT Press, Cambridge, Mass, 2009).

135. Melloni, L. & Singer, W. The explanatory gap in neuroscience. *Pontif. Acad. Sci.* **21**, 61–73 (2011).

136. Popper, K. *The Logic of Scientific Discovery*. (Routledge, 1935). doi:10.4324/9780203994627.

137. Mudrik, L., Mylopoulos, M., Negro, N. & Schurger, A. Theories of consciousness and a life worth living. *Curr. Opin. Behav. Sci.* **53**, 101299 (2023).

138. Birch, J. The search for invertebrate consciousness. *Noûs* **56**, 133–153 (2022).

139. Aru, J., Larkum, M. E. & Shine, J. M. The feasibility of artificial consciousness through the lens of neuroscience. *Trends Neurosci.* **46**, 1008–1017 (2023).

140. Chis-Ciure, R., Melloni, L. & Northoff, G. A measure centrality index for systematic empirical comparison of consciousness theories. *Neurosci. Biobehav. Rev.* **161**, 105670 (2024).

141. Pigorini, A. *et al.* Bistability breaks-off deterministic responses to intracortical stimulation during non-REM sleep. *NeuroImage* **112**, 105–113 (2015).

142. Sarasso, S. *et al.* Consciousness and Complexity during Unresponsiveness Induced by Propofol, Xenon, and Ketamine. *Curr. Biol.* **25**, 3099–3105 (2015).

143. Ferrarelli, F. *et al.* Breakdown in cortical effective connectivity during midazolam-induced loss of consciousness. *Proc. Natl. Acad. Sci.* **107**, 2681–2686 (2010).

144. Massimini, M. *et al.* Breakdown of Cortical Effective Connectivity During Sleep. *Science* **309**, 2228–2232 (2005).

145. Watakabe, A. *et al.* Local and long-distance organization of prefrontal cortex circuits in the marmoset brain. *Neuron* **111**, 2258-2273.e10 (2023).

146. Blum, L. & Blum, M. A theory of consciousness from a theoretical computer science perspective: Insights from the Conscious Turing Machine. *Proc. Natl. Acad. Sci.* **119**, e2115934119 (2022).

147. Bellet, M. E. *et al.* Prefrontal neural ensembles encode an internal model of visual sequences and their violations. 2021.10.04.463064 Preprint at https://doi.org/10.1101/2021.10.04.463064 (2021).

148. Rainer, G., Asaad, W. F. & Miller, E. K. Selective representation of relevant information by neurons in the primate prefrontal cortex. *Nature* **393**, 577–579 (1998).

149. Liu, S., Yu, Q., Tse, P. U. & Cavanagh, P. Neural Correlates of the Conscious Perception of Visual Location Lie Outside Visual Cortex. *Curr. Biol.* **29**, 4036-4044.e4 (2019).

150. Hatamimajoumerd, E., Murty, N. A. R., Pitts, M. & Cohen, M. A. Decoding perceptual awareness across the brain with a no-report fMRI masking paradigm. *Curr. Biol.* **32**, 4139-4149.e4 (2022).

151. Marti, S., King, J.-R. & Dehaene, S. Time-Resolved Decoding of Two Processing Chains during Dual-Task Interference. *Neuron* **88**, 1297–1307 (2015).

152. van Vugt, B. *et al.* The threshold for conscious report: Signal loss and response bias in visual and frontal cortex. *Science* **360**, 537–542 (2018).

153. Marti, S., Sackur, J., Sigman, M. & Dehaene, S. Mapping introspection's blind spot: Reconstruction of dual-task phenomenology using quantified introspection. *Cognition* **115**, 303–313 (2010).

154. Pack, C. C., Berezovskii, V. K. & Born, R. T. Dynamic properties of neurons in cortical area MT in alert and anaesthetized macaque monkeys. *Nature* **414**, 905–908 (2001).

155. Desimone, R., Albright, T. D., Gross, C. G. & Bruce, C. Stimulus-selective properties of inferior temporal neurons in the macaque. *J. Neurosci.* **4**, 2051–2062 (1984).

156. Moran, J. & Desimone, R. Selective Attention Gates Visual Processing in the Extrastriate Cortex. *Science* **229**, 782–784 (1985).

157. Mack, A. & Rock, I. *Inattentional Blindness*. xiv, 273 (The MIT Press, Cambridge, MA, US, 1998).

158. Simons, D. J. & Chabris, C. F. Gorillas in our midst: sustained inattentional blindness for dynamic events. *Perception* **28**, 1059–1074 (1999).

159. Sergent, C. *et al.* Cueing attention after the stimulus is gone can retrospectively trigger conscious perception. *Curr. Biol. CB* **23**, 150–155 (2013).

160. Sigman, M. & Dehaene, S. Brain Mechanisms of Serial and Parallel Processing during Dual-Task Performance. *J. Neurosci.* **28**, 7585–7598 (2008).

161. Thibault, L., van den Berg, R., Cavanagh, P. & Sergent, C. Retrospective attention gates discrete conscious access to past sensory stimuli. *PLoS ONE* **11**, (2016).

162. Xie, Y. *et al.* Geometry of sequence working memory in macaque prefrontal cortex. *Science* **375**, 632–639 (2022).

163. Kay, K., Bonnen, K., Denison, R. N., Arcaro, M. J. & Barack, D. L. Tasks and their role in visual neuroscience. *Neuron* **111**, 1697–1713 (2023).

164. Tarr, M. J. The Object Databank. *Carnegie Mellon Univ.* (1996).

165. Glover, G. H. Deconvolution of Impulse Response in Event-Related BOLD fMRI1. *NeuroImage* **9**, 416–429 (1999).

166. Pelli, D. G. The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spat. Vis.* **10**, 437–442 (1997).

167. Holdgraf, C. *et al.* iEEG-BIDS, extending the Brain Imaging Data Structure specification to human intracranial electrophysiology. *Sci. Data* **6**, 102 (2019).

168. Gramfort, A. *et al.* MNE software for processing MEG and EEG data. *NeuroImage* **86**, 446–460 (2014).

169. Li, G. *et al.* Optimal referencing for stereo-electroencephalographic (SEEG) recordings. *NeuroImage* **183**, 327–335 (2018).

170. Grossman, S. *et al.* Convergent evolution of face spaces across human face-selective neuronal groups and deep convolutional networks. *Nat. Commun.* **10**, 4934 (2019).

171. Manning, J. R., Jacobs, J., Fried, I. & Kahana, M. J. Broadband Shifts in Local Field Potential Power Spectra Are Correlated with Single-Neuron Spiking in Humans. *J. Neurosci.* **29**, 13613–13620 (2009).

172. Dale, A. M., Fischl, B. & Sereno, M. I. Cortical Surface-Based Analysis. *NeuroImage* **9**, 179–194 (1999).

173. Destrieux, C., Fischl, B., Dale, A. & Halgren, E. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *NeuroImage* **53**, 1–15 (2010).

