CONSCIOUSNESS

What a Contest of Consciousness Theories Really Proved

By ELIZABETH FINKEL

August 24, 2023

A five-year "adversarial collaboration" of consciousness theorists led to a stagy showdown in front of an audience. It crowned no winners — but it can still claim progress.

35

Myriam Wares for *Quanta Magazine*

Science routinely puts forward theories, then batters them with data till only one is left standing. In the fledgling science of consciousness, a dominant theory has yet to emerge. More than 20 are still taken seriously.

It's not for want of data. Ever since Francis

Crick, the co-discoverer of DNA's double helix, legitimized consciousness as a topic for study more than three decades ago, researchers have used a variety of advanced technologies to probe the brains of test subjects, tracing the signatures of neural activity that could reflect consciousness. The resulting avalanche of data should have flattened at least the flimsier theories by now.

Five years ago, the Templeton World Charity Foundation initiated a series of "adversarial collaborations" to coax the overdue winnowing to begin. This past June saw the results from the first of these collaborations, which pitted two high-profile theories against each other: global neuronal workspace theory (GNWT) and integrated information theory (IIT). Neither emerged as the outright winner.

The results, announced like the outcome of a sporting event at the 26th meeting of the Association for the Scientific Study of Consciousness (ASSC) in New York City, were also used to settle a 25-year bet between Crick's longtime collaborator, the neuroscientist Christof Koch of the Allen Institute for Brain Science, and the philosopher David Chalmers of New York University, who coined the term "the hard problem" to challenge the presumption that we can explain the subjective feeling of consciousness by analyzing the circuitry of the brain.

The neuroscientist Christof Koch of the Allen Institute for Brain Science deemed the mixed results of the first adversarial collaboration on consciousness to be "a victory for science."

Erik Dinnel

Onstage at NYU's Skirball Center, following interludes of rock music, a rap performance on consciousness and the presentation of the results, the neuroscientist conceded the bet to the philosopher: The neural correlates of consciousness had not yet been nailed down.

Nevertheless, Koch proclaimed, "It's a victory for science."

But was it? The event has received mixed

reviews. Some researchers point to the failure to meaningfully test the differences between the two theories. Others highlight the success of the project in driving consciousness science forward, both by delivering large, novel, skillfully executed data sets and by inspiring other contestants to engage in their own adversarial collaborations.

