**University of Milan**

## Doctoral school of
## Mind, Brain and Reasoning

**Department of Philosophy 'Piero Martinetti'**

**Doctoral Thesis**

# When lightning strikes the brain: Integrated Information Theory and the causal perspective

**Disciplinary-scientific sector BIO/09**

**Renzo Comolatti**

**PhD advisors:** **Marcello Massimini**
**Giulio Tononi**

**PhD Coordinator:** **Francesco Guala**

**Academic year 2024**

*In memoria del nonno Guido (1919-2024)*

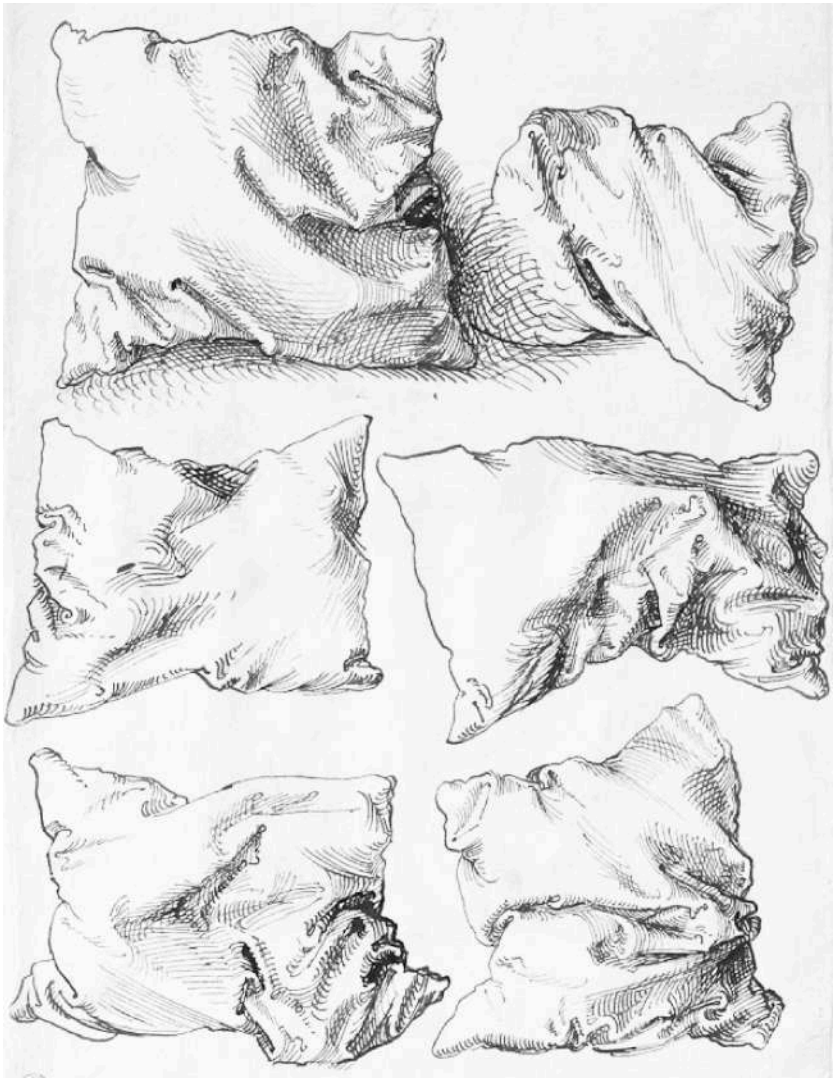
*Land and water and bird or beast, oh*

*Look at what the light did now*

*Shiny little band or golden fleece, oh*

*Look at what the light did now*

Little wings



*Pillow Studies - Albert Dürer (1493)*

# Acknowledgements

*Beautiful friends, my thought is unchangeable*
Sappho

Throughout the PhD, when I felt stuck in a project or disheartened by how much still remained to be done, I often found myself picturing this page. It wasn't so much about what would be written on it, but rather that, in those moments of anxiety or fatigue, with the contents of the dissertation in a state of obscure existence, it gave me some relief to know that at least this page was, in a sense, already there. Now, with the thesis finally (almost) done, I couldn't help but laugh when I found this page still unwritten, yet indeed here.

I'd like to start by thanking my friends from Brazil. Fred, Leo, Guili, Ya, Marcelo—anchors and boat, where I know I can return, to sail afar. Laura, Maria, Lucy, Gabi, Martha and many others—dear companions with whom I've learned so much, and made me into who I am today. Laura and Gabi taught me to treat words with care and inspired me with their writing. My friends from T22—Lucaix, Artur, Caticha, Fê, Estevão and others—who were there when all the tribulation and fun that comes with doing science began. My newest friends in Rio—Victor, Gabriel, Rafa, Luisa, and others—I'll see you all there soon. Gabriel, for the unending conversations, and to all the diagrams we still have to draw. My friends and comrades from STP and Grupão, who, in response to Malabou's question "What should we do with our brains?", I can now venture to say: to organize beyond them. To the folks at STP, for their patience and interest in having me present on IIT during two long meetings in 2020—those discussions kick-started the reflections found in the first chapter.

My heartfelt thanks to my friends and family here in Italy. To my second family from Valtellina: Mario, Fabiana and Marco, per tutto il loro affetto. To Anita, who made me feel at home since I arrived, and to our equivocal exchanges, always in search of translation—I await you for Carnival. To all my lab and conference friends I made along the way. In Milan: Michele, Simoncino, Sasha, Pigo, Simone, Mario, Eze, Silvia, and all others, who made each day there light-hearted, intriguing and uncannily funny; Gianluca, Marta, Gabriel, Letizia, Giulia, Elisabetta, who I'm immensely glad to have met in the last years of the PhD—rest assured, I'll come back to bother you all. In Madison: Tom, Will, Graham, Garrett, Joanna, Leo, Larissa, Billie, and all others whom I've had the chance to befriend and share so many stimulating discussions with; and Matteo, who I've had the joy of having by my side in this adventure through Time—it has been quite a ride. I'd like to express my gratitude to Bjørn, for inviting me to join him on my first visit to Madison—it made a difference. Last, to the fortunate encounters all across: Nao, Sina, Yuko, Manuel Baltieri, Fernando Rosas, Niccolò, João and so many others.

I'm deeply grateful to my supervisors and teachers: with them, I learned the importance of finding the questions worth pursuing, and of posing them in a way that a path towards answers could be discerned. Marcello, who taught me as much about the hard-earned beauty of experiments as about listening for the story they tell us. Giulio, who taught me as much about the compositional richness of consciousness as the sharpness of relentless reasoning. Adenauer, once again, for the generosity and guidance: for inviting me for coffee at Starbucks back in 2014, and two years later for a Master's, while encouraging me not to drop my interest in philosophy along the

way—the road would have been so much longer had we not met. Erik Hoel, for the great tutoring in the midst of the pandemics, as I ventured into the mathematics of a theoretical project for the first time.

Finally, I would like to thank my partner and my family. Julia, for her unwavering support and steadfast curiosity. She lent me eyes to keep looking around me when my head was too buried. Between the happiness of our encounters and the patience of waiting, we've had to invent ways to close the oceanic gap. And here we are: our tenacious partnership. ここ、そこ、そして他の場所. My sister, Greta, for her unbeknownst presence and our silent barter—her drawings also trace the figures and diagrams shown here. My father, with his contagious enthusiasm and his crazy inventions back when I was a kid, gave me the first image of how scientific truths emerge—one that has moved me ever since. My mother, who has shown me it's never too late to study hard and learn anew, or to find our own voice to sing out loud. Paraphrasing Henfil to his mother: you gave me the confidence to live and the security to expose myself. I wasn't afraid to be ridiculous, and I'm not afraid to risk it. Because you loved me.

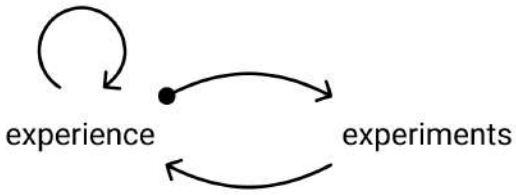
This thesis has been a bridge-building endeavor: between the first- and third-person perspectives, between theoretical practice and experimentation; between different ways to probe the brain and consciousness; a bridge across Madison and Milan, but also a bridge between Brazil and Italy, which I'm now ready to cross back.

# Contents

# Introduction

## I. Experience

*Night kneels over the sleeper*
*Where did his journey begin, where will*
*it burn through to?*
*And what does he swim for now.*
*Swim, sleeper, swim.*
Anne Carson - TV Men: The Sleeper




Sleep is nature's most extensive and longest-running experiment of consciousness[1]. Every night, we each experience our consciousness fading away until it vanishes, only to witness it reemerge in the middle of the night during a dream or the next morning upon waking. If consciousness introduces us to a world of sensations, thoughts, and desires, then dreamless sleep shows us how that world of experience can seamlessly disappear, while the world beyond continues uninterrupted as we slumber. During dreams, instead, we learn that our consciousness too, can persist, visiting unseen worlds, without the world out there following along. As sleep teaches us that the world of experience and the experience of the world[1] can part ways, it naturally prompts us to wonder how they relate and differ; where, in their mutual indifference, they might intersect; or, ultimately, whether experience arises from the world or the world from experience.

In the 17th century, as Descartes meditated on the nature of soul and matter—laying foundation for modern philosophy and erroneously pinning down the pineal gland as the site where they meet, he too would retreat each night to his private bedroom for sleep. Dreams and sleep are mentioned in all but one of Descartes' seven meditations[2], and it is said that he wrote his meditations from his bed, in the solitude of the early morning hours. Long before these distinctions had been intellectual or scientifically posed, sleep was already enacting for us, *in propria persona*, the separation between consciousness and the external world, between consciousness and what lies beyond it—something which our late philosophy and the science of consciousness have come to recognize.

If sleep did not exist, had it not evolved in nature, the scission between consciousness and the world would only rarely take place. Our lives would consist of an uninterrupted stream of consciousness tirelessly anchored to the external world from birth, only severed from it at the end of life:

[1] This fortuitous distinction came about in a conversation with Gabriel Tupinambá.

To die, to sleep—
To sleep—perchance to dream. Ay, there's the rub!
For in that sleep of death what dreams may come,
When we have shuffled off this mortal coil,
Must give us pause

(*Hamlet*, Shakespeare)

So much for the longest-running play of the Cartesian theater. Fortunately, Sleep—the interrupter—is pervasive across life and evolution. Mammals, fish, birds, flies, all find their way to sleep, and it is possible that the very first animals slept as well[3]. Newborns sleep most of the day, elders sleep less so. Sleep may intrude even while we are awake, blanking our mind and causing lapses in our attention[4].

The imperative to sleep, to pause and disconnect, may in fact have arisen as a consequence of living in interaction with our surroundings. Sleep could be the price we pay for our embeddedness in the world; the homeostatic mechanism that allows living organisms to take stock of their experiences while sensing and reacting to the environment[5]. Were we fully disconnected and self-sufficient, the evolution of sleep might not have taken place. Just as the animals on remote but abundant islands, left to themselves, tend to become larger; liberated from wordly constraints, we would be freed—or perhaps condemned—to daydream ever more, isolated in our own islands of awareness[6].

Although universal and systematic, sleep remains a rather private experiment of consciousness. The sleeper experiences firsthand the flickering of consciousness as they drift in and out dreams throughout the night. Yet, those in the room, observing the person lying with closed eyes, can never be certain whether the sleeper is awake or fast asleep, vividly dreaming or simply not present there at all.

In paradoxical insomnia, individuals misperceive their sleep, drastically underestimating the actual amount they get[7]. They believe they have been awake during the night, even though their EEG readings reveal their brain abounded with slow-waves, a hallmark of dreamless sleep and unconsciousness[8], [9]. These insomniacs are not to be confused with "phenomenal zombies", the philosopher of mind's favorite metaphysical creatures, as if the insomniacs were affirming their consciousness *while* physically unconscious. Instead, it reflects a retrospective misjudgment: the feeling of having been awake throughout the night followed by the mistaken cognitive assessment that this must have been the case. Consciousness looking back and perceiving itself there where it was not. *Cogito ergo sum*. I think... therefore I am. At night, the insomniac philosopher is visited by uncanny thoughts: do logical steps take time or space? Is there just enough temporal gap between "feeling" and "judging" for errors to slip in, or *just enough spatial distance* between "thinking" and "being", for doubt to creep in?

Curiously, it is the possibility of mistaking dreams for wakefulness, rather than unconsciousness for consciousness, that has most often haunted—and, by the same measure, fascinated—our philosophical imagination. This possibility has prompted thinkers to pause and ponder whether an ultimate criterion to distinguish dreams from reality can be found:

> Once upon a time, I, Chuang Tzu, dreamt that I was a butterfly, flitting around and enjoying myself. I had no idea I was Chuang Tzu. Then suddenly I woke up and was Chuang Tzu again. But I could not tell, had I been Chuang Tzu dreaming I was a butterfly, or a butterfly dreaming I was now Chuang Tzu?
>
> (Chuang Tzu, 4th century BC)

In the meditations, after establishing its existence with undeniable certainty, the Cogito strives to transcend itself and determine whether its consciousness of the world is as real as the world within itself. This same Cartesian drive to establish the reality of the external world from within consciousness, and subsequently to ground consciousness in that reality, also compels the solitary consciousness to find a common social ground where it can mingle with other consciousness and partake in a shared reality. After all, the fear of *derealization* goes hand in hand with that of *solipsism*. For someone who asks, "Is this real or just a dream?" is not far from doubting, "Am I the only one?" Sooner or later, the problem of the existence of external reality leads to the problem of other minds. Just as Descartes addressed the former, Husserl would later tackle the latter in his own Cartesian Meditations[10].

For the Cartesian cogito, the veridicality of its perception of the world is undermined by the possibility, raised by the exercise of radical doubt, of an evil genius systematically deceiving it. Faced with this threat, consciousness can only find acquiescence by demonstrating, from within itself, the existence of another powerful yet infinitely benevolent entity capable of preempting the actions of such a demon: God.

To Descartes, the Cogito can pass from its immanent solitude to the transcendent and omnipresent company of God for it *already* contains in ideal form that which God embodies in actuality. Our idea of God as absolute perfection and power must entail existence as its predicate—otherwise, it wouldn't be perfect, but incomplete and inconsistent. Similarly, our idea of infinity must originate from an infinite entity, given that we ourselves are finite beings. To one's surprise (and perhaps indignation), our possession of the clear and distinct ideas of God and infinity by our imperfect and finite consciousness, will logically force them into existence.

For Husserl, the existence of other minds must also be established from within the Cogito itself. However, the realm of intersubjectivity, where consciousnesses mutually recognize one another, is achieved not through the idea of God, as in Descartes' philosophy, but through an extraordinary characteristic of our bodies. Specifically, the *sui generis* "perceptual analogy" that obtains between our bodily experience and those of others. Husserl highlights that we experience our body simultaneously as an inside *and* an outside, as the phenomenal pairing of *felt body* and *seen body*. This becomes evident in the example, cherished by many phenomenologists, of looking at our hands touching one another: we feel our hands touching (from the inside) as one and the same with the hands we see moving (from the outside). The perceptual analogy thus consists in that when we see someone else's body moving and appearing like our own, we inevitably perceive it as being inhabited by a somatic inside—an "I"—that is one and the same with it. This idea is further illustrated by how a handshake, in its blunt and fleeting intimacy, perceptually reveals the presence of another person on the other side, acknowledging its existence.

On the verge of absolute skepticism, consciousness finds solace in the company of God and the infinite communication of bodies. Its finite solitude is redeemed by the encounter with infinitary bodies: other worlds

beyond itself. Like an alarm that pierces through a dream, appearing as a ticking bomb to wake us, consciousness transcends itself by discovering that its interior has been stained from the start by an outside.

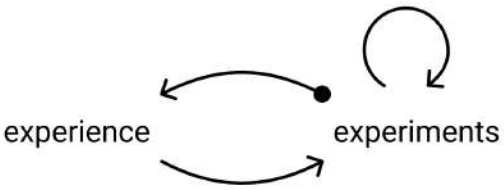
Just as in sleep, where pleasure and terror are entwined in the fabric of dreams and nightmares, the philosophical journey of the Cogito, initiated by Descartes' meditations and continued by the phenomenological tradition, seems poisoned from the outset by a peculiar kind of medicine. A *pharmakon*, which grants the indubitable certainty of one's own subjective existence at the cost of being haunted by ghosts and demons that challenge our confidence in the external world and the existence of other minds. Cartesian certainty is both a source of philosophical relief and a cause of anxiety, a stepping stone and a stumbling block, for which only an act of faith or care can provide resolution. As the sleeping philosopher awakens from his nightmare, it is the persistent familiarity of his room or the comfort of his partner's presence that soothes him back into sleep. The Guarani indigenous people of Brazil customarily share their dreams upon waking and listen to the *paje*'s interpretation as a way to keep the spirits that visited them during the night at bay [11].

We have been suggesting that the tradition of the first-person Cartesian meditation, later taken up by Husserl and many others, can be traced back to the nocturnal and lonely life of the sleeper. This connection is not intended in the sense of a psychological cause or out of neurological indulgence, but rather as what philosophers might call a condition of possibility—one halfway between the transcendental and the empirical. The Cartesian meditations are *more* than a series of thought experiments precisely because they are grounded in real experimental conundrums posed by sleep: the experience of losing and regaining consciousness, of dreaming and waking. In sum, the first-person meditations appear to be realized, dramatized, and modeled after the experience of sleeping.

Sleep—and the dreams it contains—are endogenous physiological perturbations that disturb the otherwise smooth continuity and regularity of our wakeful life. The remarkable occurrence of these natural perturbations breathes life into the distinction between consciousness (being present) and its absence (non-being), between the world of experience (the reality of dreams) and the experience of the world (wakeful reality). It is a source of wonder, as well as fear and surprise. Lying at the root of philosophy, the ancient Greeks called this *thaumázein* (in Portuguese, we call it *espanto*). These distinctions, lived through by consciousness itself, have prompted it to pause, to meditate, and to reflect—alone and within itself—on its nature, on God, and on the world. Yet, one might also be led, as Chuang Tzu or Copernicus once were, to invert the frame of reference and ponder the perspective of the world itself, with its butterflies and mischievous demons. A view from without, perhaps from nowhere, but in company. What would it take to turn an experiment *of consciousness* into an experiment *on consciousness?* To transform an intrinsic perturbation into an extrinsic one, nature into artifice, sleep into wakefulness—eyes open.

## II. Experiments

*Experimentation has many lives of its own*
Ian Hacking




In his 1958 paper "Some Mechanisms of Consciousness Discovered During Electrical Stimulation of the Brain"[12], Wilder Penfield begins by reminding us that it was through Hippocrates' contact with epileptic patients that the ancient Greek physician came to enthrone consciousness in the brain, displacing it from the heart, where it was previously believed to reside: "Men ought to know that from the brain and from the brain alone, arise our pleasures, joys, laughter and jests, as well as our sorrows, pains, griefs and tears. Through it, in particular, we think, see, hear and distinguish the ugly from the beautiful, the bad from the good, the pleasant from the unpleasant."

In that writing, Hippocrates argued for the neurological rather than divine origin of epilepsy, which was regarded at the time as the "sacred disease." He believed that attributing the disease to a divine power, simply due to its "wonderful" and violent manifestations during seizures, was an artifice to mask a lack of understanding and an inability to intervene on it. He proposed that epilepsy was due to a saturation of *phlegma* (a water-based humor) in the brain, which then overflowed to the rest of the body. His hypothesis, based on his quasi-physiological theory of the four bodily fluids, echoes the neuronal hyper-excitation and synchronization now known to occur during epileptic seizures.

Like Hippocrates, Penfield also reported on the whereabouts of consciousness through his work with epileptic patients. During the awake craniotomy surgeries he performed, patients remained fully conscious on the operating table, with their skulls open and brain exposed under local anesthesia. His goal was to delineate the boundaries of the epileptic tissue that needed to be resected, carefully avoiding the neighboring "eloquent" regions responsible for perceptual, cognitive and motor functions.

Like the swimming sleeper in Anne Carson's poem, Penfield needed orientation. As we have observed, the sleeper must traverse the waxing and waning of his consciousness on his own, drifting in and out of dreamscapes, even when someone lies awake right beside him. But to navigate the brain's myriad gyri and sulci, whose patterns are "never twice the same," Penfield, the surgeon, could rely on the company of his awake patient:

> Only thus is the cause of the attack to be found, and the surgeon's hand guided. The patient *talks* and answers the surgeon's questions while he maps out the various functional areas by applying a *gentle* electrical stimulus *here and there* on the cortex.

Not unlike a sailor carefully guiding the helm, every now and then checking the lighthouse's position and reading his compass, the surgeon triangulates between the electrode's position, the brain's visual landmarks, and the patient's report. This navigational endeavor allows Penfield to trace a unique cognitive-experiential map of his patient's brain that guides him throughout the procedure.

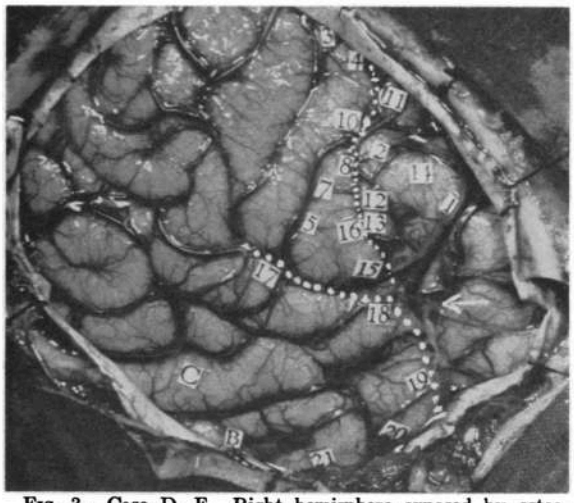


FIG. 3.—Case D. F. Right hemisphere exposed by osteoplastic craniotomy under local anesthesia. Arrow points to cortical abnormality which was more marked on the mesial surface. Numbered tickets indicate points at which electrical stimulation produced positive responses.

Much like Freud with his early patients, Penfield also witnessed an uncanny form of "talking cure." His electrical stimulation of the temporal lobe occasionally induced the "replay" of detailed memories. The report of these *reminiscences* by his patients, led him to propose a tape-recorder model of memory. Adding to the irony, these phenomena have proven difficult to replicate in modern clinical settings[13][2].

Here's Penfield's report on patient D.F., a twenty-six-year-old woman with epilepsy, in whom he performed functional mapping of the right temporal lobe prior to resection surgery:

> When the electrode was applied in gray matter on the cut face of the temporal lobe at point 23, the patient observed: "I hear some music." Fifteen minutes later, the electrode was applied to the same spot again without her knowledge. "I hear music again," she said. "It is like radio." Again and again, then, the electrode tip was applied to this point. Each time, she heard an orchestra playing the same piece of music. It apparently began at the same point and went on from verse to chorus. Seeing the electrical stimulator box, from where she lay under the surgical coverings, she thought it was a gramophone that someone was turning on from time to time. She was asked to describe the music. When the electrode was applied again, she began to hum a tune, and all in the operating room listened in astonished silence. She was obviously humming along with the orchestra at about the tempo that would be expected.

[2] In all seriousness, I believe this parallel could productively be taken far, for example, tracing how these different "talking cures" resulted in significantly different theories of consciousness and of memory (e.g. Penfield's tape-recorder and localizationism versus Freud's mystic writing-pad model of memory[14] and his stratified neuronal theory of consciousness[15]), but this will be left for another occasion.

> Other points were stimulated with no result, except at three points, quite close to 23, where the same song was reproduced.

There is a sort of wonderful suspense at play in the scene described by Penfield. The tension seems to arise from an intricate gap at the heart of it; one that divides space in two and gets redoubled in time. As we read the report, we are invited to alternate between the perspective of the patient and that of the experimenters, following their distinct streams of experiences and beliefs, hers and theirs (split time). These streams, in turn, are updated and interrupted by asynchronous surprises and realizations (split time), creating a sort of spiraling logical development.

As the electrode finds its position at point 23 on the cortical surface and the surgeon activates the current, their attention shifts to the patient's words: "I hear some music." Initially, the patient attributes the music to the stimulator box (correctly, as it turns out) but mistakes it for a gramophone, assuming that everyone in the operating room can hear it, *just as she does*. Hence, no surprise on her part yet. The experimenters, on the other hand, though having heard *what she said*, are uncertain about what her statement refers to. To register as a clinical event, they must first determine whether the effect is indeed neurological, robust, and specific to that stimulated point. Intrigued, they wait fifteen minutes. Silence. The scene starts over, and she hears it again: "It's like a radio." This isn't enough. Clinical rigor demands repeated testing, like an obsessive conductor orchestrating the same piece over and over. The electrode is applied "again and again" to point 23 and nearby points. With each session, the music resumes, "from verse to chorus," ending in the patient's report. Repetition yields knowledge: she realizes there is no gramophone, they learn point 23 plays just like one. Finally, she is asked to describe the music. When the same stimulation is applied again as at the beginning, *something*, however, has changed. The operating room now sits in "astonished silence" as the faint humming from the patient's mouth is the only index of the musical piece being played loud, clear, and perhaps majestically so—simultaneously *there* and *elsewhere*. Alas, the experimenters don't have tickets to the show in this Cartesian Theater.

There is another crucial aspect at work in this scene that deserves closer examination: the asymmetry between the power to induce experiences and the access to those experiences. On the one hand, patient D.F. has privileged access to the otherworldly experiences occurring in the operating room. Yet, without control over the stimulation box and with her brain in its most vulnerable state, she lacks the power to produce or prevent the phenomena she witnesses. On the other hand, the experimenters wield a god-like power to induce these "visions" and alterations (one can only imagine Hippocrates astonishment), all labeled on the cortical surface of the patient's brain. However, they remain largely blind to the effects of the currents they administer, relying entirely on the patient's willingness to report what she perceives as a result of the stimulation. In essence, she witnesses that which she cannot control, while they control that which they cannot witness.

This situation delineates a clear divide between what we can call the *experiential register* and the *experimental register*, personified by patient D.F. and figures like Penfield and the medical staff, respectively. The salient aspect of the experiential register is subjective *appearance*, whereas that of the experimental register is objective *control*. We use "salient" because experiences are also imbued with some level of *introspective control* (e.g., to attend, to judge, to remember), and because experiments can also be viewed as measuring devices capable of *sensing particular types of data* (e.g., electrode position, current intensity, patient reports).

The experiential register is a "first-person" perspective, insofar as what appears does so uniquely to a subject (e.g., patient D.F.). It is a partial perspective, taking place here and now; it is not accessible to everyone, and might not be available elsewhere or at another time. Conversely, the experimental register is a "third-person" perspective, composed of a conjunction of viewpoints (e.g., Penfield and his assistants), characterized by whatever remains consistently invariant across them. Experiments are often plural, as invariance—a condition for objectivity—requires repetition and variation ("again and again"). An experimental outcome exists to the extent that it is publicly verifiable and replicable under various conditions, ideally, by anyone, anywhere and at any time.

In the end of Penfield's account of this scene, what lingers is the sense that a virtuous circle took form between the patient and the experimenters in the operating room. There, where an irreconcilable gap subsisted, we also find the resolution of the scene's tension, with both the patient and the experimenters rejoicing. In fact, Penfield reports receiving a letter from patient D.F. one year after the procedure:

> Today marks a year since you operated on me, and I suppose you are wondering how I am coming along. Now to answer your questions: I heard the song right from the beginning, and you know I could remember much more of it right in the operating room. There were instruments... It was as though it were being played by an orchestra. Definitely, it was not as though I were imagining the tune to myself. I actually heard it. It is not one of my favorite songs, so I don't know why I heard that song. I finally got ahold of a copy of this piece and played it on the piano the other Sunday. Thanks again for better health.

Since Penfield's pioneering work on epilepsy surgery (now known as the Montreal procedure), a wide and heterogeneous array of experiential phenomena has been induced and mapped to different regions of the brain using electrical stimulation during awake craniotomy[16]. These phenomena range from simple visual phosphenes (primary occipital cortex), face hallucinations (fusiform gyrus)[17], and emotions (cingulate cortex)[18], to out-of-body experiences (right angular gyrus)[19] and the complex reminiscences reported by Penfield (medial temporal lobe)[20]. It's fair to say that the compilation of all these maps—infusing brain structures with experiences—wouldn't have been possible without the recurrence of that partnership between patient D.F. and Penfield, an enduring collaboration that knitted experience and experiment together across decades and operating rooms. Without patients speaking, brain surgeons would operate in the dark, strolling across the brain's moist and silent landscape of valleys and ridges—much like Leibniz exploring his imaginary Mill[3].

Nonetheless, the unity between the experiential and experimental registers remains ultimately contingent, relying on a conjunctural arrangement that may not always hold. In fact, Penfield's reporting at times alternates vigorously between the brain coordinates tied to his manual exploration of the cortical surface and the patient's accounts of their experiences:

[3] "If we imagine a machine whose structure makes it think, sense, and have perceptions, we could conceive it enlarged, keeping the same proportions, so that we could enter into it, as one enters a mill. Assuming that, when inspecting its interior, we will find only parts that push one another, and we will never find anything to explain a perception" The Monadology (1714)[21]

> When her left temporal lobe was stimulated anteriorly at point 19, she recounted, "I had a dream, I had a book under my arm. I was talking to a man. The man was trying to reassure me not to worry about the book." At a point 1 cm distant, stimulation at point 20 elicited the response: "Mother is talking to me." Fifteen minutes later, the same point was stimulated again; the patient laughed aloud while the electrode was held in place. After the electrode was withdrawn, she was asked to explain. She replied, "Well, it is kind of a long story, but I will tell you....". After an interval of time, the electrode was applied again, without warning, at point 20. The patient spoke quietly while the electrode was kept in place: "Yes, another experience," she said. "A different experience, a true experience. This man, Mr. Meerburger, he-oh well, he drinks," etc. Stimulation at 23 caused her to hear music."

With no sight of an end or glimpse of an underlying logic relating the points to the experiences or the experiences to each other, we face a vertiginous (rather than virtuous) back-and-forth between experience and experiment. We can imagine Penfield's map-making extending endlessly: the lit operating room advancing into the night, bustling with activity; Penfield maneuvering the electrode over every millimeter of the patient's cortical surface with surgical precision; patient D.F. fully absorbed in reporting any noticeable change in her experience with every electrical stimulus; and his assistants diligently recording each new stimulation point and patient report, eventually filling as many notebooks as there are folds in the brain, much like one of Lewis Carroll's stories:

> "What a useful thing a pocket-map is!" I remarked.
>
> "That's another thing we've learned from your Nation," said Mein Herr, "map-making. But we've carried it much further than you. What do *you* consider the *largest* map that would be really useful?"
>
> "About six inches to the mile."
>
> "Only *six inches*!" exclaimed Mein Herr. "We very soon got to six yards to the mile. Then we tried a *hundred* yards to the mile. And then came the grandest idea of all! We actually made a map of the country, on the scale of a *mile to the mile*!"
>
> "Have you used it much?" I enquired.
>
> "It has never been spread out, yet," said Mein Herr: "the farmers objected: they said it would cover the whole country, and shut out the sunlight! So we now use the country itself, as its own map, and I assure you it does nearly as well."
>
> (*Sylvie and Bruno Concluded,* Lewis Carroll, 1895)

Among Penfield's most emblematic explorations is that of the cortical surface just anterior to the central sulcus, where he would go on to make a striking discovery. There, electrical stimulation caused his patients to report tingling sensations in their bodies (perhaps not unlike the gentle electrical currents delivered to their brains). As he moved the electrode tip to a nearby point, the sensation likewise shifted to an adjacent location of the patient's body. From the toes up to the tongue, Penfield delineated the contours of a human-like creature lying flat on the very surface of the patient's brain—the Homunculus, as he called it[22]. As neuroscience advances and ever more detailed brain maps are created with finer electrodes, there may come a time when

scientists will stop and ponder whether they might be better off using the brain itself as its own map[23]. Leaning over the exposed wet brain while the patient remains fully awake, they might listen in astonished silence to what the Homunculus had to say.

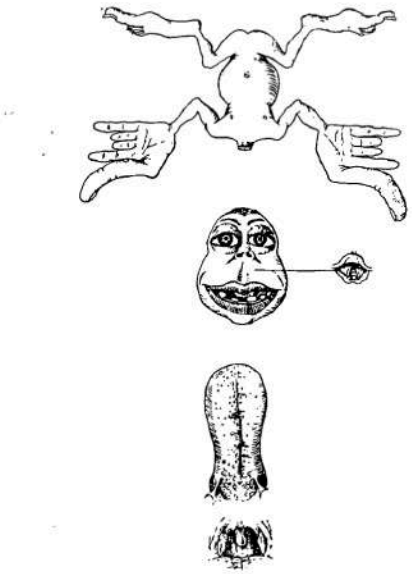


FIG. 28.—Sensory and motor homunculus. This was prepared as a visualization of the order and comparative size of the parts of the body as they appear from above down upon the Rolandic cortex. The larynx represents vocalisation, the pharynx swallowing. The comparatively large size of thumb, lips and tongue indicate that these members occupy comparatively long vertical segments of the Rolandic cortex as shown by measurements in individual cases. Sensation in genitalia and rectum lie above and posterior to the lower extremity but are not figured.


[1] G. Tononi, M. Boly, and C. Cirelli, "Consciousness and sleep," *Neuron*, vol. 0, no. 0, May 2024, doi: 10.1016/j.neuron.2024.04.011.

[2] J. Cottingham, Ed., *René Descartes: Meditations on First Philosophy: With Selections from the Objections and Replies*. Cambridge: Cambridge University Press, 2013. doi: 10.1017/CBO9781139042895.

[3] R. C. Anafi, M. S. Kayser, and D. M. Raizen, "Exploring phylogeny to find the function of sleep," *Nat. Rev. Neurosci.*, vol. 20, no. 2, pp. 109–116, Feb. 2019, doi: 10.1038/s41583-018-0098-9.

[4] T. Andrillon, J. Windt, T. Silk, S. P. A. Drummond, M. A. Bellgrove, and N. Tsuchiya, "Does the Mind Wander When the Brain Takes a Break? Local Sleep in Wakefulness, Attentional Lapses and Mind-Wandering," *Front. Neurosci.*, vol. 13, p. 949, Sep. 2019, doi: 10.3389/fnins.2019.00949.

[5] C. Cirelli and G. Tononi, "The why and how of sleep-dependent synaptic down-selection," *Semin. Cell Dev. Biol.*, vol. 125, pp. 91–100, May 2022, doi: 10.1016/j.semcdb.2021.02.007.

[6] T. Bayne, A. K. Seth, and M. Massimini, "Are There Islands of Awareness?," *Trends Neurosci.*, vol. 43, no. 1, pp. 6–16, Jan. 2020, doi: 10.1016/j.tins.2019.11.003.

[7] A. Castelnovo *et al.*, "The paradox of paradoxical insomnia: A theoretical review towards a unifying evidence-based definition," *Sleep Med. Rev.*, vol. 44, pp. 70–82, Apr. 2019, doi: 10.1016/j.smrv.2018.12.007.

[8] M. Massimini, "The Sleep Slow Oscillation as a Traveling Wave," *J. Neurosci.*, vol. 24, no. 31, pp. 6862–6870, Aug. 2004, doi: 10.1523/JNEUROSCI.1318-04.2004.

[9] G. Tononi and M. Massimini, "Why Does Consciousness Fade in Early Sleep?," *Ann. N. Y. Acad. Sci.*, vol. 1129, no. 1, pp. 330–334, 2008, doi: 10.1196/annals.1417.024.

[10] E. Husserl, *Cartesian Meditations*. Dordrecht: Springer Netherlands, 1977. doi: 10.1007/978-94-009-9997-8.
[11] M. C. N. Hotimsky, "Sonhos compartilhados: dos encontros oníricos às práticas de aconselhamento entre os Guarani Mbya," text, Universidade de São Paulo, 2022. doi: 10.11606/D.8.2022.tde-03052023-191119.
[12] W. Penfield, "SOME MECHANISMS OF CONSCIOUSNESS DISCOVERED DURING ELECTRICAL STIMULATION OF THE BRAIN," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 44, no. 2, pp. 51–66, Feb. 1958.
[13] J. Curot, F.-E. Roux, J.-C. Sol, L. Valton, J. Pariente, and E. J. Barbeau, "Awake Craniotomy and Memory Induction Through Electrical Stimulation: Why Are Penfield's Findings Not Replicated in the Modern Era?," *Neurosurgery*, vol. 87, no. 2, pp. E130–E137, Aug. 2020, doi: 10.1093/neuros/nyz553.
[14] S. Freud, "A note upon the 'mystic writing-pad,'" in *Organization and pathology of thought: Selected sources*, New York, NY, US: Columbia University Press, 1951, pp. 329–337. doi: 10.1037/10584-016.
[15] S. Freud, "Project for a scientific psychology," in *The origins of psycho-analysis: Letters to Wilhelm Fliess, drafts and notes: 1887-1902*, M. Bonaparte, A. Freud, E. Kris, E. Mosbacher, and J. Strachey, Eds., New York, NY, US: Basic Books/Hachette Book Group, 1954, pp. 347–445. doi: 10.1037/11538-013.
[16] K. C. R. Fox *et al.*, "Intrinsic network architecture predicts the effects elicited by intracranial electrical stimulation of the human brain," *Nat. Hum. Behav.*, Jul. 2020, doi: 10.1038/s41562-020-0910-1.
[17] V. Rangarajan *et al.*, "Electrical Stimulation of the Left and Right Human Fusiform Gyrus Causes Different Effects in Conscious Face Perception," *J. Neurosci.*, vol. 34, no. 38, pp. 12828–12836, Sep. 2014, doi: 10.1523/JNEUROSCI.0527-14.2014.
[18] J. Yih, D. E. Beam, K. C. R. Fox, and J. Parvizi, "Intensity of affective experience is modulated by magnitude of intracranial electrical stimulation in human orbitofrontal, cingulate and insular cortices," *Soc. Cogn. Affect. Neurosci.*, vol. 14, no. 4, pp. 339–351, May 2019, doi: 10.1093/scan/nsz015.
[19] O. Blanke, S. Ortigue, T. Landis, and M. Seeck, "Stimulating illusory own-body perceptions," *Nature*, vol. 419, no. 6904, pp. 269–270, Sep. 2002, doi: 10.1038/419269a.
[20] J. Curot *et al.*, "Memory scrutinized through electrical brain stimulation: A review of 80 years of experiential phenomena," *Neurosci. Biobehav. Rev.*, vol. 78, pp. 161–177, Jul. 2017, doi: 10.1016/j.neubiorev.2017.04.018.
[21] G. W. Leibniz, *The Monadology*. CreateSpace Independent Publishing Platform, 2016.
[22] W. Penfield and E. Boldrey, "SOMATIC MOTOR AND SENSORY REPRESENTATION IN THE CEREBRAL CORTEX OF MAN AS STUDIED BY ELECTRICAL STIMULATION," *Brain*, vol. 60, no. 4, pp. 389–443, 1937, doi: 10.1093/brain/60.4.389.
[23] M. Grasso, A. M. Haun, and G. Tononi, "Of maps and grids," *Neurosci. Conscious.*, vol. 2021, no. 2, p. niab022, Dec. 2021, doi: 10.1093/nc/niab022.
[24] L. Carroll, "Sylvie and Bruno Concluded," in *Literature and Philosophy in Nineteenth Century British Culture*, Routledge, 2024.
[25] C. Tsu, *Chuang Tsu: Inner Chapters*. New York: Random House Inc, 1974.
[26] A. Carson, *Glass, Irony and God*, Later Printing Used edition. New York: New Directions, 1995.

# Summary

Experience and experiments offer two lenses to approach the world and consciousness. In the Introduction, we have pursued the seemingly incommensurable, yet deeply imbricated paths offered by these two perspectives: the tensions (and resolutions) that take place between the two every night during sleep and exceptionally, during the awake brain surgery of epileptic patients. In parallel, we have traced how these two perspectives have aligned and diverged in philosophy and in science, personified in the figures of Descartes and Penfield. On one side, first-person experience strives to go beyond itself and touch upon the mind-independent realities of the world, objectivity and science. On the other hand, the third-person perspective offered by experiments, what Thomas Nagel called the "view from nowhere"[1], has long sought to penetrate the elusive reality of consciousness, subjectivity and ideas. Following these two perspectives, the four studies comprising the present thesis are divided into two parts.

The first part, "Experience", consists of two studies based on Integrated Information Theory (IIT), a theory of consciousness that takes phenomenal experience as its starting point, and employs first-person introspection to develop a physical account of consciousness based on causal powers and their relations [2].

The first study, *"The Many Lives of Φ: An introduction to Integrated Information Theory as a research program",* is a broad, and hopefully accessible, primer to IIT, which also aims to offer a new perspective on the theory. Rather than starting with its notorious phenomenology-to-physical approach, it first delineates its underlying explanatory framework, and then presents its core ideas (the axioms and postulates) as unique solutions to general problems faced by theories of consciousness. Additionally, it reviews the different ways and levels at which neuroscientists, theoreticians and philosophers have engaged with it, presenting IIT as an ongoing research program, rather than a monolithic theory. This chapter also sets up a framework on which we can map the other studies in the thesis (see visual summary below).

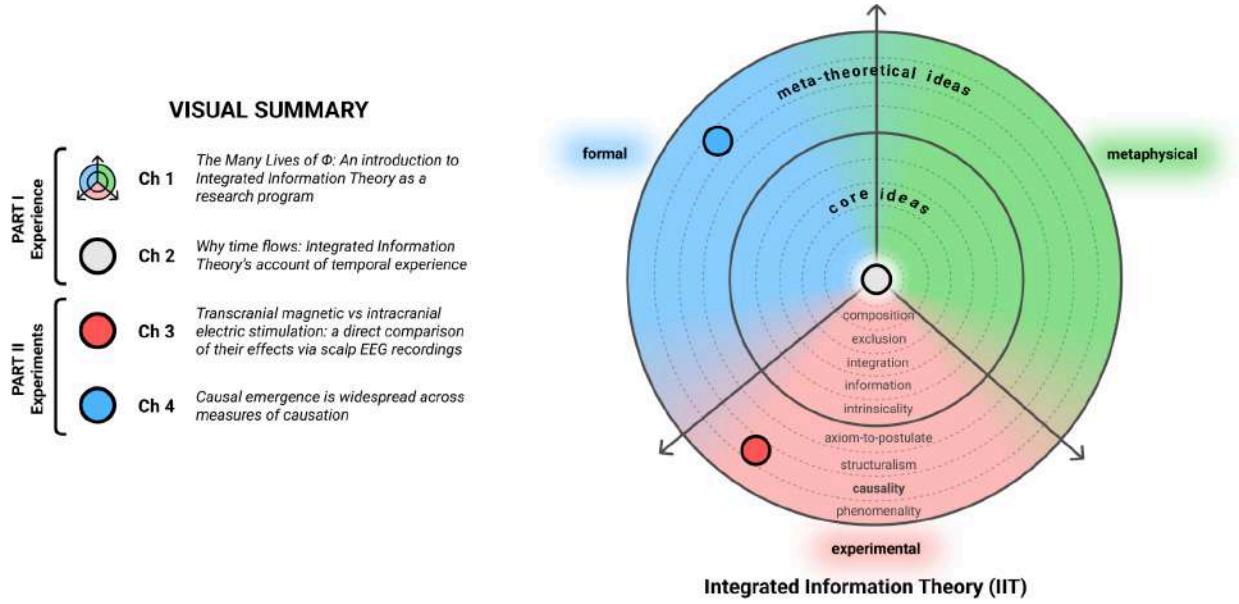
The second study, *"Why time flows: Integrated Information Theory's account of temporal experience",* is an application of IIT, in particular of its idea of composition and Φ-structures, to address the quality of consciousness, i.e. the way experiences feel like. Following previous work proposing that spatial experience corresponds to the unfolded Φ-structure of undirected 2D grids [3], this study shows that the basic properties of temporal experience can be accounted for in terms of the unfolded Φ-structure of directed 1D grids. This result leads to the prediction of directed 1D grids as the neural substrate of phenomenal time, and provides a new, principled approach to understanding the phenomenology of temporal experiences.

The second part of the thesis, "Experiments", comprises two studies, an empirical and a computational one, that employ a perturbational perspective to investigate the emergence (and breakdown) of complexity and causality across a system's different scales and states.

The first study, "*Comparing the effects of transcranial magnetic to intracranial electric stimulation with hd-EEG",* is an experimental work that compares two brain stimulation techniques used to probe the properties of brain circuits, Transcranial Magnetic Stimulation (TMS) and Intracranial Electrical Stimulation (IES). In this study, the similarities and differences between the effects of TMS and IES are for the first time compared using hd-EEG scalp recordings acquired during single-pulse stimulation delivered during wakefulness and NREM sleep. This study falls within a broader effort of aligning perturbational approaches across stimulation methods

and brain scales[4]. Beyond its methodological and neurophysiological value, this work also speaks to the idea that gauging complexity is not only a function of a system's internal make up but also correlative to the method one employs to probe the system.

The second study, *"Causal emergence is widespread across measures of causation"*, is a computational study investigating the notion of causal emergence and surveys different measures of causation in the literature. Contrary to the reductionist view that microscale descriptions contain all there is to a system, the idea of causal emergence, previously formalized based on IIT's framework[5], suggests that macroscales can "beat" the micro in terms of causal power (and information). The study identifies a convergence of in the proposed measures, quantifying causal power as a balance between two "causal primitives": determinism and non-degeneracy (also known as sufficiency and necessity), and demonstrates using simulated toy systems that causal emergence can obtain across all these measures of causation.





[1] T. Nagel, *The View From Nowhere*, Revised ed. edition. New York London: Oxford University Press, 1989.

[2] L. Albantakis *et al.*, "Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms," *PLOS Comput. Biol.*, vol. 19, no. 10, p. e1011465, de out. de 2023, doi: 10.1371/journal.pcbi.1011465.

[3] A. Haun and G. Tononi, "Why Does Space Feel the Way it Does? Towards a Principled Account of Spatial Experience," *Entropy*, vol. 21, no. 12, p. 1160, Nov. 2019, doi: 10.3390/e21121160.

[4] A. Pigorini *et al.*, "Simultaneous invasive and non-invasive recordings in humans: a novel Rosetta stone for deciphering brain activity," *J. Neurosci. Methods*, p. 110160, May 2024, doi: 10.1016/j.jneumeth.2024.110160.

[5] E. P. Hoel, L. Albantakis, and G. Tononi, "Quantifying causal emergence shows that macro can beat micro," *Proc. Natl. Acad. Sci.*, vol. 110, no. 49, pp. 19790–19795, Dec. 2013, doi: 10.1073/pnas.1314922110.

# List of publications, posters & presentations

Main publications

**Comolatti R**, Negro N, Massimini M, Rosas F. *The Many Lives of Φ: An introduction to Integrated Information Theory as a research program* (in preparation)

**Comolatti R**, Hassan G; Colombo M, D'Ambrosio S, Russo S, Casarotto S, Mikulan E, Pigorini A, Massimini M. Transcranial magnetic vs intracranial electric stimulation: a direct comparison of their effects via scalp EEG recordings (in preparation)

**Comolatti R***, Grasso M*, Tononi G. *Why does time feel flowing? Towards a principled account of temporal experience* (in preparation)

**Comolatti R**, Hoel E. *Causal emergence is widespread across measures of causation*. arXiv (2022) https://arxiv.org/abs/2202.01854

Other publications

D'Ambrosio S, Jiménez-Jiménez D, Silvennoinen K, Zagaglia S, Perulli M, Poole J, **Comolatti R**, Fecchio M, Sisodiya SM, Balestrini S. *Physiological symmetry of transcranial magnetic stimulation evoked EEG spectral features.* Human Brain Mapping (2022) https://doi.org/10.1002/hbm.26022

Marinazzo D, Roozendaal JV, Rosas FE, Stella M, **Comolatti R**, Colenbier N, Stramaglia S, Rosseel Y. *An information-theoretic approach to hypergraph psychometrics.* arXiv (2022) https://arxiv.org/abs/2205.01035

Arena A, Juel BE, **Comolatti R**, Thon S, Storm J. *Capacity for consciousness under ketamine anaesthesia is selectively associated with activity in posteromedial cortex in rats*. Neuroscience of Consciousness (2022) https://doi.org/10.1101/2021.01.22.427747

Arena A, **Comolatti R**, Thon S, Casali AG, Storm J. *General anaesthesia disrupts complex cortical dynamics in response to intracranial electrical perturbation in rats.* eNeuro (2021) https://www.doi.org/10.1523/eneuro.0343-20.2021

Posters and presentations

"The Many Lives of Phi: An introduction to Integrated Information Theory as a research program" ASSC27 Conference - Tokyo, Japan (July, 2024) [poster]

"The Many Lives of Phi: Integrated Information Theory as a multidimensional and multilayered framework" C3: Complexity, Computer, and Consciousness – Institute of Physics & Imperial College London. London, UK (Nov, 2023) [presentation]

"Why does time feel flowing?" Special Advanced Course on Consciousness – NSAS 2023. Venice, Italy (Sept, 2023) [presentation]

"Transcranial vs intracranial stimulation: similarities and differences explored by hd-EEG recordings" Mind-Brain-Body Symposium Berlin, Germany. (April, 2023) [poster]

"Transcranial vs intracranial stimulation: similarities and differences explored by hd-EEG recordings" International Brain Stimulation Conference. Lisbon, Portugal (Feb, 2023) [poster]

# Part I - Experience

# *Chapter 1*

# The Many Lives of Phi: An introduction to Integrated Information Theory as a research program[4]


## Abstract

Integrated Information Theory (IIT) stands as an ambitious framework in the field of consciousness studies, integrating novel mathematical, scientific, and philosophical ideas. Since its inception over twenty years ago, IIT has evolved significantly, expanding and refining its concepts. However, its current complexity and scope can make it challenging to attain a good grasp of its concepts and a full view of the theory. This paper offers a fresh perspective on IIT by navigating through its internal structure and mapping the research the theory has driven in various fields. We introduce IIT through a layered approach, starting from its broad explanatory framework before diving into its core ideas. Next, we examine IIT across its formal, experimental, and metaphysical dimensions, highlighting its role as a pivotal research program in consciousness science and its interactions with related disciplines such as complex systems theory, psychophysics, and philosophy of mind. By presenting IIT in this multifaceted manner, we underscore the numerous ways one can engage with the theory, whether critically or constructively, thereby paving the way for clearer and more nuanced discussions in the field. By addressing the common challenges faced by theories of consciousness, we position IIT as both a unique attempt to understand consciousness and a blueprint for future developments in the field.