174. Wang, L., Mruczek, R. E. B., Arcaro, M. J. & Kastner, S. Probabilistic Maps of Visual Topography in Human Cortex. *Cereb. Cortex* **25**, 3911–3931 (2015).

175. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).

176. Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D. & Iverson, G. Bayesian t tests for accepting and rejecting the null hypothesis. *Psychon. Bull. Rev.* **16**, 225–237 (2009).

177. Kadipasaoglu, C. M., Conner, C. R., Whaley, M. L., Baboyan, V. G. & Tandon, N. Category-Selectivity in Human Visual Cortex Follows Cortical Topology: A Grouped icEEG Study. *PLOS ONE* **11**, e0157109 (2016).

178. Niso, G. *et al.* MEG-BIDS, the brain imaging data structure extended to magnetoencephalography. *Sci. Data* **5**, 180110 (2018).

179. Appelhoff, S. *et al.* MNE-BIDS: Organizing electrophysiological data into the BIDS format and facilitating their analysis. *J. Open Source Softw.* **4**, 1896 (2019).

180. Ferrante, O. *et al.* FLUX: A pipeline for MEG analysis. *NeuroImage* **253**, 119047 (2022).

181. Taulu, S., Kajola, M. & Simola, J. Suppression of Interference and Artifacts by the Signal Space Separation Method. *Brain Topogr.* **16**, 269–275 (2003).

182. Hyvärinen, A. & Oja, E. Independent component analysis: algorithms and applications. *Neural Netw.* **13**, 411–430 (2000).

183. Dale, A. M. *et al.* Dynamic Statistical Parametric Mapping. *Neuron* **26**, 55–67 (2000).

184. Hämäläinen, M. S. & Ilmoniemi, R. J. Interpreting magnetic fields of the brain: minimum norm estimates. *Med. Biol. Eng. Comput.* **32**, 35–42 (1994).

185. Wang, J.-Z., Williamson, S. J. & Kaufman, L. Magnetic source images determined by a lead-field analysis: the unique minimum-norm least-squares estimation. *IEEE Trans. Biomed. Eng.* **39**, 665–675 (1992).

186. Engemann, D. A. & Gramfort, A. Automated model selection in covariance estimation and spatial whitening of MEG and EEG signals. *NeuroImage* **108**, 328–342 (2015).

187. Zwiers, M. P., Moia, S. & Oostenveld, R. BIDScoin: A User-Friendly Application to Convert Source Data to Brain Imaging Data Structure. *Front. Neuroinformatics* **15**, 770608 (2022).

188. Li, X., Morgan, P. S., Ashburner, J., Smith, J. & Rorden, C. The first step for neuroimaging data analysis: DICOM to NIfTI conversion. *J. Neurosci. Methods* **264**, 47–56 (2016).

189. Esteban, O. *et al.* MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. *PLOS ONE* **12**, e0184661 (2017).

190. Esteban, O. *et al.* fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nat. Methods* **16**, 111–116 (2019).

191. Gorgolewski, K. *et al.* Nipype: A Flexible, Lightweight and Extensible Neuroimaging Data Processing Framework in Python. *Front. Neuroinformatics* **5**, (2011).

192. Smith, S. M. *et al.* Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage* **23**, S208–S219 (2004).

193. Penny, W., Friston, K., Ashburner, J., Kiebel, S. & Nichols, T. *Statistical Parametric Mapping: The Analysis of Funtional Brain Images*. (Elsevier/Academic Press, Amsterdam ; Boston, 2007).

194. Hautus, M. J. Corrections for extreme proportions and their biasing effects on estimated values ofd'. *Behav. Res. Methods Instrum. Comput.* **27**, 46–51 (1995).

195. Satterthwaite, T. D. *et al.* An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. *NeuroImage* **64**, 240–256 (2013).

196. Wagenmakers, E.-J. Approximate Objective Bayes Factors From P-Values and Sample Size: The 3p√n Rule. Preprint at https://doi.org/10.31234/osf.io/egydq (2022).

197. Hershman, R., Henik, A. & Cohen, N. A novel blink detection method based on pupillometry noise. *Behav. Res. Methods* **50**, 107–114 (2018).

198. Yuval-Greenberg, S., Merriam, E. P. & Heeger, D. J. Spontaneous Microsaccades Reflect Shifts in Covert Attention. *J. Neurosci.* **34**, 13693–13700 (2014).

199. Engbert, R. & Kliegl, R. Microsaccades uncover the orientation of covert attention. *Vision Res.* **43**, 1035–1045 (2003).

200. Ferri, F. J., Pudil, P., Hatef, M. & Kittler, J. Comparative study of techniques for large-scale feature selection* *This work was suported by a SERC grant GR/E 97549. The first author was also supported by a FPI grant from the Spanish MEC, PF92 73546684. in *Machine Intelligence and Pattern Recognition* vol. 16 403–413 (Elsevier, 1994).

201. Maris, E. & Oostenveld, R. Nonparametric statistical testing of EEG- and MEG-data. *J. Neurosci. Methods* **164**, 177–190 (2007).

202. Nichols, T. E. & Holmes, A. P. Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Hum. Brain Mapp.* **15**, 1–25 (2002).

203. Cichy, R. M. & Pantazis, D. Multivariate pattern analysis of MEG and EEG: A comparison of representational structure in time and space. *NeuroImage* **158**, 441–454 (2017).

204. Mumford, J. A., Turner, B. O., Ashby, F. G. & Poldrack, R. A. Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *NeuroImage* **59**, 2636–2643 (2012).

205. Zadrozny, B. & Elkan, C. Obtaining Calibrated Probability Estimates from Decision Trees and Naive Bayesian Classifiers. *Int. Conf. Mach. Learn.* **1**, 609–616 (2001).

206. Alpaydin, E. Combined 5 × 2 cv F Test for Comparing Supervised Classification Learning Algorithms. *Neural Comput.* **11**, 1885–1892 (1999).

207. Nadeau, C. & Bengio, Y. Inference for the Generalization Error. in *Advances in Neural Information Processing Systems 12, [NIPS Conference, Denver, Colorado, USA, November 29 - December 4, 1999]* (eds. Solla, S. A., Leen, T. K. & Müller, K.-R.) 307–313 (The MIT Press, 2000).

208. Benavoli, A., Corani, G., Demšar, J. & Zaffalon, M. Time for a Change: a Tutorial for Comparing Multiple Classifiers Through Bayesian Analysis. *J. Mach. Learn. Res.* **18**, 1–36 (2017).

209. Stelzer, J., Chen, Y. & Turner, R. Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): Random permutations and cluster size control. *NeuroImage* **65**, 69–82 (2013).

210. Schönemann, P. H. A generalized solution of the orthogonal procrustes problem. *Psychometrika* **31**, 1–10 (1966).

211. Andreella, A., De Santis, R., Vesely, A. & Finos, L. Procrustes-based distances for exploring between-matrices similarity. (2023) doi:10.48550/ARXIV.2301.06164.