The Correlates of Consciousness

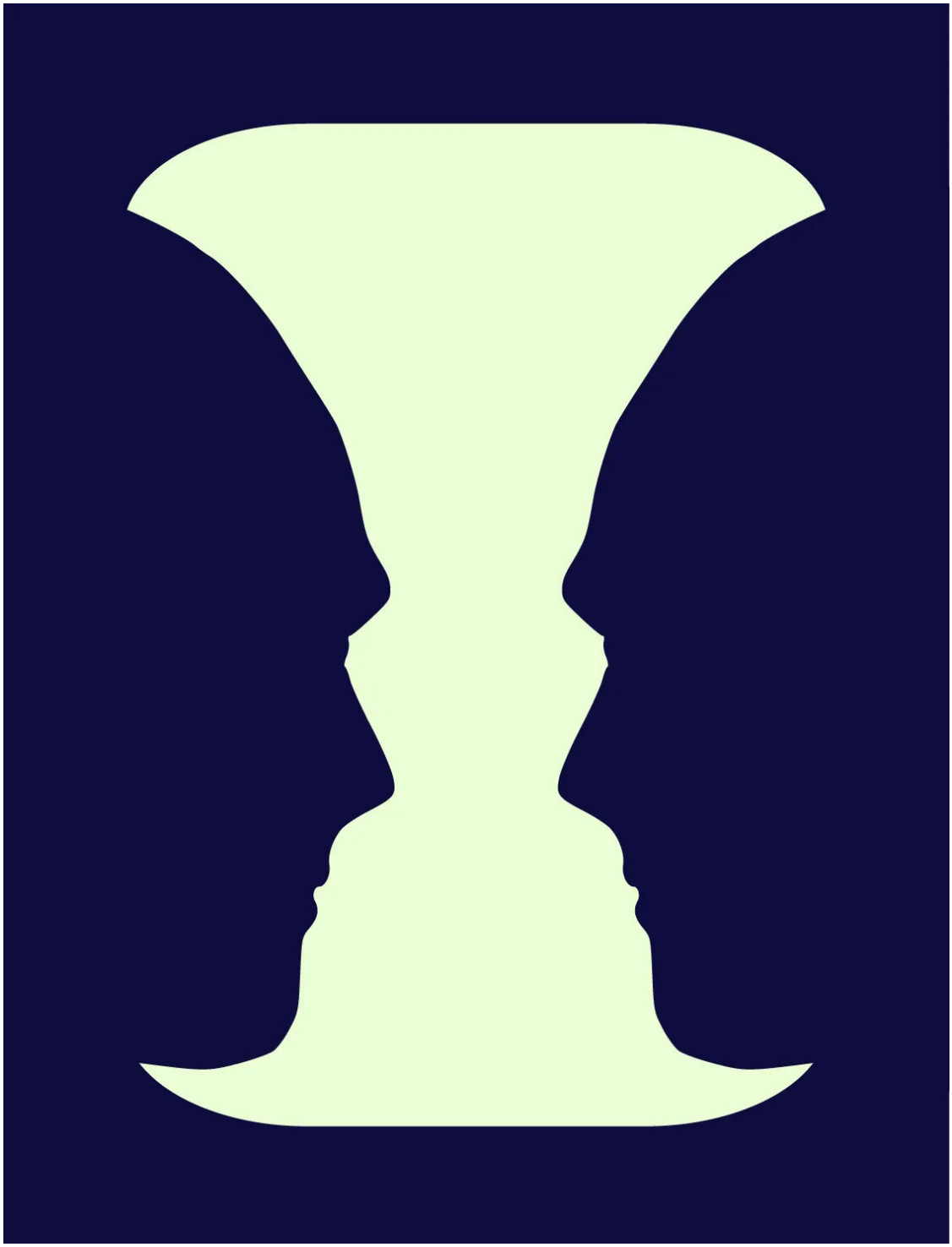
When Crick and Koch published their landmark paper "Towards a Neurobiological Theory of Consciousness" in 1990, their aim was to place consciousness — for 2,000 years the stomping ground of philosophers— onto a scientific footing. Consciousness in its entirety, they argued, was too broad and controversial a concept to serve as a starting point.

Instead, they focused on one scientifically tractable aspect of it: visual perception, which involves becoming conscious of seeing, for instance, the color red. The scientific goal was to find the circuitry that correlated with that experience, or, as they put it, the "neural correlates of consciousness."

Decoding the first stages of visual perception had already proved a fertile ground for science. Patterns of light falling on the retina send signals to the visual cortex in the back of the brain. There, upwards of 12 distinct neural modules process the signals corresponding to edges, color and movement in the images. Their output combines to build up a final dynamic picture of what we consciously see.

What clinched the usefulness of visual perception for Crick and Koch was that the

final link in that chain — consciousness — could be detached from the rest. Since the 1970s, neuroscientists have known of people with "blindsight" who have no experience of vision because of damage to their brain, yet who can navigate a room without bumping into obstacles. While they retain the ability to process an image, they're missing the ability to be conscious of it.



In this famous illusion, the brain's mechanism for producing consciousness allows us to experience the image as either a vase or two faces — but not both simultaneously.

Nevit Dilmen

All of us can experience a form of this disconnection. Consider the well-known optical illusion that can be perceived as either a vase or two faces in profile. At any moment we can see it only as one or the other. Something in how our brain processes perceptions prevents us from being conscious of both simultaneously.

Experimental psychologists can take advantage of that quirk through the phenomenon of binocular rivalry. Our brain normally has no trouble combining the slightly different, overlapping images it receives from the left and right eyes. But if the images are very different, instead of merging, they

become rivals: First one image dominates our perception, then the other. When the neuroscientist Nikos Logothetis of the Max Planck Institute for Biological Cybernetics described binocular rivalry in 1996, Crick was so excited, he proclaimed that neural correlates of consciousness would be found by the end of the 20th century. (Similar enthusiasm led to Koch's bet with Chalmers.)