[4] This chapter consists of an early version of the forthcoming article: **Comolatti, Renzo**; Negro, Niccolò; Massimini, Marcello; Rosas, Fernando E.. *The Many Lives of Phi: An introduction to Integrated Information Theory as a research program.*

# 1. Introduction

Integrated Information Theory (IIT) stands as an ambitious framework in the field of consciousness studies. Since its inception more than twenty years ago, IIT have undergone important developments which have been organized in four different versions: 1.0 (Tononi & Sporns, 2003), 2.0 (Balduzzi & Tononi, 2008), 3.0 (Oizumi et al., 2014) and 4.0 (Albantakis, Barbosa, et al., 2023). Between its earliest prefiguration (Tononi & Edelman, 1998) and its latest 4.0 version, the development of IIT has resulted in the introduction of numerous mathematical, scientific, and philosophical ideas. While enriching, the range of IIT's scope makes it challenging to grasp an overall view of the theory and a clear understanding of the complex relationships between its various aspects. This difficulty has created important misunderstandings between researchers related to interpretations of the theory and the value of its contribution.

To address this issue, this work offers a new perspective on the theory aimed at illuminating its internal structure and explanatory framework. In contrast to most presentations of IIT which depict it as a monolithic entity, here we put an emphasis on the relative independence of different aspects of the theory as well as how different scientists, theoreticians, and philosophers have engaged with it at different levels and perspectives. By doing this, this paper aims to present different ways in which the theory could be helpful for a variety of researchers without the need to embrace all its commitments. Hopefully, this paper may also function as a primer to the theory and a roadmap for those interested in finding an entry point to it.

Built upon this landscape, IIT emerges as an active and ongoing research program just as much as a theory in the process of consolidating a body of core assumptions, formal tools and experimental hypotheses. Our exposition also makes it clear how failures in one dimension or layer may not affect its validity and success in others. Finally, we hope our framework, in which IIT appears as a singular attempt to address general problems, also lays out a blueprint for other ToC, highlighting the common challenges faced by the science of consciousness as a whole.

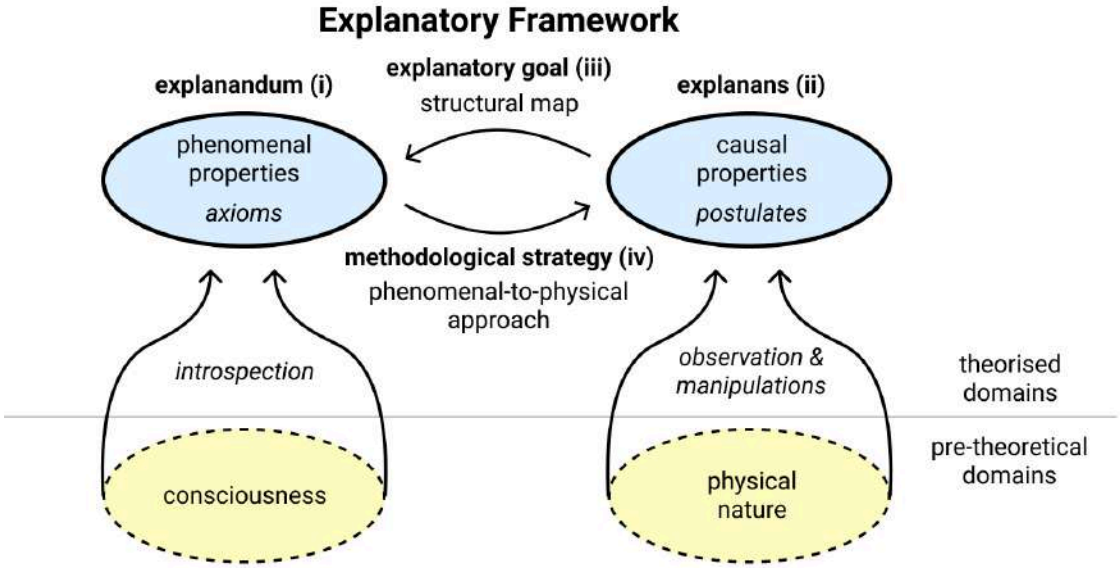
# 2. Layers

We start by constructing a scaffold of the main ideas that make up the theory, which we call **layers**. In each layer, we introduce the basic idea by pointing out the general problem that it is trying to address, and which, we argue, is faced by any theory of consciousness (ToC). Instead of directly starting from the axioms of experience and their translation into physical postulates, as IIT latest papers have done, we begin by first delineating its explanatory framework. Here, we show how IIT addresses meta-theoretical tasks of any ToC such as defining how consciousness is conceived (*explanandum*) and in which naturalistic terms it aims to scientifically account for it (*explanans*). The idea of the axiom-to-postulate approach figures as one of the layers within this explanatory arch. This preliminary work is followed by the core ideas for which the theory is recognized: intrinsicality, information, integration, composition and exclusion.

## 2.1. Meta-theoretical ideas

The **meta-theoretical ideas** address the different aspects of the general **explanatory framework** of a ToC: (i) defining consciousness and the properties that are sought to be accounted for (*explanandum*); (ii) defining

the natural domain and the properties that are taken as the basis of the explanation (*explanans*); (iii) specifying the general form of how the *explanandum* can be accounted for by the resources of the *explanans* (explanatory goal); finally, (iv) define a general strategy to build the theory (methodological strategy). We will see below that IIT address these by: (i) defining consciousness by phenomenal properties derivable from introspection, (ii) positing a physical domain based on causal properties assessed through a perturb-and-measure approach, (iii) explaining consciousness through a structural mapping between the phenomenal and the physical domain, (iv) proposing that consciousness itself can be used to guide and constrain the form of the physical properties needed to account for it.




### 2.1.1. Phenomenality and Introspection: the Explanandum

The basic premise of any ToC is that consciousness deserves to be explained. But what does one mean with consciousness and what properties are sought to be explained? To answer this, IIT begins by pointing to an everyday experience shared by all of us: consciousness is that which goes away when we fall asleep into dreamless sleep, and returns when we dream or wake up in the next morning. That there is a fundamental difference between "being there" (whether it is awake in contact with reality, or in dreams) and complete lack of experience is, for IIT, the key phenomenon that needs to be accounted for. In approaching consciousness through this lens, IIT aims at directly tackling the existence of phenomenal consciousness, of what it is like to have an experience.

After acknowledging that consciousness exists and deserves an explanation, IIT starts by employing first-person **introspection** to identify its essential properties, that is, phenomenal properties that are true of every conceivable experience. It calls these properties *axioms* of phenomenal existence.

This idea of taking phenomenology axiomatically builds on Descartes' idea that our consciousness is the sole thing whose existence we cannot doubt, and that therefore, can serve as the stepping stone for any investigation of its nature. This approach is also in line with the phenomenological tradition pioneered by Husserl, which proposes that consciousness can be captured by studying its internal structure and using introspection to extract its invariant properties (Husserl, 1977). In fact, following this car, IIT posits its "'zeroth" axiom of **phenomenal**

**existence**, which states that experience exists: that there is *something,* rather than nothing. This is followed by five axioms: intrinsicality, information, integration, exclusion and composition. These axioms (and what IIT will call physical postulates) form the very heart of the theory. But since we are currently at the theory's first layer, at its "skin", we will get back to them and examine them one by one later on, once we are equipped with more tools.

The initial foundation of IIT combines phenomenological, epistemological, and ontological aspects: consciousness is phenomenal experience and can be examined via introspection (phenomenological element); first-person introspection can provide immediate and undoubtable knowledge about consciousness (epistemological element); consciousness being the first (and only) thing we can be absolutely sure of, the existence of consciousness should be considered as prior to other types of existence (ontological element). Expressed in a weaker form, these claims lay out the assumption that not only the phenomenal aspect of consciousness should be at the center of any explanation given by a ToC, but that first-person introspection should be used to examine it .

*Objections, alternatives and replies*

Already at this early stage, there are ways one could reject the approach IIT takes to study consciousness. First, one can argue that consciousness does not really exist after all and embrace an "eliminativist" position about consciousness (Irvine et al., 2020; Ramsey, 2022). Eliminativists often claim that the idea that mental states have phenomenal properties, or *qualia*, is a misconceived notion (Dennett, 1988). As such, they think that consciousness is better left out of scientific explanations, which should entirely consist of statements involving conventional functional (Cohen & Dennett, 2011) and neurobiological (Churchland, 1994) properties. However, by opting out of explaining consciousness as such, they are then left with the problem of explaining why consciousness "appears to appear". In other words, eliminativists must face the 'meta-problem of consciousness' (D. Chalmers, 2018) of explaining how the 'illusion' of phenomenal consciousness comes about (Frankish, 2016).

Second, one can argue that even if consciousness exists and should be accounted for, consciousness is not the *feeling* of what it is like to have a certain experience but rather the *knowing* that one has such an experience (Michel, 2019). In this alternative conception, consciousness amounts to the cognitive capacity to access and manipulate mental contents, which includes attending, judging, remembering and reporting such contents. In such accounts, the real explanatory target is thus not phenomenal consciousness, but access consciousness (Block, 1995). This view often embraces a functionalist approach according to which consciousness should be studied according to what it does, and explained in terms of how parts of the brain are able to process and use information to perform functions that are salient to the organism. Scholars and scientists with this inclination will then have to explain whether phenomenal consciousness can be reduced to access consciousness (in the same way as the liquidity of water reduces to its chemical properties), or whether they think that phenomenal consciousness simply does not refer to any phenomenon in reality (in the same way as phlogiston was eliminated by our ontology once chemistry became an established science) thus falling back to the eliminativist position.

Third, one can agree with the priority of phenomenal consciousness, but question the axiomatic approach. Specifically, by pointing to the fallibility of introspection and its limited ability to probe the structure of

consciousness and of its contents (Ellia et al., 2021). In that case, introspection could mistake accidental properties for essential ones (Bayne, 2018), or conversely, take essential properties for accidental ones resulting in an incomplete enumeration of the axioms. For example, IIT leaves out the temporal character of consciousness (Singhal et al., 2022), or metacognitive and higher-order capacities (Brown et al., 2019) which figure as basic properties of consciousness in other ToC. In this case, the properties of phenomenal consciousness may have to be sought experimentally, for instance, by leveraging phenomenological reports paradigms (Petitmengin et al., 2019; Qianchen et al., 2022), until a shared intersubjective agreement is reached. Be that as it may, we can take a hint from modern mathematics, where axioms are often posited regardless of their self-evidence; rather, they are evaluated based on their consistency and the consequences they can yield down the line (Bostock, 2009).

If, however, one accepts that (i) consciousness exists; (ii) that its phenomenal character merits an explanation; and (iii) that IIT's approach can convincingly delineate the explanandum through its axiomatic foundation, then the next step is to see whether IIT employs a successful explanatory framework to account for the explanandum.

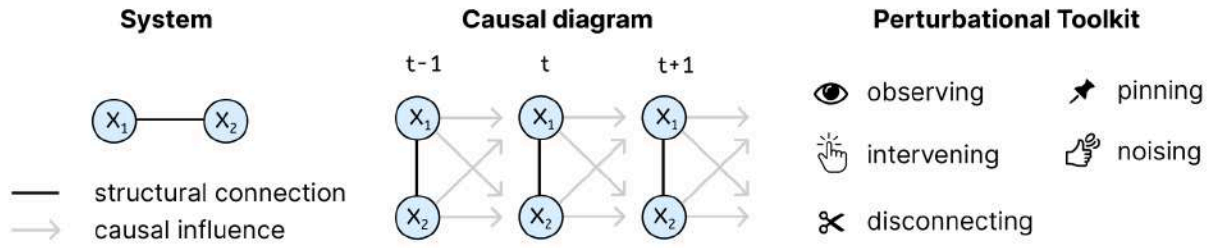
### 2.2.2. Causality and Perturbations: the Explanans

IIT proposes that a scientific explanation of consciousness should be based on a physical account rather than confined to neurobiology. Embracing a physical framework has led IIT to formalize its ideas and develop measures of consciousness that are in principle universally applicable across both biological and non-biological systems. This formalization provides a principled approach to understanding consciousness beyond our own brains, such as the question of machine consciousness (Findlay et al., *forthcoming*). This path also permits the systematic and precise study of its predictions on simulated toy systems, establishing a more grounded approach to developing its framework as well as understanding consciousness across varied domains and architectures (Albantakis, Barbosa, et al., 2023).

Traditional physics often frames its fundamental laws as equations of motion — differential equations that delineate the future evolution of a system given initial conditions. These laws are thought to capture essential regularities of the universe, which determine the unfolding of physical systems. In this view, causality is typically regarded as an auxiliary concept, effective but not fundamental to the overarching framework of physics (with some exceptions, e.g. the many-worlds interpretation of quantum mechanics).

In contrast, IIT places causality at the heart of the physical world. For IIT, the mark of **physical existence** is the capacity of entities to affect and be affected by each other: physical entities are entirely constituted by the causal powers they wield, defined as their ability "to take and make differences". For instance, neuron X wields causal powers with respect to neuron Y insofar as neuron X's ability to fire (and not fire) action potentials *makes a difference to* neuron Y's propensity to fire (and not fire) in the future, and conversely, also *takes differences from* neuron Y's propensity to having fired (and not fired) in the past. Note that producing effects (making differences) is just as important as having causes (taking differences), and that's why IIT prefers to speak about **cause-effect powers** rather than simply causal powers. And in fact, most measures and frameworks of causation are sensible to both the cause and the effect dimensions (normally conceptualized as necessity and sufficiency, or degeneracy and determinism) (Comolatti & Hoel, 2022; Pearl, 2009).

IIT uses the natural language of probabilities to express this ideas formally: the "effect of X on Y" is assessed by looking at the probability of a neuron Y firing or not (at time t+1) conditioned on neuron X having fired (at time t), $P(Y_{t+1} \mid X_t = \textit{fire})$; and the "cause of X by Y" via the probability of Y having fired or not (at time t-1) given that neuron X has fired (at time t), $P(Y_{t-1} \mid X_t = \textit{fire})$.




Correlative to this causal power ontology (Tononi, 2017), is an epistemological view in which knowledge about the physical world is operational, obtained by interacting with it through observation and manipulation. Drawing from interventionist treatments of causality, IIT argues that passive observation of a system dynamics is insufficient to infer its underlying causal relationships; in order to unveil causal powers and obtain causal knowledge one must also be able to actively perturb the system and observe the ensuing outcomes in a controlled and systematic manner (Pearl, 2009; Woodward, 2005).

In IIT, the idea that the causal powers of a physical system are gauged through a **perturbational perspective** has been formalized mathematically by the notion of the Transition Probability Matrix (TPM) of a system. The TPM encodes all the causal constraints of a system, and by extension, all its physical properties and measures in the theory can be computed from it. The TPM is obtained by perturbing the system into all possible state and assessing their transition probability into all other states, represented in a probability matrix $P(S_{t+1} \mid S_t)$.

In neuroscience, most information-theoretic measures used are usually agnostic on how the underlying probability distributions are obtained, and are often estimated from spontaneous recordings of brain activity. IIT's causal approach, instead, prescribes that since consciousness is a causal phenomenon, measures should, in principle, be computed using interventional distributions obtained from causal data.

The perturb-and-measure approach to consciousness put forward by IIT has paved the way to empirical investigations combining brain stimulation techniques (e.g. TMS) with electrophysiological recordings, to directly probe cortical circuits in different states of consciousness, including wakefulness, NREM sleep, anesthesia, as well as disorders of consciousness (Casarotto et al., 2016; Massimini, 2005; Sarasso et al., 2015). This approach has also led to the successful development of a causal measure to index consciousness, based on recordings of brain responses to stimulation – the Perturbational Complexity Index (PCI) (Casali et al., 2013) (further discussed in section 3.2).

*Objections, alternatives and replies*

IIT's physical approach to consciousness could be considered either unwarranted in principle or ad hoc. In the former, it has been proposed that consciousness is fundamentally a biological phenomenon tied to living organisms (Dreyfus, 1992; Edelman et al., 2011), a position known as biological naturalism (Searle, 1992). In the

latter, one may argue that hypotheses directly grounded in neural architecture and mechanisms are more testable and valuable, and worry that formalizing consciousness in physical terms is 'premature' given that our understanding of consciousness is limited already at the neurobiological level. A formalized framework may abstract away from these empirical realities and introduce new practical problems. For example, the computational intractability of IIT's measures to large neural networks poses significant challenges to the theory's practical applicability to real neural systems. However, there is no inherent contradiction between the formalization of a theory of consciousness and the generation of hypotheses about its neural basis. In fact, the relevance of integrated information has originated in the context of understanding the role of the thalamocortical system in dynamically integrating cortical modules for sustaining consciousness (Tononi & Edelman, 1998), and hasn't since refrained from making hypotheses based on neural substrates (see section 3.2 on Neuronal IIT). As argued by (Kanai & Fujisawa, 2023), a comprehensive theory of consciousness should be universally applicable, capable of determining consciousness in any physical system, regardless of their origin or composition. In this sense, IIT's approach can be said to be both empirically testable and universally applicable.

IIT proponents can also draw attention to the limitations of neurobiological and neurocognitive ToCs, formulated based on our knowledge of the structure and functioning of the human and mammalian brains (Dehaene et al., 2011; Edelman, 2001; Graziano & Webb, 2015). Due to their species-specific and brain-centric formulation, these theories can struggle to extrapolate their claims beyond the neural substrates they are modeled after, limiting their generality and interpretability in non-biological substrates.

Importantly, the universality of IIT does not imply that consciousness is substrate-independent. Although IIT's approach is not sensitive to specific composition properties of the substrate (e.g., if it is made of carbon or silicon), it nevertheless posits that consciousness corresponds to causal powers that must be physically implemented, not simply simulated. In this sense, IIT's approach is at odds with computational functionalism, a popular and widely endorsed view in the philosophy and the science of consciousness (Butlin et al., 2023; Cleeremans, 2005; Wiese & Friston, 2021). For IIT, since causal structure, but not computational function, supervenes on physical implementation of systems (e.g. local connectivity of neurons), the decoupling between "software" and "hardware" implied by computationalism, is preempted from the start (Findlay et al. *forthcoming*).

### 2.1.3. Structuralism: the Explanatory goal

While the description of the invariant structures of experience is the central task in the phenomenological tradition and is taken as a stepping stone for IIT, the ultimate goal of the theory is to give a scientific explanation for the properties of consciousness. This goal entails formulating an account that is consistent with the naturalized view of the world developed by science, which encompasses the laws and entities specified by physics, chemistry and biology. However, such a goal faces the challenge whereby phenomenal experience doesn't seem immediately amenable to being cast in the language of science and its disciplines. Experience – at least at first glance – seems to be fundamentally subjective, non-relational and qualitative, while science is commonly regarded to deal with objective, structural and quantitative phenomena. The existence of this gap between the 'intrinsic' character of experiences and the 'extrinsic' nature of scientific explanations (Levine, 1983), has led to

labeling the challenge of naturalizing consciousness – i.e. accounting for phenomenal experience through a scientific explanation – "the hard problem" (D. Chalmers, 2018).

In order to bridge the explanatory gap between the phenomenal and the natural domain, IIT requires that the properties identified in the explanandum domain (e.g. phenomenal properties) be mapped into a naturalized explanatory domain (e.g. physical properties). The underlying idea is that consciousness is identified through structural features, and by mapping each of these features in a domain amenable to scientific investigation, we make those phenomenal features intelligible from the third-person perspective, thus bridging the gap between the subjective and the objective (Ellia et al., 2021). At its center, IIT proposes an **explanatory identity** stating that every experience has a one-to-one correspondence to a cause-effect structure unfolded from its underlying substrate. Because of this, IIT goes beyond the search for neural correlates of consciousness, and attempts to provide a structural explanation of each aspect of consciousness based on how that aspect is translated in the explanatory domain, expressed in terms of cause-effect power (Chis-Ciure & Ellia, 2023).

IIT's proposed account of consciousness is thus formulated across two domains, the phenomenal and the physical. However, for IIT these two domains are not taken to be ontologically on par: the first-person perspective of the phenomenal domain is the domain of what exists indubitably and the domain through which all our knowledge of the world is obtained, and therefore its ontological status is primary and secured. On the other hand, the physical world is purely posited from within the phenomenal one as an explanatory device, and therefore its existence has an *operational* status that can in principle be put into question.

The final step is then to understand how these two domains are related. Given that the physical domain is a purely explanatory posit, the nature of this relationship is not primarily metaphysical, but epistemic: according to IIT, bridging the phenomenal and the physical does not amount to ontologically reducing consciousness to the physical; rather, the physical domain has an instrumental role as a means to deepening our scientific understanding of consciousness (Albantakis, Barbosa, et al., 2023; Grasso, 2018).

Alternatively, one might take a weaker stance that does not require ontological primacy but maintains that phenomenal consciousness must be accounted for in any comprehensive explanation of reality. This perspective allows for a dual-aspect approach where both physical and phenomenal domains are essential and interdependent aspects of the same underlying reality.

*Objections, alternatives and replies*

There are at least two ways to resist this explanatory framework: first, one could reject the structuralist assumption behind IIT's approach by claiming that structurally mapping the phenomenal properties of consciousness onto the physical domain is not sufficient to explain why phenomenal consciousness exists to begin with, or why experiences they feel the way they feel. In this view, IIT's structuralist strategy would not be able to capture the non-relational properties of experience, falling short of addressing the hard problem of consciousness. But it has also been argued that science in general is a structuralist endeavor, and since consciousness doesn't yield to structuralist accounts (D. J. Chalmers, 1997; Russell, 1927), this may entail that consciousness is beyond scientific understanding (E. Hoel, 2023). An alternative to this view, one could build on metaphysical positions such as panpsychism and Russellian monism to propose that consciousness is the intrinsic and non-structural substance underlying the structural entities and laws described by physics (Grasso,

2018; Mørch, 2019b). Here, rather than being "outside", consciousness would figure as nature and science very inscrutable ground.

The second objection to the IIT's strategy accepts the structuralist explanatory framework put forward by it but questions the idea that the bridge between phenomenal and physical domains should be thought in terms of an identity or an isomorphism. Mathematically, such mappings are only two among a broader typology of structure-preserving mappings. Moreover, they are among the strongest forms of morphisms, in which all structure present in one domain perfectly carries over the other domain, and vice-versa. In category theory, for example, identity and isomorphisms exist among weaker forms of mappings that are also structure-preserving (i.e. in decreasing order of how much structure they conserve across domains: equivalences, adjunctions and functors. In empirical settings where experiments are subject to material limitations and noise, precise one-to-one correspondence between the phenomenal and physical domain may be too stringent dependencies to test. Therefore, while the explanatory identity may stand as an ideal for any theory of consciousness, it can be complemented by formulating weaker forms of dependencies that can help guide the empirical testing of the theory (Tsuchiya & Saigo, 2021).

In IIT's view, the end goal of a theory of consciousness is to identify a structure-preserving mapping – ideally, an explanatory identity – between an experiential domain (the explanatory target) and a physical domain (the explanatory "source"): each property of consciousness should find a counterpart property in the explanatory domain, and each relationship between properties of consciousness should be preserved in the relationship between cause-effect powers of the physical.

### 2.1.4. Phenomenal-to-physical approach: the Methodological strategy

With the explanatory scheme in place – the properties of phenomenal consciousness (*explanandum*) are to be accounted for by being structurally mapped to a physical domain defined in causal terms (*explanans*) – a methodological strategy to build a theory and accomplish this objective remains to be specified. IIT's unique methodology proposes a sort of Copernican turn on the science of consciousness: neuroscience has so far attempted to extract consciousness out of knowledge about the brain – probing inside it in search of structures (e.g. claustrum) and mechanisms (e.g. gamma synchrony), or at particular physical phenomena (e.g. quantum effects) that could be attributed to consciousness, obtaining limited success; instead, IIT's wager proposes that we use introspective knowledge about our consciousness to guide and constrain the form that the scientific explanation must take. Phenomenology takes the center of the stage, rather than the brain. Once the basic properties of experience have been identified through introspection (its axioms), IIT translates each of them into the language of the causal powers, yielding the physical properties (its postulates) satisfied by any substrate of consciousness.

Let's take an early IIT formulation to briefly illustrates this axiom-to-postulate methodology: consciousness is both informative (i.e. each experience rules out all other possible experiences) and integrated (i.e. each experience is unified rather than an aggregate of unrelated phenomenal components); it thus follows that a physical system can only be conscious if it is – in some sense – also informative and integrated. A photodiode that distinguishes only light from dark (i.e. 1 bit), cannot specify a highly informative experience of say, multiple

intensities of light like we do (i.e. >> 1 bit), and likewise, a camera, which processes each of its pixel independently, cannot experience an unified visual field of what it photographs like we do.

Now, how should we make sense of the translation of axioms of experience into physical postulates, and what warrants this passage? Rather than being a lawful deduction, this approach utilizes abductive reasoning to bootstrap an explanation from within consciousness and build a bridge to the physical domain (N. Negro, 2023). IIT characterizes this as an "inference to a good explanation", thus better understood as an iterative process of formulating plausible hypotheses and refining them through introspective reasoning and empirical testing.

This process finds an interesting parallel in Kant's transcendental philosophy, which also introduces a 'Copernican turn' by positing that objective reality is always interpreted through subjectivity itself, not the other way (Chis-Ciure, 2022). Relatedly, Varela's neurophenomenology proposes that phenomenological insights and neuroscientific findings should mutually constraint each other, informing and refining the other through this reciprocal relationship (Varela, 1996).

In sum, IIT's methodological strategy posits consciousness as the essential constraint of its own scientific explanation, and uses reason and abduction to determine the shape of such explanation. Importantly, this strategy is iterative, in the sense that "a full explanation [of consciousness] can only be provided through a back-and-forth between the properties of a substrate, which can be explored in great detail, and the properties of experience, which can only be characterized crudely through introspection." (Albantakis, Barbosa, et al., 2023). IIT's methodological strategy to bridge consciousness to its physical substrate consists in using phenomenological axioms to guide and constrain the formulation of physical postulates. This process is abductive, iterative, and enhanced by empirical testing, aiming to translate introspective insights into scientifically testable physical principles.]

*Objections, alternatives and replies*

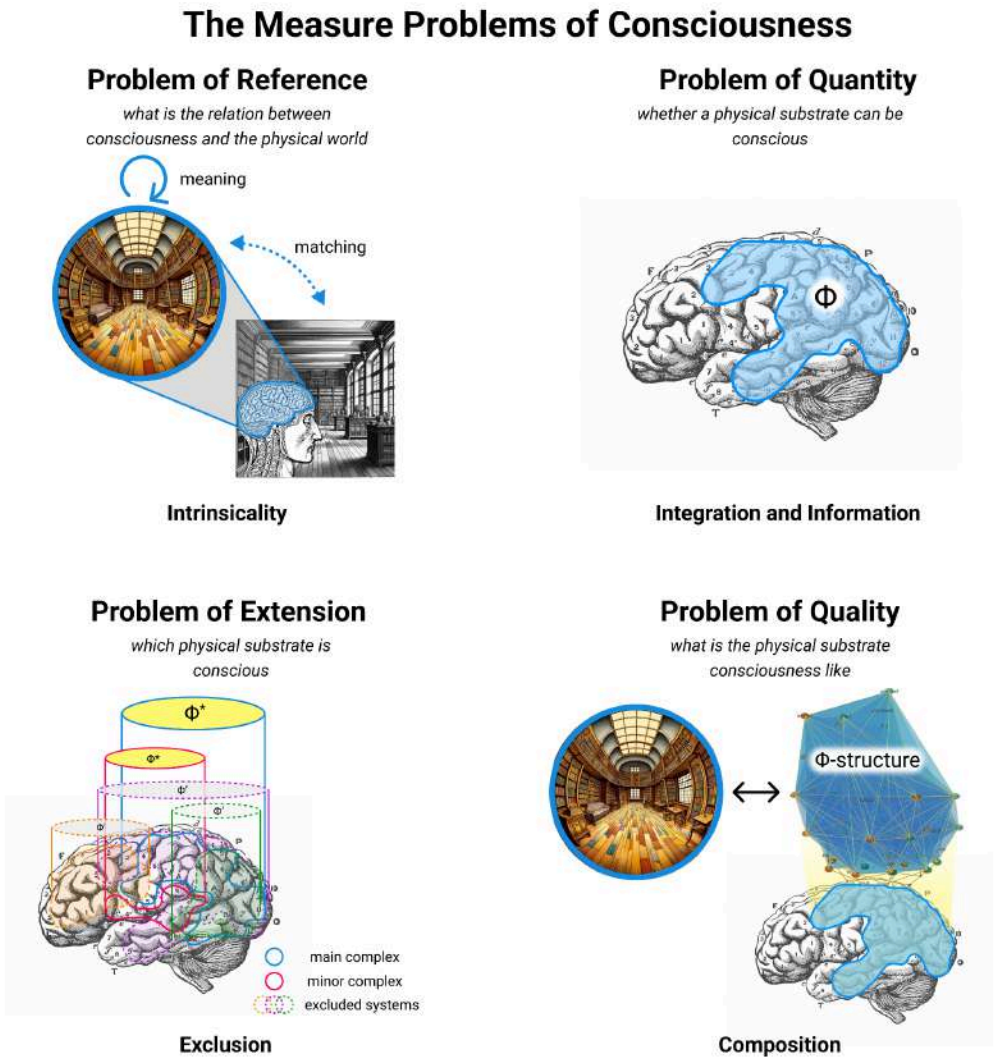
The dominant alternative to IIT's phenomenal-to-physical methodology, which has phenomenological and rationalist underpinnings, is the physical-to-phenomenal approach, which endorses instead an empiricist agenda. This strategy leverages neuroscientific knowledge and experimental work to identify candidate brain structures and mechanisms that are associated with different states (e.g. wakefulness, NREM sleep, dreaming, anesthesia, etc) and contents (e.g. faces, thoughts, etc) of consciousness. This research program based on searching the neural correlates of consciousness (NCC), made explicit in the seminal 1990 paper by Christof Koch and Francis Crick (Crick & Koch, 1990), has underpinned much of the research in consciousness studies over the past decades, driving progress in narrowing down relevant neural structures, mechanisms and markers, even though crucial debates are still ongoing (Koch et al., 2016). Given the empirical successes of the "brain-first" approach, and its alignment with the rest of neuroscience, it might be tempting to favor it over the less orthodox approach endorsed by IIT. However, assuming the NCC program is entirely successful, and provides a full list of brain regions, neuron times and neural events that correlate with consciousness, it is not clear how this would go beyond stating brute facts rather yielding a explanation of why these particular neural correlates are responsible for conscious experience and not others.

If we accept IIT's strategy, two potential challenges face the phenomenal-to-physical methodology. The first issue, which we call the **problem of underdetermination**, points out that since introspection is limited, even if phenomenological axioms constrain the form of the explanatory domain in IIT, there remains a question about whether these constraints are sufficient to fully specify the physical properties. The axioms, while foundational, could potentially be interpreted within various metaphysical frameworks, not limited to IIT's cause-effect power framework. This opens the possibility that multiple formalisms could be compatible with the axioms of consciousness. Recognizing these complexities, IIT in fact posits that the most feasible approach is to infer to the best explanation, using introspection as the guiding tool. Moreover, the dynamics of how this inference works can be made explicit by pointing out methodological assumptions and by listing specific criteria for determining the "goodness" of an explanation, introduced explicitly in IIT 4.0 (Albantakis, Barbosa, et al., 2023). This process is inherently iterative, and can be complemented and constrained by empirical findings, continually refining the alignment between phenomenological axioms and their physical postulate counterparts.

The second challenge is the **problem of chaotic inference**, which raises the possibility that slight changes in the inference of postulates can lead to significant differences in the resulting formalism and in the behavior of the theory's algorithm (e.g. the calculation of cause-effect structure and Φ). Indeed, the mathematical formulation of the physical principles derived from phenomenology involves numerous choices, each shaping the eventual behavior of the theory's algorithm and its application. Addressing this issue could involve systematically exploring how different formulations of the physical postulates and the algorithm relate to one another. For example, generalizing measures into families of related metrics, as demonstrated in (Mediano, Rosas, et al., 2019; Oizumi et al., 2016; Tegmark, 2016), can provide insights into the robustness of the theory. Additionally, making the chain of reasoning behind each inference explicit can help clarify the assumptions and logical steps involved, further stabilizing the theory's application and enhancing its explanatory power.

### 2.2. Core ideas

The **core theoretical ideas** of IIT are normally stated first as an essential phenomenal property of consciousness (axioms) and then translated into a causal property of the physical substrate (postulate). Here, we present these five ideas – intrinsicality, information, integration, exclusion and composition – as IIT's particular "solutions" to what we call the **measure problems of consciousness**. We claim that these are four general problems any ToCs should address in order to be a full theory Namely, answer: (i) what is the relation between consciousness and the external world (problem of Reference); (ii) whether and to what degree can consciousness be present in a physical substrate (problem of Quantity); (iii) what is the experience of a physical substrate of consciousness like (problem of Quality); (iv) what is the physical extension of consciousness (problem of Extension).




### 2.2.1. Intrinsicality: the problem of Reference

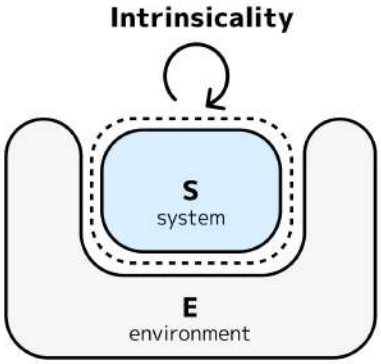
Imagine standing in a museum with a friend, looking at a painting and remarking on the vivid colors used by the artist. In our daily lives, we have an intuitive sense of being in direct contact with the world, perceiving things as they are. Conscious experiences seem not only to be about things in the world (a property known in philosophy as intentionality) but also to derive their shared meaning by referring to these external entities.
However, just as we may question whether the painting gets its meaning from what it depicts, one might question if this picture of the relationship between consciousness and the external world is adequate. We call this the problem of **reference**: where does consciousness get its meaning, and how are experiences about the world?

Neuroscience often employs the notion of mental representations to describe brain states that track features of stimuli in the world (Baker et al., 2022). However, according to IIT, this perspective implicitly assumes an external observer who can map the brain's features to those of the world (Grasso et al., 2021). IIT challenges this "extrinsic" view, asserting that consciousness is intrinsic and, therefore, its meaning must also be accounted for in intrinsic terms.

Intrinsicality asserts that consciousness exists **for itself**, making experience inherently subjective and first-person, rather than dependent on an external world or observer. That consciousness is subjective and directly accessible only to the individual having the experience doesn't entail that it cannot be accounted for in objective terms or that experiences cannot be intersubjectively validated (Ellia et al., 2021). Rather, it means that any attempt to understand consciousness in physical terms must begin from the **intrinsic perspective** of the physical system, not from the perspective of an external observer studying it.

In neuroscience, IIT's view aligns with recent critiques of notions such as neural codes or neural representations for not having inherent meaning to the neural system and instead being observer-dependent, presupposing representations and codes carry information by reference to things known to the scientist investigating the brain (Brette, 2019).

From the intrinsic perspective, the causal properties of the system must be observer-independent and cannot be arbitrary. One of the many mathematical ramifications of this idea is that causal constraints aren't estimated based on empirical probability distributions (those obtained by observing a system's dynamical evolution), since these are biased by arbitrary factors not dependent on the system (e.g., initial conditions and external conditions). Instead, to probe the causal constraints of a system in an agnostic and "fair" manner, IIT employs a uniform distribution on possible states, encoded in the system's TPM.




**Causal Marginalization**

$$\mathbf{P}(S_{t+1} \mid S_t) = \sum_e \mathbf{P}(E_t = e)\, \mathbf{P}(S_{t+1} \mid S_t, E_t = e)$$

Intrinsicality also means consciousness exists **within itself** rather than in relation to what is outside of it. Dreaming, as well as ketamine anesthesia and locked-in syndrome, provide empirical support for this view, showing that consciousness can persist independently of its connectedness to the external world, supporting the view that its contents are internally generated rather than necessarily tied to its surroundings. Physically, this entails that the causal powers of a system are evaluated within the system, according to what 'makes and takes a difference' within itself, rather than by being connected to or interacting with the environment. Mathematically, this states that all variables outside of the system (X = S\U) should not contribute directly to the system's causal properties and hence are treated as fixed background conditions via a procedure called causal marginalization.

In the philosophical debate, this aligns with an internalist view of conscious experience in which all its content and meaning are ultimately self-produced and not directly dependent on the external environment (even if its internal constitution has been shaped by it through learning and evolutionary history). This contrasts

with externalist views about the material basis of consciousness (Clark & Chalmers, 1998) and with views that claim extrinsic relationships with worldly objects are constitutive of mental content (Dretske, 1996; Lycan, 2001).

In neuroscience, embodied cognition approaches argue that consciousness is tied to the environment and bodily interactions, while some interpretations of predictive coding emphasize the brain's role in making predictions about environmental states, suggesting a strong link between consciousness and environmental interactions (Kirchhoff & Kiverstein, 2019).

### 2.3.2. Information: the problem of Quantity (I)

A central goal of a ToC is to determine whether, and to what degree, a given system has the capacity for consciousness. In fact, most ToC are concerned with explaining the presence (or absence) of consciousness in different brain states or why a presented stimulus is consciously perceived or not. ToC should be able to measure consciousness by ascertaining its **quantity**, whether by giving a binary answer ('yes' or 'no') or a real valued number. As one may already guess, IIT addresses this problem through its Φ measure, which is quantified through the joint assessment of two properties: information (I) and integration (O) together yields Φ . Let's begin with the first idea.

Phenomenally, information states that consciousness is **specific** – experience is the way it is, rather than other ways it could be. Thus, each experience is what it is by differing, in its particular way, from other possible experiences. Translating this idea to physical terms amounts to quantifying the information in a system. Information is often conceived in terms of reduction of uncertainty: it is a function of how much knowing the state of a variable rules out and constrains the probability of other states, in proportion to the size of the state space. For example, in Shannon's information theory, knowing that a rolled dice is even, rules out 3 out of 6 possible 'states' yields one bit of information while learning the exact face is 3 rules out 5 states out of 6, yielding more information (~2.56 bits).

However, this classical notion of information must be developed in order to be in accordance with IIT's ideas of causality (**2.2.2**) and intrinsicality (**3.2.1**). In order to do so, IIT gives a spin to Gregory Bateson's definition of information as "difference that makes a difference" (Bateson, 2000) reformulating it as the difference a system takes and makes to itself by constraining its state space.

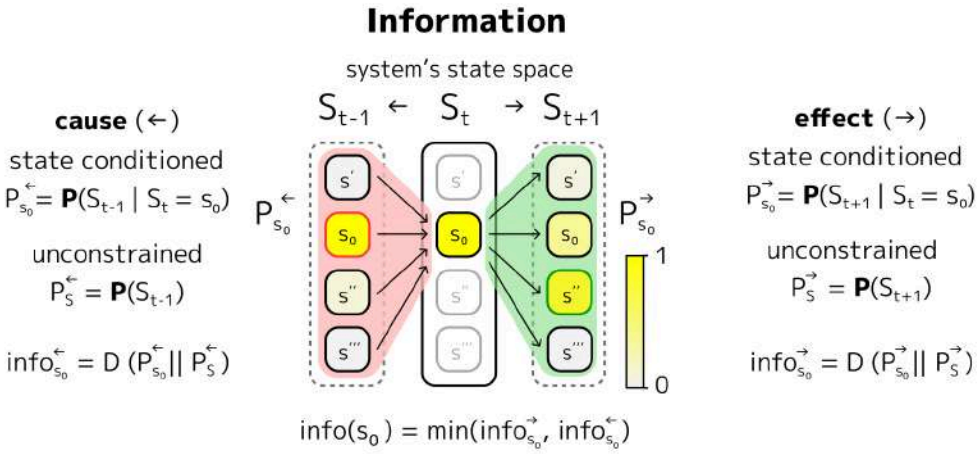
The idea that consciousness is specific is also related to the phenomenological notion that consciousness exists *here and now*, as an actual experience one is currently having, rather than as a general capacity for experiences. In physical terms, this suggests that information should be computed relative to the specific state the system is in at any given moment ($t$). In other words, in IIT information is a state-dependent quantity, measuring how a system in a current state imposes counterfactual constraints upon its states – both cause states towards its past ($t-1$) and effect states towards its future ($t+1$). In contrast, conventional quantities in information theory are not state-dependent but capacity-like quantities, i.e. computed as an average over the system's states (e.g. entropy based metrics).

In general, IIT information (I) measures – from early effective information (Tononi & Sporns, 2003) to the latest intrinsic information metric (Albantakis, Barbosa, et al., 2023) – can be framed as a distance between an

interventional distribution, encoding the causal constraints specified by the system in a state, and an unconstrained distribution, encoding the general causal constraints of the system irrespective of the state it is in.

$$info = D[P_{cs}(S \mid S = s) \,||\, P_{uncs}(S)]$$

In IIT's latest formulation, this distance gauges how a specific state constrains the state space counterfactually, by being *selective* (concentrating probability over certain states) and *informative* (deviating from chance).



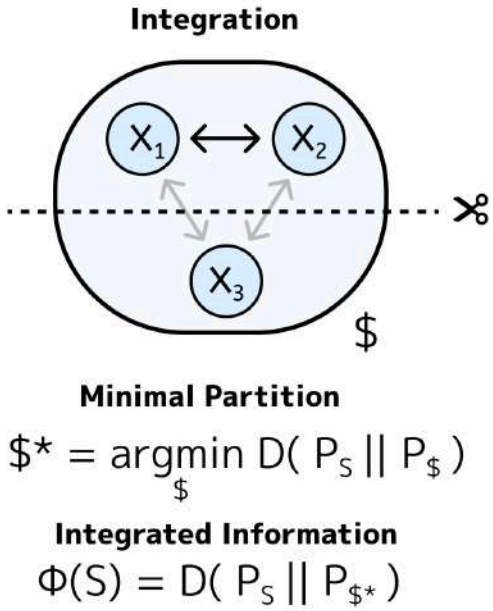

The formulation of information in causal and intrinsic terms via state-dependent counterfactuals may perhaps seem a mathematical nuance of the theory, but in reality it leads to surprising experimental predictions that challenge the commonly held notion in neuroscience that consciousness is determined by brain activity (Gidon et al., 2022). Since inactive (but not inactivated) neurons can, in principle, constrain the probability of other neurons firings just as active neurons do (e.g. inactive inhibitory neurons will likely increase the probability that synaptically connected neurons fire), IIT predicts that silent neuronal activity may directly contribute to the content of experiences.

### 2.3.3. Integration: the problem of Quantity (II)

Continuing on its path to quantify consciousness, IIT claims that information alone is not sufficient to gauge the capacity for consciousness, and requires that the information in a physical substrate is integrated, that is, irreducible to the individual parts of the system.

Phenomenally, integration states that consciousness is **unitary** – every experience is a unified *whole*, not reducible to separate phenomenal components (e.g. the left side of the visual field, from the right side; visual experiences from auditory experiences). In physical terms, integration says that the information in a system must be irreducible, meaning the system cannot be decomposed without losing information. Thus, to assess irreducibility the system is cut across its causal connections and the information lost in the process measured by taking the intrinsic distance between the (interventional) probability distribution of the intact system and the partitioned system. Among all the ways to partition a system, the partition that makes the *least* difference – the Minimum Partition (MP) – is taken as the one that measures integrated information (Φ). The principle behind this is that if there is no way to cut a system that does not lead to information being lost, then it must be

irreducible. Moreover, it is integrated to the extent of the information lost across the cut that makes the least of a difference (i.e. MP).




Experimentally, integrated information has been operationalized in the notion of neural complexity, defined as the joint presence of functional integration and differentiation in the brain. In the last decades, a remarkable amount of studies have converged on complexity-based measures as reliable markers of consciousness, using independently proposed measures across different methods and techniques (Sarasso et al., 2021) (see section 3.2).

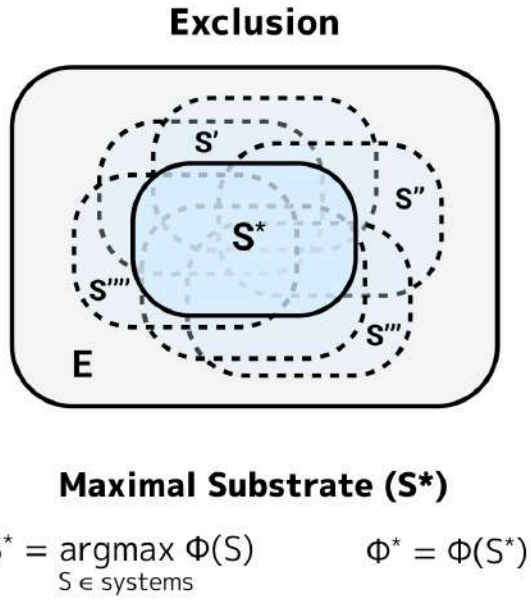
### 2.3.4. Exclusion: the problem of Extension

Integration and information together assess the degree to which a system can be conscious – the problem of quantity. Exclusion attempts to tackle instead what we call the problem of **extension**: given an experience, a ToC should be able to identify the borders of its physical substrate, across any set of (possibly overlapping) candidate substrates. Conversely, given a physical universe a ToC should be able to identify *which* subset of elements in this universe form a physical system specifying an experience, and which subsets do not.

Exclusion asserts that consciousness is **definite** – it is *this* whole experience, containing what it contains, not more, not less. In other words, consciousness has a "border" specifying a definite extension of its contents. For instance, the experience of enjoying your favorite piece of music excludes your experiencing less – listening to the music *without* any feeling of pleasure, and excludes your experiencing more – enjoying the music *and* also feeling an excruciating pain.

In physical terms, exclusion states that the substrate of consciousness is also definite: it is composed by a particular set of units (e.g. pyramidal neurons in the posterior hotzone) specifying a particular set of causes and effects, defined at a particular spatial (e.g. cortical columns) and temporal grain (e.g. 40 ms). Mathematically, exclusion proposes a **maximization principle** to identify these contours: across overlapping candidates – say, a

system and all its subsets and supersets – the substrate of consciousness is the ones that maximizes the quantity of consciousness, i.e. Φ, whereas all non-maximal candidates are excluded.




It is important to notice that exclusion realizes at the *theoretical* and algorithmic level, what the search neural correlates of consciousness attempts to do *empirically*. Applied to neural substrates, exclusion's aim coincides with the very definition of that of the NCC: identifying "the minimum neural mechanisms jointly sufficient for any one specific conscious experience" (D. J. Chalmers, 2000; Koch et al., 2016). Indeed, the scientific validation of IIT – and ToC in general – consists in the alignment of the empirical NCCs with the application of its theoretical measures.

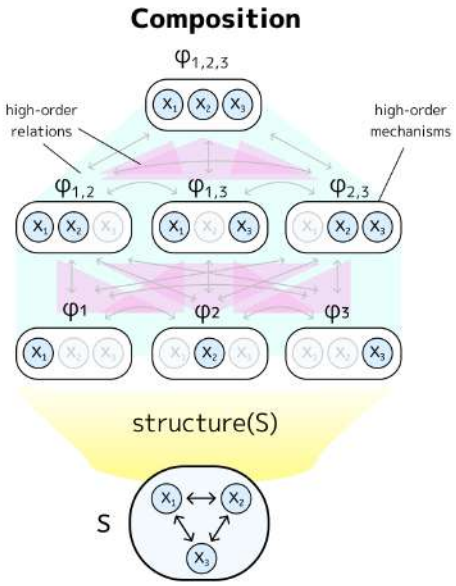
In metaphysics, the problem of the extension of consciousness at stake here parallels Unger's "problem of the many" (Unger, 1980, 2004), which poses the issue of determining which among many overlapping entities should be considered the 'true' existing entity (i.e., determining which specific collection of water droplets constitutes a cloud is akin to identifying which collection of neurons constitute a conscious experience). Proposed solutions include nihilism (denying the existence of all such entities), overpopulation (accepting multiple overlapping entities), and the principle of selection (Unger, 1980; Weatherson, 2023). The maximization principle in IIT is an example of a principle of selection, identifying which physical (or neural) substrates count as specifying a conscious experience. This principle might be questionable, but ToC that rejects IIT's exclusion postulate must still address the problem of identifying definite borders of consciousness amidst sets of overlapping candidate substrates, which is no other than giving a theoretical account of what the NCC offers in empirical terms.

### 2.3.5. Composition: the problem of Quality

Theories of consciousness have so far focused on proposing conditions for its presence or absence, or for when a presented stimulus is perceived or not. Indeed, so far we have discussed integrated information as a scalar quantity and its relation to indexing the possibility of conscious experience being present or not. But a

comprehensive theory must also account for the **quality** of consciousness, and explain why a particular experience feels the way it does, rather than some other way. Enter composition, which moves us from Φ quantification of system as integrated wholes, to Φ-structures which are set to fully characterize a system's causal structure, in order to address how consciousness in a system "feels" like in virtue of how it is structured – the problem of quality.