212. Desikan, R. S. *et al.* An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage* **31**, 968–980 (2006).

213. Ince, R. A. A. *et al.* A statistical framework for neuroimaging data analysis based on mutual information estimated via a gaussian copula: Gaussian Copula Mutual Information. *Hum. Brain Mapp.* **38**, 1541–1573 (2017).

214. Combrisson, E., Basanisi, R., Cordeiro, V. L., Ince, R. A. A. & Brovelli, A. Frites: A Python package for functional connectivityanalysis and group-level statistics of neurophysiological data. *J. Open Source Softw.* **7**, 3842 (2022).

215. McLaren, D. G., Ries, M. L., Xu, G. & Johnson, S. C. A generalized form of context-dependent psychophysiological interactions (gPPI): A comparison to standard approaches. *NeuroImage* **61**, 1277–1286 (2012).

216. Li, R. *et al.* The pulse: transient fMRI signal increases in subcortical arousal systems during transitions in attention. *NeuroImage* **232**, 117873 (2021).

217. Golan, T. *et al.* Human intracranial recordings link suppressed transients rather than 'filling-in' to perceptual continuity across blinks. *eLife* **5**, e17243 (2016).

218. Mack, A. Inattentional Blindness: Looking Without Seeing. *Curr. Dir. Psychol. Sci.* **12**, 180–184 (2003).

219. Rees, G., Russell, C., Frith, C. D. & Driver, J. Inattentional Blindness Versus Inattentional Amnesia for Fixated But Ignored Words. *Science* **286**, 2504–2507 (1999).

220. Sergent, C. & Dehaene, S. Is Consciousness a Gradual Phenomenon?: Evidence for an All-or-None Bifurcation During the Attentional Blink. *Psychol. Sci.* **15**, 720–728 (2004).

221. Shapiro, K. L., Raymond, J. E. & Arnell, K. M. The attentional blink. *Trends Cogn. Sci.* **1**, 291–296 (1997).

222. Corallo, G., Sackur, J., Dehaene, S. & Sigman, M. Limits on Introspection: Distorted Subjective Time During the Dual-Task Bottleneck. *Psychol. Sci.* **19**, 1110–1117 (2008).

223. Pashler, H. Dual-task interference in simple tasks: Data and theory. *Psychol. Bull.* **116**, 220–244 (1994).

224. Cogitate Consortium *et al. An Adversarial Collaboration to Critically Evaluate Theories of Consciousness*. http://biorxiv.org/lookup/doi/10.1101/2023.06.23.546249 (2023) doi:10.1101/2023.06.23.546249.

225. Dennett, D. C. & Kinsbourne, M. Time and the observer: The where and when of consciousness in the brain. *Behav. Brain Sci.* **15**, 183–201 (1992).

226. Wong, K. F. E. The relationship between attentional blink and psychological refractory period. *J. Exp. Psychol. Hum. Percept. Perform.* **28**, 54–71 (2002).

227. Brisson, B., Robitaille, N. & Jolicœur, P. Stimulus intensity affects the latency but not the amplitude of the N2pc. *NeuroReport* **18**, 1627 (2007).

228. Dell'Acqua, R., Jolicoeur, P., Vespignani, F. & Toffanin, P. Central processing overlap modulates P3 latency. *Exp. Brain Res.* **165**, 54–68 (2005).

229. Hesselmann, G., Flandin, G. & Dehaene, S. Probing the cortical network underlying the psychological refractory period: A combined EEG–fMRI study. *NeuroImage* **56**, 1608–1621 (2011).

230. Luck, S. J. Sources of Dual-Task Interference: Evidence From Human Electrophysiology. *Psychol. Sci.* **9**, 223–227 (1998).

231. Dux, P. E., Ivanoff, J., Asplund, C. L. & Marois, R. Isolation of a Central Bottleneck of Information Processing with Time-Resolved fMRI. *Neuron* **52**, 1109–1120 (2006).

232. Consortium, C. *et al. An Adversarial Collaboration to Critically Evaluate Theories of Consciousness*. http://biorxiv.org/lookup/doi/10.1101/2023.06.23.546249 (2023).

233. Zylberberg, A., Oliva, M. & Sigman, M. Pupil Dilation: A Fingerprint of Temporal Selection During the "Attentional Blink". *Front. Psychol.* **3**, (2012).

234. Nobre, A., Correa, A. & Coull, J. The hazards of time. *Curr. Opin. Neurobiol.* **17**, 465–470 (2007).

235. Kahneman, D. & Beatty, J. Pupil Diameter and Load on Memory. *Science* **154**, 1583–1585 (1966).

236. de Jong, R. Multiple bottlenecks in overlapping task performance. *J. Exp. Psychol. Hum. Percept. Perform.* **19**, 965–980 (1993).

237. Smith, M. C. The psychological refractory period as a function of performance of a first response. *Q. J. Exp. Psychol.* **19**, 350–352 (1967).

238. Van Selst, M., Johnston, J. C. & Shafto, M. Dual-task interference when a response is not required. in *Cognitive Science Society 19th Annual Meeting* (2002).

239. Karlin, L. & Kestenbaum, R. Effects of Number of Alternatives on the Psychological Refractory Period. *Q. J. Exp. Psychol.* **20**, 167–178 (1968).

240. Pashler, H. & Johnston, J. C. Chronometric evidence for central postponement in temporally overlapping tasks. *Q. J. Exp. Psychol. Sect. A* **41**, 19–45 (1989).

241. Ruthruff, E., Pashler, H. E. & Klaassen, A. Processing bottlenecks in dual-task performance: Structural limitation or strategic postponement? *Psychon. Bull. Rev.* **8**, 73–80 (2001).

242. Sigman, M. & Dehaene, S. Parsing a Cognitive Task: A Characterization of the Mind's Bottleneck. *PLoS Biol.* **3**, e37 (2005).

243. Niemi, P. & Näätänen, R. Foreperiod and simple reaction time. *Psychol. Bull.* **89**, 133–162 (1981).

244. Den Ouden, H. E. M., Kok, P. & De Lange, F. P. How Prediction Errors Shape Perception, Attention, and Motivation. *Front. Psychol.* **3**, (2012).

245. Hohwy, J. Attention and Conscious Perception in the Hypothesis Testing Brain. *Front. Psychol.* **3**, (2012).

246. Kok, P., Jehee, J. F. M. & de Lange, F. P. Less is more: expectation sharpens representations in the primary visual cortex. *Neuron* **75**, 265–270 (2012).

247. Sigman, M. & Dehaene, S. Dynamics of the Central Bottleneck: Dual-Task and Task Uncertainty. *PLOS Biol.* **4**, e220 (2006).

248. Bratzke, D. & Janczyk, M. Introspection about backward crosstalk in dual-task performance. *Psychol. Res.* **85**, 605–617 (2021).