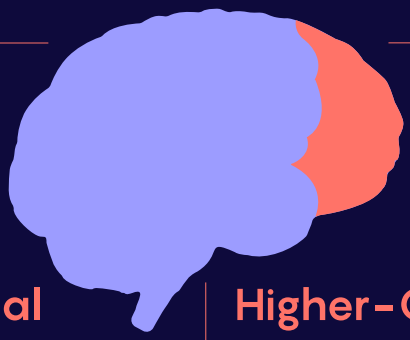
Over the last two decades, ever more sophisticated brain scanners have monitored test subjects as their perceptions were manipulated during studies of consciousness. Trickles of data have become cascades, yet rather than being washed away, theories of consciousness have multiplied.

A broad division among these many theories is that some of them, like GNWT, require the

participation of the parts of the brain that enable cognition, where we "think," while IIT and others claim that the neural correlates depend on brain areas involved in perception, where we "sense." The ideas are often casually described as "front-of-the-brain" theories versus "back-of-the-brain" theories (though the actual anatomical distinction is less cut and dried than that). This intriguing bifurcation echoes old philosophical disagreements over whether consciousness is about thinking, as in Descartes' "I think, therefore I am," or about "not thinking," as in the state experienced by a meditating yogi.

Thoughts About Consciousness

Theories about consciousness, currently numbering more than 20, can be sorted by roughly where in the brain they propose it arises. Here are some examples:

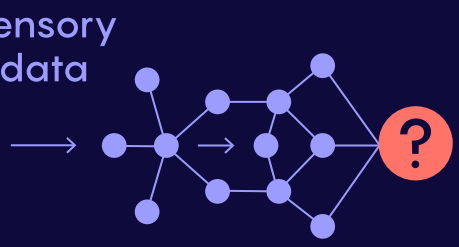


FRONT OF THE BRAIN

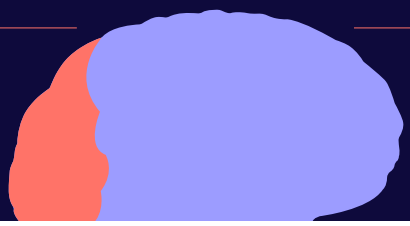
Global Neuronal Workspace Theory

Consciousness arises from the integration of signals in a "workspace" before distribution for decision-making.