Composition affirms that consciousness is **structured** – that every experience is composed of phenomenal distinctions binded by relations – and therefore, that the substrate of consciousness must equally specify a structure composed of causal distinctions binded by relations. Hence, composition simultaneously poses the problem of the quality of consciousness and addresses it by stating that this quality is structured rather than ineffably simple. In fact, the structuralist approach (discussed in section 2.1.4) presupposed this insight.




The main consequence of composition is that the steps used to compute the integrated information of a system are also applied to each of its parts, called mechanisms, and to the relation between them. In IIT, mechanisms and the causal distinctions they specify aren't just the individual "atomic" parts that make up a system (e.g., single units $X_1$, $X_2$, $X_3$), but also include all higher-order parts (e.g., pairs or groups of units, such as $X_1X_2$, $X_2X_3$, $X_1X_3$ and $X_1X_2X_3$). Likewise, relations are not restricted to being pairwise, but include all the ways first-order and high-order distinctions can relate, composing a Φ-structure – also called **cause-effect structure** – akin to high-dimensional hypergraphs.

The introduction of a compositional framework from IIT 2.0 onwards has given the theory the conceptual and mathematical resources to move beyond Φ and Φ-inspired unidimensional measures like PCI to index levels of consciousness, to using Φ-structures to tackle the contents of consciousness (e.g. space, time, objects, colors, etc). Specifically, using simulations, Φ-structures unfolded from specific architectures, namely undirected and directed grids, have been proposed as accounting for spatial and temporal experience, respectively. Moreover, initial attempts to estimate Φ-structures approximations on real neural data have shown promising results (A. M. Haun, Oizumi, et al., 2017; Leung et al., 2021; Muñoz et al., 2020).

Additionally, there has been a rise in the development of multidimensional information-theoretic metrics designed to quantify higher-order dependencies in data. Recent developments have expanded the potential of IIT's compositional framework. A notable example is the IIT-inspired Integrated Information Decomposition (ΦID) framework (Mediano, Rosas, et al., 2019), which refines the analysis of higher-order information in a system by distinguishing how different "atoms" compose into an informational structure. These advancements may suggest a compositional turn in the study of complex systems paralleled by a similar turn in treating conscious experience in structural terms (Kleiner, 2024).

| ToC | IIT |
|---|---|
| **explanatory framework** | meta-theoretical ideas |
| ***explanandum*** | phenomenality and introspection |
| ***explanans*** | causality and perturbations |
| **explanatory goal** | structuralism |
| **methodological strategy** | phenomenal-to-physical approach |
| **problems of consciousness** | core ideas |
| **reference** | intrinsicality |
| **quantity** | information |
| **quantity** | integration |
| **extension** | exclusion |
| **quality** | composition |

## 3. Dimensions

With IIT's scaffold of ideas standing, we now take a few steps back to look at it from a distance, and map the diverse research ecosystem it is embedded in, which is composed of interactions with several fields. In this context, we propose that IIT – but possibly theories of consciousness in general – can be fruitfully examined along three distinct **dimensions**: formal, experimental and metaphysical.

Each dimension is irreducible and interdependent to one another, but carves out particular objects and grammars, evaluated according to their own normative criteria, building on distinct disciplines. Importantly, these three viewpoints are taken to be irreducible but interdependent to one another, such that a theory's overall consistency and significance requires these three perspectives to be taken together.

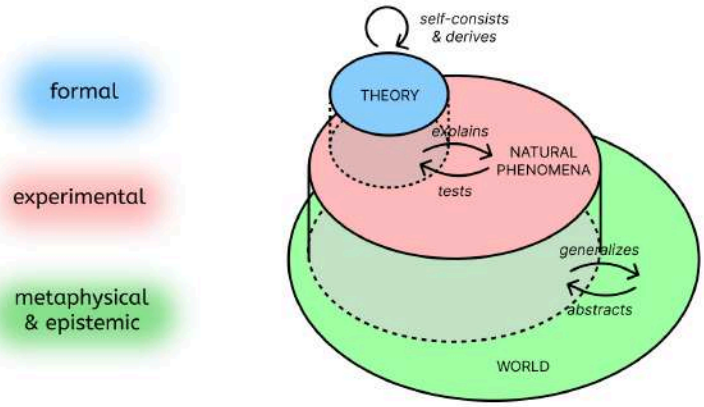
From the **formal** standpoint, a theory consists in its mathematical apparatus, encompassing a proposed set of quantities and algorithms, articulated through precise definitions and often theorems, and evaluated according to criteria inherent to mathematical practice. This includes the consistency and elegance of the framework proposed, as well as its ability to forge novel connections and extend established mathematical

results. This perspective considers the self-consistency of the theory, i.e. how the theory relates to itself taken as a formal system.

Seen from an **experimental** perspective, a theory is defined by its dual relationship to experiments, specifying the way it constrains and is constrained by them. On one side, a theory serves to interpret experimental data and account for observed phenomena. As such, the explanatory power of a theory is attested by its ability to effectively render phenomena intelligible and empirical findings interpretable, as well as pave the way for new experimental paradigms and empirical metrics. Theories must, on the other side, be sensitive to the outcome of experiments, which may support or challenge its hypotheses. A theory's testability hinges on its capacity to fail or succeed in predicting and accounting for the experimental outcomes. The interplay between a theory's predictions and explanations, and experimental phenomena is regulated by the standards of scientific practice. In sum, this perspective deals with the theory's relation with empirical phenomena made visible and intelligible through experimental methods.

Last, we can ask what entities – according to the theory – exist and how we can come to know them, thereby considering it from a **metaphysical** standpoint, comprising both *ontological* and an *epistemological* aspects. This perspective takes the theory as a conceptual system and entertains what are the metaphysical and methodological presuppositions of the theory (e.g. consciousness exist and we can know it through introspection), but also what are its metaphysical and epistemological implications (e.g. questions about ethics, free will, etc). The discipline that deals with these issues is philosophy through the exercise of reasoning. This dimension thus concerns the theory's relation to the world in general, and the way it constitutes a coherent "worldview" of what can be said to exist and of what can be known.

Distinct views of IIT emerge once the theory is 'projected' along each of the three dimensions: theoretical, scientific, and worldview IIT. **Theoretical IIT** is obtained by filtering it along the formal dimension, focusing on its mathematical structure, which emphasizes the development and refinement of the physical postulates, measures and algorithms. **Neuronal IIT**, filtered along the experimental dimension, concerns with developing and empirically testing hypotheses tied neural structures and mechanisms as well as measures that can be applied to real data. **Worldview IIT**, obtained as a projection along the metaphysical dimension, encompasses the broader philosophical implications of the theory informed by the theory's phenomenological orientation combined with its causal framework. Together, these dimensions form a comprehensive approach to understanding consciousness, ensuring that IIT remains both scientific, rigorous and sound.




| | formal | experimental | metaphysical |
|---|---|---|---|
| **discipline** | math | science | philosophy |
| **referent** | itself | natural phenomena | world in general |
| **criteria** | self-consistency | measurement | conceptual reasoning |
| **IIT** | theoretical IIT | neuronal IIT | worldview IIT |

### 3.1. Formal: Theoretical IIT

From a theoretical standpoint, IIT is distinguished by its iterative and progressive effort to formalize its conceptual framework. Alongside IIT, a few accounts of consciousness have been proposed within a mathematical setting (Chang et al., 2020). A recent trend in the development of formal frameworks and methodologies for studying consciousness can be identified (Kleiner, 2024; Tsuchiya & Saigo, 2021), also witnessed by the founding of the Association for Mathematical Consciousness Science (AMCS) in 2018. In contrast, other prominent theories of consciousness – for example, Global Workspace Theory and Higher-order theories – lack an explicit mathematical articulation of its concepts.

IIT comprehensive mathematical apparatus includes a series of quantities and algorithmic procedures, matured through the different iterations of the theory, aimed at capturing measuring consciousness in all its dimensions – quantitative, qualitative and identifying its extension –, now epitomized in the calculation of Φ, cause-effect structures and the search of the main complex.

However, since its first mathematical operationalization, IIT's measure of integrated information has had a significant impact beyond consciousness studies, extending into fields such as systems theory, physics, and mathematics. Integrated information has been reinterpreted in various theoretical settings, including information geometry (Amari et al., 2017), category theory (Tull & Kleiner, 2020), homotopy theory (Manin & Marcolli, 2022), fundamental physics (Albantakis, Prentner, et al., 2023; Barrett, 2014; Tegmark, 2015), and alternative information-theoretic settings (Seth et al., 2011a). It has also been applied to studying phase transitions and criticality in statistical physics and dynamical system theory (Aguilera, 2019; Mediano, Rosas, Farah, et al., 2022; Popiel et al., 2020), and even political theory (Apolito, 2020). IIT formalism has also been

adapted to develop original accounts of actual causation (Albantakis et al., 2019) and emergence (E. P. Hoel et al., 2013; Rosas et al., 2020).

*Refinement and stability of key notions, despite increase in theory's complexity*

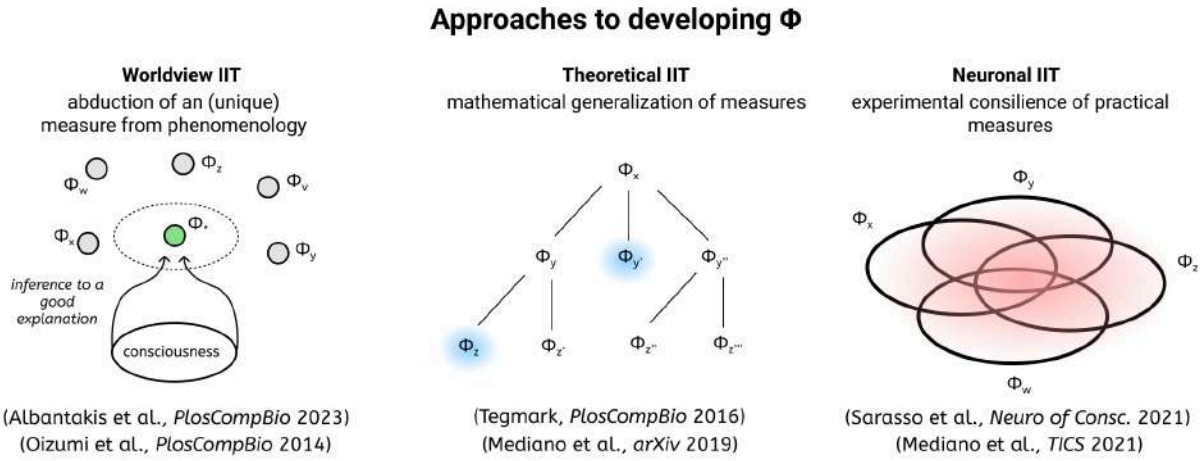
The introduction of new properties, such as composition and exclusion, beyond the founding ones of integration and information, together with the constraint that all properties be jointly satisfied, has led to a significant increase in the complexity of the theories' algorithms. The introduction of composition to the theory, needed in order to account for the structural quality of experiences, led to the initial scalar nature of Φ becoming multidimensional, i.e. a Φ-structure composed of causal distinctions and relations, which entailed computing the integrated information of both the whole system and each of its parts and relations. Likewise, the postulate of exclusion, required to identify the exact contours of the physical substrate of consciousness (and of each of its components), led to Φ's algorithmic complexity being further scaled up. To address this, theoretical research has sought to improve its computational efficiency (Hidaka & Oizumi, 2018; Kitazono et al., 2018) as well as finding bounds for Φ (Zaeemzadeh & Tononi, 2024) and well as tying it to other metrics (Marshall et al., 2016).

In spite of this increase in the theory complexity, key initial operationalizations of the theory have remained remarkably stable through its iterations. Notable examples include the notion of minimum information partitions (MIP) to quantify integration and the use of interventional distributions and its distance to unconstrained uniform distributions to quantify information, which have been introduced early on and accompanies the theory until today. Over time, these algorithms have been refined (e.g. the way to do cuts to assess integration, impose background conditions, etc), and an effort has been made to derive them from first principles. For instance, the IIT's metric has significantly evolved, from the classical Kullback-Leibler (KL) divergence, to Earth Mover's Distance (EMD) in 3.0, and finally to the 4.0 intrinsic information metric derived from first principles (Barbosa et al., 2020). All of this may suggest a stabilization of the theory's mathematical development, whereby the differences between successive versions appear to be diminishing.

*Phenomenological derivation, mathematical generalizations and practical approximations of Φ*

As we have seen, IIT mathematical development is crucially informed by phenomenology, and these guiding principles were made explicit by the introduction of the axiom-to-postulate approach in IIT 3.0 (section **2.1.5**), as well as by methodological and ontological principles introduced in IIT 4.0. In this view, the joint enforcement of all postulates is required in order to account for consciousness and ultimately to derive an universal and unique formulation for computing Φ. However, in theoretical IIT – in which IIT's metaphysical and scientific dimension is put in the background focusing on its internal mathematical structure – a range of strategies in developing and studying the idea of integrated information mathematically have been pursued, often bracketing the other principles of the theory. This has been followed by important work generalizing and unifying these measures (Mediano, Rosas, et al., 2019; Oizumi et al., 2016; Tegmark, 2016), comparing them in simulated systems (Mediano, Seth, et al., 2019; Sevenius Nilsen et al., 2019), and elaborating mathematical critiques (Barrett & Mediano, 2019), adding robustness to the whole framework. In parallel to IIT's core development Φ,

a series of alternative measures of integrated information have been proposed (Barrett & Seth, 2011; Griffith, 2014; Oizumi et al., 2016; Seth et al., 2011b).




### 3.2. Experimental: Neuronal IIT

From an experimental perspective, IIT's engagement with empirical research is pivotal in validating its theoretical propositions, with the primary aim of elucidating the neural basis of consciousness. Centered in neuroscience, Neuronal IIT focuses on the dual relationship between the theory and experiments: how the theory informs experimental design and interpretation, and how experimental outcomes validate or challenge the theory.

Here, we will focus on the heuristic value and fecundity of IIT, which has received less attention, and defer discussing its explanatory power (e.g. its explanation of why the cerebellum is not a substrate of consciousness, or why consciousness is lost in NREM sleep or anesthesia) and predictive novelty to other recent works (Ellia et al., 2021; Oizumi et al., 2014; Tononi et al., 2016). Heuristic value pertains to the theory's ability to generate new research questions and guide experimental approaches, while fecundity measures the fruitfulness of a theory in generating new insights, hypotheses, and research avenues, thereby contributing to the advancement of scientific knowledge across multiple domains.

*Fixing the explanandum: Dissociating phenomenal consciousness from other phenomena*

IIT's understanding of consciousness as intrinsically phenomenal (section **2.1.1** and **2.2.1**), has shaped a series of experimental work and debates aimed at disentangling consciousness from other brain-based phenomena. In neurology, the dissociation between (un)consciousness and (un)responsiveness has been advocated by researchers (Bayne et al., 2020; Sanders et al., 2012), with important clinical implications for the understanding, monitoring and stratification of disorders of consciousness (DoC) and anesthetized states (Sarasso et al., 2015). This debate has highlighted the limits of behavioral assessments and of sensory-motor connectedness to the environment as a criteria for evaluating the presence of consciousness (Casarotto et al., 2016; Giacino et al., 2014; Laureys et al., 2015).

In psychophysics, the IIT field has weighed in two important debates. First, in the context of content-NCC research, where the NCC have been investigated with several psychophysical paradigms based on contrasting consciously perceived and non-perceived stimuli. Aligned with IIT's view, researchers have advocated that consciousness and attention are distinct phenomena, based on the possibility of their experimental dissociation (Koch & Tsuchiya, 2007; Tsuchiya & Koch, 2011; van Boxtel et al., 2010), and have argued for the importance of controlling for attentional effects when searching for NCC, countering those who defend that attention is necessary and sufficient for a stimulus reaching conscious perception (De Brigard & Prinz, 2010; Tallon-Baudry, 2012). The debate evolved with NCC researchers pointing to the need of controlling not only for attention but any process preceding (pre-NCC, e.g. arousal, expectation) or succeeding (post-NCC, e.g. executive control, memory and report) the genuine NCC (Aru et al., 2012). This has marked the shift, advocated by IIT researchers (Koch et al., 2016), of moving toward no-report paradigms in order to mitigate these confounders (Tsuchiya et al., 2015). Similarly, when contrasting states of consciousness (e.g., wakefulness) and unconsciousness (e.g., NREM sleep) in searching for level-NCC, IIT researchers have pointed out confounders such as background conditions that enable consciousness but are not directly part of the NCC (e.g. neuromodulatory effects driven by reticular activating systems (RAS)) must also be considered, advocating for within-state paradigms. These have been implemented during sleep using a serial awakening paradigm, in which the neural activity preceding a dream report was compared with that when no dream is reported (Siclari et al., 2017).

A second debate concerns the apparent richness of visual experiences, where an influential view defends that it is in fact illusory. According to this view, consciousness in reality has a small bandwidth, limited by an attentional bottleneck, such that visual experiences actually contain just a handful of items together with some "summary statistics" (Cohen et al., 2016). In contrast to this view, it has been argued that phenomenal consciousness can "overflow" what can be consciously assessed (Block, 2011). IIT advocates have weighed in by arguing that (i) the richness of experience is being underestimated by the limited way reports are obtained in experiments (e.g. binary button presses and "see vs didn't see" questions) (A. M. Haun, Tononi, et al., 2017) and that (ii) what we see is radically different from what we can notice (A. Haun & Tononi, 2024).

This experiential richness and its neural basis has been probed experimentally by the use of both "massive report" paradigms (Qianchen et al., 2022) as well as by using no-report task-free paradigms, e.g. comparing watching movie vs scrambled versions. In the latter, neural differentiation measures have been shown to reflect the stimulus's meaningfulness (Boly et al., 2015; Mayner et al., 2022; Mensen et al., 2017).

*Coarse-Level Test of IIT: indexing level of consciousness using causality, integration and information*

The proposal linking consciousness and complexity, understood as a balance between functional integration (section **2.3.3**) and differentiation (**2.3.2**), proposed by Giulio Tononi and Gerald Edelman in 1998 (Tononi & Edelman, 1998), has been extensively validated experimentally over the past two decades, across various neuroimaging tools and quantification strategies, recently reviewed in (Sarasso et al., 2021). Given the substantial body of empirical evidence supporting the connection between brain complexity and consciousness, we can view this as a coarse-grained test of IIT, validating some of its core ideas: integration and information.

The widespread use of complexity measures to index consciousness in various states underscores the practical applicability of these theoretical principles.

IIT's idea of a perturb-and-measure approach to study consciousness (section **2.2.2**) has furthered the assessment of brain complexity with a perturbational perspective, embodied in the development of the Perturbational Complexity Index (PCI), which measures the spatiotemporal complexity of brain responses to perturbations (Casali et al., 2013; Comolatti et al., 2019). PCI is high when the brain is able to engage in deterministic interactions (causality) that are, at once, distributed among cortical areas (integrated) and differentiated in space and time (informative). Conversely, PCI is low if the brain response to the perturbation remains local (no integration) or stereotypical in space and time (no differentiation). A series of studies using PCI in different states of (un)consciousness – including wakefulness, NREM sleep, brain-injured patients, anesthesia and psilocybin – have confirmed the prediction of IIT that the loss and recovery of consciousness is associated with the breakdown and recovery of the capacity for information integration. This approach has been extended beyond its initial TMS/EEG in humans to intracranial stimulation (Comolatti et al., 2019), animal models (Arena et al., 2021; Cavelli et al., 2023) and cortical slices (D'Andola et al., 2018).

*Identifying the Extension of Consciousness: from NCC to maximal substrates of consciousness*

A central area of experimental investigation in consciousness science is the search for the NCC. In this context, currently a major debate centers on whether consciousness is primarily located in the posterior (back) or anterior (front) parts of the brain (Boly et al., 2017; Odegaard et al., 2017). IIT defends the "posterior hot zone" hypothesis, suggesting that regions in the posterior, occipital and temporal cortex are sufficient for conscious experience based on causal evidence from lesion and electrical stimulation studies, as well as studies using no-report and within-state paradigms controlling for cognitive access and background conditions (Koch et al., 2016). This contrasts with frontalist theories such as the global workspace (Mashour et al., 2020) and higher-order theories (Brown et al., 2019), which emphasize the necessary role of frontal regions and fronto-parietal connectivity.

IIT also is in position to explain why certain kinds of structures in the brain would be substrates of consciousness (e.g. cerebral cortex) and others not (e.g. cerebellum) based on their capacity to maximally integrate information (section **2.3.4**). Moreover, by studying how Φ scales for different patterns of local connectivity (e.g. grid vs random) it can identify cortical regions that are optimally connected to be maximal substrates of consciousness.

*Predicting the substrate of different qualitative kinds of experiences: space, time and objects*

IIT predicts that the quality of experience is determined by the way the Φ-structure is composed (section **2.3.5**), which is in turn dependent on the architecture of the substrate specifying it. This approach has been first demonstrated for spatial experiences, where the feeling of "extendedness" can be accounted for by the Φ-structures specified by 2D grids present in occipital and parietal cortices (A. Haun & Tononi, 2019). IIT further predicts that changes in the connectivity within these neural substrates affect the quality of the experience, even if neural activity remains unchanged.

Ongoing research aims to extend this explanatory framework to other qualitative experiences, such as time (Comolatti et al. *forthcoming*) and objects (Grasso et al. *forthcoming*). This involves identifying the specific neural architectures that support these experiences and examining how their intrinsic causal structures align with the fundamental properties of the corresponding phenomenal experiences. By systematically integrating phenomenological insights with neuroscientific evidence, IIT seeks to provide a comprehensive account of the diverse qualitative aspects of conscious experience.

### 3.3. Metaphysical: Worldview IIT

As many scientific theories, IIT comes with metaphysical assumptions and implications, namely a series of claims that, although they cannot be directly empirically tested, help drive the type of questions that scientists ask, and define the outlook on the nature of reality implied by the theory as well as our way to grasp it. This is what Popper has called "metaphysical research programme" (Popper, 1989). Given that much attention has been put on, and many criticisms have been formulated against, this specific dimension of IIT, it is important to understand its ramifications. The metaphysical and epistemological dimensions of IIT build on issues that span across philosophy of mind, ontology, and philosophy of science.

*Philosophy of mind*

In philosophy of mind, one of the fundamental questions concerns the relationship between mind and matter - the so-called "mind-body problem" (Kim, 2000) (Anthony, 2009). Many philosophers have tried to categorize IIT within the landscape of philosophical views designed to address the mind-body problem.

IIT has been associated with panpsychism (Mørch, 2019a; Tononi & Koch, 2015), emergentism (Cea, 2021; N. Negro, 2024b), Russellian Monism (Grasso, 2018), and even functionalism (Block, 2009) and illusionism (McQueen, 2019). More recently, it has been argued that the emphasis on the relationship between consciousness and intrinsic existence made explicit by the 4.0 version, as well as the emphasis on the instrumental nature of the physical, suggests an alignment between IIT and classic forms of realist idealism (Cea et al., 2023; D. Chalmers, 2019). In this view, the realism of IIT is not about the physical world, but about other experiences.

Although these categorizations are mostly interpretational, in the sense that they aim to find a place for IIT in the existing taxonomy of metaphysical views on the nature of consciousness, it seems possible to adopt a different stance, and use IIT, or parts of it, as a way to develop a metaphysical view on consciousness. For example, one could adopt IIT's formalism to formulate a version of emergentism, since the formalism is well-suited to measure the existence of a whole beyond its parts. Emergentism, however, implies some theoretical choices that will depart from IIT "proper", so one should be careful in selecting the exact theoretical construals to discard. For example, one could revise the central identity between consciousness and integrated information by positing that the identity holds only when certain contextual conditions are met (i.e., this would be a form of contextual emergence, (Bishop et al., 2022).

Another pivotal question in philosophy of mind pertains to the nature of meaning, or mental content: how and why do mental states get the contents they have? Most theories of consciousness, although not addressing

this problem directly, seem to assume a representationalist stance according to which mental content is given by the content of a neural representation of an external stimulus, and consciousness is just a specific property of the representation. In this case, consciousness is grounded on meaning. IIT takes instead a somewhat unique position (within the neuroscience of consciousness) by claiming that meaning is grounded on consciousness. In this sense, according to IIT, "the meaning is the feeling" (Mindt, 2021; N. B. Negro, 2023), a view that aligns it with phenomenal intentionality theory (Mendelovici & Bourget, 2020).

*Ontology*

The second aspect of Worldview IIT is the ontological one: As seen above, accepting that an explanation of consciousness should start from consciousness itself is not, for IIT, only an epistemological point. Instead, accepting the primacy of consciousness is in itself ontologically charged, because consciousness is taken to be the only thing whose existence we can be certain of.

This metaphysical standpoint reinforces the idea that IIT is not only a theory of consciousness, but also a theory of existence, since it claims that consciousness is intrinsic existence (i.e., it exists for itself), while physical existence is extrinsic existence (i.e., it exists for an external, conscious, observer). This aspect of Worldview IIT can have counterintuitive implications, like the conclusion that neurons in the substrate of consciousness do not really exist, given that their existence is subsumed by the Φ-max-generating whole to which they belong (Tononi et al., 2023).

A possible way to resist this type of implications is to claim that, although the epistemological primacy of consciousness is secured by the zeroth axiom (consciousness exists and is the only thing that exists with certainty), the ontological primacy does not follow. This is because the only criterion presented in defense of the ontological primacy of consciousness over the physical world is that the existence of consciousness is given with certainty and cannot be doubted, while the existence of the physical world can only be inferred. But certainty and doubts are epistemological categories, not ontological ones, and therefore the conclusion about the ontological primacy of consciousness cannot be warranted by the epistemic category of certainty (for a similar point, see (Cea et al., 2023)).

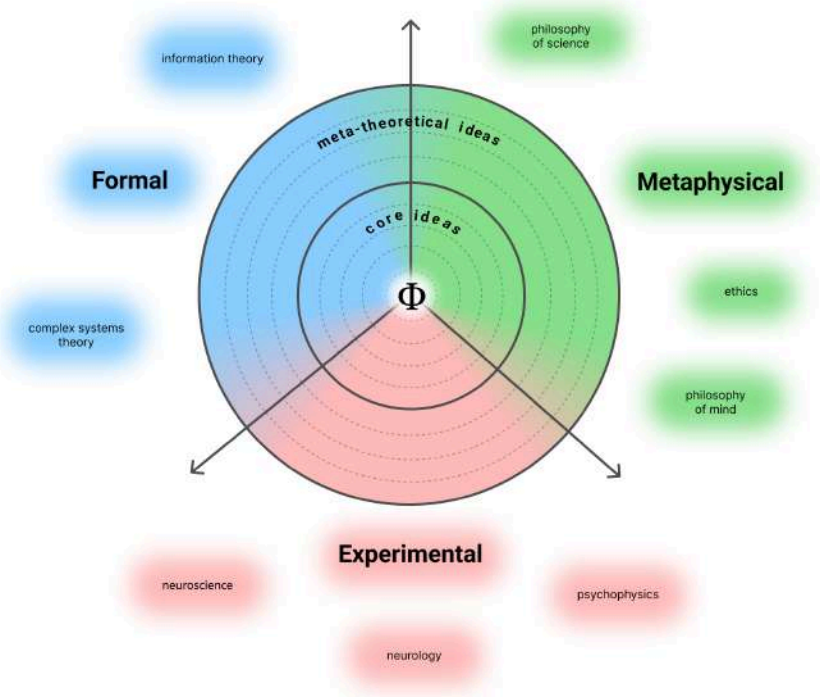
*Philosophy of science*

The third aspect of Worldview IIT impacts on philosophy of science. The standard approach in the neuroscience of consciousness is to view consciousness science as a sub-field of special sciences like biology, psychology, and neuroscience. IIT asks us to reverse this picture. If consciousness is prior to the physical world and the physical world is just posited from within consciousness, our way to grasp the physical world is dependent on our way to understand consciousness. If IIT is right, consciousness science should take the place traditionally occupied by physics: the understanding of the natural world cannot be complete without the inclusion of consciousness.

# 4. Discussion

### 4.1. IIT as a research program in the science of consciousness

In this paper we have presented IIT as a multilayered and multidimensional theory. We have navigated through its internal structure (layers) and mapped out its relation to different scientific, mathematical and philosophical research practices (dimensions). This complex theoretical architecture suggests that IIT can be a prolific and useful research programme even if one does not embrace IIT as a whole: different research projects can stem from ideas and notions derived from IIT by narrowing in on specific dimensions of IIT. This can be beneficial to various disciplines independently of the soundness of IIT as a whole. Indeed, the fact that many research groups and scholars have built upon IIT-inspired notions and construals to develop innovative research agendas testifies IIT's influence and fecundity.




In this regard, a distinction introduced in the literature distinguishes between 'strong' and 'weak' IIT (Mediano, Rosas, Bor, et al., 2022). On the one hand, strong IIT is concerned with building a comprehensive and fine-grained version of the theory, that is able to posit a central identity between consciousness and Φ-structures and derive conclusions about the nature of consciousness (and possibly reality itself). In doing that, strong IIT keeps together the formal and experimental dimensions with the worldview dimension while jointly upholding all of its core ideas. On the other hand, weak IIT focuses on developing practical measures of integrated information that can be employed in empirical research and practical applications, while being agnostic of IIT's worldview. In other words, weak IIT is content with testing a coarse-grained version of the theory, bracketing several layers of the theory (e.g. causality, composition, exclusion and so on), in order to empirically test two of its core layers, i.e. information and integration. In our framework, it becomes clear that

weak and strong IIT consist in two (out of possibly many) ways of engaging with the theory, rather than mutually exclusive approaches.

The fecundity of the IIT's research programme, in particular along its experimental dimension, speaks in its favor. Indeed, although the empirical testability and the scientific status of IIT has been vigorously questioned (Doerig et al., 2019); see Kleiner & Hoel, 2021; Negro, 2020; Tsuchiya et al., 2020 for replies, and Usher, 2021 for a comprehensive discussion), we believe that these criticisms do not sufficiently consider the multilayered nature of IIT, in which a metaphysical agenda lives together with a formal and an empirical one. On the one hand, questioning the scientific legitimacy of IIT seems to dismiss IIT's meta-theoretical layers, while on the other hand, it dismisses the fact that all scientific theories comprise non-empirical aspects, and that their assessment is based on iterative testing of how these aspects cohere with empirical predictions (Lakatos, 1978). Importantly, theories are rarely, if ever, tested as a whole. Rather, experiments typically test single predictions of the theory. The key question, then, is in which way specific low-level predictions of IIT impact the theory as a whole. Under a sophisticated falsificationist lens (Lakatos, 1978) and the multidimensional and multilayered presentation of IIT we have offered, it seems possible to hold that IIT as a whole research programme might not be directly compromised by some disconfirmatory evidence - this would be true for other theories of consciousness too (for a discussion, see N. Negro, 2024a).

At this stage, we can conclude that IIT remains an active and open research programme, which can be used as a generator of many hypotheses and sub-programmes across different fields. This is a positive virtue of IIT, but it also implies that there are many different ways for it to be wrong. We briefly turn to this issue in what follows.

### 4.2. The many ways IIT can fail – or succeed

As IIT is an ongoing research programme, it would be premature to provide a full assessment. However, we can point in some directions that can elucidate the many ways the IIT research programme could turn out to be fruitful or mistaken.

At a coarse-grain level of analysis, IIT might be seen as a precursor that might pave the way to the right future ToC, even though it bears low resemblance with it. Even ifIIT would be remembered as a wild and largely mistaken ToC, it would still be a useful and fertile research programme, functioning as a generator of hypotheses, ideas, methods, and formalisms that could prove indispensable for future consciousness science.

In this regard, IIT has already influenced consciousness science by suggesting complexity based measures for indexing consciousness (Sarasso et al., 2021) that have been adopted by researchers across research programmes (Farisco & Changeux, 2023; Frohlich et al., 2022; Sitt et al., 2014). These measures build upon the assumption that information and integration are necessary ingredients for consciousness, and it is plausible to think that these ingredients might be incorporated by future models of consciousness, even if the specific way these are modeled in the current IIT formalism will turn out to be incorrect.

A second coarse-grain notion through which IIT is already influencing consciousness science relates to the structural approach for investigating experience. Several research programmes are currently stemming from this branch of this layer of IIT (Kleiner, 2024; Prentner, 2022; Tsuchiya et al., 2016; Tsuchiya & Saigo, 2021), speaking to the originality and fertility of the theory. Further research on this area might reveal a different and

more compelling phenomenological foundation for consciousness science, as well as a different way to express the methodological apparatus for bridging experience and the physical domain (e.g., the relationship between consciousness and the physical might be cashed out in terms of isomorphism, or less strict structure-preserving mappings). Again, this would be another instance in which IIT might turn out to be incorrect, and yet hugely influential.

At a medium-grain level of analysis, IIT might result to be productive in inspiring theories that can account for many of IIT's predictions about the neural and architectural realization of consciousness, even if these predictions are incorporated in formal frameworks that diverge from IIT proper.

For example, empirical predictions of IIT, which, if corroborated, should be included in any IIT-inspired theory as a piece of evidence, are: (i) inactive neurons contribute to conscious experience; (ii) the NCC (e.g. the posterior hot zone) corresponding to high-Φ architectures (e.g. grids); (iii) Grids should support spatial experience, while directed grids should support temporal experience.

Importantly, these predictions could turn out wrong. This confirms once again IIT's empirical testability, and in case of disconfirmed predictions consciousness theorists can decide whether to hold on to the core ideas of IIT and revise some of its auxiliary assumptions, or simply abandon the research programme in favour of an alternative one (N. Negro, 2024a).

At a fine-grain level of analysis, current IIT might turn out to be a useful theory just by providing all the necessary concepts and methodology that will be further refined and widely accepted as the correct ToC. This amounts to saying that the current version of IIT 4.0 is already the right ToC, but it is just not expressed at the right level of detail.

In order to convince its detractors that this is the case, IIT needs to correctly and consistently identify the borders of the NCC by computing which brain structures maximizes Φ, which should also be expressed in a computationally tractable way to track and predict levels of consciousness in healthy adults as well as in patients with disorders of consciousness (that is, the predictive success currently exhibited by PCI should be exhibited by a more direct approximation of system Φ). Moreover, Φ-structures should precisely and accurately track any content of consciousness at any given time.

Although this is admittedly a remote possibility, the fact that IIT influences various research programmes and agendas across disciplines, providing clear ways to show not only how it could be right, but also how it could be wrong, suggests that IIT is an invaluable intellectual endeavor: this means that many research avenues deserved to be explored, in how to interpret it, how to extend it, and how to falsify it.

## Bibliography


Aguilera, M. (2019). Scaling Behaviour and Critical Phase Transitions in Integrated Information Theory. *Entropy*, *21*(12), 1198. https://doi.org/10.3390/e21121198

Albantakis, L., Barbosa, L., Findlay, G., Grasso, M., Haun, A. M., Marshall, W., Mayner, W. G. P., Zaeemzadeh, A., Boly, M., Juel, B. E., Sasai, S., Fujii, K., David, I., Hendren, J., Lang, J. P., & Tononi, G. (2023). Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms. *PLOS Computational Biology*, *19*(10), e1011465.

https://doi.org/10.1371/journal.pcbi.1011465

Albantakis, L., Marshall, W., Hoel, E., & Tononi, G. (2019). What Caused What? A Quantitative Account of Actual Causation Using Dynamical Causal Networks. *Entropy*, *21*(5), 459. https://doi.org/10.3390/e21050459

Albantakis, L., Prentner, R., & Durham, I. (2023). Computing the Integrated Information of a Quantum Mechanism. *Entropy*, *25*(3), Article 3. https://doi.org/10.3390/e25030449

Amari, S., Tsuchiya, N., & Oizumi, M. (2017). Geometry of Information Integration. *arXiv:1709.02050 [Cs, Math]*. http://arxiv.org/abs/1709.02050

Anthony, L. (2009). The Mental and the Physical. In R. L. Poidevin, S. Peter, M. Andrew, & R. P. Cameron (Eds.), *The Routledge Companion to Metaphysics*. Routledge.

Apolito, A. (2020). *The Problem of Scale in Anarchism and the Case for Cybernetic Communism*. https://c4ss.org/content/52970

Arena, A., Comolatti, R., Thon, S., Casali, A. G., & Storm, J. F. (2021). General Anesthesia Disrupts Complex Cortical Dynamics in Response to Intracranial Electrical Stimulation in Rats. *eNeuro*, *8*(4). https://doi.org/10.1523/ENEURO.0343-20.2021

Aru, J., Bachmann, T., Singer, W., & Melloni, L. (2012). Distilling the neural correlates of consciousness. *Neuroscience & Biobehavioral Reviews*, *36*(2), 737–746. https://doi.org/10.1016/j.neubiorev.2011.12.003

Baker, B., Lansdell, B., & Kording, K. P. (2022). Three aspects of representation in neuroscience. *Trends in Cognitive Sciences*, *26*(11), 942–958. https://doi.org/10.1016/j.tics.2022.08.014

Balduzzi, D., & Tononi, G. (2008). Integrated Information in Discrete Dynamical Systems: Motivation and Theoretical Framework. *PLoS Computational Biology*, *4*(6), e1000091. https://doi.org/10.1371/journal.pcbi.1000091

Banks, E. C. (2010). Neutral monism reconsidered. *Philosophical Psychology*, *23*(2), 173–187. https://doi.org/10.1080/09515081003690418

Barbosa, L. S., Marshall, W., Streipert, S., Albantakis, L., & Tononi, G. (2020). A measure for intrinsic information. *Scientific Reports*, *10*(1), Article 1. https://doi.org/10.1038/s41598-020-75943-4

Barrett, A. B. (2014). An integration of integrated information theory with fundamental physics. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00063

Barrett, A. B., & Mediano, P. A. M. (2019). The Phi Measure of Integrated Information is not Well-Defined for General Physical Systems. *Journal of Consciousness Studies*, *26*(1–2), 11–20.

Barrett, A. B., & Seth, A. K. (2011). Practical Measures of Integrated Information for Time-Series Data. *PLoS Computational Biology*, *7*(1), e1001052. https://doi.org/10.1371/journal.pcbi.1001052

Bateson, G. (2000). *Steps to an Ecology of Mind: Collected Essays in Anthropology, Psychiatry, Evolution, and Epistemology*. University of Chicago Press.

Bayne, T. (2018). On the axiomatic foundations of the integrated information theory of consciousness. *Neuroscience of Consciousness*, *2018*(1). https://doi.org/10.1093/nc/niy007

Bayne, T., Seth, A. K., & Massimini, M. (2020). Are There Islands of Awareness? *Trends in Neurosciences*, *43*(1), 6–16. https://doi.org/10.1016/j.tins.2019.11.003

Bishop, R. C., Silberstein, M., Pexton, M., Bishop, R. C., Silberstein, M., & Pexton, M. (2022). *Emergence in Context: A Treatise in Twenty-First Century Natural Philosophy*. Oxford University Press.

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, *18*(2), 227–247. https://doi.org/10.1017/S0140525X00038188

Block, N. (2009). Comparing the Major Theories of Consciousness. In M. Gazzaniga (Ed.), *The Cognitive Neurosciences IV* (pp. 1111–1123).

Block, N. (2011). Perceptual Consciousness Overflows Cognitive Access. *Trends in Cognitive Sciences*, *15*(12), 567–575. https://doi.org/10.1016/j.tics.2011.11.001

Boly, M., Massimini, M., Tsuchiya, N., Postle, B. R., Koch, C., & Tononi, G. (2017). Are the Neural Correlates of Consciousness in the Front or in the Back of the Cerebral Cortex? Clinical and Neuroimaging Evidence. *The Journal of Neuroscience*, *37*(40), 9603–9613. https://doi.org/10.1523/JNEUROSCI.3218-16.2017

Boly, M., Sasai, S., Gosseries, O., Oizumi, M., Casali, A., Massimini, M., & Tononi, G. (2015). Stimulus Set Meaningfulness and Neurophysiological Differentiation: A Functional Magnetic Resonance Imaging Study. *PLOS ONE*, *10*(5), e0125337. https://doi.org/10.1371/journal.pone.0125337

Bostock, D. (2009). *Philosophy of Mathematics: An Introduction* (1st edition). Wiley-Blackwell.

Brette, R. (2019). Is coding a relevant metaphor for the brain? *Behavioral and Brain Sciences*, *42*, e215. https://doi.org/10.1017/S0140525X19000049

Brown, R., Lau, H., & LeDoux, J. E. (2019). Understanding the Higher-Order Approach to Consciousness. *Trends in Cognitive Sciences*, *23*(9), 754–768. https://doi.org/10.1016/j.tics.2019.06.009

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S. M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2023). *Consciousness in Artificial Intelligence: Insights from the Science of Consciousness* (arXiv:2308.08708). arXiv. https://doi.org/10.48550/arXiv.2308.08708

Casali, A. G., Gosseries, O., Rosanova, M., Boly, M., Sarasso, S., Casali, K. R., Casarotto, S., Bruno, M.-A., Laureys, S., Tononi, G., & Massimini, M. (2013). A Theoretically Based Index of Consciousness Independent of Sensory Processing and Behavior. *Science Translational Medicine*, *5*(198), 198ra105-198ra105. https://doi.org/10.1126/scitranslmed.3006294

Casarotto, S., Comanducci, A., Rosanova, M., Sarasso, S., Fecchio, M., Napolitani, M., Pigorini, A., G. Casali, A., Trimarchi, P. D., Boly, M., Gosseries, O., Bodart, O., Curto, F., Landi, C., Mariotti, M., Devalle, G., Laureys, S., Tononi, G., & Massimini, M. (2016). Stratification of unresponsive patients by an independently validated index of brain complexity: Complexity Index. *Annals of Neurology*, *80*(5), 718–729. https://doi.org/10.1002/ana.24779

Cavelli, M. L., Mao, R., Findlay, G., Driessen, K., Bugnon, T., Tononi, G., & Cirelli, C. (2023). Sleep/wake changes in perturbational complexity in rats and mice. *iScience*, *26*(3), 106186. https://doi.org/10.1016/j.isci.2023.106186

Cea, I. (2021). Integrated information theory of consciousness is a functionalist emergentism. *Synthese*, *199*(1), 2199–2224. https://doi.org/10.1007/s11229-020-02878-8

Cea, I., Negro, N., & Signorelli, C. M. (2023). *The Fundamental Tension in Integrated Information Theory 4.0's*

*Realist Idealism*. PsyArXiv. https://doi.org/10.31234/osf.io/cte2q

Chalmers, D. (2018). The Meta-Problem of Consciousness. *Journal of Consciousness Studies*, *25*(9–10), 6–61.

Chalmers, D. (2019). Idealism and the Mind-Body Problem. In W. Seager (Ed.), *The Routledge Handbook of Panpsychism* (pp. 353–373). Routledge. https://philarchive.org/rec/CHAIAT-11

Chalmers, D. J. (1997). *The Conscious Mind: In Search of a Fundamental Theory* (Revised ed. edition). Oxford University Press.

Chalmers, D. J. (2000). What is a neural correlate of consciousness? In *Neural correlates of consciousness: Empirical and conceptual questions* (pp. 17–39). The MIT Press.

Chang, A. Y. C., Biehl, M., Yu, Y., & Kanai, R. (2020). Information Closure Theory of Consciousness. *Frontiers in Psychology*, *11*, 1504. https://doi.org/10.3389/fpsyg.2020.01504

Chis-Ciure, R. (2022). The transcendental deduction of Integrated Information Theory: Connecting the axioms, postulates, and identity through categories. *Synthese*, *200*(3), 236. https://doi.org/10.1007/s11229-022-03704-z

Chis-Ciure, R., & Ellia, F. (2023). Facing up to the Hard Problem of Consciousness as an Integrated Information Theorist. *Foundations of Science*, *28*(1), 255–271. https://doi.org/10.1007/s10699-020-09724-7

Churchland, P. S. (1994). Can Neurobiology Teach Us Anything about Consciousness? *Proceedings and Addresses of the American Philosophical Association*, *67*(4), 23–40. https://doi.org/10.2307/3130741

Clark, A., & Chalmers, D. (1998). The Extended Mind. *Analysis*, *58*(1), 7–19. https://doi.org/10.1093/analys/58.1.7

Cleeremans, A. (2005). Computational correlates of consciousness. *Progress in Brain Research*, *150*, 81–98. https://doi.org/10.1016/S0079-6123(05)50007-4

Cohen, M. A., & Dennett, D. C. (2011). Consciousness cannot be separated from function. *Trends in Cognitive Sciences*, *15*(8), 358–364. https://doi.org/10.1016/j.tics.2011.06.008

Cohen, M. A., Dennett, D. C., & Kanwisher, N. (2016). What is the Bandwidth of Perceptual Experience? *Trends in Cognitive Sciences*, *20*(5), 324–335. https://doi.org/10.1016/j.tics.2016.03.006

Comolatti, R., & Hoel, E. (2022). Causal emergence is widespread across measures of causation. *arXiv:2202.01854 [Physics]*. http://arxiv.org/abs/2202.01854

Comolatti, R., Pigorini, A., Casarotto, S., Fecchio, M., Faria, G., Sarasso, S., Rosanova, M., Gosseries, O., Boly, M., Bodart, O., Ledoux, D., Brichant, J.-F., Nobili, L., Laureys, S., Tononi, G., Massimini, M., & Casali, A. G. (2019). A fast and general method to empirically estimate the complexity of brain responses to transcranial and intracranial stimulations. *Brain Stimulation*, *12*(5), 1280–1289. https://doi.org/10.1016/j.brs.2019.05.013

Crick, F., & Koch, C. (1990). Towards a Neurobiological Theory of Consciousness. *Seminars in the Neurosciences*, *2*, 263–275.

D'Andola, M., Rebollo, B., Casali, A. G., Weinert, J. F., Pigorini, A., Villa, R., Massimini, M., & Sanchez-Vives, M. V. (2018). Bistability, Causality, and Complexity in Cortical Networks: An In Vitro Perturbational Study. *Cerebral Cortex*, *28*(7), 2233–2242. https://doi.org/10.1093/cercor/bhx122

De Brigard, F., & Prinz, J. (2010). Attention and consciousness. *WIREs Cognitive Science*, *1*(1), 51–59.

https://doi.org/10.1002/wcs.27

Dehaene, S., Changeux, J.-P., & Naccache, L. (2011). The Global Neuronal Workspace Model of Conscious Access: From Neuronal Architectures to Clinical Applications. In S. Dehaene & Y. Christen (Eds.), *Characterizing Consciousness: From Cognition to the Clinic?* (pp. 55–84). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-18015-6_4

Dennett, D. C. (1988). Quining Qualia. In A. J. Marcel & E. Bisiach (Eds.), *Consciousness in Contemporary Science*. Oxford University Press.

Doerig, A., Schurger, A., Hess, K., & Herzog, M. H. (2019). The unfolding argument: Why IIT and other causal structure theories cannot explain consciousness. *Consciousness and Cognition*, *72*, 49–59. https://doi.org/10.1016/j.concog.2019.04.002

Dretske, F. (1996). Phenomenal Externalism or If Meanings Ain't in the Head, Where Are Qualia? *Philosophical Issues*, *7*, 143–158. https://doi.org/10.2307/1522899

Dreyfus, H. L. (1992). *What Computers Still Can?T Do: A Critique of Artificial Reason*. MIT Press.

Edelman, G. M. (2001). *A Universe of Consciousness: How Matter Becomes Imagination* (Reprint edizione). Basic Books.

Edelman, G. M., Gally, J. A., & Baars, B. J. (2011). Biology of Consciousness. *Frontiers in Psychology*, *2*, 4. https://doi.org/10.3389/fpsyg.2011.00004

Ellia, F., Hendren, J., Grasso, M., Kozma, C., Mindt, G., Lang, J., Haun, A., Albantakis, L., Boly, M., & Tononi, G. (2021). Consciousness and the Fallacy of Misplaced Objectivity. *Neuroscience of Consciousness*, *7*(2), 1–12.

Farisco, M., & Changeux, J.-P. (2023). About the compatibility between the perturbational complexity index and the global neuronal workspace theory of consciousness. *Neuroscience of Consciousness*, *2023*(1), niad016. https://doi.org/10.1093/nc/niad016

Frankish, K. (2016). Illusionism as a Theory of Consciousness. *Journal of Consciousness Studies*, *23*(11–12), 11–39.

Frohlich, J., Chiang, J. N., Mediano, P. A. M., Nespeca, M., Saravanapandian, V., Toker, D., Dell'Italia, J., Hipp, J. F., Jeste, S. S., Chu, C. J., Bird, L. M., & Monti, M. M. (2022). Neural complexity is a common denominator of human consciousness across diverse regimes of cortical dynamics. *Communications Biology*, *5*(1), 1–17. https://doi.org/10.1038/s42003-022-04331-7

Giacino, J. T., Fins, J. J., Laureys, S., & Schiff, N. D. (2014). Disorders of consciousness after acquired brain injury: The state of the science. *Nature Reviews Neurology*, *10*(2), 99–114. https://doi.org/10.1038/nrneurol.2013.279

Gidon, A., Aru, J., & Larkum, M. E. (2022). Does brain activity cause consciousness? A thought experiment. *PLOS Biology*, *20*(6), e3001651. https://doi.org/10.1371/journal.pbio.3001651

Grasso, M. (2018). *IIT vs. Russellian Monism*. 28.