249. Bryce, D. & Bratzke, D. Introspective reports of reaction times in dual-tasks reflect experienced difficulty rather than timing of cognitive processes. *Conscious. Cogn.* **27**, 254–267 (2014).

250. Bryce, D. & Bratzke, D. The surprising role of stimulus modality in the dual-task introspective blind spot: a memory account. *Psychol. Res.* **86**, 1332–1354 (2022).

251. Bryce, D. & Bratzke, D. Are introspective reaction times affected by the method of time estimation? A comparison of visual analogue scales and reproduction. *Atten. Percept. Psychophys.* **77**, 978–984 (2015).

252. Block, N. On a confusion about a function of consciousness. *Behav. Brain Sci.* **18**, 227–247 (1995).

253. Cohen, M. A. & Dennett, D. C. Consciousness cannot be separated from function. *Trends Cogn. Sci.* **15**, 358–364 (2011).

254. Canales-Johnson, A. *et al.* Feedback information transfer in the human brain reflects bistable perception in the absence of report. *PLOS Biol.* **21**, e3002120 (2023).

255. Kleiner, M., Brainard, D. & Pelli, D. What's new in Psychtoolbox-3? (2007).

256. The MathWorks Inc. MATLAB. The MathWorks Inc. (2019).

257. R Core Team. R. R Foundation for Statistical Computing (2020).

258. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using **lme4**. *J. Stat. Softw.* **67**, (2015).

259. Lo, S. & Andrews, S. To transform or not to transform: using generalized linear mixed models to analyse reaction time data. *Front. Psychol.* **6**, (2015).

260. Van Rossum, G. & Drake Jr, F. L. Python 3 reference manual. CreateSpace (2009).

261. Gramfort, A. *et al.* MEG and EEG data analysis with MNE-Python. *Front. Neurosci.* **7**, (2013).

262. Mathôt, S. & Vilotijević, A. Methods in cognitive pupillometry: Design, preprocessing, and statistical analysis. *Behav. Res. Methods* **55**, 3055–3077 (2023).

263. Denison, R. N., Parker, J. A. & Carrasco, M. Modeling pupil responses to rapid sequential events. *Behav. Res. Methods* **52**, 1991–2007 (2020).

264. Mercier, M. R. *et al.* Evaluation of cortical local field potential diffusion in stereotactic electroencephalography recordings: A glimpse on white matter signal. *NeuroImage* **147**, 219–232 (2017).

265. Grill-Spector, K., Kushnir, T., Hendler, T. & Malach, R. The dynamics of object-selective activation correlate with recognition performance in humans. *Nat. Neurosci.* **3**, 837–843 (2000).

266. Grill-Spector, K. & Weiner, K. S. The functional architecture of the ventral temporal cortex and its role in categorization. *Nat. Rev. Neurosci.* **15**, 536–548 (2014).

267. Kanwisher, N. Functional specificity in the human brain: A window into the functional architecture of the mind. *Proc. Natl. Acad. Sci.* **107**, 11163–11170 (2010).

268. Reddy, L. & Kanwisher, N. Coding of visual objects in the ventral stream. *Curr. Opin. Neurobiol.* **16**, 408–414 (2006).

269. Fleming, S. IIT vs. GNWT and the meaning of evidence in consciousness science. *The Elusive Self* https://elusiveself.wordpress.com/2023/09/09/iit-vs-gnwt-and-the-meaning-of-evidence-in-consciousness-science/ (2023).

270. Grill-Spector, K. & Kanwisher, N. Visual Recognition: As Soon as You Know It Is There, You Know What It Is. *Psychol. Sci.* **16**, 152–160 (2005).

271. Brands, A. M. *et al.* Temporal dynamics of short-term neural adaptation across human visual cortex. *PLOS Comput. Biol.* **20**, e1012161 (2024).

272. Groen, I. I. A. *et al.* Temporal Dynamics of Neural Responses in Human Visual Cortex. *J. Neurosci.* **42**, 7562–7580 (2022).

273. Müller, J. R., Metha, A. B., Krauskopf, J. & Lennie, P. Rapid Adaptation in Visual Cortex to the Structure of Images. *Science* **285**, 1405–1408 (1999).

274. Dragoi, V., Sharma, J. & Sur, M. Adaptation-Induced Plasticity of Orientation Tuning in Adult Visual Cortex. *Neuron* **28**, 287–298 (2000).

275. Fritsche, M., Lawrence, S. J. D. & de Lange, F. P. Temporal tuning of repetition suppression across the visual cortex. *J. Neurophysiol.* **123**, 224–233 (2020).

276. Grill-Spector, K., Henson, R. & Martin, A. Repetition and the brain: neural models of stimulus-specific effects. *Trends Cogn. Sci.* **10**, 14–23 (2006).

277. Martinez-Conde, S., Macknik, S. L. & Hubel, D. H. The role of fixational eye movements in visual perception. *Nat. Rev. Neurosci.* **5**, 229–240 (2004).

278. Gawne, T. J. & Martin, J. M. Responses of Primate Visual Cortical Neurons to Stimuli Presented by Flash, Saccade, Blink, and External Darkening. *J. Neurophysiol.* **88**, 2178–2186 (2002).

279. MacEvoy, S. P., Hanks, T. D. & Paradiso, M. A. Macaque V1 Activity During Natural Vision: Effects of Natural Scenes and Saccades. *J. Neurophysiol.* **99**, 460–472 (2008).

280. McFarland, J. M., Bondy, A. G., Saunders, R. C., Cumming, B. G. & Butts, D. A. Saccadic modulation of stimulus processing in primary visual cortex. *Nat. Commun.* **6**, 8110 (2015).

281. Hasson, U., Yang, E., Vallines, I., Heeger, D. J. & Rubin, N. A Hierarchy of Temporal Receptive Windows in Human Cortex. *J. Neurosci.* **28**, 2539–2550 (2008).

282. Honey, C. J. *et al.* Slow Cortical Dynamics and the Accumulation of Information over Long Timescales. *Neuron* **76**, 423–434 (2012).

283. Herzog, M. H., Drissi-Daoudi, L. & Doerig, A. All in Good Time: Long-Lasting Postdictive Effects Reveal Discrete Perception. *Trends Cogn. Sci.* **24**, 826–837 (2020).

284. VanRullen, R. & Koch, C. Is perception discrete or continuous? *Trends Cogn. Sci.* **7**, 207–213 (2003).

285. Singhal, I. & Srinivasan, N. Time and time again: a multi-scale hierarchical framework for time-consciousness and timing of cognition. *Neurosci. Conscious.* **2021**, niab020 (2021).

286. Naccache, L. Why and how access consciousness can account for phenomenal consciousness. *Philos. Trans. R. Soc. B Biol. Sci.* **373**, 20170357 (2018).

287. Lamme, V. A. F. Towards a true neural stance on consciousness. *Trends Cogn. Sci.* **10**, 494–501 (2006).

288. Hense, A., Peters, A., Bruchmann, M., Dellert, T. & Straube, T. Electrophysiological correlates of sustained conscious perception. *Sci. Rep.* **14**, 10593 (2024).