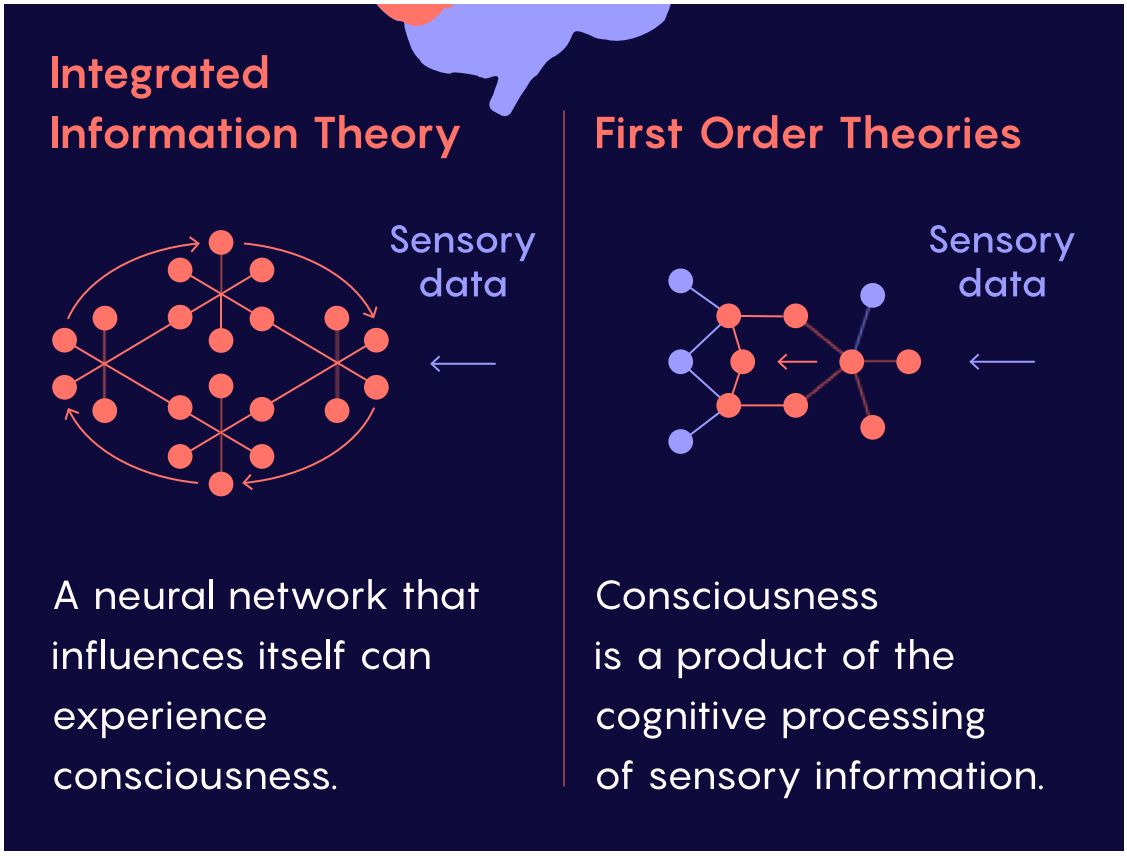
Higher-Order Theories



Consciousness involves processing that builds on cognitive representations of sensory experiences.



BACK OF THE BRAIN



Merrill Sherman/*Quanta Magazine*

To the neuroscientist Stanislas Dehaene of the Collège de France, the chief architect of GNWT, thinking is a core part of the conscious state. Referring to IIT, he told me, "It's a big difference between our theories. I don't believe in purified consciousness."

GNWT maintains that a tiny subset of the information we constantly process unconsciously is selected to pass through a bottleneck into a conscious "workspace." There, the information is integrated and broadcast to other brain areas to make it globally available for decision-making and learning. "The 'workspace' is there for a function," Dehaene said. Because decision-making and learning are responsibilities of the prefrontal cortex, the front of the brain is deemed crucial for consciousness.

The germ of the idea was originally proposed in 1988 by the psychologist Bernard Baars, now at the Society for Mind Brain Sciences, who saw an analogy to the "blackboard" of early artificial intelligence system architectures where independent programs shared information. Dehaene then tied that

conceptual template to the findings of cutting-edge neuroscience and used computational models to develop GNWT.

The neuroscientist Giulio Tononi of the University of Wisconsin, Madison, developed the integrated information theory by mathematically formulating five axioms about

consciousness.

John Maniaci/UW Health

IIT makes no analogies to AI architecture. Giulio Tononi, a neuroscientist and psychiatrist at the University of Wisconsin, Madison, developed the theory by starting with five axioms about consciousness: It is intrinsic to the entity who has it; its composition is structured; it is information-rich; it is integrated rather than reducible to components; and it is exclusive of other experiences. He then developed mathematical descriptions to fit those axioms. To Tononi and other IIT theorists, the neural structure most consistent with those mathematical descriptors is a gridlike architecture associated with sensory regions, which they've dubbed

the "hot zone."

But GNWT and IIT are only two of the theories that locate key elements of consciousness at opposite poles of the brain. There are other cognitive, front-of-the-brain concepts including several higher-order theories (HOTs) and active inference theory, and a variety of sensory, back-of-the-brain concepts such as the closely related first-order theories and localist theories.

Eliminating some of them by testing their predictions against data from living brains might seem like simplicity itself. Unfortunately, that has not turned out to be true.

Finding What They Look For

For years, researchers devised clever experiments in which test subjects reported when they became conscious of an object while psychological tricks or illusions were used to distract them. Those results often showed that the moment of conscious perception correlated with activity in the prefrontal cortex, favoring something like a GNWT or other front-of-the-brain explanation. But philosophers and experimenters began complaining that those studies could be measuring the neural activity associated with the task of reporting rather than the consciousness itself.