Graziano, M. S. A., & Webb, T. W. (2015). The attention schema theory: A mechanistic account of subjective awareness. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00500

Griffith, V. (2014). *A Principled Infotheoretic \phi-like Measure* (arXiv:1401.0978). arXiv. https://doi.org/10.48550/arXiv.1401.0978

Haun, A. M., Oizumi, M., Kovach, C. K., Kawasaki, H., Oya, H., Howard, M. A., Adolphs, R., & Tsuchiya, N. (2017). Conscious Perception as Integrated Information Patterns in Human Electrocorticography. *Eneuro*, *4*(5), ENEURO.0085-17.2017. https://doi.org/10.1523/ENEURO.0085-17.2017

Haun, A. M., Tononi, G., Koch, C., & Tsuchiya, N. (2017). Are we underestimating the richness of visual experience? *Neuroscience of Consciousness*, *2017*(1). https://doi.org/10.1093/nc/niw023

Haun, A., & Tononi, G. (2019). Why Does Space Feel the Way it Does? Towards a Principled Account of Spatial Experience. *Entropy*, *21*(12), 1160. https://doi.org/10.3390/e21121160

Haun, A., & Tononi, G. (2024). *The unfathomable richness of seeing*. OSF. https://doi.org/10.31234/osf.io/jmg35

Hidaka, S., & Oizumi, M. (2018). Fast and exact search for the partition with minimal information loss. *PLOS ONE*, *13*(9), e0201126. https://doi.org/10.1371/journal.pone.0201126

Hoel, E. (2023). *The World Behind the World: Consciousness, Free Will, and the Limits of Science*. Avid Reader Press / Simon & Schuster.

Hoel, E. P., Albantakis, L., & Tononi, G. (2013). Quantifying causal emergence shows that macro can beat micro. *Proceedings of the National Academy of Sciences*, *110*(49), 19790–19795. https://doi.org/10.1073/pnas.1314922110

Husserl, E. (1977). *Cartesian Meditations*. Springer Netherlands. https://doi.org/10.1007/978-94-009-9997-8

Irvine, E., Sprevak, M., Irvine, E., & Sprevak, M. (2020). *Eliminativism About Consciousness*. 347–370. https://doi.org/10.1093/oxfordhb/9780198749677.013.16

Kanai, R., & Fujisawa, I. (2023). *Towards a Universal Theory of Consciousness*. PsyArXiv. https://doi.org/10.31234/osf.io/r5t2n

Kim, J. (2000). *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation* (Reprint edition). Bradford Books.

Kirchhoff, M. D., & Kiverstein, J. (2019). *Extended Consciousness and Predictive Processing: A Third Wave View*. Routledge. https://doi.org/10.4324/9781315150420

Kitazono, J., Kanai, R., & Oizumi, M. (2018). Efficient Algorithms for Searching the Minimum Information Partition in Integrated Information Theory. *Entropy*, *20*(3), 173. https://doi.org/10.3390/e20030173

Kleiner, J. (2024). Towards a structural turn in consciousness science. *Consciousness and Cognition*, *119*, 103653. https://doi.org/10.1016/j.concog.2024.103653

Kleiner, J., & Hoel, E. (2021). Falsification and consciousness. *Neuroscience of Consciousness*, *2021*(1). https://doi.org/10.1093/nc/niab001

Koch, C., Massimini, M., Boly, M., & Tononi, G. (2016). Neural correlates of consciousness: Progress and problems. *Nature Reviews Neuroscience*, *17*(5), 307–321. https://doi.org/10.1038/nrn.2016.22

Koch, C., & Tsuchiya, N. (2007). Attention and consciousness: Two distinct brain processes. *Trends in Cognitive Sciences*, *11*(1), 16–22. https://doi.org/10.1016/j.tics.2006.10.012

Lakatos, I. (1978). *The Methodology of Scientific Research Programmes: Philosophical Papers* (J. Worrall & G. Currie, Eds.; Vol. 1). Cambridge University Press. https://doi.org/10.1017/CBO9780511621123

Laureys, S., Gosseries, O., & Tononi, G. (2015). *The Neurology of Consciousness: Cognitive Neuroscience and Neuropathology*. Academic Press.

Leung, A., Cohen, D., Swinderen, B. van, & Tsuchiya, N. (2021). Integrated information structure collapses with anesthetic loss of conscious arousal in Drosophila melanogaster. *PLOS Computational Biology*, *17*(2), e1008722. https://doi.org/10.1371/journal.pcbi.1008722

Levine, J. (1983). Materialism and Qualia: The Explanatory Gap. *Pacific Philosophical Quarterly*, *64*(4), 354–361. https://doi.org/10.1111/j.1468-0114.1983.tb00207.x

Lycan, W. G. (2001). The Case for Phenomenal Externalism. *Philosophical Perspectives*, *15*, 17–35.

Manin, Y., & Marcolli, M. (2022). *Homotopy Theoretic and Categorical Models of Neural Information Networks* (arXiv:2006.15136). arXiv. https://doi.org/10.48550/arXiv.2006.15136

Marshall, W., Gomez-Ramirez, J., & Tononi, G. (2016). Integrated Information and State Differentiation. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00926

Mashour, G. A., Roelfsema, P., Changeux, J.-P., & Dehaene, S. (2020). Conscious Processing and the Global Neuronal Workspace Hypothesis. *Neuron*, *105*(5), 776–798. https://doi.org/10.1016/j.neuron.2020.01.026

Massimini, M. (2005). Breakdown of Cortical Effective Connectivity During Sleep. *Science*, *309*(5744), 2228–2232. https://doi.org/10.1126/science.1117256

Mayner, W. G. P., Marshall, W., Billeh, Y. N., Gandhi, S. R., Caldejon, S., Cho, A., Griffin, F., Hancock, N., Lambert, S., Lee, E. K., Luviano, J. A., Mace, K., Nayan, C., Nguyen, T. V., North, K., Seid, S., Williford, A., Cirelli, C., Groblewski, P. A., … Arkhipov, A. (2022). Measuring Stimulus-Evoked Neurophysiological Differentiation in Distinct Populations of Neurons in Mouse Visual Cortex. *eNeuro*, *9*(1). https://doi.org/10.1523/ENEURO.0280-21.2021

McQueen, K. J. (2019). Illusionist Integrated Information Theory. *Journal of Consciousness Studies*, *26*(5–6), 141–169.

Mediano, P. A. M., Rosas, F., Carhart-Harris, R. L., Seth, A. K., & Barrett, A. B. (2019). Beyond integrated information: A taxonomy of information dynamics phenomena. *arXiv:1909.02297 [Physics, q-Bio]*. http://arxiv.org/abs/1909.02297

Mediano, P. A. M., Rosas, F. E., Bor, D., Seth, A. K., & Barrett, A. B. (2022). The strength of weak integrated information theory. *Trends in Cognitive Sciences*, *26*(8), 646–655. https://doi.org/10.1016/j.tics.2022.04.008

Mediano, P. A. M., Rosas, F. E., Farah, J. C., Shanahan, M., Bor, D., & Barrett, A. B. (2022). Integrated information as a common signature of dynamical and information-processing complexity. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *32*(1), 013115. https://doi.org/10.1063/5.0063384

Mediano, P. A. M., Seth, A. K., & Barrett, A. B. (2019). Measuring Integrated Information: Comparison of Candidate Measures in Theory and Simulation. *Entropy*, *21*(1), Article 1. https://doi.org/10.3390/e21010017

Mendelovici, A., & Bourget, D. (2020). Consciousness and Intentionality. In U. Kriegel (Ed.), *The Oxford Handbook of the Philosophy of Consciousness* (pp. 560–585). Oxford University Press. https://philarchive.org/rec/BOUCAI-3

Mensen, A., Marshall, W., & Tononi, G. (2017). EEG Differentiation Analysis and Stimulus Set Meaningfulness. *Frontiers in Psychology*, *8*, 1748. https://doi.org/10.3389/fpsyg.2017.01748

Michel, M. (2019). Consciousness Science Underdetermined: A Short History of Endless Debates. *Ergo, an Open Access Journal of Philosophy*, *6*(20200523). https://doi.org/10.3998/ergo.12405314.0006.028

Mindt, G. (2021). Not All Structure and Dynamics Are Equal. *Entropy*, *23*(9), Article 9. https://doi.org/10.3390/e23091226

Mørch, H. H. (2019a). Is the Integrated Information Theory of Consciousness Compatible with Russellian Panpsychism? *Erkenntnis*, *84*(5), 1065–1085. https://doi.org/10.1007/s10670-018-9995-6

Mørch, H. H. (2019b). The Argument for Panpsychism From Experience of Causation. In W. Seager (Ed.), *The Routledge Handbook of Panpsychism* (1st ed., pp. 269–284). Routledge. https://doi.org/10.4324/9781315717708-23

Muñoz, R. N., Leung, A., Zecevik, A., Pollock, F. A., Cohen, D., van Swinderen, B., Tsuchiya, N., & Modi, K. (2020). General anesthesia reduces complexity and temporal asymmetry of the informational structures derived from neural recordings in Drosophila. *Physical Review Research*, *2*(2), 023219. https://doi.org/10.1103/PhysRevResearch.2.023219

Negro, N. (2020). Phenomenology-first versus third-person approaches in the science of consciousness: The case of the integrated information theory and the unfolding argument. *Phenomenology and the Cognitive Sciences*. https://doi.org/10.1007/s11097-020-09681-3

Negro, N. (2023). Can the Integrated Information Theory Explain Consciousness from Consciousness Itself? *Review of Philosophy and Psychology*, *14*(4), 1471–1489. https://doi.org/10.1007/s13164-022-00653-x

Negro, N. (2024a). (Dis)confirming theories of consciousness and their predictions: Towards a Lakatosian consciousness science. *Neuroscience of Consciousness*, *2024*(1), niae012. https://doi.org/10.1093/nc/niae012

Negro, N. (2024b). Emergentist Integrated Information Theory. *Erkenntnis*, *89*(5), 1949–1971. https://doi.org/10.1007/s10670-022-00612-z

Negro, N. B. (2023). Consciousness and Content from the Perspective of the Integrated Information Theory. *Argumenta*, 1–19.

Odegaard, B., Knight, R. T., & Lau, H. (2017). Should a Few Null Findings Falsify Prefrontal Theories of Conscious Perception? *The Journal of Neuroscience*, *37*(40), 9593–9602. https://doi.org/10.1523/JNEUROSCI.3217-16.2017

Oizumi, M., Albantakis, L., & Tononi, G. (2014). From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0. *PLoS Computational Biology*, *10*(5), e1003588. https://doi.org/10.1371/journal.pcbi.1003588

Oizumi, M., Tsuchiya, N., & Amari, S. (2016). Unified framework for information integration based on information geometry. *Proceedings of the National Academy of Sciences*, *113*(51), 14817–14822. https://doi.org/10.1073/pnas.1603583113

Pearl, J. (2009). *Causality* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511803161

Petitmengin, C., Remillieux, A., & Valenzuela-Moguillansky, C. (2019). Discovering the structures of lived experience. *Phenomenology and the Cognitive Sciences*, *18*(4), 691–730. https://doi.org/10.1007/s11097-018-9597-4

Popiel, N. J. M., Khajehabdollahi, S., Abeyasinghe, P. M., Riganello, F., Nichols, E. S., Owen, A. M., & Soddu, A. (2020). The Emergence of Integrated Information, Complexity, and 'Consciousness' at Criticality. *Entropy*, *22*(3), Article 3. https://doi.org/10.3390/e22030339

Popper, K. (1989). *Quantum Theory and the Schism in Physics: From The Postscript to the Logic of Scientific Discovery* (1st edition). Routledge.

Prentner, R. (2022). *Applying category theory to the study of consciousness?* OSF. https://doi.org/10.31234/osf.io/3vhg9

Qianchen, L., Gallagher, R. M., & Tsuchiya, N. (2022). How much can we differentiate at a brief glance: Revealing the truer limit in conscious contents through the massive report paradigm (MRP). *Royal Society Open Science*, *9*(5), 210394. https://doi.org/10.1098/rsos.210394

Ramsey, W. (2022). Eliminative Materialism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2022). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2022/entries/materialism-eliminative/

Rosas, F. E., Mediano, P. A. M., Jensen, H. J., Seth, A. K., Barrett, A. B., Carhart-Harris, R. L., & Bor, D. (2020). Reconciling emergences: An information-theoretic approach to identify causal emergence in multivariate data. *PLOS Computational Biology*, *16*(12), e1008289. https://doi.org/10.1371/journal.pcbi.1008289

Russell, B. (1927). *The Analysis of Matter*. Kegan Paul.

Sanders, R. D., Tononi, G., Laureys, S., Sleigh, J. W., & Warner, D. S. (2012). Unresponsiveness ≠ Unconsciousness. *Anesthesiology*, *116*(4), 946–959. https://doi.org/10.1097/ALN.0b013e318249d0a7

Sarasso, S., Boly, M., Napolitani, M., Gosseries, O., Charland-Verville, V., Casarotto, S., Rosanova, M., Casali, A. G., Brichant, J.-F., Boveroux, P., Rex, S., Tononi, G., Laureys, S., & Massimini, M. (2015). Consciousness and Complexity during Unresponsiveness Induced by Propofol, Xenon, and Ketamine. *Current Biology*, *25*(23), 3099–3105. https://doi.org/10.1016/j.cub.2015.10.014

Sarasso, S., Casali, A. G., Casarotto, S., Rosanova, M., Sinigaglia, C., & Massimini, M. (2021). Consciousness and complexity: A consilience of evidence. *Neuroscience of Consciousness*, *niab023*. https://doi.org/10.1093/nc/niab023

Searle, J. R. (1992). *The Rediscovery of the Mind*. Bradford Books.

Seth, A. K., Barrett, A. B., & Barnett, L. (2011a). Causal density and integrated information as measures of conscious level. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *369*(1952), 3748–3767. https://doi.org/10.1098/rsta.2011.0079

Seth, A. K., Barrett, A. B., & Barnett, L. (2011b). Causal density and integrated information as measures of conscious level. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *369*(1952), 3748–3767. https://doi.org/10.1098/rsta.2011.0079

Sevenius Nilsen, A., Juel, B. E., Marshall, W., & Storm, J. F. (2019). *Evaluating Approximations and Heuristic Measures of Integrated Information* [Preprint]. MATHEMATICS & COMPUTER SCIENCE. https://doi.org/10.20944/preprints201904.0077.v1

Siclari, F., Baird, B., Perogamvros, L., Bernardi, G., LaRocque, J. J., Riedner, B., Boly, M., Postle, B. R., & Tononi, G. (2017). The neural correlates of dreaming. *Nature Neuroscience*, *20*(6), 872–878.

https://doi.org/10.1038/nn.4545

Singhal, I., Mudumba, R., & Srinivasan, N. (2022). In search of lost time: Integrated information theory needs constraints from temporal phenomenology. *Philosophy and the Mind Sciences*, *3*. https://doi.org/10.33735/phimisci.2022.9438

Sitt, J. D., King, J.-R., El Karoui, I., Rohaut, B., Faugeras, F., Gramfort, A., Cohen, L., Sigman, M., Dehaene, S., & Naccache, L. (2014). Large scale screening of neural signatures of consciousness in patients in a vegetative or minimally conscious state. *Brain: A Journal of Neurology*, *137*(Pt 8), 2258–2270. https://doi.org/10.1093/brain/awu141

Tallon-Baudry, C. (2012). On the Neural Mechanisms Subserving Consciousness and Attention. *Frontiers in Psychology*, *2*. https://doi.org/10.3389/fpsyg.2011.00397

Tegmark, M. (2015). Consciousness as a State of Matter. *Chaos, Solitons & Fractals*, *76*, 238–270. https://doi.org/10.1016/j.chaos.2015.03.014

Tegmark, M. (2016). Improved Measures of Integrated Information. *PLOS Computational Biology*, *12*(11), e1005123. https://doi.org/10.1371/journal.pcbi.1005123

Tononi, G. (2017). Integrated Information Theory of Consciousness: Some Ontological Considerations. In S. Schneider & M. Velmans (Eds.), *The Blackwell Companion to Consciousness* (pp. 621–633). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119132363.ch44

Tononi, G., Albantakis, L., Boly, M., Cirelli, C., & Koch, C. (2023). *Only what exists can cause: An intrinsic view of free will* (arXiv:2206.02069). arXiv. https://doi.org/10.48550/arXiv.2206.02069

Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: From consciousness to its physical substrate. *Nature Reviews Neuroscience*, *17*(7), 450–461. https://doi.org/10.1038/nrn.2016.44

Tononi, G., & Edelman, G. M. (1998). Consciousness and Complexity. *Science*, *282*(5395), 1846–1851. https://doi.org/10.1126/science.282.5395.1846

Tononi, G., & Koch, C. (2015). Consciousness: Here, there and everywhere? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1668), 20140167. https://doi.org/10.1098/rstb.2014.0167

Tononi, G., & Sporns, O. (2003). Measuring information integration. *BMC Neuroscience*, 20.

Tsuchiya, N., Andrillon, T., & Haun, A. (2020). A reply to "the unfolding argument": Beyond functionalism/behaviorism and towards a science of causal structure theories of consciousness. *Consciousness and Cognition*, *79*, 102877. https://doi.org/10.1016/j.concog.2020.102877

Tsuchiya, N., & Koch, C. (2011). *Relationship between selective visual attention and visual consciousness* (B. E. Rogowitz & T. N. Pappas, Eds.; p. 78650X). https://doi.org/10.1117/12.881465

Tsuchiya, N., & Saigo, H. (2021). A relational approach to consciousness: Categories of level and contents of consciousness. *Neuroscience of Consciousness*, *2021*(2), niab034. https://doi.org/10.1093/nc/niab034

Tsuchiya, N., Taguchi, S., & Saigo, H. (2016). Using category theory to assess the relationship between consciousness and integrated information theory. *Neuroscience Research*, *107*, 1–7. https://doi.org/10.1016/j.neures.2015.12.007

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

https://doi.org/10.1016/j.tics.2015.10.002

Tull, S., & Kleiner, J. (2020). Integrated Information in Process Theories. *arXiv:2002.07654 [Quant-Ph]*. http://arxiv.org/abs/2002.07654

Unger, P. (1980). The Problem of the Many. *Midwest Studies In Philosophy*, *5*(1), 411–468. https://doi.org/10.1111/j.1475-4975.1980.tb00416.x

Unger, P. (2004). The Mental Problems of the Many. In D. Zimmerman (Ed.), *Oxford Studies in Metaphysics, Vol. 1* (pp. 195–222). Oxford: Clarendon Press.

Usher, M. (2021). Refuting the unfolding-argument on the irrelevance of causal structure to consciousness. *Consciousness and Cognition*, *95*, 103212. https://doi.org/10.1016/j.concog.2021.103212

van Boxtel, J. J. A., Tsuchiya, N., & Koch, C. (2010). Consciousness and Attention: On Sufficiency and Necessity. *Frontiers in Psychology*, *1*. https://doi.org/10.3389/fpsyg.2010.00217

Varela, F. J. (1996). Neurophenomenology: A methodological remedy for the hard problem. *Journal of Consciousness Studies*, *3*(4), 330–349.

Weatherson, B. (2023). The Problem of the Many. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2023). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2023/entries/problem-of-many/

Wiese, W., & Friston, K. J. (2021). Examining the Continuity between Life and Mind: Is There a Continuity between Autopoietic Intentionality and Representationality? *Philosophies*, *6*(1), Article 1. https://doi.org/10.3390/philosophies6010018

Woodward, J. (2005). *Making Things Happen: A Theory of Causal Explanation* (1st edition). Oxford University Press.

Zaeemzadeh, A., & Tononi, G. (2024). *Upper bounds for integrated information* (arXiv:2305.09826). arXiv. https://doi.org/10.48550/arXiv.2305.09826

# *Chapter 2*
# Why does time feel flowing?
# Towards a principled account of temporal experience[5]


## Abstract

Time flows, or at least the time of experience. Can we provide an objective account of why experience, confined to the short window of the conscious present, encompasses a succession of moments that slip away from *now* to *then*—an account of why time feels flowing? Integrated Information Theory (IIT) aims to account for both the presence and quality of consciousness in objective, physical terms. Given a substrate's architecture and current state, the formalism of IIT allows one to unfold the cause–effect power of the substrate in full, yielding a cause–effect structure. According to IIT, this accounts in full for the presence and quality of experience, without any additional ingredients. In previous work, we showed how unfolding the cause–effect structure of non-directed grids, like those found in many posterior cortical areas, can account for the way space feels—namely, *extended*. Here we show that unfolding the cause–effect structure of directed grids can account for how time feels—namely, *flowing*. First, we argue that the conscious present is experienced as flowing because it is composed of *phenomenal distinctions* (*moments*) that are directed, and these distinctions are related in a way that satisfies *directed inclusion*, *connection*, and *fusion*. We then show that directed grids, which we conjecture constitute the substrate of temporal experience, yield a cause–effect structure that accounts for these and other properties of temporal experience. In this account, the experienced present does not correspond to a process unrolling in "clock time," but to a cause–effect structure specified by a system in its current state: time is a structure, not a process. We conclude by outlining similarities and differences between the experience of time and space, and some implications for the neuroscience, psychophysics, and philosophy of time


## 1. Introduction

Why does an experience feel the way it does? On a morning walk, you hear the notes of the bird's song succeed one another and see it fly across the blue expanse of the sky. Can we provide a scientific account for the quality of consciousness, including the feeling of why time feels flowing, space feels extended, and the sky feels blue?

Integrated Information Theory (IIT) (Albantakis et al., 2023; Oizumi et al., 2014; Tononi, 2004, 2008) aims to do just this: to provide a principled and comprehensive account of phenomenal properties in physical terms. First, it identifies the essential properties of consciousness—those that are true of every conceivable experience. It then formulates these properties in physical, operational terms. In doing so, IIT provides the tools to identify the substrate of consciousness (a *complex*) and unfold its cause–effect power (the *cause–effect structure* it specifies, composed of *causal distinctions* and *relations*). According to IIT, the cause–effect structure specified by a given

[5] This chapter corresponds to the near finalized manuscript of the forthcoming article: **Comolatti, Renzo***; Grasso, Matteo*; Tononi, Giulio. *Why does time feel flowing? Towards a principled account of temporal experience*. *These authors contributed equally to this work

substrate in its current state is sufficient—without additional ingredients—to fully account for the quality (content) and quantity of experience. In a previous paper, we showed how the cause–effect structures specified by undirected grids can account for the feeling of extendedness that characterizes spatial experiences (Haun & Tononi, 2019). In this paper, we employ the formalism of IIT to account for why time feels flowing.

The time under inquiry here is not what is measured by clocks ("clock time") but the subjective time of experience, the feeling of a short window of a conscious "present," composed of moments that succeed one another, which slide by from "now" to "then" and vanish into the past, and may include a feeling of what will come "next."

We start from the basic phenomenology of time and characterize its fundamental properties: a temporal experience is a kind of phenomenal structure, called a *phenomenal flow*, composed of distinctions and relations characterized by *directedness*. We then propose an account of phenomenal flow in physical terms, where "physical" is understood in a purely operational sense (manipulations and observations on a substrate), yielding a transition probability matrix (TPM). Specifically, we show that a certain kind of substrate—namely, a directed grid—specifies a cause–effect structure that can account in full for the phenomenal properties of temporal experience.

## 2. Phenomenology of time

Like the feeling of spatial extendedness, the feeling of temporal flow is not only pervasive in our experience but also partially penetrable. In other words, unlike, say, the feeling of color or pain, we can partially dissect the basic structure of phenomenal time through introspection (Haun & Tononi, 2019), even though its fleeting nature makes it more difficult to characterize than phenomenal space.

Below, we highlight some fundamental features of phenomenal time that we intend to account for. Consider the temporal phenomenology of hearing a melody—say, the first few notes of Beethoven's Fifth Symphony (Fig. 1), abstracting away from the phenomenal qualities of sound (we might be listening to another melody, or to speech, or even to silence, and phenomenal time would still be flowing).

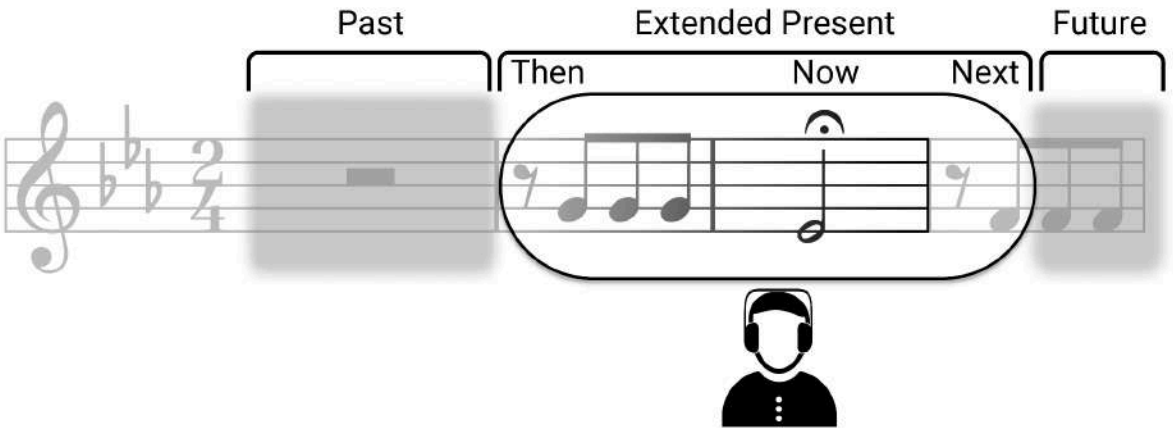



**Figure 1.** A depiction of temporal experience. The bubble indicates the content of a single experience, whose content is represented, for convenience, by the musical score—a few bars of Beethoven's Symphony no. 5. The experience can be triggered by clicking the play button at this page until the note E is perceived (the fourth note). The *present* is experienced as *extended*, i.e. as having a duration (e.g., the four notes played). Moments within the extended

present (associated with notes, pauses, and their combinations) are more vivid towards the *now* (e.g., the E note still sounding) and progressively fade towards the *then* (e.g., the three G notes just played). The silence before the first note (grayed out on the left) has vanished from experience into what we call the past, although it may be summoned within experience by recalling it. The extended present may include a feeling of what will come next (the upcoming notes), typically less vividly. Beyond that lays what we call the future.

First, our experience comprises an *extended present*, structured by phenomenal distinctions, called *moments*, which are related in a special way. The present is extended in the sense that we hear the melody, rather than a single note: our experience contains the note we heard just *now* together with a few other notes we heard just *then* (moments ago, but still *present* in our consciousness), and may contain the feeling of what will come *next* (the note that we will hear in a moment, already present in our experience, albeit less vividly). On the other hand, we do not experience notes outside the present, whether in the future, beyond the next (Fig. 1, right), or in the past, beyond the then (Fig. 1, left). The notion of the extended or "specious" present was popularized by William James as "the short duration of which we are immediately and incessantly sensible" (James, 1890), borrowing from E. Robert Kelly (Clay, 1882).

Second, the present is structured by phenomenal relations that make it feel directed, yielding a sense of flow: Within the present, we experience moments, such as those corresponding to individual notes or pauses, that appear to "flee away," directed towards what we call the past. The moments that compose the present are bound together by directed relations that order them, as we shall see, according to inclusion, connection, and fusion. These relations yield an experience of succession, rather than a succession of experiences (James, 1890).

Since the experience of the extended present does not necessarily include the feeling of what will come next, our account will mainly focus on the flow of time from the now to the then—what in the philosophical literature has been referred to as "retention" (Husserl, 1991). Nonetheless, in the Discussion we will show how this approach can also account for the experience of "protention", occasionally extending to what we feel will come next.

## 2.1. Moments

Let us dissect the phenomenology of temporal flow in more detail and introduce some nomenclature (Fig. 2A). As already mentioned, the phenomenal distinctions that compose phenomenal time are called *moments*. Moments can be as short as an instant (the shortest moment one can phenomenally resolve), as long as the entire present, or anything in between. They can be close to the conscious *now*, to the conscious *then*, or anytime in between. There are various estimates about the duration of the conscious present, from a few hundred milliseconds up to three seconds of clock time (Dainton, 2023; Pöppel, 2009), and of the grain of a conscious instant, typically a few tens of milliseconds (Herzog et al., 2016; White, 2018). But the moments composing the conscious present—short and long, now and then—only feel like moments flowing in time owing to the *relations* that bind them together. In what follows, we argue that four fundamental properties, characterizing moments and their relations, are necessary and sufficient for the experience of time: *directedness*, *directed inclusion*, *directed connection*, and *directed fusion* (Fig. 2B).

## 2.2. Directedness

Moments are directed, each of them pointing away from itself—flowing away from the now and towards the

then. Directedness implies that there is a fundamental asymmetry between the now and the then. For this reason, we represent the phenomenology of moments in time by arrows (pointing away from the now towards the then).

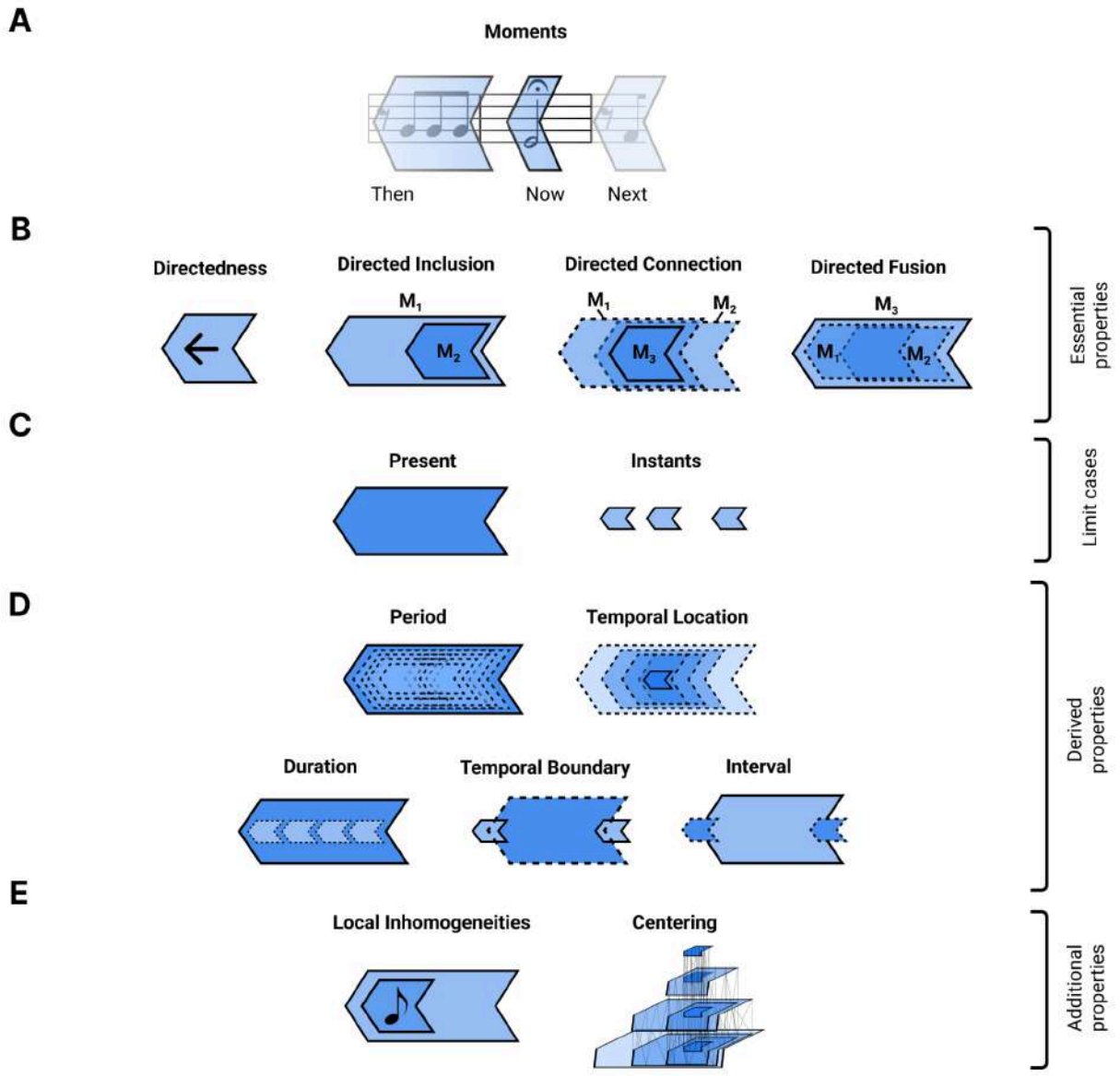



**Figure 2.** Phenomenology of temporal experience. (A) The distinctions composing the phenomenal structure of temporal experience are called moments (blue block arrows). (B) The fundamental relations of the structure of temporal experience are directedness and directed inclusion, connection, and fusion: (i) Moments are directed; they point away from themselves. (ii) Moments include and are included by other moments and do so in a directed manner, either forward (towards the now, in the case depicted) or backward (towards the then). (iii) Moments connect when they partially overlap each other in a directed way, such that one is the predecessor and the other the successor, and their overlap is also a moment. (iv) Moments that connect also fuse with one another such that their union is also a moment that includes them in a directed way and nothing else. (C) Limit cases of moments that satisfy only some of the fundamental properties of temporal experience: The present is the "total" moment; it includes all other moments but does not connect nor fuse up with other moments. By contrast, instants are the shortest moments that are phenomenally resolvable in experience; they are included by all other moments but do not connect nor fuse down. (D) Derived properties of temporal experience: The period covered by a moment is given by the

set of moments it includes, its location by the moments that include it, and its duration by the number of instants it includes. The boundary of a moment is the set of shortest moments (a predecessor and a successor) that connect to it; conversely, the interval between two moments consists of the shortest moment that connects to both of them. (E) Additional properties: Local inhomogeneities are moments that stand out because of a difference in local qualities (say, a tone that breaks the silence or a pause that breaks a sound). Centering refers to the feeling that we are anchored to the now, placed towards the end of the extended present. The period preceding the now towards the then is experienced past, the period succeeding the now towards the next is experienced future. The hierarchy of distinctions and relations corresponding, for example, to feelings of self and agency, is strongly bound to the now, enhancing its vividness and centering the temporal experience.

## 2.3. Directed inclusion

For any moment, there are always other moments which include it or are included by it. By virtue of being directed, inclusion can be of two kinds: moments can be included towards the now (*forward inclusion*) or towards the then (*backward inclusion*). Forward inclusion is such that the included moment is a subset of the including moment but is aligned on the latest instant on which they both overlap: they "share their ending"—the last instant towards the now they both include. Similarly, backward inclusion is such that the included moment is a subset of the including moment and is aligned on the earliest instant on which they both overlap: they "share their beginning"—the first instant towards the then they both include. Directed inclusion captures the fact that every moment feels nested within the structure of the present, encompassing a certain period (determined by the moments it includes) and having a temporal location within the present (determined by the moments that include it).

## 2.4. Directed connection

For any moment we can always find *predecessor* moments that overlap it partially and asymmetrically towards the then, and *successor* moments that overlap it partially and asymmetrically towards the now. Directed connection is asymmetric because there is an intrinsic ordering within *phenomenal* time: a moment that is connected to a successor moment cannot be its successor but only its predecessor, and a moment connected to its predecessor cannot be its predecessor, only its successor. When two moments overlap, there is always another moment that covers exactly their overlap—the *connecting* moment (or *connection*). This third moment is included by both in a directed way, such that the connecting moment is *forward-included* by the predecessor and *backward-included* by the successor. For any two overlapping moments, one can always find a moment they connect onto (*directed connection down*). Moreover, every moment is also the connection of two overlapping moments, such that it is included by both of them and covers their overlap (*directed connection up*). Directed connection accounts for the directed ordering of moments within the present according to relations of *succession* and *predecession*.

## 2.5. Directed fusion

For any moment, one can always find another connected moment with which it *fuses*, such that together they compose a third moment that includes both of them in a directed way (either *backward* or *forward*) and coincides with their union (*directed fusion up*). Every moment is also the fusion of two connected moments (*directed fusion down*), one towards the now and the other towards the then, such that it includes both of them and coincides with their union. Fusion accounts for the *fullness* of the present—that phenomenal time is not

fragmented.

## 2.6. Derived properties

The fundamental properties just described apply to all moments, with the exception of the "total" moment, corresponding to the conscious present, which includes all other moments but neither connects nor fuses up with other moments. Likewise, instants, the finest-grain moments one can phenomenally resolve, do not connect and do not fuse down, being only included by other moments (Fig. 2C).

The fundamental properties of temporal flow are sufficient to derive other phenomenal properties of experienced time (Fig. 2D). Thus, the *period* covered by a moment can be defined as the set of all moments it includes, while its *temporal location* is the set of all moments that include it, placing it uniquely within the present and among other moments. The *duration* of a moment can be characterized as the number of instants it includes. The *boundary* of a moment is the set of the shortest moments that are directedly connected to its beginning and its end—that is, its shortest predecessor and successor. To note, the present itself (the "total" moment) is not experienced as having a boundary that marks its beginning and its end. Even so, the present is definite because it has a limit: it starts with the now and does not extend into the future, and it ends with the then, not extending into the past. Finally, the *interval* between any two moments is the shortest moment that connects to both of them.

## 2.7. Inhomogeneities and centeredness

Some other phenomenal properties tightly bound to the experience of temporal flow should also be mentioned (Fig. 2E). Within the present, one or more moments can stand out because of an inhomogeneity in local properties. These are properties, such as sound or touch, that are not in themselves temporal but are typically experienced as bound to time. For instance, a sudden sound may pierce the silence (e.g., the first note in Fig. 2A), or a sudden pause interrupting a droning noise. These inhomogeneities highlight particular moments in the flowing present, without disrupting its flow, only warping it locally and often capturing our attention. But time flows in perfect silence too, say during the expressive pauses at the end of Sibelius's fifth symphony, or throughout the provocative emptiness of John Cage's piece 4'33.

While experienced time always flows from the now to the then, another prominent phenomenal property is that we usually feel *centered* in the now (rather than in the then or in the middle of the present): when we hear a sound, it suddenly appears in the now, we experience it as the "latest" and "most vivid" event we are aware of, and we feel that whatever happened before (but is still present in the experience) is less vivid or faded compared to it. For example, in hearing Beethoven's Fifth Symphony, the latest note is present in "full sound," while the group of three notes in the previous bar is still present but fainter (Fig. 2A). The now is the moment that is typically bound to actions: we typically feel that, when we act, we are doing so from the "now" rather than from the "then."

# 3. Methods

We now proceed to lay out an account of the subjective properties of temporal flow in objective, operational

terms, according to the principles of IIT. This means that phenomenal distinctions and relations that compose temporal flow—the moments bound by directedness, directed inclusion, directed fusion, and directed connection—must have a correspondent in the cause–effect structure specified by the substrate of temporal experience in the brain, in the causal distinctions that compose it and the way they relate.

Below, we first briefly summarize the IIT formalism (for a complete presentation see Albantakis et al. (2023) and the "Integrated Information Wiki," (2024)). Next, we apply the formalism to unfold the cause–effect power to a directed grid—the kind of substrate that, we conjecture, can support the experience of time.

## 3.1. Unfolding cause–effect structures

IIT starts by applying the postulates of *existence*, *intrinsicality*, *information*, *integration*, and *exclusion* and identifying a maximum of *system integrated information* ($\varphi_s$) over the units of a substrate (Marshall et al., 2023). By IIT a substrate of consciousness, or *complex*, must be such a maximum. Next, in line with the postulate of *composition*, the cause–effect power of the complex is unfolded in full, yielding its *cause–effect structure*. By IIT, the causal distinctions and relations that compose the cause–effect structure account for the content of the corresponding experience, with no additional ingredients. Here we focus on unfolding the cause–effect structure specified by a directed grid, assuming that the grid is part of a larger complex. Every system subset that satisfies IIT's postulates of physical existence (except for composition, which does not apply to the components themselves) specifies a causal *distinction*. A distinction consists of a *mechanism* (a subset of units in a state) that specifies a *cause purview* and an *effect purview* (each a subset of units in a state). Overlaps among causes and/or effects of one or more distinctions specify *causal relations*. (Fig. 3A, top left).

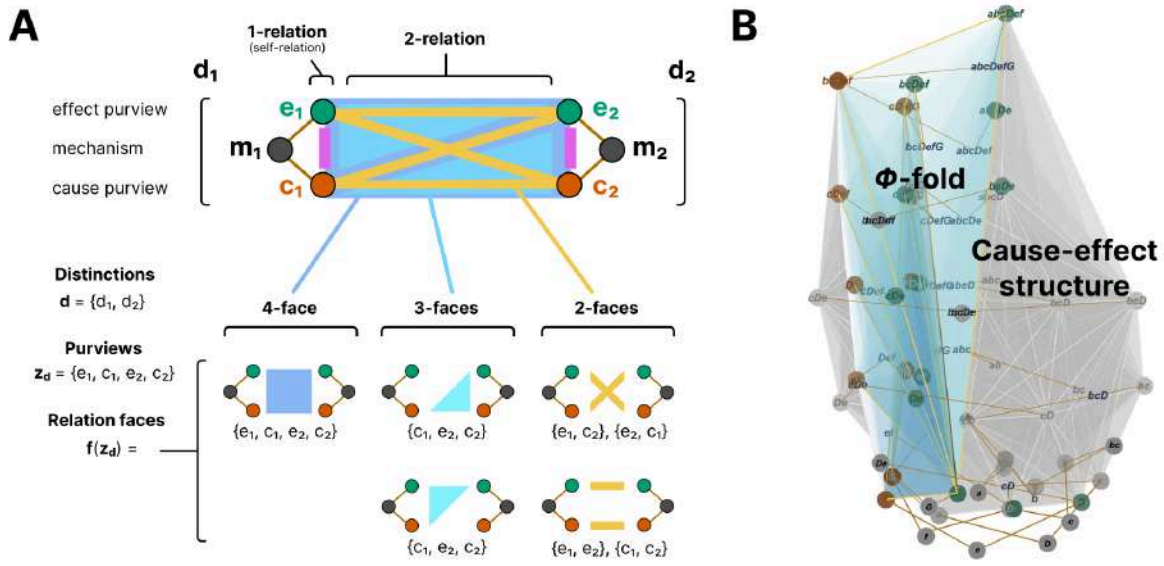



**Figure 3**. A cause–effect structure and its components. (A) A cause–effect structure is composed of causal distinctions and relations. Distinctions are specified by irreducible mechanisms (subsets of the substrate's units, indicated by black circles) linking a cause and an effect purview (red and green circles, respectively) over subsets of the substrate's units. Causal relations obtain when there is a congruent (same unit state) overlap between two or more purviews specified by one or more distinctions. The degree of a relation is the number of distinctions involved in the overlap, while the degree of a relation face is the number of purviews contributing to it. Second-degree faces (or 2-faces) are depicted as edges (yellow and magenta), and higher-degree faces are depicted as surfaces (blue). Shown are two generic distinctions, with their 1- and 2-relations and the 2-, 3- and 4-relation faces they could have. (B) Causal

distinctions and relations compose a cause–effect structure (depicted in gray), from which sub-structures (or $\Phi$-folds, in blue) can be isolated. For a full description of distinctions, relations, and cause–effect structures and how to compute them using the formalism of IIT 4.0, see Albantakis et al. (2023) and the "Integrated Information Wiki" (2024).

### 3.1.1. Distinctions

A mechanism specifies a causal distinction if (i) it has cause–effect power (existence postulate), that is, it can take and make a difference with respect to itself or other units; (ii) it has cause–effect power within the system (intrinsicality postulate); (iii) its cause–effect power is specific (information postulate), that is, being in its specific current state, it selects a state for its purviews (the one with maximal *intrinsic information* $\boldsymbol{ii}$ on the input side for cause purview and on the output side for effect purview), and this state is congruent with the *cause–effect state* selected by the complex as a whole; (iv) its cause–effect power is irreducible (integration postulate), that is, the distinction's *integrated information* ($\boldsymbol{\varphi_d}$) (the amount of $\boldsymbol{ii}$ lost by partitioning mechanism and purview) is positive and the minimum across all partitions for a candidate purview; and (v) the amount of $\boldsymbol{\varphi_d}$ it specifies is maximal across other candidate purviews (exclusion postulate). In sum, a causal distinction comprises a mechanism linking a cause purview and an effect purview, and has an associated $\boldsymbol{\varphi_d}$ value.

### 3.1.2. Relations

Causal relations capture the way in which causal distinctions are bound together within a cause–effect structure. There is a relation if cause and effect purviews overlap congruently (i.e., they specify the same state) over a subset of their units (Fig. 3A). The purviews specified by a set of distinctions can overlap in different ways, depending on whether the overlap involves causes, effects, or both, and on the number of purviews that overlap. Each of the purview overlaps in a relation is called a *face*. The unit in the overlap are the *face purview*, with a corresponding face irreducibility value $\boldsymbol{\varphi_f}$. The union of the face purviews constitutes the *relation purview*. The relation irreducibility value ($\boldsymbol{\varphi_r}$) is calculated by unbinding one distinction at a time and finding the one that makes the least difference. This is calculated by multiplying the average $\boldsymbol{\varphi_d}$ per unique purview unit by the size of the overlap across all faces (the number of units in the relation purview) and taking the minimum value across distinctions in the relation.

A relation that binds $n$ distinctions is called an $n^{\text{th}}$-degree relation (or $n$-relation for short) and a face that binds $k$ purviews within a relation is called a $k^{\text{th}}$-degree face (or $k$-face for short). For instance, given two distinctions $\boldsymbol{d} = [d_1, d_2]$, each with a cause and an effect purview, we have a set of four purviews $\boldsymbol{z_d} = [e_1, c_1, e_2, c_2]$ (where $e$ and $c$ stand for effect and cause purview, respectively; Fig. 3A). There are nine potential relation faces across the two distinctions: one 4-face involving all four purviews; four 3-faces involving three purviews (either two effects and one cause, or two causes and one effect); and four 2-faces involving two purviews (either a cause and an effect, two causes, or two effects). Finally, there are two potential 1-relations (a self-relation between the cause and effect of each distinction).

Together, distinctions and relations compose the cause–effect structure (Fig. 3B, in gray). Here we will limit our analysis to 2-relations and their underlying set of faces, in particular 2-faces, which are sufficient to characterize the cause–effect structure corresponding to temporal flow.

### 3.1.3. Contexts

To study the contribution of individual distinctions, it is useful to decompose the cause–effect structure into sub-structures or *Φ-folds*. A relevant sub-structure is the distinction's *context*—the set of relations bound to it (Fig. 3B, in blue). Through its context, a distinction is related to a set of other distinctions within the cause–effect structure. More specifically, the *purview context* is the set of relations involving a distinction's purview (either cause of effect). In accounting for the properties of temporal phenomenology in terms of properties of the cause–effect structure, we will present the correspondence both at the "local" level of relation faces between pairs and triples of distinctions (restricting our analysis to 2-relations among them) and at the "global" level of *Φ*-folds corresponding to relation contexts.

## 3.2. Causal model of the substrate: a directed 1D grid of binary units

Which brain mechanisms and regions may support the experience of the flow of time is not known. Here we conjecture that the experience of time is supported by brain regions harboring connectivity patterns resembling directed grids. Such connectivity may be found, for instance, within the auditory cortex.

The substrate model employed in this paper is a 1D grid, assumed to be part of a larger complex, comprising seven probabilistic units *AbcDefg* with binary state (-1, or OFF, indicated with lowercase, and +1, or ON, indicated with uppercase; Fig. 4A). This is considered as a "macro" state, corresponding to an interval of the order of 30 milliseconds or so of clock time (see below). Each unit has a self-connection (weight of $w = 0.3$), a stronger outgoing lateral connection ($w = 0.6$) to one of its two neighboring units, and a weaker outgoing lateral connection ($w = 0.1$) to the other neighboring unit. Each unit also receives a feedforward input from a sensory interface (*input array*; Fig. 4A, bottom), assumed to be outside the complex, also comprising seven units. The input array not only provides bottom-up inputs that drive the activation of the units of the 1D grid, but works as a delay line, such that activation percolates sequentially from unit $A'$ to unit $G'$.

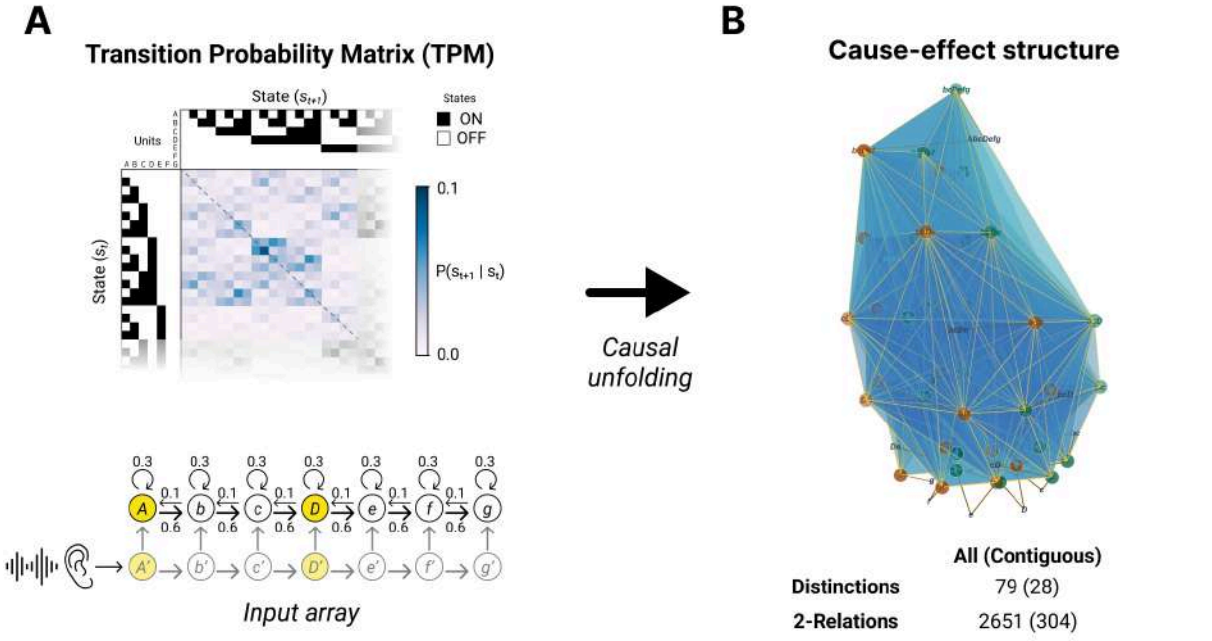


**Figure 4**. Substrate model of a directed 1D grid and its cause–effect structure. (A) Below, the substrate consisting of a directed 1D grid with seven probabilistic units, where the binary state is represented by -1 (OFF) in lowercase and +1 (ON) in uppercase. Each unit is characterized by a self-connection weight of $w = 0.3$, an outgoing lateral connection with a weight of $w = 0.6$ to one neighboring unit, and with a lesser weight of $w = 0.1$ to the other neighbor. The substrate is assumed to be part of a larger complex. Outside the complex is an input array conveying sensory input. The input array functions as a delay line, percolating activations from the ear from unit *A′* to *G′*. The input array also drives the activation state of the directed grid, which "endorses" its driven state through its self-and lateral connections that undergo short-term plasticity. Above, the associated transition probability matrix (TPM) which contains all information needed to unfold the cause–effect structure of the substrate model. Each state $s_t$ (rows) can transition to a state $s_{t+1}$ (columns) with probability $P(s_{t+1} | s_t)$. The binary states are represented as blocks (+1 as black, -1 as white) and only the first twenty states are shown. (B) Unfolded cause–effect structure of the seven-unit directed grid. Each distinction consists of a mechanism (black units) linked by brown lines to its cause (red units) and its effect (green units). Lower-order distinctions are depicted towards the bottom and higher-order distinctions towards the top. Only 1st- and 2nd-degree relations are plotted, with 2nd-degree faces depicted as edges (yellow) and higher-degree faces depicted as surfaces (blue).