289. Schwartz, D. Experimentum Crucis/Instantia Crucis in the Seventeenth Century. in *Encyclopedia of Early Modern Philosophy and the Sciences* (eds. Jalobeanu, D. & Wolfe, C. T.) 1–5 (Springer International Publishing, Cham, 2020). doi:10.1007/978-3-319-20791-9_61-1.

290. Negro, N. (Dis)confirming theories of consciousness and their predictions: towards a Lakatosian consciousness science. *Neurosci. Conscious.* **2024**, niae012 (2024).

291. Stanford, K. Underdetermination of Scientific Theory. in *The Stanford Encyclopedia of Philosophy* (eds. Zalta, E. N. & Nodelman, U.) (Metaphysics Research Lab, Stanford University, 2023).

292. Harding, S. *Can Theories Be Refuted?: Essays on the Duhem-Quine Thesis*. (Springer Science & Business Media, 2012).

293. Lakatos, I. & Musgrave, A. *Criticism and the Growth of Knowledge: Volume 4: Proceedings of the International Colloquium in the Philosophy of Science, London, 1965*. (Cambridge University Press, 1970).

294. Musgrave, A. & Pigden, C. Imre Lakatos. in *The Stanford Encyclopedia of Philosophy* (eds. Zalta, E. N. & Nodelman, U.) (Metaphysics Research Lab, Stanford University, 2023).

295. Lau, H. & Rosenthal, D. Empirical support for higher-order theories of conscious awareness. *Trends Cogn. Sci.* **15**, 365–373 (2011).

296. Yarkoni, T. The generalizability crisis. *Behav. Brain Sci.* **45**, e1 (2022).

297. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).

298. Lepauvre, A. *et al.* iEEG-data-release. (2024).

299. Musgrave, A. Method or Madness? in *Essays in Memory of Imre Lakatos* (eds. Cohen, R. S., Feyerabend, P. K. & Wartofsky, M. W.) 457–491 (Springer Netherlands, Dordrecht, 1976). doi:10.1007/978-94-010-1451-9_27.

300. Crupi, V. Confirmation. in *The Stanford Encyclopedia of Philosophy* (ed. Zalta, E. N.) (Metaphysics Research Lab, Stanford University, 2021).

301. Olcese, U. *et al.* Accelerating research on consciousness: An adversarial collaboration to test contrasting predictions of the Integrated Information Theory and Predictive Processing accounts of consciousness. (2021) doi:10.17605/OSF.IO/35RHX.

Nederlandse Samenvatting

In het veld van het bewustzijnsonderzoek worden tal van theorieën voorgesteld die onverenigbare mechanistische verklaringen bieden voor de neurale basis van bewustzijn. Deze theorieën ontwikkelen zich vaak parallel, waarbij elke theorie eigen empirisch bewijs verzamelt zonder interactie met of uitdaging van alternatieve perspectieven. Deze fragmentatie is symptomatisch voor methodologische beperkingen van de traditionele bottom-up benadering, die steunt op controversiële experimentele operationalisaties. Verschillende operationalisaties kunnen per ongeluk verschillende fenomenen volgen, die door onderzoekers als "bewustzijn" worden geïnterpreteerd. Dit leidt tot inconsistente bevindingen en belemmert de vooruitgang naar een verenigde wetenschappelijke verklaring van bewustzijn.

Erkennend dat dit tot stilstand heeft geleid, verschuift mijn proefschrift van de bottom-up benadering door te focussen op rigoureuze tests van bestaande theorieën over bewustzijn. Ik hanteer een noodzaak-dissociatiebenadering, met als doel gevallen te identificeren waarin bewustzijn optreedt in afwezigheid van de mechanismen die door een theorie worden voorgesteld, om zo de noodzakelijkheid van deze mechanismen te testen. Deze aanpak omzeilt de controversiële voorwaarden die nodig zijn bij de contrastieve benadering en maakt het mogelijk om theorieën te testen op nieuwe experimentele gronden.

Ik voerde twee experimenten uit die gericht waren op de temporele dynamiek van bewustzijn en de bijbehorende neurale activiteit tijdens de presentatie van stimuli. In de eerste studie maakte ik gebruik van het krachtige raamwerk van adversariële samenwerking om de voorspellingen te testen van twee toonaangevende theorieën van bewustzijn—de Geïntegreerde Informatietheorie (IIT) en de Global Neuronal Workspace Theory (GNWT)—met betrekking tot de verwachte neurale dynamiek bij aanhoudende visuele ervaringen. Deze samenwerking zorgde voor een onbevooroordeelde test van de voorspellingen van de theorieën en leidde tot belangrijke uitdagingen voor beide.

In de tweede studie onderzocht ik meer direct de temporele dynamiek van bewuste ervaring in verband met de presentatiecondities van het eerste experiment, waarbij bleek dat deze dynamiek sterk kan afwijken van intuïtieve verwachtingen. De bevindingen suggereren dat bewuste toegang tot stimuli mogelijk vluchtiger is dan eerder werd gedacht, wat een verfijning van ons huidige begrip van bewustzijn oplevert.

De resultaten van beide studies tonen aan dat de noodzaak-dissociatiebenadering effectief bijdraagt aan de vooruitgang van het veld door de beperkingen van bestaande theorieën over bewustzijn te identificeren en deze theorieën te verfijnen. De falsificatie van bepaalde voorspellingen leidde tot de ontwikkeling van nieuwe hypothesen en opende nieuwe empirische wegen om bewustzijn te onderzoeken en ons begrip ervan te verfijnen.

Concluderend illustreert dit proefschrift dat rigoureuze tests van theorieën over bewustzijn aanzienlijke vooruitgang in het veld bevorderen. Door te identificeren waar theorieën tekortschieten en hun evolutie aan te moedigen, verbeteren we zowel de theoretische modellen als ons begrip van bewustzijn zelf, en komen we dichter bij een verenigde wetenschappelijke verklaring. Deze benadering onderstreept het belang van het overstijgen van de traditionele bottom-up benadering door daarnaast te focussen op theorie-testing en verfijning, om de huidige impasse te doorbreken en de ontwikkeling van progressieve onderzoeksprogramma's te bevorderen.

English Summary

In the field of consciousness research, numerous theories propose incompatible mechanistic explanations for the neural underpinnings of consciousness. These theories often progress in parallel, each accumulating its own supporting empirical evidence without engaging with or challenging alternative perspectives. This fragmentation is symptomatic of methodological limitations inherent to the traditional bottom-up approach that relies on controversial experimental operationalization. Different operationalization may mistakenly track different phenomena, each interpreted by researchers as "consciousness", leading to inconsistent findings and hindering progress toward a unified scientific explanation of consciousness.