"No-report" paradigms were therefore developed as a workaround. A popular one involved binocular rivalry. If a leftward-moving face is shown to a test subject's left eye and a rightward-moving house to their right

eye, their conscious perception will flip between the two images. Researchers can identify the perceived image without a report by tracking which way the eyes are moving. Data at the time suggested that in these no-report paradigms, the signal for conscious perception localized to the back of the brain.

Yet theorists were rarely persuaded by any of the experiments and data. In a 2016 review, the IIT camp dismissed the report-based experiments as methodologically flawed. The debate continued in 2017 with dueling articles in the *Journal of Neuroscience.* In one of them, Hakwan Lau, now at the Riken Center for Brain Science in Japan, and his colleagues offered the riposte that no-report paradigms were themselves rife with confounding variables.

A further complication was that the

experimental results depended on the type of brain recording technique used. That's not surprising since each technology provides a different lens into the brain. Functional magnetic resonance imaging (fMRI), for instance, tracks blood flow and offers good spatial resolution but is too sluggish to keep up with the pace of chatter among the neurons. Magnetoencephalography (MEG), on the other hand, tracks brain chatter but has poorer spatial resolution. It also makes a difference whether researchers are measuring signal strength in specific locations of the brain or analyzing patterns over wider areas.

The upshot was that despite the wealth of experimental data gathered to study the correlates of consciousness, the uncertainties gave theorists room to claim that the data supported their preferred explanations.

The neuroscientist Liad Mudrik (left) of Tel Aviv University and her doctoral student Itay Yaron (right) have compiled evidence that the goal of using experimental studies to test theories of consciousness is often thwarted by biases that creep into the design of the experiments.

Sophie Kelly

Liad Mudrik, a Tel Aviv University neuroscientist, believes that part of the problem lies in the way the studies were (and often continue to be) designed. A recent survey by her doctoral student Itay Yaron looked at more than 400 published consciousness experiments and found that it was largely possible to predict which theory would be supported solely on the basis of the design of the experiment, without knowing anything about the results.

Adversarial Collaboration

Five years ago, Dawid Potgieter, the head of the special programs section of the Templeton World Charity Foundation, was astonished to find that there were still so many viable theories about consciousness. He felt the time was ripe to do something about it.

Koch suggested a head-to-head contest, which had sometimes been used to settle controversies in physics. There were precedents in psychology too. In the 1980s, the psychology researcher Dan Kahneman of Princeton University coined the term "adversarial collaboration" to describe exercises in which scientists with opposing views jointly developed experiments. By working together, they could smooth out disagreements over goals and methodology that might undermine the conclusions of the work. (Kahneman hit on this approach while resolving a theoretical feud he was having with his psychologist colleague and wife, Anne Treisman.)

Dawid Potgieter, acting on behalf of the Templeton World Charity Foundation, convened a workshop five years ago to develop

adversarial collaborations for testing theories of consciousness.

—

Templeton World Charity Foundation

Potgieter was keen to try. In March 2018, he and Koch hosted a weekend workshop at the Allen Institute in Seattle for 14 participants. It included three theorists — Dehaene, Tononi and Lau, who champions HOTs — as well as Chalmers and two other philosophers, four psychologists, two neuroscientists, a neurologist and Potgieter as a representative of the Templeton Foundation. Their charge was to collaboratively design new experiments to iron out all the past wrinkles and discriminate cleanly between the theories.

Three of the psychologists — Mudrik, Lucia

Melloni of the Max Planck Institute and Michael Pitts of Reed College in Portland — already had a history of challenging theories of consciousness. "At some point, I think Giulio suggested, 'Why don't the three of you lead the project?'" Pitts recalled. "We had no idea what we were in for. It's consumed our lives."

Over the next nine months, discussions continued. The theorists drilled down into their theories and offered new predictions — one of the novel contributions of the collaboration. Mudrik was impressed by the willingness of the adversaries to negotiate. "It takes a lot of courage; you're putting your neck on the line," she said.

The team came up with two experimental designs for disentangling the predictions of IIT and GNWT. They never came up with

predictions that were different enough to disentangle GNWT and HOTs, so HOTs were left for a different adversarial collaboration involving Lau and the NYU philosopher Ned Block, who champions first order theories.