The 1D grid (*A* to *g*) does not percolate activity patterns on its own, but "endorses" the activity macro state driven by the input array through an activation function that is the combination of two sub-functions (see Mayner et al. (*in preparation*) for details on their implementation). The first function $f_1(x_k, s_k)$ assures that grid units are reliably turned ON and OFF if the feedforward sensory input is ON and OFF, respectively. If the unit's current state $s_k$ differs from the sensory driving input $x_k$, the unit's state flips. The second function determines the state of each grid unit as a function of the inputs it receives through its lateral and self-connections. Each unit implements a sigmoid function of an input state $\boldsymbol{I}^*$ parametrized by the current state of the unit itself $s_k$, the connection weight $w_k$, and the current state of its input units $\boldsymbol{I}$:

$$\sigma(I^*; s_k, w_k, I) = \frac{1}{(1+\exp\left[-s_k \sum_{j=1}^{|I|} I_j w_{k,j} I_j^*\right]}$$

This makes connections to a unit that is ON (+1) effectively excitatory, and connections to a unit that is OFF (-1) effectively inhibitory. The state-dependent nature of this function ensures that each unit's state is endorsed by the lateral connections by adjusting the effective sign of the input to the unit (assumed to be mediated by short-term plasticity, see Mayner et al. (*in preparation*)). The two functions are combined to obtain the probability of a unit turning ON by taking the one that deviates maximally from chance (i.e., the "maximally selective" one):

$$Pr(k = ON) = argmax_{p \in \{f_1(x_k, s_k), \sigma(I^*; s_k, w_k, I)\}} |p - 0.5|$$

As a result, the macro state of the grid is driven by the sensory input array, while simultaneously allowing the units to endorse their current state by rapidly adjusting the strength of their intrinsic connections (at a faster time scale than that of the units' macro state).

## 4. Results

According to IIT, the properties of the cause–effect structure specified by the substrate of consciousness account in full for the phenomenal properties of the experience of time. We must thus establish a correspondence between the fundamental properties of temporal experience and the properties of the cause–effect structure unfolded from directed grids (Fig. 4).

## 4.1. Moments

The phenomenal distinctions composing the experience of time are moments. In physical terms, these correspond to causal distinctions specified by first- or higher-order mechanisms of directed grids. Out of 127 possible mechanisms for a 7-unit grid, 79 are irreducible and therefore specify causal distinctions (Fig. 4B). Nearly a third of the distinctions are specified by contiguous units, as depicted in Fig. 5 (right), and these will be the focus for the account below.

## 4.2. Directedness

Phenomenally, moments are characterized by directedness: each moment points away from itself. In the cause–effect structure specified by directed grids, this corresponds to distinctions whose cause and effect do not overlap or overlap only partially and asymmetrically: each purview always has at least one element that is not included in the other, with the causes leaning towards the *now* and the effects towards the *then* (Fig. 5). For instance, distinction *bc* has a cause over *b* and an effect over *c*. Distinction *cDef* has a cause over *cDe* and an effect over *Def*. This results in further asymmetries in the way causes and effects relate to the rest of the structure.

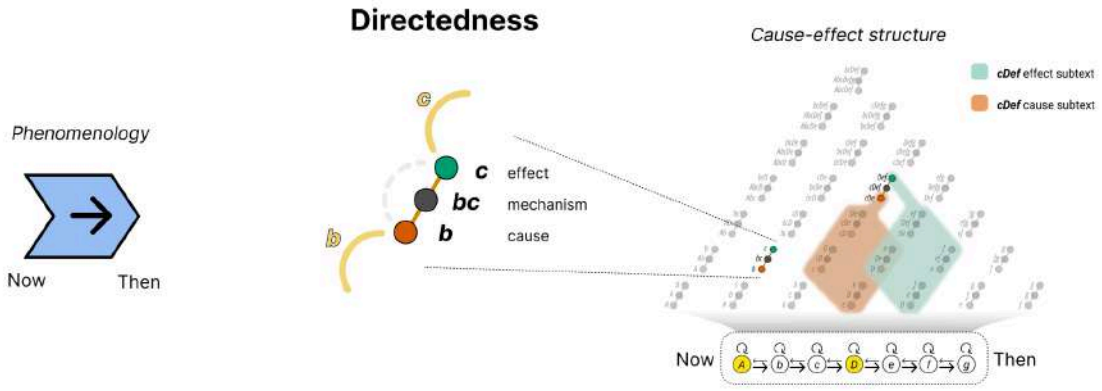



**Figure 5.** Directedness. Phenomenal distinctions in temporal experience are called moments. Moments are fundamentally directed, pointing away from themselves (left panel). Moments are the basic "building blocks" of phenomenal flow. In physical terms, they correspond to the causal distinctions specified by the seven-unit directed grid (right) in state *AbcDefg* (as before, ON units are represented with uppercase and OFF units with lowercase). Causal distinctions comprise the mechanism (black) linked to its cause (red) and effect purviews (green). The directedness of moments corresponds to causal distinctions that are also directed: causes and effects are misaligned asymmetrically, such that each contains elements not contained in the other, with causes leaning towards the *now* (left direction) and effects towards the *then* (right direction). For example (center panel), the distinction *bc* has *b* as its cause and *c* as its effect. Thus, the cause of *bc* can relate to distinctions over unit *c* while its effect *c* cannot, whereas its effect can relate to other distinctions over unit *c* which its cause *c* cannot. Directedness applies to all other distinctions and can also be seen in terms of the contexts of the distinctions (right panel). For example, distinction *cDef* is directed such that its cause subtext (i.e. the distinctions, highlighted in red, whose purviews are included in its cause *cDe*) is different from its effect subtext (i.e. the distinctions, highlighted in green, whose purviews are included in its effect *Def*). Note: to align phenomenology to the orientation of the directed grid (which receives the "latest" inputs from the left), from here on we flip the convention used in the previous figures and depict the *now* on the left and the *then* on the right.

## 4.3. Directed inclusion

Phenomenally, directed inclusion captures the fact that every moment includes and is included by other moments in a directed way, both towards the now (forward inclusion) and towards the then (backward inclusion).

The cause–effect structure specified by a directed grid has properties that account for phenomenal directed inclusion, because its distinctions include and are included by other distinctions in a directed way (Fig. 6). Specifically, for distinctions that qualify as moments, there is another distinction that includes them (one that overlaps them fully, non-mutually, and asymmetrically), and there is another distinction that they include (one that is fully, non-mutually, and asymmetrically overlapped by them). For each including distinction, there is one distinction that is forward-included and another that is backward-included by it. Forward inclusion is when the elements of both the cause and effect purviews of the included distinction are a subset of the elements constituting the *cause* purview of the including distinction (it is called "forward" because the included distinction is on the side of the now). For example, distinction *bcDef* forward-includes distinction *bcDe* because the elements of *bcDe*'s cause and effect are included in the elements of *bcDef*'s cause (Fig. 6A, center). Backward inclusion is when the elements of both the cause and the effect purviews of the included distinction are a subset of the elements constituting the *effect* purview of the including distinction (it is called "backward" because the included distinction is on the side of the then). For example, distinction *bcDef* backward-includes distinction *cDef* because the elements of *cDef*'s cause and effect are included in the elements of *bcDef*'s effect (Fig. 6B, center).

Directed inclusion is true not only for individual pairs of distinctions but also for their relations with the rest of the structure—that is, with their *context* (see above). The *subtext* and *supertext* of a distinction (or its purviews) can be defined as the set of distinctions (or purviews) included by it and including it, respectively (Haun & Tononi, 2019). A distinction is forward-included when its subtext is fully included in the cause subtext of a larger distinction, but it is only partially included in the larger distinction's effect subtext (Fig. 6A, right), and *vice versa* for backward inclusion (Fig. 6B, right).

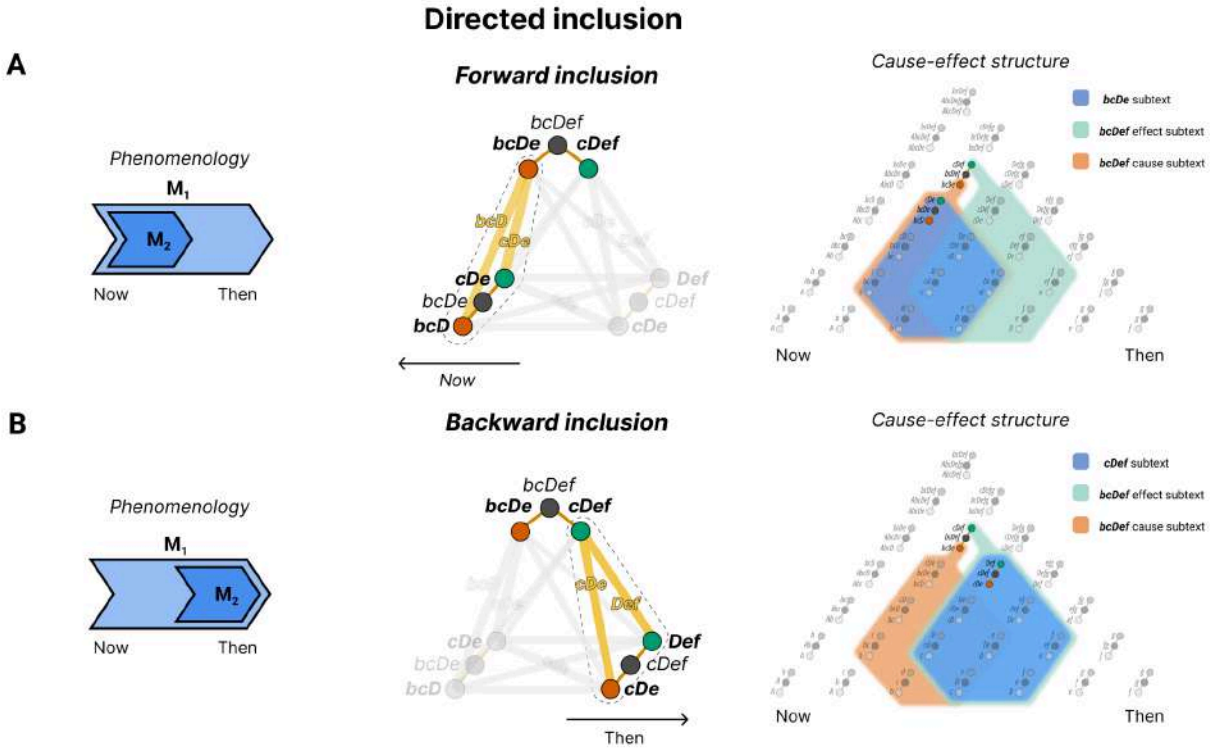


**Figure 6.** Directed inclusion. In temporal experience, moments include and are included by other moments, which can occur towards the *now* (forward inclusion, panel A left) or towards the *then* (backward inclusion, panel B left). In the cause–effect structure, directed inclusion corresponds to a distinction including other distinctions (both their cause and effect) aligned on their cause (forward inclusion, panel A center) or on their effect (backward inclusion, panel B center). This is reflected by the presence of two 2-faces within the 2-relation binding the cause (or effect) of the including distinction to both the cause and effect of the included distinction (center, top and bottom). In the example, distinction *bcDe* is forward-included by distinction *bcDef* because *bcDe*'s cause (*bcD*) and effect (*cDe*) are included in distinction *bcDef*'s cause (*bcDe*) (panel A middle), while distinction *cDef* is backward-included because its cause (*cDe*) and effect (*Def*) are included in distinction *bcDef*'s effect (*cDef*) (panel B middle). This relation of directed inclusion is also reflected at the level of the context of distinctions. In forward inclusion (backward inclusion), the subtext of the included distinction is fully included in the subtext of the cause (or effect) of the including distinction (right, panel A/B bottom). The subtext of a distinction (shaded regions in the cause–effect structure) consists of all distinctions whose purviews it includes (via its cause and/or effect purviews). In the example, *bcDe*'s subtext is included in *bcDef*'s cause subtext only (top right), illustrating that *bcDe* is forward-included by *bcDef*, while *cDef*'s subtext is included in *bcDef*'s effect subtext only (bottom right), illustrating that *cDef* is backward-included by *bcDef*.

## 4.4. Directed connection

Phenomenally, directed connection captures the fact that every moment has a predecessor moment that overlaps it partially and asymmetrically towards the then, and a successor moment that overlaps it partially and asymmetrically towards the now, and that the overlaps are also moments. This applies to all moments except for the ones starting in the now, which only have predecessors and no successors, and the ones ending in the then, which only have successors and no predecessors.

The cause–effect structure specified by a directed grid has properties that account for phenomenal directed connection because of the way its causal distinctions overlap asymmetrically with other distinctions (Fig. 7A). For each distinction that qualifies as a moment in the cause–effect structure, there is another distinction that overlaps it partially and asymmetrically, and there is another distinction that is included by both. Directed connection is asymmetric because the effect of one distinction overlaps the cause of the other in a way that is different from how the effect of the other distinction overlaps its cause. For instance, in Fig. 7A (center), distinction *bcDe*'s effect (*cDe*) overlaps distinction *cDef*'s cause (*cDe*) fully (over units *cDe*), whilst distinction *cDef*'s effect (*Def*) overlaps distinction *bcDe*'s cause (*bcD*) only partially (over unit *D*). Moreover, their overlap is also a distinction (*cDe*). This imposes a natural ordering between the two distinctions, such that *bcDe* succeeds *cDef* or, equivalently, *cDef* precedes *bcDe*. Note that distinctions are directed such that effects are towards the then and causes towards the now. This is a consequence of the connectivity of the substrate, and accounts for the feeling that what is experienced in the now (cause) flows towards the then (effect).

Directed connection also applies to each distinction's context. For example, the subtext of the connection distinction coincides with the cause subtext of the distinction that forward-includes it and with the effect subtext of the one that backward-includes it (Fig. 7A, right).

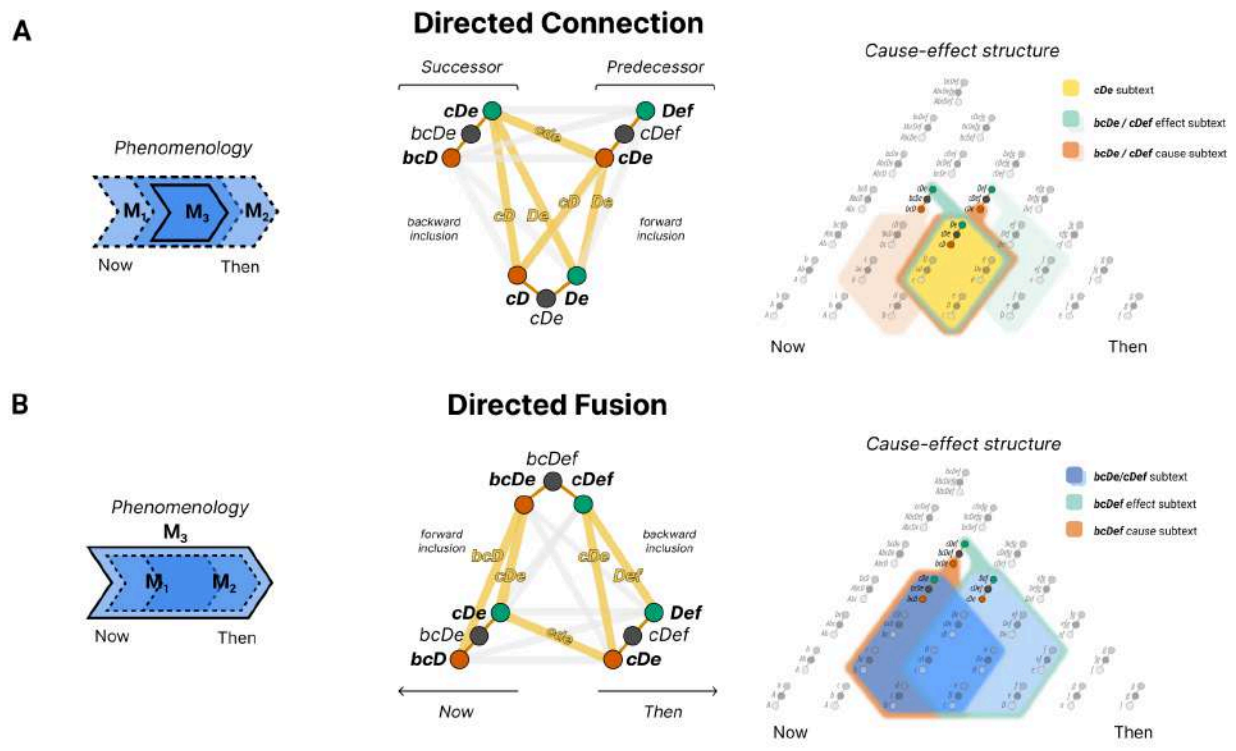



**Figure 7**. Directed connection and directed fusion. (A) Directed connection. Phenomenally, moments overlap partially, their overlap is directed (one feels more towards the *now* and one more towards the *then*), and their overlap is always a moment. Similarly, in the cause–effects structure, causal distinctions connect in a directed way: the effect of the distinctions closer to the now overlaps the cause of the other distinction closer to the then, in a way that is different from how the effect of the second distinction overlaps its cause. In the example (top center), distinction *bcDe*'s effect (*cDe*) overlaps distinction *cDef*'s cause (*cDe*) fully (over *cDe*), whilst distinction *cDef*'s effect (*Def*) overlaps distinction *bcDe*'s cause (*bcD*) only partially (over *D*). Moreover, their overlap is also a distinction (*cDe*) (their "connection"), which is included by them in a directed manner (backward and forwards). At the level of the context (top right), the intersection of the subtexts of the two connected distinctions (here, *bcDe* and *cDef*) coincides with the distinction subtext of their connection (*cDe*). (B) Directed fusion. Phenomenally, each moment is composed of two or more connected moments, and each moment together with other connected moments fuse to compose another moment. In the cause–effect structure, this corresponds to the fact that when distinctions connect they always fuse: for every distinction (e.g., *bcDe*) there is another distinction that includes that distinction (through either backward or forward inclusion, e.g., *bcDef*) plus another connected distinction (e.g., *cDef*) such that the union of the purview elements of the including distinction is equivalent to the union of the purview elements of the included distinctions. At the level of the context, the union of the subtexts of the two fusing distinctions (e.g., *bcDe* and *cDef*) coincides with the distinction subtext of their fusion (*bcDef*), with the fusion's cause subtext coinciding with the subtext of the distinction that is forward-included (*bcDe*), and the fusion's effect subtext coinciding with the subtext of the distinction that is backward-included (*cDef*).

## 4.5. Directed fusion

Phenomenally, directed fusion expresses how each moment is composed of two or more connected moments (which are its fusion down), and each moment together with other connected moments can fuse to compose another moment (which is their fusion up).

The cause–effect structure specified by a directed grid can account for phenomenal directed fusion (Fig. 7B). Specifically, for each distinction that qualifies as a moment (say, *bcDe*), there is another distinction (*bcDef*) that includes both it (through either backward or forward inclusion) and another distinction connected to it (*cDef*), such that the union of the purview units of the including distinction (*bcDef*) coincides with the union of the purview units of the included distinctions (fusion up) (Fig. 7B, center). Similarly, each distinction qualifying as

a moment includes one distinction (through either backward or forward inclusion) plus another distinction connected to it, such that the union of the purview units of the including distinction coincides with the union of the purview units of the included distinctions (fusion down).

Similar considerations apply to the context of the fusing distinctions (Fig. 7B, right). At the level of distinction contexts, the union of the subtexts of two fusing distinctions (e.g., *bcDe* and *cDef*) coincides with the distinction subtext of their fusion (*bcDef*). Moreover, the fusion's cause subtext coincides with the subtext of the distinction that is forward-included (*bcDe*), and the fusion's effect subtext coincides with the subtext of the distinction that is backward-included (*cDef).*

## 4.6. Flow

Phenomenally, the flow of time from the now to the then within the extended present can be understood as a structure composed of distinctions, or moments, that capture the fundamental properties of directedness (pointing away from themselves) and relations of directed inclusion, connection, and fusion. As we have seen, the cause–effect structure unfolded form a directed grid is composed of distinctions that are directed, include and are included in a directed way, connect in a directed way, and fuse in a directed way (Fig. 8). A cause–effect structure that satisfies these properties, called a *flow*, can therefore account for the fundamental phenomenal properties underlying the feeling of time flowing. As shown below, a flow can also account for phenomenal properties of temporal experiences that are derived from the fundamental ones.

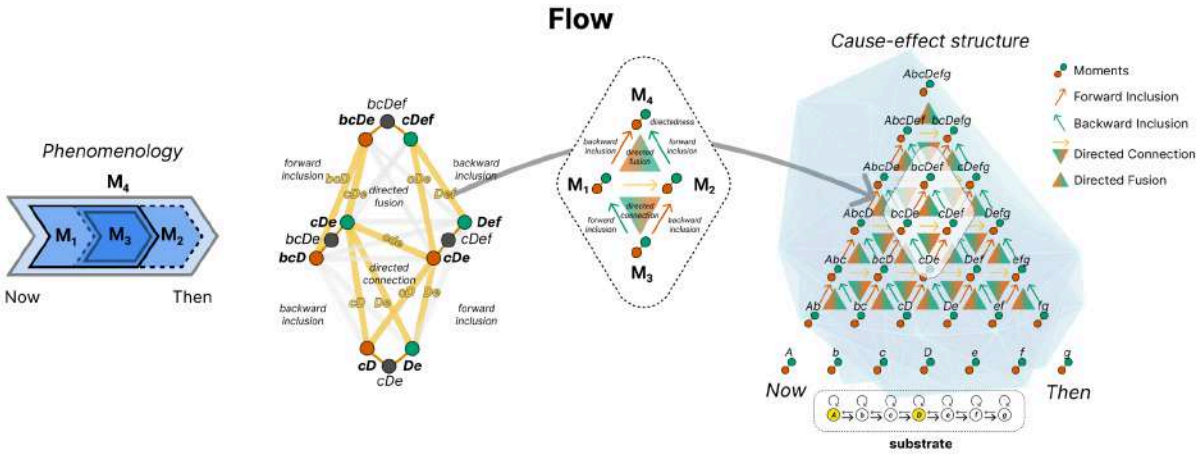



**Figure 8.** Flow. The phenomenal properties of temporal flow—namely directedness, directed inclusion, directed connection, and directed fusion (left)—correspond to properties of the cause–effect structure unfolded from a directed 1D grid (right). This is exemplified by four causal distinctions and the relations that bind them (second from left). All four distinctions (*bcDe, cDef, bcDef and cDe*) are directed, with their causes and effects not aligned. Distinction *bcDef* forward-includes distinction *bcDe* towards the *now* (since its cause includes *bcDe*'s purviews) and backward-includes distinction *cDef* towards the *then* (since its effect includes *cDef*'s purviews). Similarly, distinction *cDe* is forward-included by distinction *cDef* and backward-included by distinction *bcDe*. Distinction *bcDe* also has a partial asymmetric overlap with *cDef* (since *bcDe*'s effect fully overlaps *cDef*'s cause, but not the other way around), and they both connect on distinction *cDe* by backward-including it (in the case of distinction *bcDe*) and forward-including it (in the case of distinction *cDef*). Moreover, distinction *bcDe* and *cDef* fuse into distinction *bcDef*, being forward- and backward-included by it, respectively, such that the union of their purviews coincides with the union the purviews of *bcDef*. Taken together, the four distinctions satisfy the fundamental properties of temporal flow. This also holds for the other distinctions that compose the cause–effect structure, which can thus be considered a *flow*. The third panel from left summarizes the relations of directed inclusion, connection, and fusion as they apply between four distinctions corresponding to moments ($M_1$ through $M_4$), and the right-most panel shows how they apply between contiguous distinctions throughout the cause–effect structure (for simplicity only the label of the mechanisms are shown).

## 4.7. Derived properties

The phenomenal properties of temporal periods, temporal locations, durations, boundaries, and intervals, can be accounted for in physical terms by considering sub-structures of the cause–effect structure specified by a directed grid. The *period* of time picked out by a moment corresponds to the set of distinctions included by the corresponding distinction (its *subtext*; Fig. 9A, top left). Conversely, the temporal *location* of a moment is the set of all distinctions that fully include it (its *supertext*; Fig. 9A, top right). The *duration* of a moment is accounted for by the number of smallest distinctions (instants) included by a given distinction (Fig. 9A, bottom). The *boundary* of a moment is the set of distinctions with shortest duration that are in a relation of directed connection with it (Fig. 9B, left). An *interval* can be defined as the shortest moment that separates two moments—that is, the smallest distinction that connects to two distinctions (Fig. 9B, right).

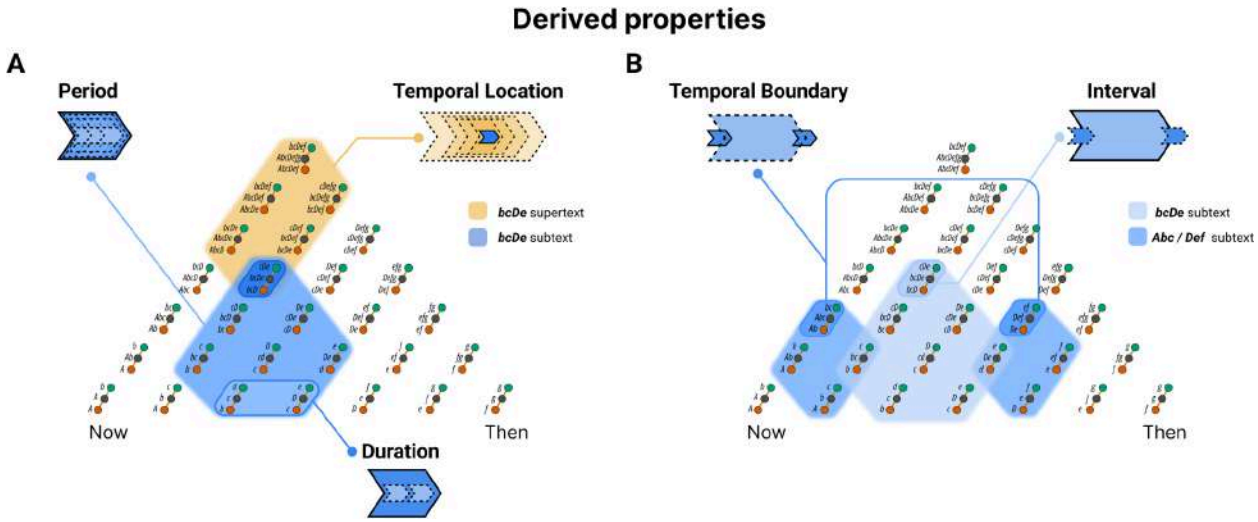



**Figure 9.** Derived properties: period, temporal location, duration, temporal boundary, and interval, and their correspondence in the cause–effect structure. (A) The period picked out by a moment is the set of distinctions included by it (its subtext, blue shading). The temporal location of a moment is the set of distinctions that fully include it (its supertext, yellow shading). The duration of a moment is the set of smallest distinctions (instants) included by it (blue contour). (B) The boundary of a moment is the set of smallest distinctions that connect to it (indicated in dark blue). The interval between two moments is the shortest moment that connects the two distinctions (indicated in light blue).

## 4.8. Inhomogeneities and centering

*Local inhomogeneities* within the flow of phenomenal time can occur whenever one experiences, for example, a sound that breaks the silence, or a pause in a series of tones. The activation or deactivation of specific units within a directed grid, accompanied by the interplay between higher-level and lower-level mechanisms connected to directed grids (Fig. 10A), can result in the local warping of phenomenal flow that "stands out" in its corresponding cause–effect structure (not shown). Even when warped, the cause–effect structure unfolded from a directed grid retains the fundamental properties that characterize a flow. As indicated in the figure, local qualities such as pitch, loudness, and timbre, would be accounted for by the sub-structures supported by neuronal "cliques" associated with the directed grid. Similarly, configurations of low-level features and invariants such as tones would be contributed by the convergent/divergent connectivity among higher-level areas.

As already mentioned, the experience of time flowing is typically characterized by the feeling that we are

*centered* in the now, with moments flowing away from it and towards the then. Moreover, the now is typically experienced more vividly than the then. A plausible explanation for these phenomenal features is that the neural mechanisms at the now terminus of the directed grid may be more densely connected to neural mechanisms in higher-level areas that eventually drive action (Fig. 10A). A denser connectivity implies a much larger number of causal relations. This would not only make adaptive sense, but would also account for the greater vividness of the now (see Albantakis et al. (2023) and Haun & Tononi (2019) for an account of vividness in terms of the number and irreducibility of distinctions and relations). It is also plausible that the neural substrate of sensory modalities characterized by shorter delays may serve to align experience across slower modalities, and to place the "now" of perception just before that of action.

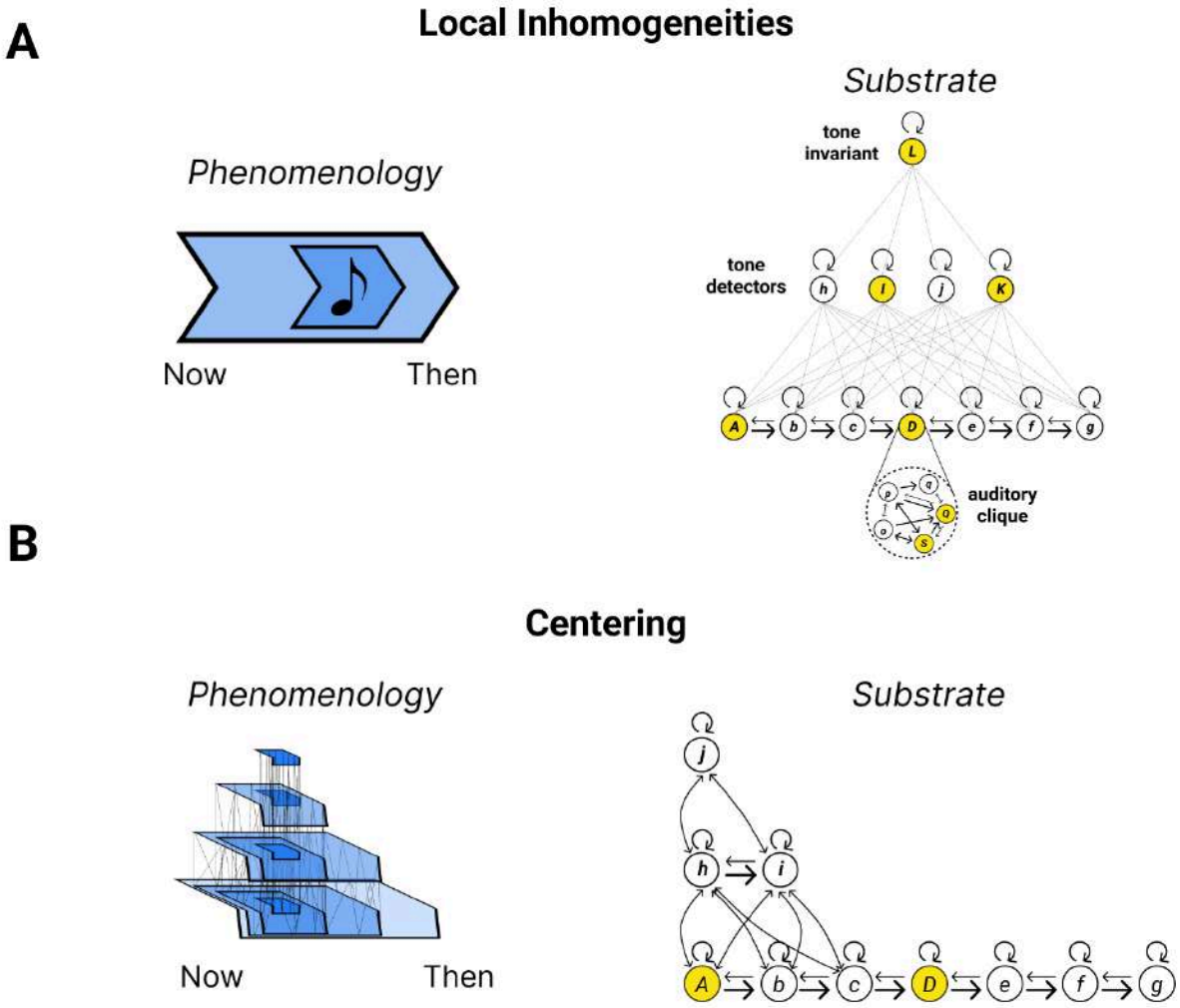



**Figure 10.** Local inhomogeneities and centering in the *now*. (A) Phenomenally, some moment may "stand out" and locally disrupt the flow of time, as when we hear a sudden sound or pause (left). This may be accounted for activation or deactivation of specific units within a directed grid, accompanied by the interplay between higher-level and lower-level mechanisms connected to directed grids. Locally, distinctions and relations would be altered, resulting in a local thickening and warping of the cause–effect structure, which does not disrupt the global flow of time. Note that the "local quality" of the sound or pause would be accounted for by local mechanisms (auditory cliques) and associated sub-structures embedded at every locale of the directed grid. (B) Phenomenally, moments flow away from the *now* towards the *then*, and we feel centered in the now, which also feels more vivid (left). This may be accounted for by denser connections between the now terminus of directed grids to higher-level areas involved with agency, corresponding to a much larger

number of relations binding the now with the rest of the cause–effect structure (right).

## 5. Discussion

According to IIT, all properties of an experience can be accounted for in physical terms by corresponding properties of the cause–effect structure unfolded from a substrate in its current state. The unfolding procedure is based on IIT's principles and its five postulates (intrinsicality, information, integration, exclusion, and composition), which capture in causal terms the *essential* properties of every conceivable experience. According to the theory, no additional ingredients are needed to account for the *accidental* properties of specific experiences, such as the feeling of spatial extension, of temporal flow, of objects binding general concepts with particular features, of local qualities such as color or sound, and so on. These accidental properties should be accounted for by corresponding properties of the cause–effect structure specified by a neural substrate depending on its connectivity and current activity pattern.

This paper aims to show how the IIT framework can be employed to account for the experience of temporal flow. Just as most of our conscious life is "painted" on the "canvas" of experienced space, much of it is "played" on the "track" of experienced time. The conscious present is confined between the *now* and the *then*, occasionally including a *next* that extends beyond the now. It is composed of moments, short and long, some closer to the now and some to the then. Moments are directed, pointing away from themselves, and overlap through relations of directed inclusion, connection, and fusion, to yield the feeling of flow.

As demonstrated here, a substrate such as a directed grid supports a cause–effect structure that can account for the fundamental properties of temporal flow: its units specify causal distinctions (moments) whose cause and effect overlap in a directed manner, yielding causal relations of directed inclusion, connection, and fusion. From these fundamental properties, other properties of temporal experience can be derived, such as the period occupied by a moment, its temporal location within the present and with respect to the now and the then, its duration, its boundary, and the interval between it and other moments. The results exemplify an *explanatory identity* (Albantakis et al., 2023) between the properties of temporal experience and those of the flow structure specified by directed grids.

### 5.1. Temporal flow as a directed structure

A central aspect of IIT's account is that the experience of time flowing corresponds to a *directed structure*, rather than to a process that actually "flows" in clock time. This is illustrated in Fig. 11 (left). The interval of clock time depicted is ~10 seconds, longer than the duration of the conscious present—assumed here, for convenience, to be ~210 milliseconds. The portion of the arrow of clock time corresponding to the "clock past" (i.e., all events that have already happened) is dashed, the "clock now" is indicated with a thicker thick, and the portion corresponding to the "clock future" (which has not happened yet) is dotted. Clock time can be assumed to tick at much faster resolution (say ~1 picosecond, not depicted) than instants of experienced time, assumed here to last for ~30 milliseconds of clock time (compatibly with experimental evidence discussed in section 4.13). Each instant corresponds to a "macro" state of the directed grid (Marshall et al., 2024).

In the figure, the foreground illustrates a 1D directed grid (units $A$ to $g$) in its current macro state, with units $A$ and $D$ ON and all other units OFF. A macro state, as explained in (Hoel et al., 2016; Marshall et al., 2018;

Marshall et al., 2024), is the intrinsic update grain of the units of a complex—the grain at which, from the complex's intrinsic perspective, the value of $\varphi_s$ is maximized. If we assume that the intrinsic macro units may be neurons and their intrinsic update grain 30 milliseconds, the macro state of the units in Fig. 11 would extend backwards for 30 milliseconds from the clock "now." Because of the way the delay line driving the directed grid is organized, the macro state of the seven-unit grid preserves a trace of what happened over 210 milliseconds of clock time—in this case, that two notes were played (in different colors on the score) over a track of silence.

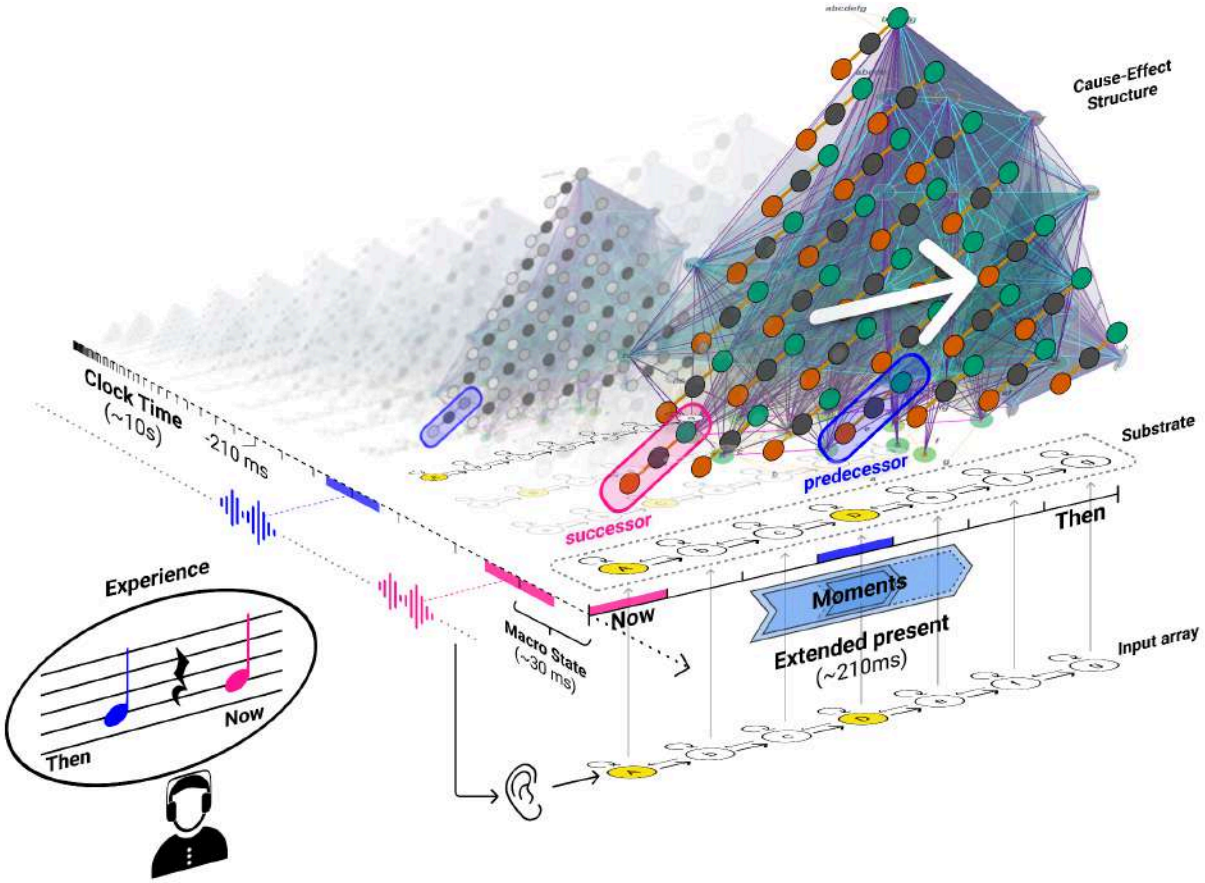



**Figure 11.** Phenomenal time and clock time. The axis representing ~10 seconds of clock time shows the occurrence of two sound waves (blue and pink, lasting ~30 milliseconds) and separated by an absence of waves. The bubble at the bottom left represents a subject experiencing an extended present containing two tones triggered by the sound waves and some pauses of silence around them. Orthogonal to clock time, the figure shows a directed grid in its current macro state and the cause–effect structure unfolded from it. This is assumed to account in full for the feeling on an extended present and the flow of time. The macro state of the directed grid is driven by an input array conveying auditory inputs, which functions as a delay line that preserves a trace of occurrences lasting for ~210 milliseconds of clock time. The cause–effect structure can thus keep track of occurrences over ~210 milliseconds of clock time, with a short delay due to neural transmission and activation. The moments that coexist within the extended present are bound by relations of directedness, directed inclusion, connection, and fusion, that yield a feeling of flow from the now to the then. The latest note (pink) is experienced in the now, preceded by the earlier note (blue) receding towards the then. The figure also shows a few cause–effect structures unfolded from macro states of the directed grid associated with earlier "ticks" of clock time. These are faded to indicate that they are not actual.

As illustrated in the figure, the grid in its current macro state supports a cause–effect structure composed of a multitude of directed distinctions and relations that order it according to directed inclusion, connection, and fusion. In this way, the cause–effect structure can account for a conscious present that feels extended in time, flowing from the phenomenal *now* to the phenomenal *then*. Furthermore, a substrate defined over a macro state of, say, 30 milliseconds of clock time, can support an experience capturing a longer interval of clock time, say

210 milliseconds or longer (depending on the number of units in the grid). This has the obvious advantage that contents triggered by a sequence of inputs can be bound together within a single experience, say that of a melody or a spoken phrase, while preserving their ordering and direction.

The figure also shows a few cause–effect structures (in gray) preceding the current one. In principle, a new structure would be is specified over a macro state at every micro update ("tick") of clock time. However, because neurons update their macro state at a much coarser grain than the ticks of clock time, cause–effect structures succeeding one another over many consecutive ticks of clock time will be identical or nearly so (and so will the corresponding experiences).

## 5.2. Flexible matching between intrinsic temporal flow and extrinsic clock time

In a brain well adapted to its environment, one would expect that the flow of experience, say the succession of notes in a melody, will match well enough, with a short delay and proper ordering, the sequence of stimuli sampled in clock time. However, this matching can be somewhat flexible, allowing for some "editing" and "extrapolating" of the "track" of experienced time. There are several instances, in auditory psychophysics (Herzog et al., 2020), language perception (Rönnberg et al., 2019), music perception (Juslin & Västfjäll, 2008), and motion perception (Shimojo, 2014), where stimuli occurring later can affect the experience triggered by stimuli occurring earlier. Such "postdictive" effects can be naturally accommodated within the present framework. For example, top-down connections from higher level areas may affect the activation of units towards the "then" terminus of directed grids in lower-level areas.

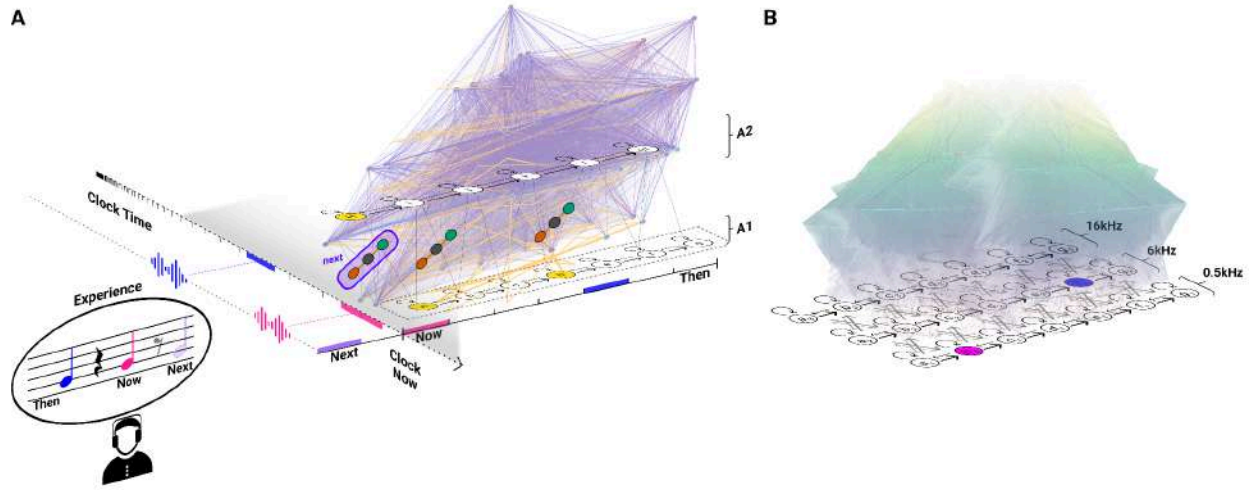



**Figure 12.** Further aspects of temporal experiences and their substrate. A) The extended present may include the experience of what will happen next, in addition to the experience of what happened between the *now* and the *then*. Possible mechanisms supporting an experienced future may involve directed grids at higher levels in a sensory hierarchy (here, A2) whose substrate extends beyond the *now* at lower levels (here, A1). Units in A2 may be activated endogenously by "imagining" what might be heard next (for example, the purple note on the music score in the bottom left). The extended present would then map a longer interval of clock time that comprises possible future occurrences (grey shaded area projected onto the clock time axis). (B) The substrate of the extended present, at every hierarchical level, is assumed to be not one grid, but an array of directed grids interacting through lateral connections. In auditory areas, for example, each grid in the array may comprise units selective for different frequency bands.

As illustrated in Fig. 12A, units in directed grids at higher levels in the auditory hierarchy may also specify moments that succeed the "now" and extend towards the "next." These "extrapolations" would be experienced,

typically less vividly, as upcoming occurrences, say as the expected next note in a known melody (faded purple note on the music score). The adaptive matching of the intrinsic temporal flow and extrinsic clock time might hold, leading to priming and confirmation effects, or it might be violated by what actually happens next, potentially accounting for various illusions (Eagleman, 2008; Merchant et al., 2013) as well as for desired effects in music (Huron, 2006; Vuust et al., 2022).

Finally, Fig. 12B illustrates that, at a minimum, the simplified account presented here should be expanded by considering 2D arrays of directed grids interacting through lateral connections. For example, each directed grid might correspond to a different frequency band in the tonotopic organization of auditory cortex (Saenz & Langers, 2014).

## 5.3. Similarities and differences between the experience of time and space

The feeling of time as an extended present, as analyzed here, bears many similarities with the feeling of space as an extended canvas. Yet time also feels "flowing," unlike space. As previously proposed (Haun & Tononi, 2019), the experience of space can be dissected into countless distinctions, called *spots*, which compose a spatial *extension* through relations of reflexivity, inclusion, connection, and fusion (Fig. 13A). Phenomenally, instead of being directed like moments, spots are *reflexive*, in the sense that they point to themselves. Because they are reflexive rather than directed, spatial distinctions include, connect with, and fuse with one another in a non-directed way. Furthermore, experienced space is typically 2D (or 3D), rather than 1D.