Recognizing this stalemate, my thesis shifts from the bottom-up approach by focusing on rigorous testing of existing theories of consciousness. I adopt a necessity dissociation approach, aiming to identify cases where consciousness occurs in the absence of the mechanisms proposed by a theory to test the necessity of these mechanisms. This approach circumvents the controversial conditions required by the contrastive approach and enables testing theories on novel experimental grounds.

I conducted two experiments centered on the temporal dynamics of consciousness and the associated neural activity during stimulus presentation. In the first study, I relied on the powerful framework of adversarial collaboration to test the predictions of two leading theories of consciousness—the Integrated Information Theory (IIT) and the Global Neuronal Workspace Theory (GNWT)—regarding the neural dynamics expected with persistent visual experiences. This collaborative approach allowed for unbiased testing of theories' predictions and led to significant challenges for both.

In the second study, I investigated more directly the temporal dynamics of conscious experience associated with the presentation conditions of the first experiment, revealing that these dynamics might differ strikingly from intuitive expectations. The findings suggest that conscious access to stimuli may be more transient than previously thought, yielding a refinement of our current understanding of consciousness.

The results of both studies demonstrate that the necessity dissociation approach effectively advances the field by identifying the limitations of existing theories of consciousness and prompting their refinement. The falsification of certain predictions led to the development of new hypotheses, opening novel empirical avenues for exploring consciousness and refining our understanding thereof.

In conclusion, this thesis illustrates that rigorous testing of theories of consciousness fosters significant progress in the field. By identifying where theories fall short and encouraging their evolution, we enhance both the theoretical models and our understanding of consciousness itself, moving closer to a unified scientific explanation. This approach underscores the importance of moving beyond the traditional bottom-up approach by focusing in addition on theory testing and refinement to break through the current stalemate and foster the evolution of progressive research programs.

Research data management

The research presented in this thesis followed the applicable laws and ethical guidelines. Research data management was conducted according to the FAIR principles (Findable, Accessible, Interoperable, Reusable). The paragraphs below specify this in detail and provide access information to the data.

ETHICS

This thesis uses data from human participants. The experiment presented in Chapter 2 was approved by the institutional committees of each data collecting labs (Comprehensive Epilepsy Center at New York University Langone Health, Brigham and Women's Hospital, Boston Children's Hospital (Harvard), and University of Wisconsin School of Medicine and Public Health, Centre for Human Brain Health of the University of Birmingham, the Center for MRI Research of Peking University, Yale Magnetic Resonance Research Center and at the Donders Centre for Cognitive Neuroimaging of Radboud University Nijmegen). All volunteers and patients provided oral and written informed consent before participating in the study. Epilepsy patients were also informed that clinical care was not affected by participation in the study. The Experiment presented in Chapter 3 was approved by the Ethics Council of the Max Planck Society (No. 2017_12). Participants provided written informed consent before the study. Both study procedures were carried out in accordance with the Declaration of Helsinki.

This research was supported by the Templeton World Charity Foundation (TWCF0389) and the Max Planck Society.

FINDABLE, ACCESSIBLE

Data, code, and research documentation were shared on openly accessible platforms. The data collected in Chapter 2 are archived on hard disk drives (HDD) at the Max Planck Institute for empirical institute and can be downloaded as data bundles or through the XNAT database platform. The data are accompanied by extensive machine-readable metadata, enabling programmatic queries based on various attributes of the data and subjects to maximize accessibility and findability of the data. The data collected in Chapter 3 are available on the OSF platform. All code and documentation necessary for the replication of published results are available on GitHub. All data will remain available for at least 5 years after termination of the studies.

Chapter	Resource	Link	License
2	Data	https://www.arc-cogitate.com/data-release	MIT License
2	Analysis code	https://github.com/Cogitate-consortium/cogitate-msp1	MIT License
2	Experiment code	https://github.com/Cogitate-consortium/Experiment1	MIT License
3	Data	https://osf.io/krjh7	MIT License
3	Analysis code	https://github.com/ncc-brain/Reconstructed_time_analysis	MIT License
3	Experiment code	https://github.com/ncc-brain/Reconstructed_time_experiment	MIT License

INTEROPERABLE, REUSABLE

The raw data for Chapter 2 and 3 are available in their raw format following the BIDS conventions, to ensure interoperability and reusability. All codes used for analysis and experiment have been documented extensively and we provide readme files to instruct users how to use our data. In addition, I have created code and notebooks to illustrate how to use the iEEG data of our project to ensure reusability of the data for other scientific enquiries and serve as educational resources for researchers. These resources can be found here: https://github.com/Cogitate-consortium/iEEG-data-release

PRIVACY

The privacy of all participants has been warranted by using pseudonymized subject codes. Linking pseudonymized codes to personal data is not possible, as all keys Key files were deleted after finalization of the projects presented in chapter 2 and 3. Personal identifiable information was removed from all files and headers and MRI-data from Chapter 2 were defaced before being shared.

List of publications

Lepauvre, A., & Melloni, L. (2021). **The search for the neural correlate of consciousness: Progress and challenges.** *Philosophy and the Mind Sciences*, *2*. Doi: https://doi.org/10.33735/phimisci.2021.87

Cogitate Consortium, Ferrante, O[1], Gorska-Klimowska, U.[1], Henin, S.[1], Hirschhorn, R.[1], Khalaf, A.[1], Lepauvre, A.[1], Liu, L.[1], Richter, D.[1], Vidal, Y.[1], Bonacchi, N., Brown, T., Sripad, P., Armendiz, M., Bendtz, K., Ghafari, T., Hetenyi, D., Jeschke, J., Kozma, C., Mazumder, D. R., Montenegro, S., Seedat, A., Sharafeldin, A., Yang, S., Baillet, S., Chalmers, D. J., Cichy, R. M., Fallon, F., Panagiotaropoulos, T. I., Blumenfeld, H., de Lange, F. P., Devore, S., Jensen, O., Kreiman, G., Luo, H, Boly, M., Dehaene, S., Koch, C., Tononi, G., Pitts, M., Mudrik, L., Melloni, L. (2023). **An adversarial collaboration to critically evaluate theories of consciousness**, *BioRxiv*. doi: https://doi.org/10.1101/2023.06.23.546249

Melloni, L., Mudrik, L., Pitts, M., Bendtz, K., Ferrante, O., Gorska, U., Hirschhorn, R., Khalaf, A., Kozma, C., Lepauvre, A., Liu, L., Mazumder, D., Richter, D., Zhou, H., Blumenfeld, H., Boly, M., Chalmers, D. J., Devore, S., Fallon, F., de Lange, F. P., Jensen, O., Kreiman, G., Luo, H., Panagiotaropoulos, T. I., Dehaene, S., Koch, C., Tononi, G. (2023) An adversarial collaboration protocol for testing contrasting predictions of global neuronal workspace and integrated information theory. *PLoS One*, *18*(2), e0268577. doi: https://doi.org/10.1371/journal.pone.0268577