Tononi was particularly keen on the design of the first GNWT-versus-IIT experiment. Since tasks had created such a wrinkle in past experiments, it would iron them out by varying the tasks to see how that affected conscious perception.

The test subjects would be presented with a series of varied images, such as faces, clocks and letters of the alphabet in different fonts. They would see each image for 0.5 to 1.5 seconds. At the beginning of each series, two specific images would be defined as targets (say, the face of a woman and a vintage clock),

and participants were given the reporting task of pressing a button if they saw either of them. Other faces and objects in the images would therefore be task-relevant (because they fell into the same categories as the targets), but no report was required. Other types of images in the series, such as alphabet letters and meaningless symbols, would be task-irrelevant. The test was run repeatedly with different targets in the series so that each set of stimuli could be tested as both task-relevant and task-irrelevant. State-of-the-art brain signal decoders would correlate neural firing patterns with what the subjects were seeing.

GNWT predicted that the brain patterns corresponding to conscious perceptions of objects would be similar whether a task was involved or not. The brain decoders should be able to identify a distinctive signal associated

with a target image regardless of the task. Moreover, it ought to be possible to detect the "ignition signal" of a new conscious percept entering the brain's workspace as well as an "off signal" clearing it.

IIT, on the other hand, predicted that the brain patterns of consciousness would vary with the tasks, because carrying out a task would involve the prefrontal cortex and perception stripped of a task would not. This "pure" form of consciousness would only require the sensory hot zone at the back of the brain. The connectivity and duration of the signals for consciousness of an image would match the duration of the visual stimulus.

The psychology researcher Daniel Kahneman of Princeton University believes strongly in the value of adversarial collaborations for advancing science, even though he finds that the results rarely change the minds of the adversaries.

Roger Parkes/Alamy Stock Photo

Dehaene favored the second experiment, which also involved the comprehensive decoding of brain patterns. Test subjects would be

randomly exposed to faces and objects flashed on a screen while they played a distracting Tetris-like video game. Shortly after an image was shown, the game would stop and the subject would be asked whether they saw it. Dehaene preferred this design because it offered a more clear-cut contrast between conscious and unconscious mental states, which he considered essential to getting unambiguous data on the correlates of consciousness.

Because Kahneman was so familiar with adversarial collaborations, he mentored the three project leaders. But he also warned them that, in his experience, adversaries don't change their minds after seeing the results of their collaborations. Instead, when faced with an inconvenient result, "their IQ leaps 15 points" as they invent ways to accommodate

the new, conflicting data, he said.

Mixed Results, With No Losers

The researchers set to work performing the experiments suggested by the workshop team. The GNWT-versus-IIT experiment that Tononi liked best, which tested with different levels of tasks, finished up first. It was carried out in two different labs using fMRI, MEG and intracranial electroencephalography. In all, six theory-neutral labs and 250 test subjects participated.

On the evening of June 23, an excited audience gathered at NYU to learn the outcome of that experiment. Writ large on a giant screen, the results were shown on a chart marked by red and green highlights, as though the researchers were reporting on a steeplechase

with three types of hurdles.

The first hurdle checked how well each theory decoded the categories of the objects that the subjects saw in the presented images. Both theories performed well here, but IIT was better at identifying the orientation of objects.

The second hurdle tested the timing of the signals. IIT predicted sustained, synchronous firing in the hot zone for the duration of the conscious state. While the signal was sustained, it did not remain synchronous. GNWT predicted an "ignition" of the workspace followed by a second spike when the stimulus disappeared. Only the initial spike was detected. In the on-screen scoring for the NYU audience, IIT pulled ahead.

The third hurdle concerned overall

connectivity across the brain. GNWT scored better than IIT here, largely because some analyses of the results supported GNWT predictions while the signals across the hot zone were not synchronous.

Both theories were challenged by the results. But in the final tally on screen at the event, IIT scored more green highlights than GNWT, and the audience responded as though a victor had been crowned. Melanie Boly of the University of Wisconsin, Madison, a supporter of IIT, was buoyed enough by the outcome to declare onstage: "The results corroborate IIT's overall claim that posterior cortical areas are sufficient for consciousness, and neither the involvement of [the prefrontal cortex] nor global broadcasting are necessary."