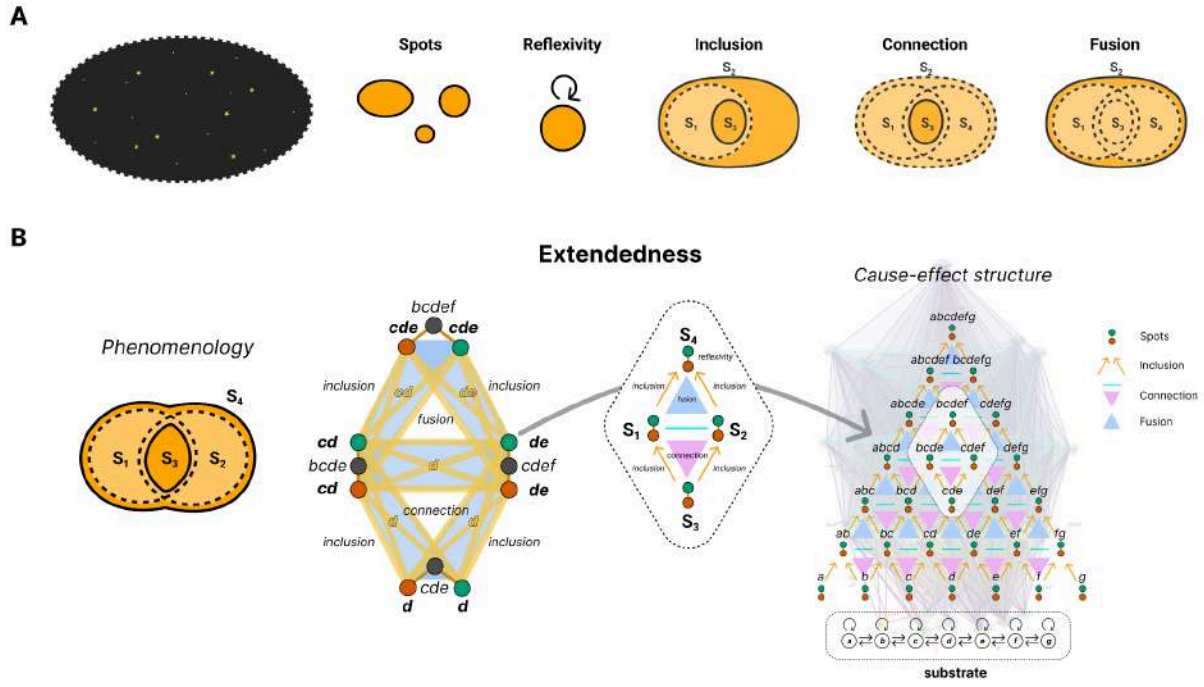



**Figure 13.** IIT's account of spatial experience *vs*. temporal experience. (A) Phenomenology of spatial experience and its fundamental properties. The experience of (visual) space is characterized by countless phenomenal distinctions, called spots, bound through relations of reflexivity, inclusion, connection, and fusion (all non-directed). (B) These fundamental properties of space find correspondence in the properties of the cause–effect structure unfolded from non-directed grids (right). Non-directed grids specify distinctions that are reflexive, each specifying a cause and an effect that fully overlap and that relate through non-directed inclusion, connection, and fusion (second from left). This also holds for the other distinctions that compose the cause–effect structure, which can thus be considered an *extension*. The third panel from left summarizes the relations of non-directed inclusion,

connection, and fusion as they apply between four distinctions corresponding to spots ($S_1$ through $S_4$), and the right-most panel shows how they apply between contiguous distinctions throughout the cause–effect structure (for simplicity only the label of the mechanisms are shown).

In physical terms, the similarities and differences between space and time can be accounted for by a different kind of neural substrate: non-directed 2D (or 3D) grids for space, and arrays of directed 1D grids for time. Crucially, a non-directed grid specifies causal distinctions that are reflexive—having cause and effect over the same elements (usually a subset of the mechanism elements)—rather than directed, with cause and effect over different elements, as is the case for directed grids (Fig. 13B). It follows from reflexivity that the inclusion, connection, and fusion relations are also non-directed. In other words, the reflexivity of spatial distinctions—the fact that their cause and effect purviews coincide—guarantees that any overlap with other spatial distinctions will be symmetrical over their cause and effect sides.

Spatial and temporal experiences are also remarkably similar with respect to further properties that can be derived from their fundamental ones. The region occupied by a spot, its location within the extension of space and with respect to its borders, its size, its boundary, and the distance from other spots are the non-directed analog of the period occupied by a moment, its temporal location within the present and with the now and the then, its duration, its boundary, and the interval between it and other moments. Similarly, inhomogeneities in local qualities can highlight particular spots that locally warp the extendedness of space, just as they can highlight particular moments in the present, without disrupting its flow. And, just as time can flow silently, space can be completely empty and still feel extended. Finally, just as we feel centered in the now temporally, we typically feel centered in the middle spatially, in both cases the natural starting point for action.

## 5.4. Introspection as an essential but limited tool for dissecting the phenomenal structure of temporal flow

Introspection is the indispensable starting point for the analysis of experience. As a first attempt to account for the quality of consciousness in physical terms, we focused on spatial extendedness precisely because the experience of space is not just pervasive, but also highly penetrable through introspection, largely thanks to the power and flexibility of spatial attention (Haun & Tononi, 2019). Temporal flow is also pervasive and partially introspectable, though less easily so than spatial extendedness. This is presumably because the fleeting nature of time does not lend itself to being steadily grasped by attention, which is deployed sequentially and with limited speed. Introspection is also selective in the contents of experience it can access, likely because it depends on the limited ability of top-down connections to increase the excitability of specific subsets of neurons (Ellia et al., 2021; Haun & Tononi, 2019).

Nonetheless, as testified by venerable traditions in temporal phenomenology, the fundamental structural properties of temporal experience remain more penetrable by introspection than those of a musical chord, a color, or a smell (for more on the role and limitations of introspection, see Ellia et al. (2021) and Haun & Tononi (2019)). As shown here, we can rely on introspection to characterize the directedness of moments and their relations of directed inclusion, connection, and fusion—as well as many derived properties such as durations and intervals. This allowed us to demonstrate a systematic correspondence between the phenomenal properties of temporal flow and the physical properties of cause–effect structures specified by directed grids.

This correspondence is assumed to hold when we cease introspecting because phenomenal time flows, just as phenomenal space envelops us, whether we pay attention to it or not.

Beyond this, the power and reliability of introspection are clearly limited. For example, while introspection clearly reveals that the present is extended, precisely estimating its duration is no easy task, and psychophysical results differ depending on the criteria employed. Thus, William James thought the specious present could last as long as 12 seconds (James, 1890). Others placed it at ~3 seconds based on criteria such as the ability to impose a subjective rhythm to uniform auditory stimuli, to precisely estimate intervals, and so on (Montemayor & Wittmann, 2014; Pöppel, 2009). On the other hand, using tachistoscopic presentations of stimuli to assess "that stretch of change which is apprehended as a unit and which is the object of a single mental act of apprehension" has led to an estimate of 750 milliseconds (Albertazzi, 1996). Some have suggested even shorter durations, down to 300 milliseconds (Dainton, 2000; Strawson, 2009) (see Dainton (2023) and White (2017) for critical reviews).

The duration of instants has been estimated indirectly by assessing temporal order thresholds (the shortest inter-stimulus interval under which two sequential stimuli are perceived as simultaneous (Brecher, 1932; Hirsh & Sherrick, 1961; Kanabus et al., 2002)) and flicker fusion thresholds (the shortest inter-stimulus interval under which flickering stimuli are perceived as continuous (Andrews et al., 1996; Curran & Wattis, 1998)). The results yield a range of 10–60 milliseconds depending on the paradigm employed (Elliott & Giersch, 2016; Herzog et al., 2016; Pöppel, 1997a, 1997b; VanRullen & Koch, 2003; White, 2018).

## 5.5. Directed grids in the brain as the substrate of temporal experience

In previous work, we proposed that the neural substrate of the feeling of spatial extendedness is provided by non-directed 2D grids, connected hierarchically and in parallel to constitute a dense 3D lattice (Haun & Tononi, 2019; Tononi, 2014). This kind of substrate is ubiquitous in posterior cortex, and its relevance for the experience of space—both visual space and body space—is supported by clinical and neurophysiological evidence (Heinzle et al., 2011; Salin & Bullier, 1995; Sereno & Huang, 2014; Wang et al., 2015).

Here we conjectured that arrays of directed grids constitute the neural substrate of the feeling of temporal flow. However, little is known about the presence and location of such directed grids in the brain. According to IIT, the substrate of specific aspects of experience must be a subset of units within the *main complex*—the overall substrate of consciousness. This implies that the relevant directed grids must constitute, together with the rest of the complex, a substrate that is maximally irreducible. Moreover, one would expect that such grids should be closely connected to the neural substrate of modalities, such as sound, speech, and music, that are tightly bound to temporal flow.

Based on such considerations, directed grids supporting the experience of temporal flow might be located, for example, in portions of posterior cortex specialized for sound, speech, and music perception. There is substantial evidence indicating that the overall substrate of consciousness is primarily localized to posterior and central cortical regions (Boly et al., 2017; Koch et al., 2016; Siclari et al., 2017). Moreover, it is well established that hearing sound, speech, and music depends on specialized portions of cortex connected to primary auditory cortex (Hickok & Poeppel, 2015; Norman-Haignere et al., 2015). We therefore hypothesize that within such regions, one should be able to identify arrays directed grids serving as delay lines as well as substrates for the experience of flow. Specific details of the local connectivity would be responsible for local phenomenal qualities

typically bound to temporal flow, such as pitch, timbre, and loudness. (In a similar way, the details of the local connectivity in non-directed 2D grids would contribute to local phenomenal properties of spatial extendedness, such as hue, saturation, and brightness for visual space).

We also expect that the overall experience of temporal flow should be supported by multiple directed grids distributed across many areas of the main complex, at multiple levels. Convergent/divergent connections across hierarchically organized areas will support relations that bind, say, phonemes with syllables and words within a spoken sentence (Hickok & Poeppel, 2015). Lateral connections may further support the binding of temporal contents across submodalities, say, between speech and music (Janata, 2015; Janata et al., 2002), or even across modalities. Temporal aspects of experience may also be bound to spatial aspects, say, when experiencing visual motion between adjacent spatial locations. It is possible that areas such as V5, which plays a critical role in the perception of patterned motion (Albright, 1984; Clifford & Ibbotson, 2002), may be organized such that non-directed and directed grids may intertwine.

On the other hand, elsewhere in the brain neurons that do not belong to the main complex may be capable of representing temporal order without contributing to experience. For example, endogenous circadian "clocks" allow the brain, and specifically the suprachiasmatic nucleus of the hypothalamus, to keep track of the time of day and appropriately regulate various bodily functions unbeknownst to us (Roenneberg, 2012). Similarly, some brainstem neurons can detect microsecond intervals between the arrival of sounds at the two ears, intervals of which we are unaware (though they may contribute indirectly to our awareness of sound location through their effects on neurons in posterior cortex (Grothe et al., 2010)). The contribution of brain regions often considered as "organs of succession," such as the cerebellum, the basal ganglia, and the hippocampus, is more complex. For example, neurons in the hippocampus may subserve the memory of temporal order (Eichenbaum, 2014) as well as cognitive maps (O'Keefe & Nadel, 1978). However, the anatomical organization of the hippocampal formation is very different from that of posterior cortex, making it less likely to be part of the substrate of consciousness. Lesion data also indicate that, while the hippocampal formation is critical for supporting functions such as episodic memory and imagination, it may not directly contribute specific conscious contents (Postle, 2016). With respect to time, lesion studies in humans and rats show that hippocampal lesions do not impair estimating and recalling distances and durations, but rather impair mostly the ability to remember the sequential order of events (Buzsáki & Tingley, 2018; Dede et al., 2016; Fortin et al., 2002; Maguire et al., 2006).

## 5.6. Some tests and predictions

Besides providing a principled account of the subjective feeling of time flowing in objective, physical terms, the current proposal lays the foundation for experimental tests. However, it should be recognized that such tests are made more challenging by our uncertainty concerning the neural substrate of temporal experience.

The most general prediction concerns the substrate of the experience of an extended present and the sense of time flowing away from now to then. As proposed here, this substrate should correspond to a single macro state (lasting, say, ~30 milliseconds) of arrays of directed grids within the main complex, rather than to a sequence of neuronal events covering the duration of the extended present in clock time.

Another prediction is that the duration of the extended present should be proportional to the number of

macro units constituting a directed grid. Thus, everything else being equal, a grid with more units should support temporal experiences that encompass a longer stretch of clock time, with potential adaptive advantages. Units at higher levels in the sensory hierarchy (and beyond) would then be able to learn concepts that span over longer stretches, in line with the observation of longer temporal receptive fields in higher level (Hasson et al., 2008).

Yet another prediction is that the duration of phenomenal instants should be compatible with the grain of the macro states of the units constituting directed grids. According to IIT, this is given by the time interval (in clock time) yielding maximal $\varphi_s$ for the main complex (Albantakis et al., 2023; Marshall et al., 2024). For macro units such as neurons, this would likely be determined by the time constants at which synaptic and cellular mechanisms ensure maximal causal efficacy.

As already mentioned, the present framework is in principle well poised to accommodate several empirical observations that imply some "editing" of the neural traces left by a sequence of stimuli (Hogendoorn, 2022), see also (Libet et al., 1979). A related prediction is that artificial activation of grid units near the "now" terminus should result in perceiving a stimulus as occurring now, while the activation of grid units near the "then" terminus should result in perceiving a stimulus as having occurred earlier.

The IIT framework further predicts that modulation of synaptic strength or of the excitability of neurons in directed grids should induce changes in the properties of phenomenal flow regardless of activity levels (Haun & Tononi, 2019). Such modulations could account, for instance, for the slowing or speeding up of time caused by strong emotions, deep meditation, or drugs (Coull et al., 2011; Droit-Volet & Meck, 2007; Kramer et al., 2013; Sewell et al., 2013; Wackermann et al., 2008).

## 5.7. Time: cognitive mechanisms and phenomenal properties

The investigation of neural mechanisms of time perception and temporal processing has been an active area of research for decades (Kononowicz et al., 2018). Psychophysical paradigms have focused on interval estimation (Grondin, 2010; Tsao et al., 2022), temporal integration (Herzog et al., 2020; Lerner et al., 2011; Norman-Haignere et al., 2022), and time illusions (Eagleman, 2008; Merchant et al., 2013). For example, subjects may be asked to assess interval durations verbally or by reproducing target intervals. Several mechanistic and computational models have been developed to account for psychophysical results (Hass & Durstewitz, 2016; Muller & Nobre, 2014), based, for example, on ramping activations models (Wittmann, 2013), neural oscillations (Matell & Meck, 2000; VanRullen & Koch, 2003), and population state dynamics (Paton & Buonomano, 2018; Tsao et al., 2022). In parallel, neurophysiological studies have investigated neural correlates of temporal processing (Nani et al., 2019; Rao et al., 2001). Neurons tracking intervals and sequences, at varying time scales, have been reported in the hippocampus (Buzsáki & Tingley, 2018; Eichenbaum, 2014), basal ganglia (Buhusi & Meck, 2005), the cerebellum (Ivry & Spencer, 2004), supplementary motor area (Ferrandez et al., 2003; Macar et al., 2006), entorhinal cortex (Tsao et al., 2018), and frontal and parietal cortex (Hayashi & Ivry, 2020; Hayashi et al., 2018). As already mentioned, there are cellular and system-level mechanisms involved in tracking circadian time (Roenneberg & Merrow, 2003).

These findings are critical for characterizing how the brain "represents" clock time (Hogendoorn, 2022) and employs these representations for motor control, memory, and cognitive functions. However, the framework

presented here differs from cognitive and computational paradigms both with respect to what it tries to explain (the *explanandum*) and to how it tries to do so (the *explanans*). The explanandum is not so much the cognitive capacity to discriminate and report the objective duration of stimuli (in clock time) but rather the subjective properties of temporal experience as assessed through introspection (Ellia et al. 2021). In this respect, the present work parallels some proposals in consciousness research that have attempted to directly address the temporal quality of conscious experiences (Bogotá & Djebbara, 2023; Piper, 2019; Varela, 1999; Wiese, 2017). Furthermore, the explanans is not so much the nature of the neural "representation" of temporal features of stimuli (Hass & Durstewitz, 2016; Ivry & Spencer, 2004; Wittmann, 2013) or of how experienced time maps and represents clock time (Herzog et al., 2020; Herzog et al., 2016; Hogendoorn, 2022; Northoff & Zilio, 2022). Instead, it is a one-to-one correspondence between the subjective, phenomenal properties of temporal experiences and objective, physical properties of the cause–effect structure unfolded from a certain kind of substrate.

## 5.8. IIT and philosophical approaches to time

There has been a remarkable lack of recognition that the "extendedness" of spatial experiences is as much in need of explanation as the blueness of blue and the painfulness of pain (for a few exceptions, see James (1879), Kant et al. (1998), and Lotze (1884)). One reason may be that space is generally assumed to exist physically "out there," so experienced space may pass for a mapping or "representation" that does not require further explanations.

It is less obvious, however, that time is flowing "out there," as indicated by the diversity of positions in both philosophy and physics. According to "eternalism," all times are equally real, similar to the modern conception of a block universe of space-time. The "growing-block" universe grants existence to the past but not the future (Broad, 1923; Tooley, 1997). For the "moving spotlight" model, on the other hand, a window of actual present relentlessly advances over a block universe (Skow, 2015). Finally, "presentism" assumes that physical time, if it exists at all, can only exist for an instant. Therefore, the extendedness of experienced time, if not time itself, can only exist as a construct "in the mind" (Augustine & Chadwick, 2009). In fact, several philosophers, including Kant, Husserl, and Bergson, as well as contemporary investigators (Kent & Wittmann, 2021; Northoff & Zilio, 2022; Singhal & Srinivasan, 2024), have considered time as a basic ingredient of consciousness. Along these lines, some influential phenomenological models of temporal experiences have been developed and refined (Dainton, 2000, 2012; McTaggart, 1908).

Specifically, *retentional* models explicitly propose that experiences of temporal flow do not have temporal extension but are characterized by a feeling of succession (rather than a succession of feelings, James, 1890). Thus, at every moment, in addition to the feeling of now, or "primal impression," we would also experience fainter "retentions" of past moments (and "protentions" of moments to come, (Husserl, 1991)). *Extensional* models assume instead that experienced time unrolls over an extended interval of clock time (Dainton, 2008, 2012). Thus, a conscious present that feels half a second long would unroll over an equivalent interval of clock time. Successive moments within the interval are considered as parts of a whole bound by "diachronic" relations of succession, yielding a sense of immanent flow. Finally, *cinematic* and *snapshot* models assume that all there is is a succession of experiences—a series of phenomenal "snapshots"—supported by a series of discrete physical

events (for example, (Arstila, 2023; Prosser, 2017)). The feeling of flow would then be merely an illusion.

Where does IIT's account stand? Temporal flow is certainly not an "illusion" but a property of experience in need of a physical account (as also recognized by (Singhal et al., 2022; Singhal & Srinivasan, 2021, 2024)). Even so, flow is not an essential property of consciousness, because while pervasive, it is not true of every conceivable experience (unlike intrinsicality or integration). Indeed, experiences devoid of temporal content are not only conceivable, but they have long been reported, for example, during deep meditation (experiences of "pure presence," see Boly et al. (2024)) and under the effect of psychedelic drugs (Wittmann, 2015).

IIT partly agrees with cinematic and snapshot models in assuming that a new temporal experience comes into being at every "tick" of the clock, existing over a short interval of clock time (say, 30 milliseconds). However, IIT goes beyond such models by identifying each temporal experience with a directed cause–effect structure, which accounts for why a single experience feels like a succession of moments.

IIT also captures the intuition behind retentional approaches that an experience supported by a macro state corresponding to a short interval of clock time can contain within itself the duration of the entire conscious present (corresponding, say, to ~210 milliseconds or more of clock time), ordered according to a feeling of succession. However, retentional approaches only describe the phenomenology of succession, and only some aspects of it, without dissecting its relational structure or suggesting a physical correspondent for it.

IIT also captures the intuition behind extensional approaches that the experience of time must be structured by relations of succession and is characterized by a sense of flow. However, extensional approaches do not further characterize directed relations phenomenally, nor do they provide a physical correspondent that would account for them. Moreover, it is unclear what it would mean for temporal parts to overlap physically across clock time. This last point highlights a critical aspect of IIT's physical conception of relations. In IIT, relations are defined in causal terms (an overlap of causes and/or effects over the same units in the same state) and are intrinsic to a system (as well as unitary and definite, as per the postulates of integration and exclusion). Extensional approaches, if they attempt to characterize temporal relations at all, do so in non-causal, extrinsic terms—from the point of view of an observer who already knows what temporal flow feels like and who understands what a label such as "diachronic" should mean.

## 5.9. Conclusions

This paper employed the framework of IIT to (i) identify the fundamental phenomenal distinctions and relations that characterize the experience of temporal flow and (ii) formulate them operationally in terms of causal distinctions and relations specified by a certain type of substrate—namely, directed grids. The results presented here illustrate how the cause–effect structure unfolded from a directed grid can account for the properties of experienced time. They thus exemplify the explanatory identity proposed by IIT between phenomenal, subjective properties and physical, objective properties of causal structures, as already shown for spatial extendedness (Haun & Tononi, 2019).

To permit the systematic unfolding of cause–effect structures, the substrates employed in this paper were necessarily small (seven units with near-neighbor connections). Even so, the present examples provide a principled illustration of the kinds of distinctions and relations required to account for experienced time—an extended present composed of moments of various duration, ordered through relations of directedness and

directed inclusion, connection, and fusion, which flows away from now to then. Conceiving of the flow of time as a cause–effect structure specified by a directed grid in its current macro state, which can be "edited" dynamically through multiple neural mechanisms, offers a template to address various aspects of temporal psychophysics, temporal illusions, speech and language perception. Changes in connectivity within directed grids may also explain the slowing or quickening of time caused by strong emotions, deep meditation, or drugs.

As with the experience of space, a full account of temporal experiences—of how temporal flow is bound with both hierarchically invariant concepts and local features, and with local qualities belonging to different modalities—will require the unfolding of larger neural substrates and an adequate understanding of their anatomy and physiology. Ultimately, however, only a structural explanation can account in physical terms for the way time feels, rather than presupposing it. For example, a directed delay line can only serve to represent time if one already knows what time means and feels like. But to feel temporally extended, the ordering of moments within the present must be established by causal distinctions and relations composing a cause–effect structure intrinsic to a system, one that means what it means absolutely, rather than by reference to external clocks.

This conclusion is very much in line with Augustine's original insight—that time is in the mind. But it adds that the mind—or rather every experience in the stream of an individual consciousness—is an extraordinarily rich structure. It is a structure that contains time, space, objects, thoughts, and everything else that exists intrinsically—for itself.

## Acknowledgements

We thank Larissa Albantakis, Andrew Haun, and Jeremiah Hendren for their feedback and for many helpful discussions. We also thank Julia Thompson for her comments. This project was made possible through support from Templeton World Charity Foundation (nos. TWCF0216 and TWCF0526). The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of Templeton World Charity Foundation.

## Bibliography

Albantakis, L., Barbosa, L., Findlay, G., Grasso, M., Haun, A. M., Marshall, W.,...Tononi, G. (2023). Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms. *PLOS Computational Biology*, *19*(10), e1011465. https://doi.org/10.1371/journal.pcbi.1011465

Albertazzi, L. (1996). Comet tails, fleeting objects and temporal inversions. *Axiomathes*, *7*(1), 111-135. https://doi.org/10.1007/BF02357202

Albright, T. D. (1984). Direction and orientation selectivity of neurons in visual area MT of the macaque. *Journal of Neurophysiology*, *52*(6), 1106-1130. https://doi.org/10.1152/jn.1984.52.6.1106

Andrews, T. J., White, L. E., Binder, D., & Purves, D. (1996). Temporal events in cyclopean vision. *Proceedings of the National Academy of Sciences*, *93*(8), 3689-3692. https://doi.org/10.1073/pnas.93.8.3689

Arstila, V. (2023). Explanation in theories of the specious present. *Philosophical Psychology*, 1-24. https://doi.org/10.1080/09515089.2023.2241501

Augustine, S., & Chadwick, H. (2009). *Confessions* (1st edition ed.). Oxford University Press.

Bogotá, J. D., & Djebbara, Z. (2023). Time-consciousness in computational phenomenology: a temporal analysis of active inference. *Neuroscience of Consciousness*, *2023*(1), niad004. https://doi.org/10.1093/nc/niad004
Boly, M., Massimini, M., Tsuchiya, N., Postle, B. R., Koch, C., & Tononi, G. (2017). Are the Neural Correlates of Consciousness in the Front or in the Back of the Cerebral Cortex? Clinical and Neuroimaging Evidence. *The Journal of Neuroscience*, *37*(40), 9603-9613. https://doi.org/10.1523/JNEUROSCI.3218-16.2017
Boly, M., Smith, R., Borrego, G. V., Pozuelos, J. P., Alauddin, T., Malinowski, P., & Tononi, G. (2024). Neural correlates of pure presence. In: bioRxiv.
Brecher, G. A. (1932). Die Entstehung und biologische Bedeutung der subjektiven Zeiteinheit, — des Momentes. *Zeitschrift für vergleichende Physiologie*, *18*(1), 204-243. https://doi.org/10.1007/BF00338160
Broad, C. D. (1923). *Scientific Thought*.
Buhusi, C. V., & Meck, W. H. (2005). What makes us tick? Functional and neural mechanisms of interval timing. *Nature Reviews Neuroscience*, *6*(10), 755-765. https://doi.org/10.1038/nrn1764
Buzsáki, G., & Tingley, D. (2018). Space and Time: The Hippocampus as a Sequence Generator. *Trends in Cognitive Sciences*, *22*(10), 853-869. https://doi.org/10.1016/j.tics.2018.07.006
Clay, E. R. (1882). *The Alternative: A Study in Psychology*. Macmillan.
Clifford, C. W. G., & Ibbotson, M. R. (2002). Fundamental mechanisms of visual motion detection: models, cells and functions. *Progress in Neurobiology*, *68*(6), 409-437. https://doi.org/10.1016/S0301-0082(02)00154-5
Coull, J. T., Morgan, H., Cambridge, V. C., Moore, J. W., Giorlando, F., Adapa, R.,...Fletcher, P. C. (2011). Ketamine perturbs perception of the flow of time in healthy volunteers. *Psychopharmacology*, *218*(3), 543-556. https://doi.org/10.1007/s00213-011-2346-9
Curran, S., & Wattis, J. P. (1998). Critical flicker fusion threshold: a useful research tool in patients with Alzheimer's disease. *Human Psychopharmacology: Clinical and Experimental*, *13*(5), 337-355. https://doi.org/10.1002/(SICI)1099-1077(199807)13:5<337::AID-HUP7>3.0.CO;2-P
Dainton, B. (2000). *Stream of Consciousness: Unity and Continuity in Conscious Experience*. Routledge.
Dainton, B. (2008). Sensing Change. *Philosophical Issues*, *18*(1), 362-384. https://doi.org/10.1111/j.1533-6077.2008.00152.x
Dainton, B. (2012). Time and Temporal Experience. In A. Bardon (Ed.), *The Future of the Philosophy of Time* (pp. 123-148). Routledge.
Dainton, B. (2023). Temporal Consciousness. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Spring 2023 ed.). Metaphysics Research Lab, Stanford University.
Dede, A. J. O., Frascino, J. C., Wixted, J. T., & Squire, L. R. (2016). Learning and remembering real-world events after medial temporal lobe damage. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(47), 13480-13485. https://doi.org/10.1073/pnas.1617025113
Droit-Volet, S., & Meck, W. H. (2007). How emotions colour our perception of time. *Trends in Cognitive Sciences*, *11*(12), 504-513. https://doi.org/10.1016/j.tics.2007.09.008
Eagleman, D. M. (2008). Human time perception and its illusions. *Current opinion in neurobiology*, *18*(2), 131. https://doi.org/10.1016/j.conb.2008.06.002

Eichenbaum, H. (2014). Time cells in the hippocampus: a new dimension for mapping memories. *Nature Reviews Neuroscience*, *15*(11), 732-744. https://doi.org/10.1038/nrn3827

Ellia, F., Hendren, J., Grasso, M., Kozma, C., Mindt, G., Lang, J.,...Tononi, G. (2021). Consciousness and the Fallacy of Misplaced Objectivity. *Neuroscience of Consciousness*, *7*(2), 1-12.

Elliott, M. A., & Giersch, A. (2016). What Happens in a Moment. *Frontiers in Psychology*, *6*, 1905. https://doi.org/10.3389/fpsyg.2015.01905

Ferrandez, A. M., Hugueville, L., Lehéricy, S., Poline, J. B., Marsault, C., & Pouthas, V. (2003). Basal ganglia and supplementary motor area subtend duration perception: an fMRI study. *NeuroImage*, *19*(4), 1532-1544. https://doi.org/10.1016/s1053-8119(03)00159-9

Fortin, N. J., Agster, K. L., & Eichenbaum, H. B. (2002). Critical role of the hippocampus in memory for sequences of events. *Nature Neuroscience*, *5*(5), 458-462. https://doi.org/10.1038/nn834

Grondin, S. (2010). Timing and time perception: A review of recent behavioral and neuroscience findings and theoretical directions. *Attention, Perception, & Psychophysics*, *72*(3), 561-582. https://doi.org/10.3758/APP.72.3.561

Grothe, B., Pecka, M., & McAlpine, D. (2010). Mechanisms of Sound Localization in Mammals. *Physiological Reviews*, *90*(3), 983-1012. https://doi.org/10.1152/physrev.00026.2009

Hass, J., & Durstewitz, D. (2016). Time at the center, or time at the side? Assessing current models of time perception. *Current Opinion in Behavioral Sciences*, *8*, 238-244. https://doi.org/10.1016/j.cobeha.2016.02.030

Hasson, U., Yang, E., Vallines, I., Heeger, D. J., & Rubin, N. (2008). A Hierarchy of Temporal Receptive Windows in Human Cortex. *Journal of Neuroscience*, *28*(10), 2539-2550. https://doi.org/10.1523/JNEUROSCI.5487-07.2008

Haun, A., & Tononi, G. (2019). Why Does Space Feel the Way it Does? Towards a Principled Account of Spatial Experience. *Entropy*, *21*(12), 1160.

Hayashi, M. J., & Ivry, R. B. (2020). Duration-selectivity in right parietal cortex reflects the subjective experience of time. *The Journal of Neuroscience*, JN-RM-0078-0020. https://doi.org/10.1523/JNEUROSCI.0078-20.2020

Hayashi, M. J., van der Zwaag, W., Bueti, D., & Kanai, R. (2018). Representations of time in human frontoparietal cortex. *1*(1), 233-233. https://doi.org/10.1038/s42003-018-0243-z

Heinzle, J., Kahnt, T., & Haynes, J.-D. (2011). Topographically specific functional connectivity between visual field maps in the human brain. *NeuroImage*, *56*(3), 1426-1436. https://doi.org/https://doi.org/10.1016/j.neuroimage.2011.02.077

Herzog, M. H., Drissi-Daoudi, L., & Doerig, A. (2020). All in Good Time: Long-Lasting Postdictive Effects Reveal Discrete Perception. *Trends in Cognitive Sciences*, S1364661320301704. https://doi.org/10.1016/j.tics.2020.07.001

Herzog, M. H., Kammer, T., & Scharnowski, F. (2016). Time Slices: What Is the Duration of a Percept? *PLOS Biology*, *14*(4), e1002433. https://doi.org/10.1371/journal.pbio.1002433

Hickok, G., & Poeppel, D. (2015). Chapter 8 - Neural basis of speech perception. In M. J. Aminoff, F. Boller, & D. F. Swaab (Eds.), *Handbook of Clinical Neurology* (Vol. 129, pp. 149-160). Elsevier.

Hirsh, I. J., & Sherrick, C. E. (1961). Perceived order in different sense modalities. *Journal of Experimental Psychology*, *62*, 423-432. https://doi.org/10.1037/h0045283

Hoel, E. P., Albantakis, L., Marshall, W., & Tononi, G. (2016). Can the macro beat the micro? Integrated information across spatiotemporal scales. *Neuroscience of Consciousness*, *2016*(1), niw012. https://doi.org/10.1093/nc/niw012

Hogendoorn, H. (2022). Perception in real-time: predicting the present, reconstructing the past. *Trends in Cognitive Sciences*, *26*(2), 128-141. https://doi.org/10.1016/j.tics.2021.11.003

Huron, D. B. (2006). *Sweet Anticipation: Music And the Psychology of Expectation*. Mit Pr.

Husserl, E. G. (1991). *On the Phenomenology of the Consciousness of Internal Time (1893-1917). Translated by John Barnett Brough*. Kluwer Academic Publishers.

Integrated Information Wiki. (2024). *Center for Sleep and Consciousness, University of Wisconsin–Madison*. https://www.iit.wiki

Ivry, R. B., & Spencer, R. M. C. (2004). The neural representation of time. *Current Opinion in Neurobiology*, *14*(2), 225-232. https://doi.org/10.1016/j.conb.2004.03.013

James, W. (1879). The Spatial Quale. *Journal of Speculative Philosophy*, *13*(1), 64-87.

James, W. (1890). *The Principles of Psychology*. Henry Holt and Co.

Janata, P. (2015). Chapter 11 - Neural basis of music perception. In M. J. Aminoff, F. Boller, & D. F. Swaab (Eds.), *Handbook of Clinical Neurology* (Vol. 129, pp. 187-205). Elsevier.

Janata, P., Birk, J. L., Van Horn, J. D., Leman, M., Tillmann, B., & Bharucha, J. J. (2002). The Cortical Topography of Tonal Structures Underlying Western Music. *Science*, *298*(5601), 2167-2170. https://doi.org/10.1126/science.1076262

Juslin, P. N., & Västfjäll, D. (2008). Emotional responses to music: the need to consider underlying mechanisms. *The Behavioral and Brain Sciences*, *31*(5), 559-575; discussion 575-621. https://doi.org/10.1017/S0140525X08005293

Kanabus, M., Szelag, E., Rojek, E., & Pöppel, E. (2002). Temporal order judgement for auditory and visual stimuli. *Acta Neurobiologiae Experimentalis*, *62*(4), 263-270. https://doi.org/10.55782/ane-2002-1443

Kant, I., Guyer, P., & Wood, A. W. (1998). *Critique of Pure Reason*. Cambridge University Press.

Kent, L., & Wittmann, M. (2021). *Time Consciousness: The Missing Link in Theories of Consciousness*. https://psyarxiv.com/56mvg/

Koch, C., Massimini, M., Boly, M., & Tononi, G. (2016). Neural correlates of consciousness: progress and problems. *Nature Reviews Neuroscience*, *17*(5), 307-321. https://doi.org/10.1038/nrn.2016.22

Kononowicz, T. W., van Rijn, H., & Meck, W. H. (2018). Timing and Time Perception: A Critical Review of Neural Timing Signatures Before, During, and After the To-Be-Timed Interval. *1*, 1-38. https://doi.org/10.1002/9781119170174.epcn114

Kramer, R. S. S., Weger, U. W., & Sharma, D. (2013). The effect of mindfulness meditation on time perception. *Consciousness and Cognition*, *22*(3), 846-852. https://doi.org/10.1016/j.concog.2013.05.008

Lerner, Y., Honey, C. J., Silbert, L. J., & Hasson, U. (2011). Topographic Mapping of a Hierarchy of Temporal

Receptive Windows Using a Narrated Story. *Journal of Neuroscience*, *31*(8), 2906-2915. https://doi.org/10.1523/JNEUROSCI.3684-10.2011

Libet, B., Wright, E. W., Feinstein, B., & Pearl, D. K. (1979). Subjective referral of the timing for a conscious sensory experience: a functional role for the somatosensory specific projection system in man. *Brain: A Journal of Neurology*, *102*(1), 193-224. https://doi.org/10.1093/brain/102.1.193

Lotze, H. (1884). *Lotze's system of philosophy*. Oxford : Clarendon Press.

Macar, F., Coull, J., & Vidal, F. (2006). The supplementary motor area in motor and perceptual time processing: fMRI studies. *Cognitive Processing*, *7*(2), 89-94. https://doi.org/10.1007/s10339-005-0025-7

Maguire, E. A., Woollett, K., & Spiers, H. J. (2006). London taxi drivers and bus drivers: A structural MRI and neuropsychological analysis. *Hippocampus*, *16*(12), 1091-1101. https://doi.org/10.1002/hipo.20233

Marshall, W., Albantakis, L., & Tononi, G. (2018). Black-boxing and cause-effect power. *PLOS Computational Biology*, *14*(4), e1006114. https://doi.org/10.1371/journal.pcbi.1006114

Marshall, W., Findlay, G., Albantakis, L., & Tononi, G. (2024). From micro to macro units: a mathematical framework for identifying the causal grain of a system from its intrinsic perspective. In: bioRxiv.

Marshall, W., Grasso, M., Mayner, W. G. P., Zaeemzadeh, A., Barbosa, L. S., Chastain, E.,...Tononi, G. (2023). System Integrated Information. *Entropy*, *25*(2), 334. https://doi.org/10.3390/e25020334

Matell, M. S., & Meck, W. H. (2000). Neuropsychological mechanisms of interval timing behavior. *BioEssays*, *22*(1), 94-103. https://doi.org/10.1002/(SICI)1521-1878(200001)22:1<94::AID-BIES14>3.0.CO;2-E

Mayner, W., Juel, B. E., & Tononi, G. Meaning, perception, and matching: quantifying how the structure of experience matches the environment. *in preparation*.

McTaggart, J. E. (1908). The Unreality of Time. *Mind, New Series*, *17*(68), 457-474.

Merchant, H., Harrington, D. L., & Meck, W. H. (2013). Neural Basis of the Perception and Estimation of Time. *Annual Review of Neuroscience*, *36*(Volume 36, 2013), 313-336. https://doi.org/10.1146/annurev-neuro-062012-170349

Montemayor, C., & Wittmann, M. (2014). The Varieties of Presence: Hierarchical Levels of Temporal Integration. *Timing & Time Perception*, *2*(3), 325-338. https://doi.org/10.1163/22134468-00002030

Muller, T., & Nobre, A. C. (2014). Perceiving the passage of time: neural possibilities: Time's neural arrow(s). *Annals of the New York Academy of Sciences*, *1326*(1), 60-71. https://doi.org/10.1111/nyas.12545

Nani, A., Manuello, J., Liloia, D., Duca, S., Costa, T., & Cauda, F. (2019). The Neural Correlates of Time: A Meta-analysis of Neuroimaging Studies. *Journal of Cognitive Neuroscience*, *31*(12), 1796-1826. https://doi.org/10.1162/jocn_a_01459

Norman-Haignere, S., Kanwisher, N. G., & McDermott, J. H. (2015). Distinct Cortical Pathways for Music and Speech Revealed by Hypothesis-Free Voxel Decomposition. *Neuron*, *88*(6), 1281-1296. https://doi.org/10.1016/j.neuron.2015.11.035

Norman-Haignere, S. V., Long, L. K., Devinsky, O., Doyle, W., Irobunda, I., Merricks, E. M.,...Mesgarani, N. (2022). Multiscale temporal integration organizes hierarchical computation in human auditory cortex. *Nature Human Behaviour*, *6*(3), 455-469. https://doi.org/10.1038/s41562-021-01261-y

Northoff, G., & Zilio, F. (2022). Temporo-spatial Theory of Consciousness (TTC) – Bridging the gap of neuronal activity and phenomenal states. *Behavioural Brain Research*, *424*, 113788. https://doi.org/10.1016/j.bbr.2022.113788

O'Keefe, J., & Nadel, L. (1978). *The Hippocampus as a Cognitive Map* (0 edition ed.). Oxford University Press.

Oizumi, M., Albantakis, L., & Tononi, G. (2014). From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0. *PLoS Computational Biology*, *10*(5), e1003588. https://doi.org/10.1371/journal.pcbi.1003588

Paton, J. J., & Buonomano, D. V. (2018). The Neural Basis of Timing: Distributed Mechanisms for Diverse Functions. *Neuron*, *98*(4), 687-705. https://doi.org/10.1016/j.neuron.2018.03.045

Piper, M. S. (2019). Neurodynamics of time consciousness: An extensionalist explanation of apparent motion and the specious present via reentrant oscillatory multiplexing. *Consciousness and Cognition*, *73*, 102751. https://doi.org/10.1016/j.concog.2019.04.006

Postle, B. R. (2016). Chapter 21 - The Hippocampus, Memory, and Consciousness. In S. Laureys, O. Gosseries, & G. Tononi (Eds.), *The Neurology of Conciousness (Second Edition)* (pp. 349-363). Academic Press.

Prosser, S. J. (2017). Rethinking the Specious Present. In I. Phillips (Ed.), *The Routledge Handbook of Philosophy of Temporal Experience* (pp. 146-156).

Pöppel, E. (1997a). A hierarchical model of temporal perception. *Trends in Cognitive Sciences*, *1*(2), 56-61. https://doi.org/10.1016/S1364-6613(97)01008-5

Pöppel, E. (1997b). The Brain's Way to Create "Nowness". In H. Atmanspacher & E. Ruhnau (Eds.), *Time, Temporality, Now: Experiencing Time and Concepts of Time in an Interdisciplinary Perspective* (pp. 107-120). Springer.

Pöppel, E. (2009). Pre-semantically defined temporal windows for cognitive processing. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *364*(1525), 1887-1896. https://doi.org/10.1098/rstb.2009.0015

Rao, S. M., Mayer, A. R., & Harrington, D. L. (2001). The evolution of brain activation during temporal processing. *Nature Neuroscience*, *4*(3), 317-323. https://doi.org/10.1038/85191

Roenneberg, T. (2012). *Internal Time: Chronotypes, Social Jet Lag, and Why You're So Tired*. Harvard University Press.

Roenneberg, T., & Merrow, M. (2003). The Network of Time: Understanding the Molecular Circadian System. *Current Biology*, *13*(5), R198-R207. https://doi.org/10.1016/S0960-9822(03)00124-6

Rönnberg, J., Holmer, E., & Rudner, M. (2019). Cognitive hearing science and ease of language understanding. *International Journal of Audiology*, *58*(5), 247-261. https://doi.org/10.1080/14992027.2018.1551631

Saenz, M., & Langers, D. R. M. (2014). Tonotopic mapping of human auditory cortex. *Hearing Research*, *307*, 42-52. https://doi.org/10.1016/j.heares.2013.07.016 (Human Auditory NeuroImaging)

Salin, P. A., & Bullier, J. (1995). Corticocortical connections in the visual system: structure and function. *Physiological Reviews*, *75*(1), 107-154. https://doi.org/10.1152/physrev.1995.75.1.107

Sereno, M. I., & Huang, R.-S. (2014). Multisensory maps in parietal cortex. *Current Opinion in Neurobiology*,

*24*(1), 39-46. https://doi.org/10.1016/j.conb.2013.08.014

Sewell, R. A., Schnakenberg, A., Elander, J., Radhakrishnan, R., Williams, A., Skosnik, P. D.,...D'Souza, D. C. (2013). Acute effects of THC on time perception in frequent and infrequent cannabis users. *Psychopharmacology*, *226*(2), 401-413. https://doi.org/10.1007/s00213-012-2915-6

Shimojo, S. (2014). Postdiction: its implications on visual awareness, hindsight, and sense of agency. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00196

Siclari, F., Baird, B., Perogamvros, L., Bernardi, G., LaRocque, J. J., Riedner, B.,...Tononi, G. (2017). The neural correlates of dreaming. *Nature Neuroscience*, *20*(6), 872-878. https://doi.org/10.1038/nn.4545

Singhal, I., Mudumba, R., & Srinivasan, N. (2022). In search of lost time: Integrated information theory needs constraints from temporal phenomenology. *Philosophy and the Mind Sciences*, *3*. https://doi.org/10.33735/phimisci.2022.9438

Singhal, I., & Srinivasan, N. (2021). Time and time again: a multi-scale hierarchical framework for time-consciousness and timing of cognition. *Neuroscience of Consciousness*, *2021*(2), niab020. https://doi.org/10.1093/nc/niab020

Singhal, I., & Srinivasan, N. (2024). Just one moment: Unifying theories of consciousness based on a phenomenological "now" and temporal hierarchy. *Psychology of Consciousness: Theory, Research, and Practice*, No-Pagination Specified-No Pagination Specified. https://doi.org/10.1037/cns0000393

Skow, B. (2015). *Objective Becoming*. Oxford University Press UK.

Strawson, G. (2009). *Selves: An Essay in Revisionary Metaphysics*. Oxford University Press.

Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience*, *5*(1), 42. https://doi.org/10.1186/1471-2202-5-42

Tononi, G. (2008). Consciousness as Integrated Information: a Provisional Manifesto. *The Biological Bulletin*, *215*(3), 216-242. https://doi.org/10.2307/25470707

Tononi, G. (2014). Why Scott should stare at a blank wall and reconsider (or, the conscious grid). *Shtetl-Optimized*.

Tooley, M. (1997). *Time, Tense, and Causation*. Oxford University Press.

Tsao, A., Sugar, J., Lu, L., Wang, C., Knierim, J. J., Moser, M.-B., & Moser, E. I. (2018). Integrating time from experience in the lateral entorhinal cortex. *Nature*, *561*(7721), 57-62. https://doi.org/10.1038/s41586-018-0459-6

Tsao, A., Yousefzadeh, S. A., Meck, W. H., Moser, M.-B., & Moser, E. I. (2022). The neural bases for timing of durations. *Nature Reviews Neuroscience*, *23*(11), 646-665. https://doi.org/10.1038/s41583-022-00623-3

VanRullen, R., & Koch, C. (2003). Is perception discrete or continuous? *Trends in Cognitive Sciences*, *7*(5), 207-213. https://doi.org/10.1016/S1364-6613(03)00095-0

Varela, F. J. (1999). The Specious Present: A Neurophenomenology of Time Consciousness. *Stanford University Press*, *Naturalizing phenomenology: Issues in contemporary phenomenology and cognitive science*, 41.

Vuust, P., Heggli, O. A., Friston, K. J., & Kringelbach, M. L. (2022). Music in the brain. *Nature Reviews Neuroscience*, *23*(5), 287-305. https://doi.org/10.1038/s41583-022-00578-5

Wackermann, J., Wittmann, M., Hasler, F., & Vollenweider, F. X. (2008). Effects of varied doses of psilocybin on time interval reproduction in human subjects. *Neuroscience Letters*, *435*(1), 51-55. https://doi.org/10.1016/j.neulet.2008.02.006

Wang, L., Mruczek, R. E. B., Arcaro, M. J., & Kastner, S. (2015). Probabilistic Maps of Visual Topography in Human Cortex. *Cerebral Cortex (New York, N.Y.: 1991)*, *25*(10), 3911-3931. https://doi.org/10.1093/cercor/bhu277

White, P. A. (2017). The three-second "subjective present": A critical review and a new proposal. *Psychological Bulletin*, *143*(7), 735-756. https://doi.org/10.1037/bul0000104

White, P. A. (2018). Is conscious perception a series of discrete temporal frames? *Consciousness and Cognition*, *60*, 98-126. https://doi.org/10.1016/j.concog.2018.02.012

Wiese, W. (2017). Predictive Processing and the Phenomenology of Time Consciousness: A Hierarchical Extension of Rick Grush?s Trajectory Estimation Model. In T. Metzinger & W. Wiese (Eds.), *Philosophy and Predictive Processing*.

Wittmann, M. (2013). The inner sense of time: how the brain creates a representation of duration. *Nature Reviews Neuroscience*, *14*(3), 217-223. https://doi.org/10.1038/nrn3452

Wittmann, M. (2015). Modulations of the experience of self and time. *Consciousness and Cognition*, *38*, 172-181. https://doi.org/10.1016/j.concog.2015.06.008

# Part II - Experiments

# *Chapter 3*

# Transcranial magnetic vs intracranial electric stimulation: a direct comparison of their effects via scalp EEG recordings[6]


## Abstract

*Background*

Single-pulse Transcranial Magnetic Stimulation (TMS) and Intracranial Electrical Stimulation (IES) are two widely used methodologies to assess cortical excitability and connectivity. However, a direct comparison of their effects under a common read-out has not been performed.

*Objective*

This study aims to fill this gap by using high-density scalp EEG to examine the neurophysiological impacts of TMS and IES. We analyze the amplitude, spectral, and spatiotemporal features of TMS- and IES-evoked potentials as well as the biophysical characteristics of their estimated electrical fields.

*Methods*

The dataset consisted of TMS evoked potentials recorded using hd-EEG acquired from healthy subjects (n=22) and IES evoked potentials recorded from drug-resistant epileptic patients (n=31) during wakefulness, and in smaller dataset, also during NREM sleep (n=12).

*Results*

We found that IES evoked EEG responses are slower and larger amplitude than those elicited by TMS, which in turn display higher spatiotemporal complexity and minimal suppression of high-frequency activity. Moreover, the estimated electric field revealed that the stimulation delivered by IES is strong and focal, that by TMS is weak and widespread. Despite these differences in amplitude, complexity spectral and biophysical features, TEPs and IEPs exhibited consistent state-dependent changes across wakefulness and NREM sleep.

*Conclusion*

Our findings highlight important differences and similarities in the neural responses elicited by TMS and IES, which are valuable for interpreting the literature and aligning non-invasive and invasive approaches. They also offer new insights into the mechanisms of cortical responses to stimulation under various stimulation parameters and brain states.


## 1. Introduction

Single-pulse cortical stimulation in combination with electrophysiological recordings offers a unique window to explore the input-output properties of cortical neurons and their large-scale interactions from a causal perspective. This perturb-and-measure approach has been widely employed to investigate cortical plasticity, excitability and connectivity. In humans, single-pulse direct cortical perturbations have been implemented in multiple ways, including non-invasive methods - such as Transcranial Magnetic Stimulation (TMS), in which magnetic pulses are delivered through the scalp - and invasive methods - such as Intracranial Electrical

[6] This chapter corresponds to the near finalized manuscript of the forthcoming article: **Comolatti, Renzo**; Hassan, Gabriel; Colombo, Michele; D'Ambrosio, Sasha; Russo, Simone; Casarotto, Silvia; Mikulan, Ezequiel; Pigorini, Andrea; Massimini, Marcello. *Transcranial magnetic vs intracranial electric stimulation: a direct comparison of their effects via scalp EEG recordings.*

Stimulation (IES), in which electric pulses are delivered through surgically implanted epi- or intra-cortical electrodes. TMS and IES have been similarly applied to explore brain responses across different physiological and pathological conditions. For example, both TMS and IES have been employed to probe cortical excitability (Bonato et al., 2006; Casali et al., 2010; Keller et al., 2018; Parmigiani et al., 2022), effective connectivity (Keller et al., 2014; Momi et al., 2021; Morishima et al., 2009; Trebaul et al., 2018) and to map cortical (Lemaréchal et al., 2022; Matsumoto et al., 2004a, 2007; Ozdemir et al., 2020) and subcortical (Russo et al., 2024) networks. Both techniques have also been employed to study altered states of consciousness such as NREM sleep (Massimini, 2005; Pigorini et al., 2015; Usami et al., 2019), anesthesia (Ferrarelli et al., 2010; Sarasso et al., 2015; Zelmann et al., 2023), severe brain injury (Casarotto et al., 2016; Mofakham et al., 2021; Rosanova et al., 2012) and in other pathological conditions such as epilepsy (Valentín et al., 2002; Valentin et al., 2008) and Parkinson's disease (Casarotto et al., 2019; Dale et al., 2022).