Grassi, F., Kulke, L., Lepauvre, A., & Schacht, A. (2024). Relevance acquisition through motivational incentives: Modeling the time-course of associative learning and the role of visual features. *Imaging Neuroscience*, *2*, 1-20. doi: https://doi.org/10.1162/imag_a_00162

Lepauvre, A., Melloni, L., Hirschhorn, R., Mudrik, L., & Bendtz, K. (2024). A standardized framework to test event-based experiments. *Behavior Research Methods* (2024): 1-17. doi: https://doi.org/10.3758/s13428-024-02508-y

Seedat, A.[1], Lepauvre, A.[1], Jeschke, J.[1], Gorska-Klimowska, U.[1], Armendariz, M., Bendtz, K., Henin, S., Hirschhorn, R., Brown, T., Jensen, E., Kozma, C., Mazumder, D., Montenegro, S., Yu, L., Bonacchi, N., Sripad, P., Taheriyan, F., Devinsky, O., Dugan, P., Doyle, W., Flinker, A., Friedman, D., Lake, W., Pitts, M., Mudrik, L., Boly, M., Devore, S., Kreiman, G., Melloni, L. (2024). Open multi-

center iEEG dataset with task probing conscious visual perception. Submitted at Nature Scientific data

Lepauvre, A. [1], Engeser, M., Dehaene, M., Melloni, L. (2024). **Investigating the timing of conscious experience using a dual-task paradigm and quantified introspection**. Submitted at eLife

Zeidman, P., Lepauvre, A., Melloni, L., Friston, K. (2024). Variational Representation Similarity Analysis (vRSA) for EEG/MEG, in preparation

[1] Shared first authorship

Included in this thesis

Acknowledgements

It is often said that science is a lonely enterprise, tackling tough problems for years on one's own. If this statement is true in some cases, I have learned over the years that this is definitely not the kind of science I want to do. I was lucky enough to see that there are other ways to do science. I never felt alone throughout the years, and I have several people to thank for that.

Lucia, you are the first I ought to thank. When I started working as a lab manager in your lab, I had not realized how lucky I had just been. At the time, I already knew my one passion was consciousness research, but I knew little about how large the field already was. Fortunately, I had just landed in one of the labs that was about to embark on one of the most exciting and ambitious projects in the field. Not only that, you decided to give me the chance to actually work on that project, and for that, I am eternally grateful. It has now been four hectic years since then, and a lot has happened. Through this, we learned to know each other; you recognized my qualities and my flaws and guided me to get the best out of both, even when it took me a long time to realize it. And of course, this thesis would never have been what it is without the invaluable scientific input you have provided me over the years. I thank you for your patience, but also for your relentlessness in pushing me beyond my fears and convincing me that I can achieve even what I thought I could not.

Floris, I would like to thank you for your help and guidance throughout my PhD. In all our interactions, you provided insightful and wise comments, both from a scientific standpoint and from a career standpoint. Through your advice, I was able to navigate the difficulties of conducting a PhD in ways that I do not believe would have been possible otherwise. You made me realize that while my work in the Cogitate project was important, it is essential for me to conduct independent research and develop my own research agenda. Your encouragement and support in this regard have been invaluable to my growth as a scientist. One insight of yours stands out above all: “perfect is the enemy of good.”

Catherine, I cannot thank you enough for your honesty and your commitment to the role of advisor. You were always direct when it came to constructive feedback, and your guidance helped me overcome numerous challenges. You were an ideal advisor, and your only care and concern were that my work be as valuable and successful as possible, for which I am extremely grateful.

Part of the reason I never felt alone is, of course, because I was engaged in a project as large as Cogitate. However, having many people around is not enough to avoid feeling alone; that takes nice and caring people. I was blessed to work with a group that excelled not only in its scientific achievements but also in its interpersonal qualities. We have been through a lot, good and bad, but always together. There are several colleagues I would like to thank in particular, whose support was not only sufficient but also necessary for the completion of this work. **Rony**, you have been a great collaborator and partner in extinguishing fires. Your encyclopedic knowledge about consciousness, as well as your scientific relentlessness, is something I can only aspire to. I like to think we complement each other, and I will take any chance I get to write another paper with you. **Yamil**, the commander-in-chief of the Cogitate infantry, what would we have done without you? Your bravery and wisdom were instrumental to our success. To me personally, you were a mentor of sorts, always willing to spare time to provide scientific insights on my various projects and share your experience and knowledge of the academic world. **Tanya**, I am also not sure what we would have done without you. You were the best project manager we could have wished for, always having the best interest of everyone at heart, making sure we saw things through. Your kindness and dedication also made this thesis possible. **Oscar**, your dedication was impressive to witness, and the amount of data you have collected speaks for itself. Thank you for being a great colleague to design analysis with and for your willingness to spare some time to provide feedback on my papers. **Katarina**, you were a great teacher and friend, and I am glad I got to figure out the intricacies of iEEG data analysis together with you. Your dedication to getting to the bottom of any problem you encountered was formative in shaping me as a scientist. **Simon**, your support and advice propelled me towards a level of expertise in intracranial recordings that I doubt I would otherwise have reached. **Praveen**, thanks for your advice on coding practices. I was luckier than most because, in the end, it is you and the data release team (**Niccolo, Fatemeh, Dip**) who did most of the work to ensure that our data were properly documented, adhered to the BIDS standards, and went beyond by creating such rich and detailed metadata. **David R.**, I am glad to have worked with you. I have fond memories of getting on calls at impossible hours in the middle of the corona pandemic to make sure that the cryptic parameters of the Eyelink and Psychtoolbox in the Donders fMRI lab were correct. You guided me on how to find my way through the Donders PhD program, and for that, I am thankful. **Liad and Mike**, thank you for your efforts in steering the massive ship we were, together with Lucia and for always listening what I and others had to say, no matter what it was. I also have to thank all the members of the iEEG team, without whom this thesis would not have been possible. Setting up the experiment remotely during the pandemic was not easy, but it is thanks to you, **Jay, Stephanie and Alia**,

and your willingness to test the experiment countless times that it was possible. And **Alia**, thanks for coordinating our efforts in collecting the data and pushing for the publication of a Scientific data paper, to make sure our work gets the recognition it deserves. Thanks to you, **Sasha, Gabriel, Urszula, Marcello, and David M.,** for all the discussions and advice on how to ensure we acquired high-quality data and for your guidance through the analysis.

Through my involvement in the Cogitate and Lucia's support, I had the privilege to work together with some of the most brilliant neuroscientists of our times. Despite their busy schedules, they were also incredibly generous with their time, sharing their views, ideas, and theories with us. I has been a tremendous learning experience. **Stan,** your support and supervision in both studies presented in this thesis was invaluable. Your creativity and proficiency in experimental design was awe-inspiring to witness. Your deep understanding of consciousness and precise thinking pushed me to challenge my assumptions and convictions, making me a better scientist in the process. **Giulio and Melanie**, thank you for sparing time to teach us about your theory, helping me grasping its intricacies. It was instrumental to understand how unique this approach is to study conscious experience and how much there is to learn. I would like to thank **Peter and Karl**, for giving me the unique opportunity to spend two months at your amazing institute, to learn about Bayesian inference, opening my eyes to the value of this framework to pursue my research agenda. While I did not have the chance to apply this knowledge in the studies presented in this thesis, it will play a central role in the next steps of my scientific career.