The neuroscientist Stanislas Dehaene of the Collège de France, who developed the global neuronal workspace theory, considers thinking to be a core part of consciousness.

Abaca Press / Alamy Stock Photo

When Dehaene took to the stage, he did not admit defeat either. "I've decided to follow the advice of Dan Kahneman," he quipped. He professed to be happy because "the most interesting part of this experiment was the task-irrelevant stimuli." The question was whether they would indicate the ignition of a conscious percept in the frontal brain. "The answer is yes!" he said.

Later, Dehaene suggested to me that the hurdles for IIT were set lower than those for his theory. "There was no real test of the complex mathematical core of [IIT]," he said. And as Block noted in his remarks that night, the finding that there was support for the

back-of-the-brain theories does not specifically support IIT.

Notwithstanding the slightly higher number of green marks scored by IIT, the project leaders themselves are adamant that there was no winner. "These results confirm some predictions of IIT and GNWT, while substantially challenging both theories," they wrote in a paper describing the results posted on the biorxiv.org preprint server.

Just as Kahneman predicted, the adversaries explained away the discrepancies. Boly argued that the failure to detect sustained synchrony in the hot zone "may be due to sampling limitations." Dehaene suspected that no "off" signal was detected because the subjects allowed their minds to wander. "My claim is that consciousness became decoupled from the

stimulus," he said.

Boly and Dehaene now await the results of the second experiment, involving the Tetris-like game distraction. Those results won't be available until next year.

The Pace of Progress

So has science been advanced? Not everyone thinks so.

Some researchers, such as Olivia Carter, a psychologist at the University of Melbourne and past president of the ASSC, think the two theories were too far apart for their predictions to be meaningfully compared. "My personal feeling is they are testing totally different things," she said. "IIT is focusing on phenomenal content, and GNWT is much more

interested in working memory and attention."

That assessment seems apt. Yet it's also frustrating, given that a dispositive comparison was the stated purpose of the adversarial collaboration in the first place. If it's a victory for science, it seems like a qualified one.

The Monash University philosopher Jakob Hohwy, who is part of another Templeton-funded adversarial collaboration, sees it differently. "This goes to the philosophy of science," he said. He points out the field is still divided over such fundamentals as the definition of consciousness, whether it is closer to thinking or feeling, and even whether self-reported results truly confound the data. For Hohwy, this kind of collaborative effort is the way to move forward. "We will find out as

we go along in exactly this type of adversarial collaboration," he said.

Others, like the computational neuroscientist Megan Peters of the University of California, Irvine, bristled at media coverage that reported the results as a two-horse race between GNWT and IIT rather than a field with multiple contenders. Instead of focusing on winners and losers, Peters said, it's important to see that science advances by learning from each experimental hurdle. (Having attended the proceedings that night, however, I can attest that the event was arranged to resemble a sporting event.)

Still, Peters remains a fan of adversarial collaborations. During the Covid-19 lockdown, she was inspired by the Templeton process to help organize a series of workshops hosted by

the Cognitive Computational Neuroscience conference. In these "generative adversarial collaborations," researchers engaged in robust debate. "Watching the teams chew on stuff was instructive," she said.

The first adversarial collaboration on consciousness may not have succeeded in winnowing out any theories from the field. But it did force theorists to make more tangible predictions, and it made experimentalists work out new techniques. "The findings of the collaboration remain extremely valuable," wrote the University of Sussex neuroscientist Anil Seth in a commentary after the June event. "They will push forward the development of both IIT and [GNWT] — and other theories of consciousness, too — by providing new constraints and new explanatory targets."

For Melloni, the fact that the adversaries have not changed their minds does not detract from the value of the process. "As Kahneman says, people don't change their mind, yet the way they react to the challenges makes their theory progress or degenerate," she said. "If the latter, [then] over time the theory 'dies' and scientists abandon it."

Corrections added August 24, 2023:

A missing paragraph describing the adversarial collaborations arising from the workshop, including one for testing HOTs and first order theories, was restored. Also, some details of the descriptions of the experiments that the adversarial collaboration developed for testing GNWT against IIT were clarified.