TMS and IES share fundamental characteristics: they are both of causal nature and bypass sensory and subcortical pathways, directly activating cortical neurons. However, the two stimulation techniques rely on different biophysical principles that may differentially shape the intensity and spatial extent of the stimulating field, leading to potentially divergent responses. Directly comparing the electrophysiological effects of TMS and IES is thus key to interpret the current literature, align non-invasive and invasive approaches, and design future experiments. Yet, such a direct comparison has been so far hindered by the lack of a comparable read-out, as the cortical effects of TMS and IES have been previously recorded at different levels: scalp EEG in the first case (Massimini, 2005; Paus et al., 2001) and epi- or intra-cranial EEG in the second case (Matsumoto et al., 2004b; Trebaul et al., 2018).

In the present study, we directly compare for the first time the amplitude, spectral and spatio-temporal features of TMS-evoked potentials (TEPs) and IES-evoked potentials (IEPs) at the common scale of high-density scalp EEG recordings. Our dataset comprises 90 TEPs recorded from 22 healthy subjects and 228 IEPs acquired in 31 epileptic patients undergoing presurgical evaluation. Recordings were performed during wakefulness, and in a subset also during sleep, while TMS and IES were delivered with stimulation parameters used in typical research and clinical protocols.

By maintaining scalp EEG as the common read-out for both TMS and IES, we found major differences in the magnitude and nature of their cortical effects. Seen from the scalp, IES elicits responses that are slower and up to an order of magnitude larger than TMS. In spite of major differences in amplitude, complexity and spectral features, TEPs and IEPs undergo changes that are consistent across brain states. These results highlight differences and commonalities between the electrophysiological effects of TMS and IES that are relevant for aligning non-invasive and invasive measurement and provide novel insight on the mechanisms of cortical responses to direct perturbations across different stimulation parameters and global brain states.

## 2. Material and Methods

### *Participants, data acquisition and preprocessing*

TMS-EEG

The TMS-EEG dataset included in the present study comprises 90 sessions collected from 22 healthy awake subjects (4.1±1.5 sessions per subject), in addition to 12 paired sessions collected during NREM sleep from 12 participants. Each TMS session consisted of a minimum of 200 pulses (mean±std=230±27) per area per subject administered with an inter-stimulus interval randomly jittered between 2000 and 2300 ms. Biphasic pulses lasting 230μs were delivered using a focal figure-of-eight coil at estimated E-field intensities of approximately 120 V/m (Fig. 1A). EEG data were recorded using a TMS-compatible 64-channel amplifier (Nexstim Ltd.)

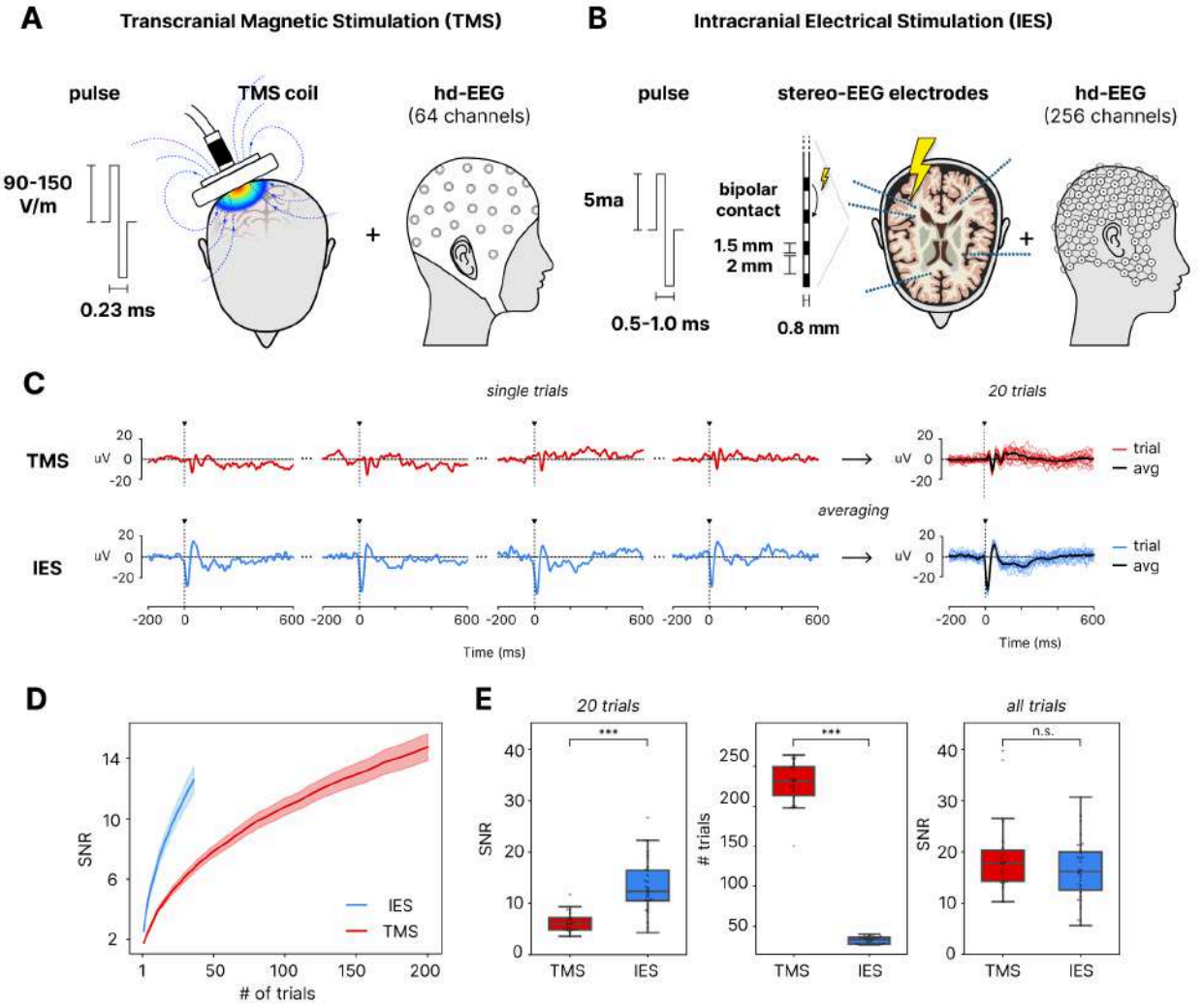



**Fig. 1. Experimental setup of Transcranial Magnetic Stimulation (TMS) and Intracranial Electrical Stimulation (IES) using hd-EEG and quantification of signal-to-noise ratio.** (A) Schematic of the TMS-EEG procedure performed in healthy subjects depicting the biphasic pulse delivered with the TMS coil to the scalp, resulting in a magnetic field intensity between 90-150 V/m over a duration of 0.23 ms, and the recording setup with a 64-channel hd-EEG. (B) Illustration of the IES-EEG setup used on epileptic patients undergoing presurgical evaluation comprising the intracranially implanted stereotactic electrodes (SEEG) which deliver biphasic pulse of 0.5-1.0 ms duration through bipolar contacts (separated by 2mm) at constant intensity of 5mA, alongside the simultaneous recording setup with a 256-channel hd-EEG. (C) Scalp EEG traces from single trials of TMS (top) and IES (bottom) for the strongest responding channel, showing time-locked potentials elicited by the single-pulses (black triangle). To the right, the evoked average (black trace) of 20 single trials (colored traces) depicting the build up of signal-to-noise (SNR) through trial-averaging. (D) Relationship between SNR and the number of trials utilized for trial-averaging, obtained by bootstrapping procedure. (E) Boxplots depicting SNR comparisons for TMS

and IES after 20 trials (left), the total number of trials in each recording (middle), and SNR from all trials (right). Statistical significance is denoted by asterisks (n.s.=p>0.05, * = p<0.05, ** = p<0.01, *** = p<0.001)

TMS-EEG data were processed similarly to (Rosanova et al., 2018). The stimulation artifact was first removed and hd-EEG data was high-pass filtered at 0.5 Hz, splitted into epochs and bad trials and channels were rejected by visual inspection. Epochs were re-referenced to the average reference and baseline corrected. After Independent Component Analysis (ICA) was applied to remove EMG and EOG activity, the signal was low-pass filtered at 45 Hz and down-sampled to 1000 Hz (see Supplementary Materials for further details on data acquisition and preprocessing).

IES-EEG

The IES-EEG dataset was obtained from patients undergoing intracranial monitoring for pre-surgical evaluation of drug-resistant epilepsy (Cossu et al., 2005). A total of 231 sessions were acquired from 31 subjects during wakefulness (7.5±3.4 sessions per subject; previously published data available at https://osf.io/wsgzp/ (Parmigiani et al., 2022)). In a subset of 13 patients, 54 paired sessions were also acquired during NREM sleep (8.3 ±4.4 paired sessions per subject). Each session consisted of circa thirty bipolar biphasic pulses (mean±std=32±4.2) of 500 and 1000 μs duration delivered at intervals ranging from 1 to 5 seconds. Pulses were delivered through stereotactically implanted (SEEG) intracerebral electrodes at 5 mA intensity between adjacent contact pairs. Recordings were simultaneously conducted using 256 channels high-density scalp EEG (Fig. 1B).

IES-EEG data were processed using a pipeline analogous to that employed for TMS-EEG data as detailed in (Parmigiani et al., 2022). First, channels and trials contaminated by noise, muscle activity or spontaneous interictal epileptic discharges were rejected using a semi-automatic procedure, manually verified by an expert electrophysiologist. Next, the stimulation artifact was removed and data were band-pass filtered (0.5-45 Hz) and epoched. Finally, trials were re-referenced to the average and baseline corrected and ICA was applied to remove EOG and residual EMG activity.

### *Data Analysis*

Signal-to-noise ratio (SNR) measures signal strength time-locked to stimulation with respect to the pre-stimulus background activity. SNR was calculated as the square root of the ratio of average power between the early response (0 to 80 ms) and the baseline (-300 to -5 ms), at the channel with the highest power in the initial 80 ms. The influence of the number of trials on the SNR was evaluated using a bootstrap method: for a given number of $k$ trials, SNR was computed on 100 surrogate responses formed by randomly selecting $k$ trials without replacement.

The global mean field power (GMFP) estimates the overall power evoked at each time point across all channels and corresponds to the spatial standard deviation of the signal. The GMFP was then baseline corrected to quantify the power of the evoked response exceeding the power in the spontaneous baseline activity.

The Perturbational Complexity Index state-transition ($PCI^{ST}$) was used to assess the spatiotemporal complexity of the evoked potentials (Comolatti et al., 2019). $PCI^{ST}$ gauges the ability of thalamocortical circuits to engage in complex causal interactions, by jointly quantifying the spatial diversity and temporal differentiation of brain responses (Casali et al., 2013). $PCI^{ST}$ was computed over the 0-600 ms response window with remaining

parameters following (Comolatti et al., 2019). Code utilized for calculations is available at github.com/renzocom/PCIst.

We evaluated the modulation of high-frequency (≥ 20Hz) EEG oscillations induced by the stimulation using the event-related spectral perturbation (ERSP) method (Grandchamp & Delorme, 2011) following procedure used in (Pigorini et al., 2015; Rosanova et al., 2018). We computed the high-frequency power (HFp) of each channel as the averaged time course of the significant (bootstrap method; $\alpha \leq 0.01$, 500 permutations) instantaneous high-frequency power over the interval between 120 and 220 ms. From the distribution of HFp across channels we computed three metrics to assess the suppression of high-frequency: (i) the *extent* of suppression was measured as the percentage of suppressing (HFp < 0) channels (%Ch HFsup); (ii) its maximal *intensity* as the minimum HFp value across all channels (max HFsup); and (iii) its *total* amount as the integral of HFp < 0 across channels normalized by the total number of channels (total HFsup).

### *Statistical analysis*

Statistical analyses were performed in the following manner. For each subject, metrics were first averaged across sessions and then tested across groups using t-tests. Specifically, differences between TMS and IES, which involve different numbers of subjects, were assessed using Welch's t-test. Conversely, comparisons between wakefulness and NREM sleep, within stimulation method, were conducted using a paired Student's t-test, correcting for multiple comparisons using the False Discovery Rate (FDR) method.

### *Simulation of electric field*

E-fields are instantaneous estimates of the electrical potential gradients induced in the brain tissue by the respective stimulation devices–the magnetic coil in TMS and the bipolar contacts of the SEEG electrodes for IES. The TMS and IES E-fields were computed with the finite element method (FEM) using a realistic volume conductor model based on the MNI152 template (Fonov et al., 2009). Conductivities were set at 0.14 S/m for white matter and 0.33 S/m for gray matter (Vorwerk et al., 2014). TMS E-field was calculated using SimNIBS 4.0 (Saturnino et al., 2019) using a 70 mm figure-of-eight coil template (specifically, MagVenture MC-B70 for its similarity to the Nexstim coil used in the experiments), and the IES E-field with LeadDBS 3.0 using a template of the same electrode used in the clinical IES protocol (Neudorfer et al., 2023). Simulations were conducted at standard (120 V/m for TMS and 5mA for IES) and high intensities (160/Vm and 10mA, respectively). The TMS coil was placed over the crown of a cortical gyrus with the field oriented perpendicular to it to maximize the cortical efficacy (Aberra et al., 2020). Intracranial bipolar contacts for IES stimulation were positioned in the gyrus aligned with the TMS coil's normal projection. The E-field peak strength was defined as the 99.9% percentile of the field strength distribution to avoid potential outliers in the FEM simulation (Aberra et al., 2020). The spatial decay of the E-field was calculated by finding, for every E-field threshold value, the farthest distance in 3D anatomical coordinates that was still above it.

## 3. Results

We compared the effects of single-pulse TMS and IES by first analyzing scalp hd-EEG responses registered during wakefulness. For both TMS and IES, stimulation parameters followed the typical settings used to effectively elicit cortical responses in experimental and clinical settings within safe operational ranges (for example (Casarotto et al., 2016; Sarasso et al., 2020) for TMS, and (Parmigiani et al., 2022; Trebaul et al., 2018; Usami et al., 2019) in the case of IES; see Methods for details).

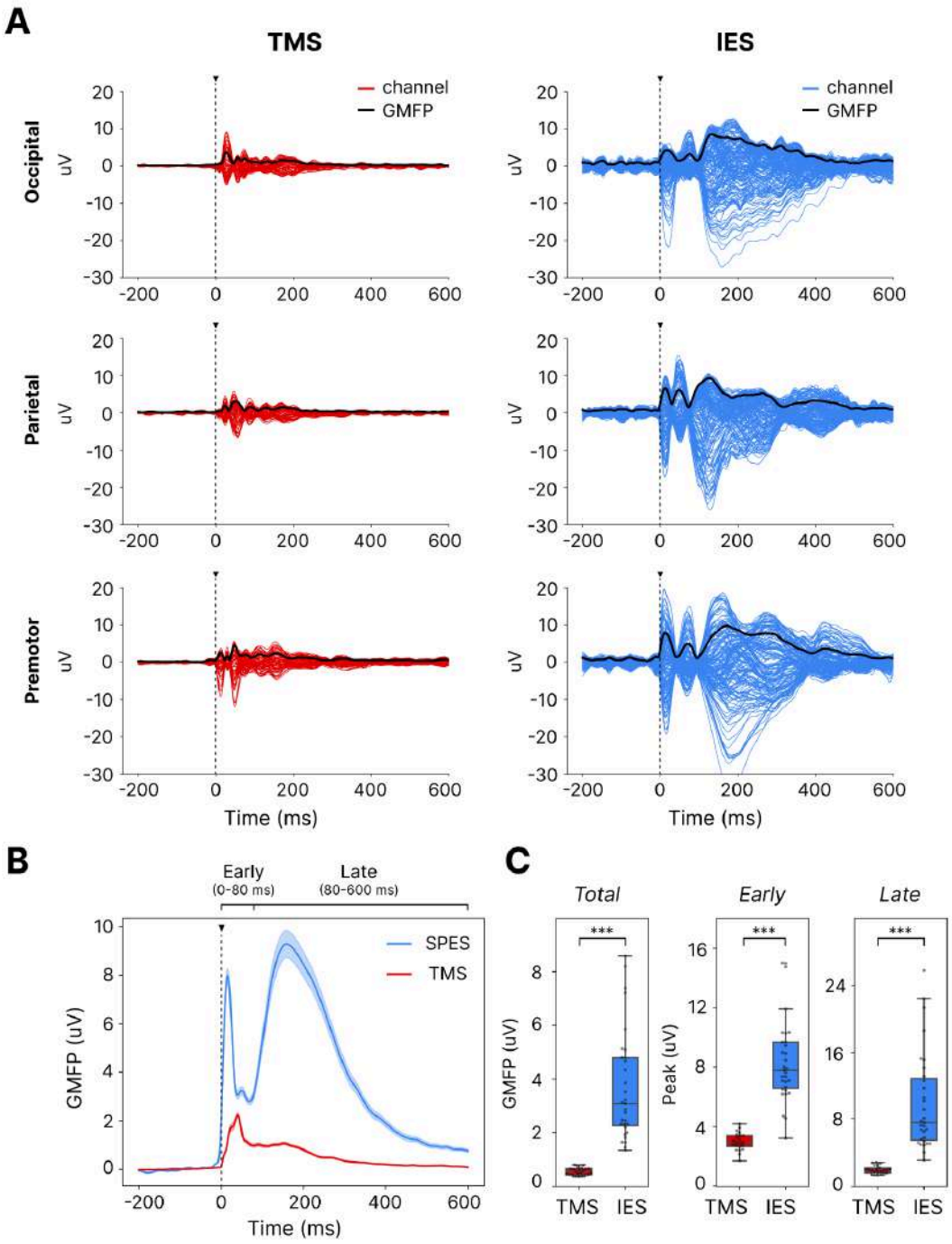


**Fig. 2. GMFP comparisons between TMS and IES Evoked Potentials.** (A) Butterfly plots of representative trial-averaged TEPs and IEPs from occipital, parietal, and premotor brain regions. The colored traces depict the evoked response across EEG channels, with the GMFP overlaid in black, illustrating the power evoked globally across the EEG channels over time. (B) Grand averages of the baseline corrected GMFP time course for TEP (red) and IEP (blue) and their respective standard errors (light shadow), illustrating the differences in evoked power over time, segmented into early (0-80 ms) and late (80-600 ms) intervals. (C) Boxplots compare the total GMFP across the full response interval (0-600 ms), and the peak GMFP amplitudes for the early (0-80 ms), and late (80-600 ms) intervals. Statistical significance is denoted by asterisks (n.s.=p>0.05, * = p<0.05, ** = p<0.01, *** = p<0.001).

### *IES evokes EEG responses with higher signal-to-noise ratio than TMS*

As displayed in Fig. 1C, both TMS and IES induced time-locked potentials that were visible from the scalp EEG at the single trial level and reproducible from trial to trial. Once averaged, both TMS and IES responses yielded evoked potentials with high signal-to-noise ratio (SNR) (Fig. 1C, right). Nonetheless, TEPs were less prominent with respect to the pre-stimulus baseline than IEPs. The SNR of IEPs increased more rapidly as a function of the number of trials averaged than TEPs (Fig. 1D). After averaging twenty trials, IEPs showed a twofold difference in SNR with respect to TEPs (mean±std, TMS=6.3±2.0, IES=13±5.1, $p=2.0x10^{-7}$) (Fig. 1E, left). This difference was offset by the larger number of trials collected in the case of TMS (TMS=230±27, IES=32±4.2, $p=1.06x10^{-39}$) (Fig. 1E, middle), resulting in comparable SNR that did not differ significantly once all trials were considered (SNR, TMS=19±7.3, IES=16±5.9, p=0.12) (Fig 1E, right).

### *IES evokes EEG responses that are larger than those evoked by TMS*

At comparable SNR, trial-averaged evoked potentials revealed both commonalities as well as noticeable differences between TMS and IES. Fig. 2A depicts, in a common scale, the butterfly plots of representative trial-averaged TEPs and IEPs from different brain regions (occipital, parietal and premotor) overlaid by the respective global mean field power (GMFP) (black traces). Both stimulation methods evoked high-amplitude and long lasting potentials, displaying a composite response with non-trivial spatiotemporal activation profiles, which varied across stimulation sites. Nonetheless, the global power of IEPs was strikingly larger than that of TEPs, as illustrated by the direct comparison between the grand averages of the GMFP evoked by the two stimulation modalities (Fig 2B). Significant differences were found when considering the average GMFP across the whole time course (mean±std, TMS=0.55±0.13 uV, IES=3.3±2.3 uV, $p=2.8x10^{-9}$) (Fig. 2C, left), as well as when considering separately the peak amplitudes of the early (0-80 ms, TMS=3.0±0.58 uV, IES=8.4±2.8 uV, $p=1.1x10^{-11}$) (Fig. 2C, middle) and late (80-600 ms, TMS=2.0±0.46 uV, IES=9.9±5.6 uV, $p=5.8x10^{-8}$) components of the GMFP (Fig 2C, right).

### *The estimated electric field is weak and widespread for TMS, strong and focal for IES*

The common read-out of scalp EEG recordings revealed a substantial difference in the magnitude of the overall response evoked by TMS and IES. We next investigated this finding in light of the differences in the strength and spatial extent of the electric fields (E-fields) generated in the cortex by the two stimulation modalities (see Methods for details). Figure 3A depicts the simulated E-field of a parietal cortex stimulation

(Brodmann area 7) at standard stimulation intensity for TMS (Fig. 3A, *top*) and IES (Fig. 3A, *bottom*) in a common scale.

The E-field pattern of the two stimulation differed substantially: the E-field computed for IES showed a high intensity profile concentrated at the stimulation site, whereas the E-field computed for TMS was weaker but more widespread. Specifically, the maximum E-field value of IES was more than an order of magnitude greater than that of TMS (peak of E-field, TMS=120 V/m, IES=1560 V/m). On the other hand, the E-field induced by TMS was significantly less focal, exhibiting a more gradual spatial decay (Fig. 3B). For example, at 90 V/m – the typical E-field strength of the resting motor threshold in TMS (Fecchio et al., 2017) – TMS E-field was above it at 33.6 mm from the target site whereas this region was limited to a radius of 6.3 mm, almost one fifth, in the case of IES. These differences were preserved across stimulation intensities (Fig. 3B) and sites.

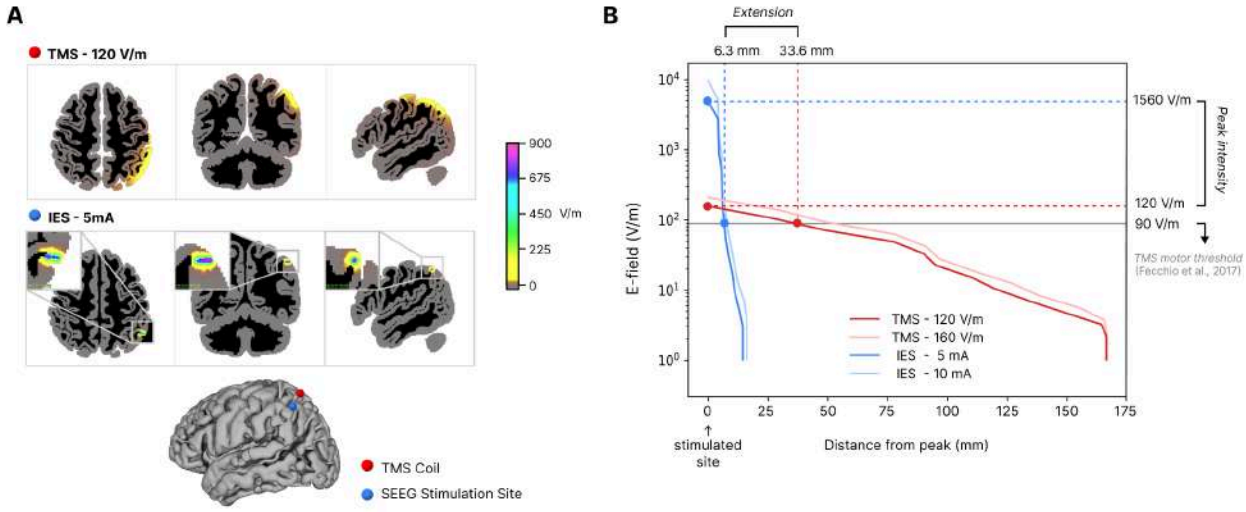



**Fig. 3. Simulated electric fields induced by TMS and IES.** (A) Simulated E-fields induced during standard intensity stimulation to parietal cortex by TMS and IES on a MNI152 template. The upper panels show E-fields for IES at 5mA, highlighting the focused high amplitude area close to the stimulation bipolar contacts. The lower panels depict the more diffuse and weaker E-field for TMS at 120 V/m stimulation intensity. The placement of the TMS coil (red) and SEEG stimulation site (blue) are indicated on a 3D brain model (bottom). (B) Spatial decay of E-field with distance from the stimulated site for TMS (red) and IES (blue), for both standard and high stimulation intensity. The horizontal dashed lines indicate maximum E-field strength (intensity), and vertical dashed lines indicate the E-field strength at an operational threshold of 90 V/m (gray line) (extension).

### *The amplitude of IEPs and TEPs differs in wakefulness but becomes more similar in NREM sleep*

Having explored the differences in the magnitude of the response and the different E-field patterns generated by IES and TMS, we moved to investigate the effects of changes between brain states. To this aim, we analyzed a cohort of paired TEPs and IEPs recorded during wakefulness and NREM sleep.

During wakefulness, TMS elicited faster EEG components with a richer spatiotemporal profile as compared to IES, as indicated by the higher number of peaks in the GMFP (mean±std, TMS=5.9±0.73, IES=4.6±0.63, $p=6.0x10^{-9}$) (Fig. 4A). Upon falling asleep, the EEG response to TMS became larger and slower, and thus more similar to the one triggered by IES, as substantiated by the analysis of the GMFP time course (Fig. 4B). The average GMFP of both TEPs and IEPs was higher during NREM sleep than during wakefulness

($TMS_{wake}$=0.50±0.35 uV, $TMS_{sleep}$=3.8±2.9 uV, p=3.4x$10^{-3}$; $IES_{wake}$=4.4±3.3 uV, $IES_{sleep}$=6.3±3.2 uV, p=3.8x$10^{-4}$). Interestingly, the GMFP of TEPs during NREM sleep resembled the GMFP profile observed in IEPs in both wakefulness and NREM sleep, characterized by an early peak and a slow sustained late activation. Indeed, the total GMFP of TEPs recorded during NREM sleep was not significantly different from the GMFP of IEPs recorded during both wakefulness (p=0.64) and NREM sleep (p=0.07).

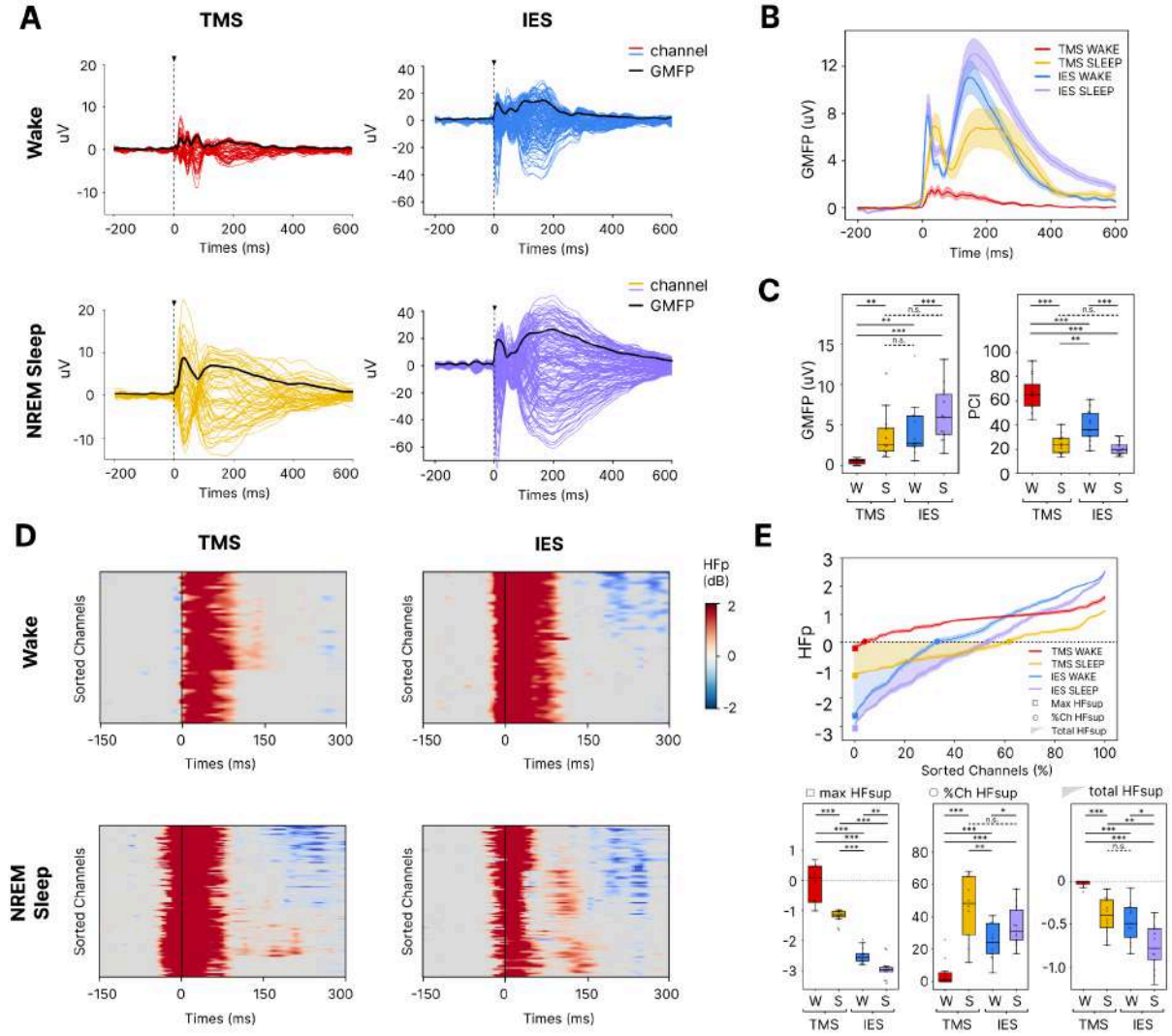



**Fig. 4. State-Dependent Changes in TMS and IES Evoked Responses.** (A) Butterfly plots of representative TMS (left column) and IES (right column) evoked potentials during wakefulness (top) and NREM sleep (bottom). Each trace represents a single EEG channel response with the GMFP overlaid as a black line, highlighting the differences in waveform and complexity between states and stimulation types. (B) GMFP time course for TMS and IES during wakefulness and NREM sleep, with shaded areas indicating standard deviation. These illustrate changes in GMFP amplitude across brain states. (C) Boxplots of GMFP and PCI, quantifying the evoked potential's amplitude and complexity, respectively, for TMS and IES during wakefulness (W) and NREM sleep (S). (D) High-frequency power (>20Hz) time courses across EEG channels during wakefulness and NREM sleep for TMS and IES, with blue indicating suppression (negative dB values) and red indicating an increase (positive dB values). Channels are sorted by their respective high-frequency power averaged across 120-220 ms (HFp). (E) Distributions of HFp sorted across channels for TMS and IES during wakefulness and NREM sleep (traces depict average and standard errors). Highlighted are the extent (circle, %Ch HFsup), maximum degree (square, Max HFsup) and overall (shaded area below zero, Total HFsup) suppression of high-frequency for each condition. (F) Boxplots show the metrics

derived from the HFp distribution across channels: the proportion of channels showing suppression (%Ch HFsup), the maximal suppression across all channels (max HFsup), and the integral of HFp suppression across channels normalized by total number of channels (total HFsup), across conditions and stimulation types. These metrics gauge the extent, intensity, and overall suppression of HFp across channels, comparing TMS and IES during wakefulness and NREM sleep. Statistical significance is denoted by asterisks (n.s.=p>0.05, * = p<0.05, ** = p<0.01, *** = p<0.001).The spatiotemporal complexity of TEPs and IEPs differs but shows consistent changes across brain states

We then assessed the richness of the spatial and temporal profile of the evoked responses by computing the Perturbational Complexity Index (PCI), a measure of spatiotemporal complexity of evoked responses (Casali et al., 2013; Comolatti et al., 2019). During wakefulness, PCI values were systematically higher for TEPs than IEPs (p=8.9x10-5) reflecting the richer spatiotemporal profile observed in the TEPs and the slower and more stereotypical features of IEPs. In line with previous findings (Comolatti et al., 2019), however, PCI revealed state-dependent changes in complexity for both TMS and IES, being consistently higher in wakefulness than in NREM sleep (mean±std, TMSwake=66±14, TMSsleep=25±7.8, p=4.1x10-6; IESwake=39±13, IESsleep=21±5.3, p=8.9x10-5). Notably, the PCI computed on IEPs during wakefulness remained consistently above the values obtained in NREM sleep not only within IES, but also with respect to the values obtained with TMS (p=4.1x10-3). Lastly, consistent with the observed convergence of TEP and IEP's waveform, the difference in PCI values between TEPs and IEPs was non-significant during NREM sleep (p=0.22).

***Unlike TMS, IES induces suppression of high-frequency activity also during wakefulness***

Last, we analyzed the modulation of high-frequency power (>20Hz, HFp) in the responses as a proxy for the suppression or increase of neuronal activity induced by TMS and IES (Cash et al., 2009; Mukovski et al., 2007). Fig. 4D displays the high-frequency time courses of each channel sorted by their respective HFp, for representative signals of each condition (see panel A). In the case of TMS, an initial increase of neuronal activity was followed by a suppression of high-frequency power only in NREM sleep but not in wakefulness, as previously reported (Fecchio et al., 2017; Rosanova et al., 2018). Conversely, the IES-induced initial activation was invariably followed by a clear-cut suppression of neuronal activity as indexed by negative HFp values in many channels both in wakefulness and NREM sleep (Pigorini et al., 2015). The distribution of HFp across EEG sensors depicted in Fig. 4E shows that, while HFp remained positive in most channels in the case of TMS during wakefulness, IES induced suppression of high-frequency in about 30% of channels during wakefulness. During NREM, instead, both TMS and IES induced a suppression of high-frequency in most channels.

The extension (%Ch HFsup), intensity (maxCh HFsup) and overall amount (total HFsup) of high-frequency suppression was significantly stronger during NREM sleep compared to wakefulness for both TMS and IES (%Ch HFsup, TMS $p=5.5x10^{-5}$, IES $p=3.8x10^{-2}$; max HFsup, TMS $p=4.3x10^{-4}$, IES $p=3.7x10^{-3}$; total HFsup, TMS $p=1.1x10^{-4}$, IES $p=1.3x10^{-2}$) (Fig. 4F). During wakefulness, TMS responses displayed little to no suppression of high-frequency (mean±std, total HFsup=-0.024±0.038), both in terms of its extension across channels (%Ch HFsup=4.7±7.6) and maximal intensity (max HFsup=-0.097±0.65). In contrast, IES responses exhibited a significant amount of high-frequency suppression during wakefulness, comparable to the total high-frequency suppression observed in NREM sleep for TMS (total HFsup, $IES_{wake}$=-0.52±0.39,

$TMS_{sleep}$=-0.41±0.2, p=0.39). Taken together, these results suggest that while IES induces significant suppression of high-frequencies in both wakefulness and NREM sleep, TMS does so only during NREM sleep.

## 4. Discussion

### *Divergence between IEPs and TEPs during wakefulness: the role of stimulation parameters*

During wakefulness, IES evoked global EEG responses were up to an order of magnitude larger than those triggered by TMS (Fig.1). Moreover, while TMS evoked multiple waves of recurrent activation and negligible suppression of high-frequency activity across channels, the initial deflections triggered by IES were rapidly followed by a large negative slow wave associated with a suppression of high-frequency oscillations. As demonstrated by intracranial and extracranial studies in animals and humans, large EEG negative waves associated with high-frequency suppression correspond to the occurrence of a silent period (OFF-period) in cortical neurons (Cash et al., 2009; Mukovski et al., 2007). This tendency of cortical neurons to fall into a silent OFF-period after an initial activation reflects activity-dependent mechanisms such as Na+/Ca++ dependent K+ currents (Cattani et al., 2023; Compte et al., 2003) and/or active inhibition (Timofeev et al., 2001). The differential effects of IES and TMS can be partially explained by the distinct E-field patterns generated by each stimulation technique (Fig. 3) and by their interactions with activity-dependent mechanisms. Indeed, the IES electric field is up to five times more focal than that induced by TMS and, at the typical stimulation parameters used in research and clinical settings, over ten times more intense. A parsimonious explanation is that, unlike the weaker and more diffuse input provided by TMS (Opitz et al., 2011; Siebner et al., 2022), the intense and highly-focused E-field (Astrom et al., 2015; McIntyre et al., 2004) induced by IES strongly recruits active inhibition, resulting in an OFF period which tends to curtail recurrent excitatory activity during wakefulness (Hajnal et al., 2024; Hao et al., 2016).

### *Convergence between IEPs and TEPs upon falling asleep: the role of neuromodulation*

Upon falling asleep, we found an attenuated difference between TEPs and IEPs, since in this state also TMS triggered a positive deflection followed by a large negative component associated with suppression of high-frequency activity. During NREM sleep, changes in neuromodulation are known to increase both Na+/Ca++ dependent K+ currents (Compte et al., 2003) and inhibition (Timoveev et al., 2001) in cortical circuits. Under this condition of increased adaptation, even a relatively weaker cortical activation such as that represented by TMS may trigger an OFF-period (Cattani et al., 2023), thus converging towards the response elicited by IES.

In this way, changes in neuromodulation upon falling asleep may reduce the range of local activation levels that cortical circuits can withstand. During wakefulness, in the presence of low adaptation/inhibition levels, local cortical circuits plunge into an OFF-period only after the powerful input represented by IES but can withstand the relatively weaker input induced by TMS, resulting in multiple waves of recurrent activation. During NREM sleep, increased adaptation/inhibition mechanisms dramatically reduce this window and both inputs lead to slow responses associated with suppressed activity. This finding is compatible with results from computer simulation (Cattani et al., 2023) and with the general notion that the dynamic range of cortical

circuits and their information capacity is reduced during NREM sleep (Tononi et al., 2016; Tononi & Massimini, 2008). In this vein, it is worth noting that in pathological conditions associated with increased adaptation and inhibition, such as stroke and traumatic brain injury, TMS can evoke local sleep-like cortical responses during wakefulness that are similar to those evoked by IES (Rosanova et al., 2018; Sarasso et al., 2020; Tscherpel et al., 2020).

### *Aligning IES and TMS*

The suppression of high-frequency cortical activity associated with IES observed during wakefulness did not seem to obliterate the emergence of physiological cortical interactions beyond the immediate area of stimulation – as evidenced by the higher PCI values recorded during wakefulness compared to NREM sleep. The present findings help reconcile the apparently conflicting views on IES. While some studies used IES as a tool to probe global network connectivity (Entz et al., 2014; Keller et al., 2014; Lemaréchal et al., 2022), others emphasize the role of inhibition (Hajnal et al., 2024), which may hinder signal propagation across transsynaptic chains of activation (Borchers et al., 2012). Our findings suggest that IEPs during wakefulness are characterized by a hybrid pattern of cortical dynamics wherein a strong local inhibition does not necessarily prevent the build-up of recurrent interactions beyond the stimulated site. This is in line with previous observations from intracranial stereo-EEG recordings, which suggest that during wakefulness IES triggers a slow response associated with high-frequency suppression in contacts adjacent to the stimulated site, whereas remote contacts show more complex responses (Pigorini et al., 2015). The finding present that IEPs exhibited state-dependent modulation between wakefulness and NREM sleep consistent with that observed in TEPs, suggest that both stimulation techniques can induce large-scale responses that are informative about the global state of thalamocortical networks (Comolatti et al., 2019; Fecchio et al., 2021).

The present comparison is also relevant to align and interpret responses to direct cortical stimulation across species and experimental models. In a growing body of literature, IEPs have been used as a tool to replicate TMS-EEG results during wakefulness, sleep and anesthesia in rats (Arena et al., 2020; Cavelli et al., 2023), mice (Cavelli et al., 2023; Claar et al., 2023) and cortical slices (D'Andola et al., 2018). These works have shown differences in the profile of local activation, wherein slow responses and suppressed activity are often present also during wakefulness, as well as consistent changes across global states. The possibility that IES may saturate adaptation and inhibition mechanisms also in these models, warrant some caution when drawing parallels with TEPs. To tackle this issue, it would be interesting to explore the parameter space of intracranial stimulation to attenuate the recruitment of activity-dependent mechanisms and match the effects of TMS. This could potentially be obtained, for example, by lowering stimulation intensity while increasing the number of trials, and/or by diluting the E-field within a larger volume by stimulating across contacts farther apart instead of adjacent ones (Hays et al., 2023; Paulk et al., 2022). Such exploration would also provide key empirical data to better understand the general activation mechanisms of TMS and their exquisite sensitivity to local (Sarasso et al. 2020; Tscherpel et al., 2020) and global changes in the state of cortical circuits (Casarotto et al. 2016).

In sum, besides helping align non-invasive and invasive stimulation methods and their respective findings through a common read-out, our study sheds light on fundamental input-output properties of cortical circuits across different activation parameters and brain states.

## Acknowledgements

This work was supported by the European Union's Horizon 2020 Framework Program for Research and Innovation under the Specific Grant Agreement No. 945539 (Human Brain Project SGA3), by the Tiny Blue Dot Foundation, by the European Research Council (ERC-2022-SYG - 101071900 - NEMESIS). S.D. is supported by the Italian Ministry of Health – RicercaCorrente 2024.

## Bibliography

Comolatti, R., Pigorini, A., Casarotto, S., Fecchio, M., Faria, G., Sarasso, S., Rosanova, M., Gosseries, O., Boly, M., Bodart, O., Ledoux, D., Brichant, J.-F., Nobili, L., Laureys, S., Tononi, G., Massimini, M., & Casali, A. G. (2019). A fast and general method to empirically estimate the complexity of brain responses to transcranial and intracranial stimulations. *Brain Stimulation*, *12*(5), 1280–1289. https://doi.org/10.1016/j.brs.2019.05.013Fecchio, M., Russo, S., Parmigiani, S., Mazza, A., Viganò, A., Casali, A. G., Comolatti, R., Mikulan, E., Massimini, M., & Rosanova, M. (2021). Spatiotemporal specificity of TMS-evoked potentials versus sensory evoked potentials. *Brain Stimulation: Basic, Translational, and Clinical Research in Neuromodulation*, *14*(6), 1688. https://doi.org/10.1016/j.brs.2021.10.320

Integrated Information Wiki. (2024). *Center for Sleep and Consciousness, University of Wisconsin–Madison*. https://www.iit.wiki

Aberra, A. S., Wang, B., Grill, W. M., & Peterchev, A. V. (2020). Simulation of transcranial magnetic stimulation in head model with morphologically-realistic cortical neurons. *Brain Stimulation*, *13*(1), 175–189. https://doi.org/10.1016/j.brs.2019.10.002

Arena, A., Comolatti, R., Thon, S., Casali, A. G., & Storm, J. F. (2020). *General anaesthesia disrupts complex cortical dynamics in response to intracranial electrical stimulation in rats* [Preprint]. Neuroscience. https://doi.org/10.1101/2020.02.25.964056

Astrom, M., Diczfalusy, E., Martens, H., & Wardell, K. (2015). Relationship between neural activation and electric field distribution during deep brain stimulation. *IEEE Transactions on Bio-Medical Engineering*, *62*(2), 664–672. https://doi.org/10.1109/TBME.2014.2363494

Bonato, C., Miniussi, C., & Rossini, P. M. (2006). Transcranial magnetic stimulation and cortical evoked potentials: A TMS/EEG co-registration study. *Clinical Neurophysiology*, *117*(8), 1699–1707. https://doi.org/10.1016/j.clinph.2006.05.006

Borchers, S., Himmelbach, M., Logothetis, N., & Karnath, H.-O. (2012). Direct electrical stimulation of human cortex—The gold standard for mapping brain functions? *Nature Reviews Neuroscience*, *13*(1), 63–70. https://doi.org/10.1038/nrn3140

Casali, A. G., Casarotto, S., Rosanova, M., Mariotti, M., & Massimini, M. (2010). General indices to characterize the electrical response of the cerebral cortex to TMS. *NeuroImage*, *49*(2), 1459–1468. https://doi.org/10.1016/j.neuroimage.2009.09.026

Casali, A. G., Gosseries, O., Rosanova, M., Boly, M., Sarasso, S., Casali, K. R., Casarotto, S., Bruno, M.-A., Laureys, S., Tononi, G., & Massimini, M. (2013). A Theoretically Based Index of Consciousness Independent of Sensory Processing and Behavior. *Science Translational Medicine*, *5*(198), 198ra105-198ra105. https://doi.org/10.1126/scitranslmed.3006294

Casarotto, S., Comanducci, A., Rosanova, M., Sarasso, S., Fecchio, M., Napolitani, M., Pigorini, A., G. Casali, A., Trimarchi, P. D., Boly, M., Gosseries, O., Bodart, O., Curto, F., Landi, C., Mariotti, M., Devalle, G., Laureys, S., Tononi, G., & Massimini, M. (2016). Stratification of unresponsive patients by an independently validated index of brain complexity: Complexity Index. *Annals of Neurology*, *80*(5), 718–729. https://doi.org/10.1002/ana.24779

Casarotto, S., Turco, F., Comanducci, A., Perretti, A., Marotta, G., Pezzoli, G., Rosanova, M., & Isaias, I. U. (2019). Excitability of the supplementary motor area in Parkinson's disease depends on subcortical damage. *Brain Stimulation*, *12*(1), 152–160. https://doi.org/10.1016/j.brs.2018.10.011

Cash, S. S., Halgren, E., Dehghani, N., Rossetti, A. O., Thesen, T., Wang, C., Devinsky, O., Kuzniecky, R., Doyle, W., Madsen, J. R., Bromfield, E., Eross, L., Halasz, P., Karmos, G., Csercsa, R., Wittner, L., & Ulbert, I. (2009). The Human K-Complex Represents an Isolated Cortical Down-State. *Science*, *324*(5930), 1084–1087. https://doi.org/10.1126/science.1169626

Cattani, A., Galluzzi, A., Fecchio, M., Pigorini, A., Mattia, M., & Massimini, M. (2023). Adaptation Shapes Local Cortical Reactivity: From Bifurcation Diagram and Simulations to Human Physiological and Pathological Responses. *eNeuro*, *10*(7). https://doi.org/10.1523/ENEURO.0435-22.2023

Cavelli, M. L., Mao, R., Findlay, G., Driessen, K., Bugnon, T., Tononi, G., & Cirelli, C. (2023). Sleep/wake changes in perturbational complexity in rats and mice. *iScience*, *26*(3), 106186. https://doi.org/10.1016/j.isci.2023.106186

Claar, L. D., Rembado, I., Kuyat, J. R., Russo, S., Marks, L. C., Olsen, S. R., & Koch, C. (2023). *Cortico-thalamo-cortical interactions modulate electrically evoked EEG responses in mice* [Preprint]. elife. https://doi.org/10.7554/eLife.84630.1

Comolatti, R., Pigorini, A., Casarotto, S., Fecchio, M., Faria, G., Sarasso, S., Rosanova, M., Gosseries, O., Boly, M., Bodart, O., Ledoux, D., Brichant, J.-F., Nobili, L., Laureys, S., Tononi, G., Massimini, M., & Casali, A. G. (2019). A fast and general method to empirically estimate the complexity of brain responses to transcranial and intracranial stimulations. *Brain Stimulation*, *12*(5), 1280–1289. https://doi.org/10.1016/j.brs.2019.05.013

Compte, A., Sanchez-Vives, M. V., McCormick, D. A., & Wang, X.-J. (2003). Cellular and Network Mechanisms of Slow Oscillatory Activity (<1 Hz) and Wave Propagations in a Cortical Network Model. *Journal of Neurophysiology*, *89*(5), 2707–2725. https://doi.org/10.1152/jn.00845.2002

Cossu, M., Cardinale, F., Castana, L., Citterio, A., Francione, S., Tassi, L., Benabid, A. L., & Lo Russo, G. (2005). Stereoelectroencephalography in the presurgical evaluation of focal epilepsy: A retrospective analysis of 215 procedures. *Neurosurgery*, *57*(4), 706–718; discussion 706-718.