Of course, I also have to thank all the members of the NCC group for making it such a nice lab to work in and for tolerating my typing loudness over the years. **Zefan**, it is only after you arrived that I realized I was missing something—a friend with whom to have daily informal interactions to discuss the subject that fascinates us both. Thank you for your open-mindedness and for all the great conversations we have had over the years. You helped me clarify so many of the concepts floating around in my mind, shared so many relevant papers that I have lost count, and made this thesis so much better. Thanks for always accepting to read my manuscripts and comfort my insecurities. My only regret is that I did not end my post-corona seclusion earlier to benefit longer from your friendship. **Darinka**, your advice has been invaluable, not only in helping me define my own research agenda but also in guiding me toward scientific independence. Your feedback on my manuscripts was always insightful, highlighting strengths that needed emphasis and weaknesses that needed addressing. You helped me develop my research beyond the collective endeavor of Cogitate, and your insights on both research and career development have had a profound impact

on how I approach my work. I am deeply grateful for your support. **Micha**, thank you for your amazing work on the study presented in Chapter 3. You were a great student to supervise, but also someone to learn from. Your sharpness and precise thinking brought the project to a level I would not have achieved alone. I can only hope that my supervision kept up with your talent. **Qian**, thank you for being such a talented and helpful colleague. Your pedagogic qualities, paired with your breadth of knowledge of neuroscience, helped my work tremendously. Thanks for showing me the ropes of contributing to open-access toolboxes, making me a much better open scientist. **Jan**, of course, thanks for all the discussions we had over the past few months. When I first started writing this thesis, a single discussion we had one morning helped me shape the thesis to its current form, and recurring interactions helped me refine it much further. I can only imagine what I would have achieved had you been there since the beginning. And, of course, thanks to everyone who came and went over the years for creating such a dynamic and inspiring scientific environment (ordered by whose name my eyes landed on first in our group chat): **Piermatteo, Bene, Qiyuan, Jiyun, Diana G., Deysha, Ece, Kyle, Yuranny, Arion, and William**. And thank you of course to all the labs staff for providing incredible support in recruiting participants and collecting data over the years.

Of course, I have a world outside academia, and they too were instrumental in the completion of this thesis. I have to thank my parents (**Beatrice** and **Christian**) for never suggesting that any career path was too far beneath or beyond me. You have raised me with the belief that what I ought to do is what I like to do, free of any moral judgments. Your unwavering support and the values you have instilled in me have allowed me to pursue my passion, and I could not have asked for a better foundation in life. You both do something better than most—simply listen, and sometimes that is the most helpful thing in the world. I also need to mention my grandmothers (**Cecile** and **Yvette**) and my grandfathers (**Francois** and **André,** whom we lost along the way), as this PhD probably means even more to them than it does to me—though my grandmothers still don't quite understand how someone can be a doctor without writing prescriptions. Their pride make this accomplishment all the more special for me. **Diana K.**, I also would not have made it without you. Your influence grounded me, making me realize that as much as my work is important to me, the world does not revolve around which part of the brain is activated when we look at horrid computer-generated faces, and that my world will not end when I fail at something. In times of hardship, you helped me take the necessary distance to realize that the best I can do is to do the best I can—nothing more. And, of course, thanks for supporting me and my obsessions over the years. A tremendous thank you to all my friends and family for supporting me over the years. Discussing with you all has helped me realize that

as important as it is to do good research, communicating it well is equally important. After all, what is the point of doing it if we cannot even tell a nice story about it? I will not name you all by fear of forgetting someone important. But I need to point out that you, **Simeon**, would not have been included in the list, because I would never forgive myself for not making this joke to conclude this thesis.

Thanks everyone for making this work possible.

Donders Graduate School

For a successful research Institute, it is vital to train the next generation of scientists. To achieve this goal, the Donders Institute for Brain, Cognition and Behaviour established the Donders Graduate School in 2009. The mission of the Donders Graduate School is to guide our graduates to become skilled academics who are equipped for a wide range of professions. To achieve this, we do our utmost to ensure that our PhD candidates receive support and supervision of the highest quality.

Since 2009, the Donders Graduate School has grown into a vibrant community of highly talented national and international PhD candidates, with over 500 PhD candidates enrolled. Their backgrounds cover a wide range of disciplines, from physics to psychology, medicine to psycholinguistics, and biology to artificial intelligence. Similarly, their interdisciplinary research covers genetic, molecular, and cellular processes at one end and computational, system-level neuroscience with cognitive and behavioural analysis at the other end. We ask all PhD candidates within the Donders Graduate School to publish their PhD thesis in de Donders Thesis Series. This series currently includes over 700 PhD theses from our PhD graduates and thereby provides a comprehensive overview of the diverse types of research performed at the Donders Institute. A complete overview of the Donders Thesis Series can be found on our website: https://www.ru.nl/donders/donders-series

The Donders Graduate School tracks the careers of our PhD graduates carefully. In general, the PhD graduates end up at high-quality positions in different sectors, for a complete overview see https://www.ru.nl/donders/destination-our-former-phd. A large proportion of our PhD alumni continue in academia (>50%). Most of them first work as a postdoc before growing into more senior research positions. They work at top institutes worldwide, such as University of Oxford, University of Cambridge, Stanford University, Princeton University, UCL London, MPI Leipzig, Karolinska Institute, UC Berkeley, EPFL Lausanne, and many others. In addition, a large group of PhD graduates continue in clinical positions, sometimes combining it with academic research. Clinical positions can be divided into medical doctors, for instance, in genetics, geriatrics, psychiatry, or neurology, and in psychologists, for instance as healthcare psychologist, clinical neuropsychologist, or clinical psychologist. Furthermore, there are PhD graduates who continue to work as researchers outside academia, for instance at non-profit or government organizations, or in pharmaceutical companies. There are also PhD graduates who work in education, such as teachers in high school, or as lecturers in higher education. Others continue in a wide range of positions, such as policy advisors, project managers, consultants,

data scientists, web- or software developers, business owners, regulatory affairs specialists, engineers, managers, or IT architects. As such, the career paths of Donders PhD graduates span a broad range of sectors and professions, but the common factor is that they almost all have become successful professionals.

For more information on the Donders Graduate School, as well as past and upcoming defences please visit:

http://www.ru.nl/donders/graduate-school/phd/

9 789465 150550 >

MAX PLANCK INSTITUTE
FOR EMPIRICAL AESTHETICS

DONDERS INSTITUTE

Radboud University

Radboudumc