Dale, J., Schmidt, S. L., Mitchell, K., Turner, D. A., & Grill, W. M. (2022). Evoked potentials generated by deep brain stimulation for Parkinson's disease. *Brain Stimulation*, *15*(5), 1040–1047. https://doi.org/10.1016/j.brs.2022.07.048

D'Andola, M., Rebollo, B., Casali, A. G., Weinert, J. F., Pigorini, A., Villa, R., Massimini, M., & Sanchez-Vives, M. V. (2018). Bistability, Causality, and Complexity in Cortical Networks: An In Vitro Perturbational Study. *Cerebral Cortex*, *28*(7), 2233–2242. https://doi.org/10.1093/cercor/bhx122

Entz, L., Tóth, E., Keller, C. J., Bickel, S., Groppe, D. M., Fabó, D., Kozák, L. R., Erőss, L., Ulbert, I., & Mehta, A. D. (2014). Evoked effective connectivity of the human neocortex. *Human Brain Mapping*, *35*(12), 5736–5753. https://doi.org/10.1002/hbm.22581

Fecchio, M., Pigorini, A., Comanducci, A., Sarasso, S., Casarotto, S., Premoli, I., Derchi, C.-C., Mazza, A., Russo, S., Resta, F., Ferrarelli, F., Mariotti, M., Ziemann, U., Massimini, M., & Rosanova, M. (2017). The spectral features of EEG responses to transcranial magnetic stimulation of the primary motor cortex depend on the amplitude of the motor evoked potentials. *PLOS ONE*, *12*(9), e0184910. https://doi.org/10.1371/journal.pone.0184910

Ferrarelli, F., Massimini, M., Sarasso, S., Casali, A., Riedner, B. A., Angelini, G., Tononi, G., & Pearce, R. A. (2010). Breakdown in cortical effective connectivity during midazolam-induced loss of consciousness. *Proceedings of the National Academy of Sciences*, *107*(6), 2681–2686. https://doi.org/10.1073/pnas.0913008107

Fonov, V., Evans, A., McKinstry, R., Almli, C., & Collins, D. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, *47*, S102. https://doi.org/10.1016/S1053-8119(09)70884-5

Grandchamp, R., & Delorme, A. (2011). Single-Trial Normalization for Event-Related Spectral Decomposition Reduces Sensitivity to Noisy Trials. *Frontiers in Psychology*, *2*. https://doi.org/10.3389/fpsyg.2011.00236

Hajnal, B., Szabó, J. P., Tóth, E., Keller, C. J., Wittner, L., Mehta, A. D., Erőss, L., Ulbert, I., Fabó, D., & Entz, L. (2024). Intracortical mechanisms of single pulse electrical stimulation (SPES) evoked excitations and inhibitions in humans. *Scientific Reports*, *14*(1), 13784. https://doi.org/10.1038/s41598-024-62433-0

Hao, Y., Riehle, A., & Brochier, T. G. (2016). Mapping Horizontal Spread of Activity in Monkey Motor Cortex Using Single Pulse Microstimulation. *Frontiers in Neural Circuits*, *10*. https://doi.org/10.3389/fncir.2016.00104

Hays, M. A., Kamali, G., Koubeissi, M. Z., Sarma, S. V., Crone, N. E., Smith, R. J., & Kang, J. Y. (2023). Towards optimizing single pulse electrical stimulation: High current intensity, short pulse width stimulation most effectively elicits evoked potentials. *Brain Stimulation*, *16*(3), 772–782. https://doi.org/10.1016/j.brs.2023.04.023

Keller, C. J., Honey, C. J., Mégevand, P., Entz, L., Ulbert, I., & Mehta, A. D. (2014). Mapping human brain networks with cortico-cortical evoked potentials. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1653), 20130528. https://doi.org/10.1098/rstb.2013.0528

Keller, C. J., Huang, Y., Herrero, J. L., Fini, M. E., Du, V., Lado, F. A., Honey, C. J., & Mehta, A. D. (2018). Induction and Quantification of Excitability Changes in Human Cortical Networks. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *38*(23), 5384–5398. https://doi.org/10.1523/JNEUROSCI.1088-17.2018

Lemaréchal, J.-D., Jedynak, M., Trebaul, L., Boyer, A., Tadel, F., Bhattacharjee, M., Deman, P., Tuyisenge, V., Ayoubian, L., Hugues, E., Chanteloup-Forêt, B., Saubat, C., Zouglech, R., Reyes Mejia, G. C., Tourbier, S., Hagmann, P., Adam, C., Barba, C., Bartolomei, F., … F-TRACT consortium. (2022). A brain atlas of axonal and synaptic delays based on modelling of cortico-cortical evoked potentials. *Brain*, *145*(5), 1653–1667. https://doi.org/10.1093/brain/awab362

Massimini, M. (2005). Breakdown of Cortical Effective Connectivity During Sleep. *Science*, *309*(5744), 2228–2232. https://doi.org/10.1126/science.1117256

Matsumoto, R., Nair, D. R., LaPresto, E., Bingaman, W., Shibasaki, H., & Lüders, H. O. (2007). Functional connectivity in human cortical motor system: A cortico-cortical evoked potential study. *Brain*, *130*(1), 181–197.

Matsumoto, R., Nair, D. R., LaPresto, E., Najm, I., Bingaman, W., Shibasaki, H., & Lüders, H. O. (2004a). Functional connectivity in the human language system: A cortico-cortical evoked potential study. *Brain*, *127*(10), 2316–2330. https://doi.org/10.1093/brain/awh246

Matsumoto, R., Nair, D. R., LaPresto, E., Najm, I., Bingaman, W., Shibasaki, H., & Lüders, H. O. (2004b). Functional connectivity in the human language system: A cortico-cortical evoked potential study. *Brain*, *127*(10), 2316–2330.

McIntyre, C. C., Mori, S., Sherman, D. L., Thakor, N. V., & Vitek, J. L. (2004). Electric field and stimulating influence generated by deep brain stimulation of the subthalamic nucleus. *Clinical Neurophysiology*, *115*(3), 589–595. https://doi.org/10.1016/j.clinph.2003.10.033

Mofakham, S., Fry, A., Adachi, J., Stefancin, P. L., Duong, T. Q., Saadon, J. R., Winans, N. J., Sharma, H., Feng, G., Djuric, P. M., & Mikell, C. B. (2021). Electrocorticography reveals thalamic control of cortical dynamics following traumatic brain injury. *Communications Biology*, *4*(1), 1–10. https://doi.org/10.1038/s42003-021-02738-2

Momi, D., Ozdemir, R. A., Tadayon, E., Boucher, P., Shafi, M. M., Pascual-Leone, A., & Santarnecchi, E. (2021). Network-level macroscale structural connectivity predicts propagation of transcranial magnetic stimulation. *NeuroImage*, *229*, 117698. https://doi.org/10.1016/j.neuroimage.2020.117698

Morishima, Y., Akaishi, R., Yamada, Y., Okuda, J., Toma, K., & Sakai, K. (2009). Task-specific signal transmission from prefrontal cortex in visual selective attention. *Nature Neuroscience*, *12*(1), 85–91. https://doi.org/10.1038/nn.2237

Mukovski, M., Chauvette, S., Timofeev, I., & Volgushev, M. (2007). Detection of active and silent states in neocortical neurons from the field potential signal during slow-wave sleep. *Cerebral Cortex (New York, N.Y.: 1991)*, *17*(2), 400–414. https://doi.org/10.1093/cercor/bhj157

Neudorfer, C., Butenko, K., Oxenford, S., Rajamani, N., Achtzehn, J., Goede, L., Hollunder, B., Ríos, A. S., Hart, L., Tasserie, J., Fernando, K. B., Nguyen, T. A. K., Al-Fatly, B., Vissani, M., Fox, M., Richardson, R. M., van Rienen, U., Kühn, A. A., Husch, A. D., … Horn, A. (2023). Lead-DBS v3.0: Mapping deep brain stimulation effects to local anatomy and global networks. *NeuroImage*, *268*, 119862. https://doi.org/10.1016/j.neuroimage.2023.119862

Opitz, A., Windhoff, M., Heidemann, R. M., Turner, R., & Thielscher, A. (2011). How the brain tissue shapes the electric field induced by transcranial magnetic stimulation. *NeuroImage*, *58*(3), 849–859.

https://doi.org/10.1016/j.neuroimage.2011.06.069

Ozdemir, R. A., Tadayon, E., Boucher, P., Momi, D., Karakhanyan, K. A., Fox, M. D., Halko, M. A., Pascual-Leone, A., Shafi, M. M., & Santarnecchi, E. (2020). Individualized perturbation of the human connectome reveals reproducible biomarkers of network dynamics relevant to cognition. *Proceedings of the National Academy of Sciences*, *117*(14), 8115–8125. https://doi.org/10.1073/pnas.1911240117

Parmigiani, S., Mikulan, E., Russo, S., Sarasso, S., Zauli, F. M., Rubino, A., Cattani, A., Fecchio, M., Giampiccolo, D., Lanzone, J., D'Orio, P., Vecchio, M. D., Avanzini, P., Nobili, L., Sartori, I., Massimini, M., & Pigorini, A. (2022). Simultaneous stereo-EEG and high-density scalp EEG recordings to study the effects of intracerebral stimulation parameters. *Brain Stimulation: Basic, Translational, and Clinical Research in Neuromodulation*, *15*(3), 664–675. https://doi.org/10.1016/j.brs.2022.04.007

Paulk, A. C., Zelmann, R., Crocker, B., Widge, A. S., Dougherty, D. D., Eskandar, E. N., Weisholtz, D. S., Richardson, R. M., Cosgrove, G. R., Williams, Z. M., & Cash, S. S. (2022). Local and distant cortical responses to single pulse intracranial stimulation in the human brain are differentially modulated by specific stimulation parameters. *Brain Stimulation*, *15*(2), 491–508. https://doi.org/10.1016/j.brs.2022.02.017

Paus, T., Sipila, P. K., & Strafella, A. P. (2001). Synchronization of Neuronal Activity in the Human Primary Motor Cortex by Transcranial Magnetic Stimulation: An EEG Study. *Journal of Neurophysiology*, *86*(4), 1983–1990. https://doi.org/10.1152/jn.2001.86.4.1983

Pigorini, A., Sarasso, S., Proserpio, P., Szymanski, C., Arnulfo, G., Casarotto, S., Fecchio, M., Rosanova, M., Mariotti, M., Lo Russo, G., Palva, J. M., Nobili, L., & Massimini, M. (2015). Bistability breaks-off deterministic responses to intracortical stimulation during non-REM sleep. *NeuroImage*, *112*, 105–113. https://doi.org/10.1016/j.neuroimage.2015.02.056

Rosanova, M., Casali, A., Bellina, V., Resta, F., Mariotti, M., & Massimini, M. (2009). Natural Frequencies of Human Corticothalamic Circuits. *Journal of Neuroscience*, *29*(24), 7679–7685. https://doi.org/10.1523/JNEUROSCI.0445-09.2009

Rosanova, M., Fecchio, M., Casarotto, S., Sarasso, S., Casali, A. G., Pigorini, A., Comanducci, A., Seregni, F., Devalle, G., Citerio, G., Bodart, O., Boly, M., Gosseries, O., Laureys, S., & Massimini, M. (2018). Sleep-like cortical OFF-periods disrupt causality and complexity in the brain of unresponsive wakefulness syndrome patients. *Nature Communications*, *9*(1), 4427. https://doi.org/10.1038/s41467-018-06871-1

Rosanova, M., Gosseries, O., Casarotto, S., Boly, M., Casali, A. G., Bruno, M.-A., Mariotti, M., Boveroux, P., Tononi, G., Laureys, S., & Massimini, M. (2012). Recovery of cortical effective connectivity and recovery of consciousness in vegetative patients. *Brain*, *135*(4), 1308–1320. https://doi.org/10.1093/brain/awr340

Russo, S., Claar, L., Marks, L., Krishnan, G., Furregoni, G., Zauli, F. M., Hassan, G., Solbiati, M., d'Orio, P., Mikulan, E., Sarasso, S., Rosanova, M., Sartori, I., Bazhenov, M., Pigorini, A., Massimini, M., Koch, C., & Rembado, I. (2024). *Thalamic feedback shapes brain responses evoked by cortical stimulation in mice and humans* (p. 2024.01.31.578243). bioRxiv. https://doi.org/10.1101/2024.01.31.578243

Sarasso, S., Boly, M., Napolitani, M., Gosseries, O., Charland-Verville, V., Casarotto, S., Rosanova, M.,

Casali, A. G., Brichant, J.-F., Boveroux, P., Rex, S., Tononi, G., Laureys, S., & Massimini, M. (2015). Consciousness and Complexity during Unresponsiveness Induced by Propofol, Xenon, and Ketamine. *Current Biology*, *25*(23), 3099–3105. https://doi.org/10.1016/j.cub.2015.10.014

Sarasso, S., D'Ambrosio, S., Fecchio, M., Casarotto, S., Viganò, A., Landi, C., Mattavelli, G., Gosseries, O., Quarenghi, M., Laureys, S., Devalle, G., Rosanova, M., & Massimini, M. (2020). Local sleep-like cortical reactivity in the awake brain after focal injury. *Brain*, *143*(12), 3672–3684. https://doi.org/10.1093/brain/awaa338

Saturnino, G. B., Puonti, O., Nielsen, J. D., Antonenko, D., Madsen, K. H., & Thielscher, A. (2019). SimNIBS 2.1: A Comprehensive Pipeline for Individualized Electric Field Modelling for Transcranial Brain Stimulation. In S. Makarov, M. Horner, & G. Noetscher (Eds.), *Brain and Human Body Modeling: Computational Human Modeling at EMBC 2018* (pp. 3–25). Springer International Publishing. https://doi.org/10.1007/978-3-030-21293-3_1

Siebner, H. R., Funke, K., Aberra, A. S., Antal, A., Bestmann, S., Chen, R., Classen, J., Davare, M., Di Lazzaro, V., Fox, P. T., Hallett, M., Karabanov, A. N., Kesselheim, J., Beck, M. M., Koch, G., Liebetanz, D., Meunier, S., Miniussi, C., Paulus, W., ... Ugawa, Y. (2022). Transcranial magnetic stimulation of the brain: What is stimulated? – A consensus and critical position paper. *Clinical Neurophysiology*, *140*, 59–97. https://doi.org/10.1016/j.clinph.2022.04.022

Timofeev, I., Grenier, F., & Steriade, M. (2001). Disfacilitation and active inhibition in the neocortex during the natural sleep-wake cycle: An intracellular study. *Proceedings of the National Academy of Sciences*, *98*(4), 1924–1929. https://doi.org/10.1073/pnas.98.4.1924

Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: From consciousness to its physical substrate. *Nature Reviews Neuroscience*, *17*(7), 450–461. https://doi.org/10.1038/nrn.2016.44

Tononi, G., & Massimini, M. (2008). Why Does Consciousness Fade in Early Sleep? *Annals of the New York Academy of Sciences*, *1129*(1), 330–334. https://doi.org/10.1196/annals.1417.024

Trebaul, L., Deman, P., Tuyisenge, V., Jedynak, M., Hugues, E., Rudrauf, D., Bhattacharjee, M., Tadel, F., Chanteloup-Foret, B., Saubat, C., Reyes Mejia, G. C., Adam, C., Nica, A., Pail, M., Dubeau, F., Rheims, S., Trébuchon, A., Wang, H., Liu, S., ... David, O. (2018). Probabilistic functional tractography of the human cortex revisited. *NeuroImage*, *181*, 414–429. https://doi.org/10.1016/j.neuroimage.2018.07.039

Tscherpel, C., Dern, S., Hensel, L., Ziemann, U., Fink, G. R., & Grefkes, C. (2020). Brain responsivity provides an individual readout for motor recovery after stroke. *Brain: A Journal of Neurology*, *143*(6), 1873–1888. https://doi.org/10.1093/brain/awaa127

Usami, K., Korzeniewska, A., Matsumoto, R., Kobayashi, K., Hitomi, T., Matsuhashi, M., Kunieda, T., Mikuni, N., Kikuchi, T., Yoshida, K., Miyamoto, S., Takahashi, R., Ikeda, A., & Crone, N. E. (2019). The neural tides of sleep and consciousness revealed by single-pulse electrical brain stimulation. *Sleep*, *42*(6), zsz050. https://doi.org/10.1093/sleep/zsz050

Valentín, A., Anderson, M., Alarcón, G., Seoane, J. J. G., Selway, R., Binnie, C. D., & Polkey, C. E. (2002). Responses to single pulse electrical stimulation identify epileptogenesis in the human brain in vivo.

*Brain*, *125*(8), 1709–1718. https://doi.org/10.1093/brain/awf187

Valentin, A., Arunachalam, R., Mesquita-Rodrigues, A., Garcia Seoane, J. J., Richardson, M. P., Mills, K. R., & Alarcon, G. (2008). Late EEG responses triggered by transcranial magnetic stimulation (TMS) in the evaluation of focal epilepsy. *Epilepsia*, *49*(3), 470–480. https://doi.org/10.1111/j.1528-1167.2007.01418.x

Vorwerk, J., Cho, J.-H., Rampp, S., Hamer, H., Knösche, T. R., & Wolters, C. H. (2014). A guideline for head volume conductor modeling in EEG and MEG. *NeuroImage*, *100*, 590–607. https://doi.org/10.1016/j.neuroimage.2014.06.040

Zelmann, R., Paulk, A. C., Tian, F., Balanza Villegas, G. A., Dezha Peralta, J., Crocker, B., Cosgrove, G. R., Richardson, R. M., Williams, Z. M., Dougherty, D. D., Purdon, P. L., & Cash, S. S. (2023). Differential cortical network engagement during states of un/consciousness in humans. *Neuron*, *111*(21), 3479-3495.e6. https://doi.org/10.1016/j.neuron.2023.08.007

# *Chapter 4*

# Causal emergence is widespread across measures of causation[7]

[7] This chapter corresponds to the article **Comolatti, Renzo**; Hoel, Erik. *Causal emergence is widespread across measures of causation*. arXiv (2022) https://arxiv.org/abs/2202.01854

# CAUSAL EMERGENCE IS WIDESPREAD ACROSS MEASURES OF CAUSATION

**Renzo Comolatti**
University of Milan
Milan, MI, Italy
renzo.com@gmail.com

**Erik Hoel***
Allen Discovery Center
Tufts University
Medford, MA, USA
erik.hoel@tufts.edu

February 7, 2022

**ABSTRACT**

Causal emergence is the theory that macroscales can reduce the noise in causal relationships, leading to stronger causes at the macroscale. First identified using the effective information and later the integrated information in model systems, causal emergence has been analyzed in real data across the sciences since. But is it simply a quirk of these original measures? To answer this question we examined over a dozen popular measures of causation, all independently developed and widely used, and spanning different fields from philosophy to statistics to psychology to genetics. All showed cases of causal emergence. This is because, we prove, measures of causation are based on a small set of related "causal primitives." This consilience of independently-developed measures of causation shows that macroscale causation is a general fact about causal relationships, is scientifically detectable, and is not a quirk of any particular measure of causation. This finding sets the science of emergence on firmer ground, opening the door for the detection of intrinsic scales of function in complex systems, as well as assisting with scientific modeling and experimental interventions.

## 1 Introduction

While causation has historically been a subject of philosophical debate, work over the last few decades has shown that metaphysical speculations can be put aside in favor of mathematical formalisms [1]. Indeed, causation is referenced universally throughout the sciences without metaphysical commitments, and mathematical treatments of causation come from diverse scientific fields like psychology and statistics [2]. E.g., in the neurosciences, people have used a number of measures of causation to track the result of experimental interventions [3, 4, 5, 6]. However, due to this plethora of measures of causation, one might argue there is subjectivity in terms of what counts as a cause or not, since a particular scientist might prefer one measure over another.

Here we offer a way around this problem by showing that popular measures of causation are mathematically related, behave very similarly under many conditions, and are sensitive to the same fundamental properties. Indeed, all the measures we examined turned out to be based on a small set of what we dub *causal primitives*. By showing how over a dozen measures of causation are grounded in the same primitives, we reveal there is widespread consilience in terms of what constitutes a strong or weak cause (or more generally, a strong or weak causal relationship). This research obviates the need to arrive at a lone measure of causation that researchers must universally agree upon, but rather reveals a sphere of viable measures with significant overlap (much like the definitions of "complexity" in complex systems science [7]). By focusing on the agreement between a family of well-accepted and closely-related measures, we can move on to understanding other causal phenomena.

One such important phenomena is causal emergence, which is when a causal relationship is stronger at the macroscale [8]. While at first counterintuitive, causal emergence is grounded in the fact that macroscales can lead to noise reduction in causal relationships. Broadly, this noise is synonymous with uncertainty, which can come from different sources,

*Corresponding author

and macroscale models can reduce or minimize this error. In such cases, universal reduction is unworkable, since such reduction would "leave some causation on the table," even though the macroscale supervenes (is fixed by) its underlying microscale. Note that claims of emergence are not metaphysical speculations. They have real consequences. For example, emergent macroscale models are more useful to intervene on and understand the system in question with [9]; causal emergence can reveal the intrinsic scales of function in opaque non-engineered systems where the scale of interest is unknown, like in gene regulatory networks [10]; it can also be used to find partitions of directed graphs and is more common in biological networks vs. technological networks [11]; it has revealed novel groupings of cellular automata rules [12]; causal emergence has been used to identify macrostates in timeseries data using artificial neural networks [13]; there's even some evidence that evolution selects for causal emergence, possibly because macroscales that are causally-emergent have been shown to be more robust to knock-outs and attacks [14]. Such questions are relevant across the sciences, e.g., there are fundamental questions about what scale is of most importance in brain function [15, 16, 17] that only a scientific theory of emergence can resolve; indeed, causal emergence might explain the spatiotemporal scale of consciousness in the brain [18, 19].

However, evidence for causal emergence has previously been confined to a small set of measures: first, the effective information [8, 20, 11], and then later, the integrated information [18, 21]. Both these measures, grounded in information theory, are designed to capture subtly different aspects of causation. Yet they are related mathematically and involve similar background assumptions. Because of this, some have criticized the results of the measures, pointing to how interventions are performed (e.g., perhaps effective information requiring a maximum-entropy intervention distribution means it's somehow invalid or assumptive [22]), as well as the meaning of effective information in general (e.g., perhaps it is somehow merely capturing "explanatory" causation rather than real causation [23]). Meanwhile, the integrated information has been criticized for being one of many possible measures [24, 25], and unsubstantiated from its axioms [26]. While there are counterarguments to these specific criticisms of info-theoretic accounts of causation, it is a reasonable question whether causal emergence is a general phenomenon or some highly peculiar quirk of these measures and background assumptions, as this would limit its relevancy significantly.

There are already some reasons to think causal emergence is indeed a broader phenomenon. For example, recent evidence has indicated that the synergistic and unique information component of the mutual information can be greater at macroscales (while the redundant information component is lower) [27], and there have been other causal emergence-based approaches to the partial information decomposition as well [28, 29].

Here we provide evidence for widespread generality of causal emergence as a phenomenon. We show that across a dozen popular historical measures of causation from different fields, causal emergence universally holds true under many different conditions and assumptions as to how the measures are applied. That is, instances of emergent macroscale causation can be detected by the majority of independent measures of causation—at least, all of those that we considered. The widespread nature of causal emergence is because most measures of causation are based on a small set of primitives: specifically, sufficiency and necessity, along with their generalizations (which we provide here) of determinism and degeneracy, respectively. All these causal primitives can be improved at a macroscale. Therefore, all the measures also demonstrate causal emergence (indeed, we find that effective information is the most conservative measure of those we analyzed). This is all despite the fact that macroscales are simply dimension-reductions of microscales. So while two scales may both be valid descriptions of a system, one may possess stronger causation (the interpretation of which, whether as more causal work, information, or explanation, depends on the measure of causation itself). Yet causal emergence is not trivially universal either. It is system-dependent: in many cases, specifically those without any uncertainty in microscale system dynamics, causal reduction dominates.

First, in Section 2, we define causal primitives along with the formal language of cause and effect we will use throughout. In Section 3, we overview twelve independently-proposed measures of causation (several of which end up being identical, as we show). In Section 4, we highlight how the behavior of the measures is based on causal primitives using a simple bipartite Markov chain model. In Section 5, we directly compare macroscales to microscales across all the causal measures using the bipartite model, and find widespread evidence for causal emergence across all the measures.

In the Discussion, we overview how the consilience of causation we've revealed can provide a template for an objective understanding of causation, and discuss the beginnings of the scientific subfield of emergence.

## 2 Formalizing causation and causal primitives

First, a note on terminology. We must use a general enough one that it can incorporate a number of different notions of causation from different fields. Therefore, we focus on a given a space $\Omega$, i.e., the set of all possible occurrences. In this space, we can consider causes $c \in \Omega$ and effects $e \in \Omega$, where we assume causes $c$ to precede effects $e$, so that we also speak of a set of causes $C \subseteq \Omega$ and of effects $E \subseteq \Omega$.

As we will later be applying these measures in Markov chains, we can consider the space $\Omega$ to be a state-space and $c$ or $e$ as states. The set of causes and effects can be related probabilistically via transition probabilities $P(e \mid c)$, which specifies the probability of obtaining a candidate effect $e$, given that a candidate cause $c$ actually occurred.

As we will see, in order to gauge causation, we will have to evaluate counterfactuals of $c$, and consider the probability of obtaining the effect $e$ given that $c$ didn't occur. We will write this probability $P(e \mid C\backslash c)$, where $C\backslash c$ stand for the complement of $c$, by which we mean the probability of $e$ given that any cause in $C$ could have produced $e$ except for $c$. Note that although conventionally written $P(e)$ we will write $P(e \mid C)$ to underscore the following notion: namely, that to meaningfully talk about $P(e \mid C)$ (and $P(e \mid C\backslash c)$), a further distribution over $C$ must be specified. That is:

$$P(e \mid C) = \sum_{c \in C} P(c)P(e \mid c)$$

where there is some assumption of a distribution $P(C)$. This assumption is necessary because, unlike terms like $P(e \mid c)$ which are stated in some transition probability matrix (TPM) or system description, terms like $P(c)$ or $P(e \mid C\backslash c)$ need to be explicitly defined (e.g., what is the distribution of the effects when $c$ *didn't* occur?). In Section 4.2 we overview how $P(C)$ itself is defined via an intervention distribution, which is necessary to specify for the application of measures of causation, although not their definitions. Therefore, we simply assume a particular $P(C)$ is defined for the following measures of causation. Note that in examining counterfactual probabilities like $P(e \mid C\backslash c)$ it implies that $P(C)$ is restricted to exclude $c$ and normalized.

### 2.1 Formalizing sufficiency and necessity

Causation should be viewed not as an irreducible single relation between a cause and an effect but rather as having two dimensions: that of sufficiency and necessity [1, 30].

For any cause $c$, we can always ask, on one hand, how sufficient $c$ is for the production of an effect $e$. A sufficient relation means that whenever $c$ occurs, $e$ also follows (Figure 1A, red region). Separably, we can also ask how necessary $c$ is to bring about $e$, that is, whether there are different ways then through $c$ to produce $e$ (Figure 1A, blue region). Yet these properties are orthogonal: a cause $c$ may be sufficient to produce $e$, and yet there may be other ways to produce $e$. Similarly, $c$ may only sometimes produce $e$, but is the only way to do so.

We refer to sufficiency and necessity as *causal primitives*. This is because, as we will show, popular measures of causation generally put these two causal primitives in some sort of relationship (like a difference or a ratio). This ensures such measures are mathematically quite similar, indeed, sometimes unknowingly identical.

First we must define the primitives formally. To start, we associate the sufficiency of the cause $c$ to the probability:

$$suff(e, c) = P(e \mid c)$$

which is 1 when $c$ is fully sufficient to produce $e$. This allows for degrees of sufficiency (e.g., a cause might bring about its effect only some of the time), which is important because many measures of causation rely on probability raising or difference making.

Comparably, the necessity of the cause for the effect we associate with the probability:

$$nec(e, c) = 1 - P(e \mid C\backslash c)$$

which gives "the probability of not-e given the probability of not-c." Necessity is 1 when $c$ is absolutely necessary for $e$. In such cases there is no other candidate cause but $c$ that could produce $e$. Note that some definition of counterfactuals needs to be made explicit for the calculation of necessity, unlike sufficiency (more on this in later sections, where possible counterfactuals are represented as performable interventions).

### 2.2 Determinism and degeneracy as generalizations of sufficiency and necessity

The two causal primitives of sufficiency and necessity each have a generalization. These are the determinism and degeneracy coeffients [8]. Specifically, the determinism coefficient is a generalized notion of sufficiency, while the degeneracy coefficient is a generalized notion of necessity. These generalizations will prove useful in two ways: a) they provide a more general version of the original primitive, and b) some measures of causation are based off of determinism and degeneracy instead of sufficiency and necessity.

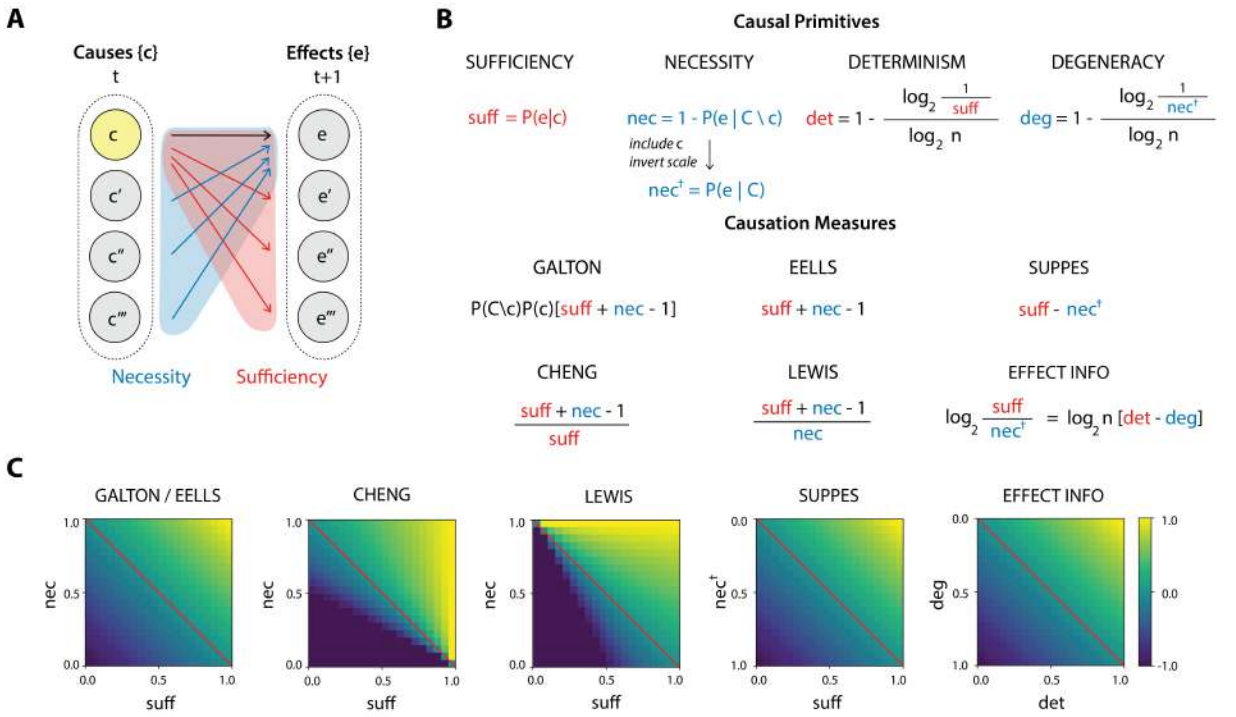



Figure 1: **Causal primitives and causation measures.** (A) Schematic representation of causation as a relation between occurrences (or events) connecting a set of causes to a set of effects. Each individual candidate cause $(c, c', ...)$ and candidate effect $(e, e', ...)$ is depicted in a circle, while their sets $C$ and $E$ are marked as the enclosing dotted line. Causes and effects are assumed to be temporally ordered, with the former preceding the latter, hence are indexed at a time $t$ and a time $t + 1$, respectively. Given a pair of a candidate cause $c$ and a candidate effect $e$, the relation between $c$ and $e$ can be analysed in terms of the causal primitives of sufficiency and necessity. On one hand, one can assess whether $c$ is *sufficient* to bring about $e$, or whether $c$ can instead transition to other effects in $E$ (region shaded red); on the other hand, one can ask whether $c$ is *necessary* for $e$ to obtain, or instead whether other causes in $C$ could also produce $e$ (region shaded blue). (B) The functional dependence of the causation measures on the causal primitives is highlighted (sufficiency and determinism in red, necessity and degeneracy in blue). On the top are the formulas of the causal primitives and on the bottom, the formulas of some of the causation measures written in terms of the causal primitives (measures like the Bit-flip and Lewis' closest possible world are not shown because they rely on additional structure, e.g. distances between occurrences). (C) Behavior of the causation measures for different combinations of the causal primitives. Heatmaps show causation as a function of the causal primitives (using $n = 2$). For Suppes and effect information the y axis is inverted to highlight the similarity with the other measures.

We can define the determinism as the opposite of noise (or randomness), that is, the certainty of causal relationships. Specifically, it is based on the entropy of the probability distribution of the effects of a cause:

$$H(e \mid c) = \sum_{e \in E} P(e \mid c) \log_2 \frac{1}{P(e \mid c)}$$

This entropy term is zero if a cause has a single effect with $P = 1$, and the entropy is maximal, i.e. $\log_2 n$, if a cause has a totally random effect. We therefore define the determinism of a cause $c$ to be $\log_2(n) - H(e \mid c)$. Note that this is different than the mere sufficiency, although is also based on the sufficiency $P(e \mid c)$. To see their difference, let us consider a system of four states $\Omega = \{a, b, c, d\}$, wherein state $a$ transitions to the other states $b$, $c$, or $d$, and also back to itself, $a$, with probability $1/4$ each. The average sufficiency of $a$'s transitions would be $1/4$. However, the determinism of $a$ would be zero, since there is no difference between $a$ and randomly generating the next state of the system.

Unlike sufficiency, the determinism is a property of a cause, not a particular transition (although the contribution of each transition to the determinism term can be calculated). And unlike sufficiency, the determinism term is influenced

by the number of considered possibilities. Generally, we normalize the term to create a determinism coefficient that ranges, like the sufficiency, between 0 (fully random) and 1 (fully deterministic), for a given cause:

$$det(c) = 1 - \frac{H(e \mid c)}{\log_2 n}$$

And with this in hand, we can define a determinism coefficient for individual transitions as:

$$det(e, c) = 1 - \frac{\log_2 \frac{1}{P(e|c)}}{\log_2 n}$$

as well as a system-level determinism coefficient:

$$det = \sum_{c \in C} P(c)\ det(c) = \sum_{e \in E,\ c \in C} P(e, c)\ det(e, c) = 1 - \frac{\sum_{c \in C} P(c)\ H(e \mid c)}{\log_2 n}$$

Degeneracy is the generalization of necessity. It is zero when no effect has a greater probability than any other (assuming an equal probability across the full set of causes). Degeneracy is high if certain effects are "favored" in that more causes lead to them (and therefore those causes are less necessary). It is also based on an entropy term:

$$H(e \mid C) = \sum_{e \in E} P(e \mid C) \log_2 \frac{1}{P(e \mid C)}$$

and the degeneracy coefficient of an individual effect is given by:

$$deg(e) = 1 - \frac{\log_2 \frac{1}{P(e|C)}}{\log_2 n}$$

while the system-level degeneracy coefficient is:

$$deg = \sum_{e \in E} P(e \mid c)\ deg(e) = 1 - \frac{H(e \mid C)}{\log_2 n}$$

## 3 Measures of causation are based on causal primitives

In the following section, we demonstrate how the basic causal primitives of sufficiency and necessity or their generalized forms of determinism and necessity underlie the independent popular measures of causation we examined.

### 3.1 Humean constant conjunction

One of the earliest and most influential approaches to a modern view of causation was David Hume's regularity account. Hume famously defined a cause as "an object, followed by another, and where all the objects, similar to the first, are followed by objects similar to the second" [31]. In other words, causation stems from patterns of succession between events [32].

Overall, the "constant conjunction" of an event $c$ followed by an event $e$, would lead us to *expect* $e$ once observing $c$, and therefore infer $c$ to be the cause of $e$. There are a number of modern formalisms of this idea. Here we follow Judea Pearl, who interprets Hume's notion of "regularity of succession" as amounting to what we today call correlation between events [1]. This can can be formalized as the observed statistical covariance between a candidate cause $c$ and effect $e$:

$$Cov(X, Y) = E(XY) - E(X)E(Y)$$

If we substitute the indicator function $X_c$ (and $Y_e$), which is 1 if $c$ (respectively $e$) occurs and 0 otherwise, in the equation above we obtain:

$$\begin{aligned} Cov(X_c, Y_e) &= P(c,e) - P(c)P(e) \\ &= P(c)P(e \mid c) - P(c)[P(c)P(e \mid c) + P(\bar{c})P(e \mid C\backslash c)] \\ &= P(e \mid c)P(c)[1 - P(c)] + P(c)P(C\backslash c)P(e \mid C\backslash c) \\ &= P(e \mid c)P(c)P(C\backslash c)] + P(c)P(C\backslash c)P(e \mid C\backslash c) \\ &= P(c)P(C\backslash c)[P(e \mid c) - P(e \mid C\backslash c)]) \end{aligned}$$

Where we used the fact that $P(e \mid C))$ can be decomposed into two weighted sums, i.e. over $c$ and over $C\backslash c$. Following other's nomenclature [2] we call this the "Galton measure" of causal strength, since it closely resembles the formalism for heredity of traits in biology, and also is a form of the statistical co-variance:

$$CS_{Galton}(e,c) = P(c)P(C\backslash c)[P(e \mid c) - P(e \mid C\backslash c)] = P(c)P(C\backslash c)[suff(e,c) + nec(e,c) - 1]$$

It's worth noting that such a regularity account can be stated in ways that involve causal primitives, as can be seen above.

### 3.2 Eells's measure of causation as probability raising

Ellery Eells proposed that a condition for $c$ to be a cause of $e$ is that the probability of $e$ in the presence of $c$ must be higher than its probability in its absence: $P(e \mid c) > P(e \mid C\backslash c)$ [33]. This can be formalized in a measure of causal strength as the difference between the two quantities:

$$CS_{Eells} = P(e \mid c) - P(e \mid C\backslash c) = suff(e,c) + nec(e,c) - 1$$

When $CS_{Eells} < 0$ the cause is traditionally said to be a negative or preventive cause [32], or in another interpretation, such negative values should not be considered a cause at all [34].

### 3.3 Suppes's measure of causation as probability raising

Another notion of causation as probability raising was defined by Patrick Suppes, a philosopher and scientist [35]. Translated into our formalism, his measure is:

$$CS_{Suppes}(c,e) = P(e \mid c) - P(e \mid C) = suff(e,c) - nec^{\dagger}(e)$$

The difference between the $CS_{Eells}$ and $CS_{Suppes}$ measures involves a shift from measuring how causally *necessary* $c$ is for $e$—whether it can be produced by other causes than $c$—to assessing how *degenerate* is the space of ways to bring $e$ about. Both are valid measures, and in fact turn out to be equivalent in some contexts [36].

Note that we can extend the conditional probability $P(e \mid C\backslash c)$ to $P(e \mid C)$, including $c$ itself. If so, we are considering whether $e$ can be produced not just in the absence of $c$, but all the ways, including via $c$ itself, that $e$ can occur. Therefore, another version can be defined as:

$$CS_{Suppes_{II}}(c,e) = \frac{P(e \mid c)}{P(e \mid C)}$$

### 3.4 Cheng's causal attribution

Patricia Cheng has proposed a popular psychological model of causal attribution, where reasoners go beyond assessing pure covariation between events to estimate the "causal power" of a candidate cause producing (or preventing) an effect [37]. In this account, the causal power of $c$ to produce $e$ is given by:

$$CS_{Cheng}(c,e) = \frac{P(e \mid c) - P(e \mid C\backslash c)}{1 - P(e \mid C\backslash c)} = \frac{suff(e,c) + nec(e,c) - 1}{nec(e,c)}$$

Cheng writes: "The goal of these explanations of $P(e \mid c)$ and $P(e \mid C\backslash c)$ is to yield an estimate of the (generative or preventive) power of $c$...." While originally proposed as a way to estimate causes from data based off of observables,

it's worth noting that, in our application of this measure, we have access to the real probabilities given by the transition probability matrix $P(e \mid c)$, and the measure therefore yields a true assessment of causal strength, not an estimation.

### 3.5 Lewis's counterfactual theory of causation

Another substantive and influential account of causation based on counterfactuals was given by philosopher David Lewis [38]. Lewis defines a cause as if given events $c$ and $e$ took place, $c$ can be said to be a cause of $e$ if it is the case that if $c$ hadn't occurred, then $e$ would not have occurred. Lewis also extended his theory for "chancy worlds", where $e$ can follow from $c$ probabilistically [39].

Following [2] we formalize Lewis's causal strength as the ratio:

$$\frac{P(e \mid c)}{P(e \mid C \backslash c)}$$

This definition is also known as "relative risk:" "it is the risk of experiencing $e$ in the presence of $c$, relative to the risk of $e$ in the absence of $c$ " [2]. This measure can be normalized to obtain a measure ranging from -1 to 1 using the mapping $p/q \rightarrow (p-q)/p$ as:

$$CS_{Lewis}(c,e) = \frac{P(e \mid c) - P(e \mid C \backslash c)}{P(e \mid c)} = \frac{suff(e,c) + nec(e,c) - 1}{suff(e,c)}$$

Again we see that Lewis's basic notion, once properly formalized, is based on the comparison of a small set of causal primitives. Note that this definition doesn't rely on a specification of a particular possible world. In other work, Lewis specifies that the counterfactual not-$c$ is taken to be the closest possible world where $c$ didn't occur. That notion, which specifies a rationale for how to calculate the counterfactual, is formalized in Section 3.7.

### 3.6 Judea Pearl's measures of causation

If our claim for consilience in the study of causation is true, then authors should regularly rediscover previous measures. Indeed, this is precisely what occurs. Consider Judea Pearl, who in his work on causation has defined the previous measures $CS_{Eells}$, $CS_{Lewis}$, and $CS_{Cheng}$ (in some of these terms apparently knowingly, in others not).

Within his structural model semantics framework [1], he defines the "probability of necessity" as the counterfactual probability that $e$ would not have occurred in the absence of $c$, given that $c$ and $e$ did in fact occur, which in his notation is written as PN $= P(\bar{e}_{\bar{c}} \mid c, e)$ (where the bar stands for the complement operator, i.e. $\bar{c} = C \backslash c$). Meanwhile, he defines the "probability of sufficiency" as the capacity of $c$ to produce $e$ and is defined as the probability that $e$ would have occurred in the presence of $c$, given that $c$ and $e$ didn't occur: PS $= P(e_c \mid \bar{c}, \bar{e})$.

Finally, both aspects are combined to measure both the sufficiency and the necessity of $c$ to produce $e$ as PNS $= P(e_c, \bar{e}_{\bar{c}})$, such that the following relation holds: PNS $= P(e,c)$PN $+ P(\bar{e},\bar{c})$PS.

Under conditions of exogeneity of $c$ relative to $e$ (which renders Pearl's counterfactual $P(e_c)$, i.e. the causal effect of $c$ on $e$, computable from $P(e \mid c)$) and monotonicity of $e$ relative to $c$ (which roughly means $c$ does not prevent $e$ from happening), the measures are given by:

$$\text{PNS} = P(e \mid c) - P(e \mid C \backslash c)$$

$$\text{PN} = \frac{P(e \mid c) - P(e \mid C \backslash c)}{P(e \mid c)}$$

$$\text{PS} = \frac{P(e \mid c) - P(e \mid C \backslash c)}{1 - P(e \mid C \backslash c)}$$

where we can recognize that PNS $= CS_{Eells}$, PN $= CS_{Lewis}$ and PS $= CS_{Cheng}$ [2]. That is, Pearl independently recreated previous measures.

Yet we find Pearl's terminology confusing. Therefore, we reserve terms like "probability of sufficiency" to mean the original sufficiency, $P(e \mid c)$, and likewise for the probability of necessity $1 - P(e \mid C \backslash c)$. We also continue to refer to the measures by the names of their respective original authors to further distinguish them and preserve origin credit.

Overall this consilience should increase our confidence that measures based on the combinations of causal primitives are good candidates for assessing causation.

### 3.7 Closest possible world causation

As state previously, David Lewis traditionally gives a counterfactual theory of causation wherein the counterfactual is specified as the closest possible world where $c$ didn't occur [38]. In order to formalize this idea, we need to add further structure beyond solely probability transitions. That is, such a measurement requires a notion of distance between possible states of affairs (or "worlds"). One simple way to do is use binary state labels of states to induce a metric using the Hamming distance [40], which is the number of bit flips needed to change one binary string into the other. In this way we induce a metric in a state-space so that we can define Lewis notion of a closest possible world:

$$D_H(x, y) = \sum_i^N |x_i - y_i|$$

where $x$ and $y$ are two state labels with $N$ binary digits (e.g. $x = 0001$ and $y = 0010$, $N = 4$, such that $D_H(x, y) = 2$). With such a distance notion specified the counterfactual taken as the "closest possible world" where $c$ didn't occur is given by:

$$\bar{c}_{CPW} = \min_{c'} D_H(c, c')$$

And with this in hand, we can define another measure based closely on Lewis's account of causation as reasoned about from a counterfactual of the closest possible world:

$$CS_{Lewis\ CPW} = \frac{P(e \mid c) - P(e \mid \bar{c}_{CPW})}{P(e \mid c)}$$

### 3.8 Bit-flip measures

Another measure that relies on a notion of distance between states is the idea of measuring the amount of difference created by a minimal change in the system. For instance, the outcome of flipping of a bit from some local perturbation. In [41] such a measure is given as "the average Hamming distance between the perturbed and unperturbed state at time $t + 1$ when a random bit is flipped at time $t$". While originally introduced with an assumption of determinism, here we extend their measure to non-deterministic systems as:

$$CS_{bit-flip}(e, c) = \frac{1}{N} \sum_i^N \sum_{e' \in E} P(e' \mid c_{[i]}) D_H(e, e')$$

where $c_{[i]}$ correspond to the state where the $i^{th}$ bit is flipped (e.g., if $c = 000$, then $c_{[3]} = 001$).

### 3.9 Actual causation and the effect information

Recently a framework was put forward [34] for assessing actual causation on dynamical causal networks, using information theory. According to this framework, a candidate cause must raise the probability of its effect compared to its probability when the cause is not specified (again, we see similarities to previous measures). The central quantity is the *effect information*, given by:

$$ei(c, e) = \log_2 \frac{P(e \mid c)}{P(e \mid C)} = \log_2 n[det(e, c) - deg(c)]$$

Note that the effect information is actually just the log of $CS_{Suppes_{II}}$, again indicating consilience as measures of causation are re-discovered by later authors. It is also the individual transition contribution of the previously defined "effectiveness" given in [8].

The effect information is thus on one hand a bit-measure version of the probabilistic Suppes measure, and on the other an non-normalized difference between degeneracy and determinism.

### 3.10 Effective information

The effective information ($EI$) was first introduced by Giulio Tononi and Olaf Sporns as a measure of causal interaction, in which random perturbations of the system are used in order to go beyond statistical dependence [42]. It was rediscovered without reference to prior usage and called "causal specificity" [43].

The effective information is simply the expected value of the effect information over all the possible cause-effect relationships of the system:

$$EI = \sum_{e \in E, c \in C} P(e,c) ei(c,e) = \log_2 n[det - deg]$$

As a measure of causation, $EI$ captures how effectively (deterministically and uniquely) causes produce effects in the system, and how selectively causes can be identified from effects [8].

Effective information is an assessment of the causal power of $c$ to produce $e$ – as measured by the *effect information* – for all transitions between possible causes and possible effects, considering a maximum-entropy intervention distribution on causes (the notion of an intervention distribution is discussed in Section 3.4). More simply, it is the non-normalized difference between the system's determinism and degeneracy. Indeed, we can normalize the effective information by its maximum value, $\log_2 n$, to get the *effectiveness* of the system:

$$eff = det - deg = \frac{EI}{\log_2 n}$$

### 3.11 Summary

Across every measure of causation we examined the two primitives (sufficiency and necessity), or alternatively their generalized forms (determinism and degeneracy), are explicitly put in some relationship, often that of a difference or ratio or trade-off (Figure 1, panels B and C). The only measure that lacked an explicitly obvious basis in causal primitives was the bit-flip measure, but as a measure of sensitivity to perturbation it seems likely there is some basis or relationship (we did not seek out a decomposition).

We are not the first to point out that causation has two dimensions: for instance, Judea Pearl [1] states: "Clearly, some balance must be struck between the necessary and the sufficient components of causal explanation." Also J. L. Mackie, although not proposing a quantitative measure of causal strength, famously considers both a necessity and a sufficiency aspect in his proposal of a INUS condition that causes should satisfy, namely being an (i)nsufficient but (n)ecessary part of a condition which is itself (u)nnecessary but (s)ufficient for an effect to occur [44]. However, to our knowledge this is the first time a full set of popular measures has been assessed in this light, and so we state it explicitly: substantial consilience in measures of causation indicates we should expect measures of causal strength to be based on *both* causal primitives.

## 4 Measures of causation are sensitive to causal primitives

### 4.1 Model system

In order to examine the behavior of measures of causation presented in the previous section, we make use of a simple model. It was chosen because it allows us to parametrically vary the causal primitives of determinism ($det$) and degeneracy ($deg$) in order to see how the measures of causation change under uncertainty. We make use of a simple bipartite Markov chain model where the microstates of the system oscillate back and forth between two groups. What is important to keep in mind is that this model is a) bipartite, and b) that we can vary these bipartite connections to either increase the determinism (increasing the average probability of state transition closer to $p = 1$) or the degeneracy (increasing the overlap of state transitions, such that transitions cluster in their targets). This allows us to apply the measures of causation under different amounts of uncertainty and different types of uncertainty (like indeterminism vs. degeneracy) and later to also examine causal emergence in such regimes as well. A detailed description of the bipartite model, as well as how we vary these parameters can be found in the Supplementary Information Section 7.1. See Figure 2 for a visual representation of the system state-space and transition probability matrix (panel A) and of the different regimes of model architecture we examine (panel C). The code used for calculating the measures of causation as well as assessing causal emergence on the bipartite Markov chain model is available at `https://github.com/renzocom/causal_emergence`.

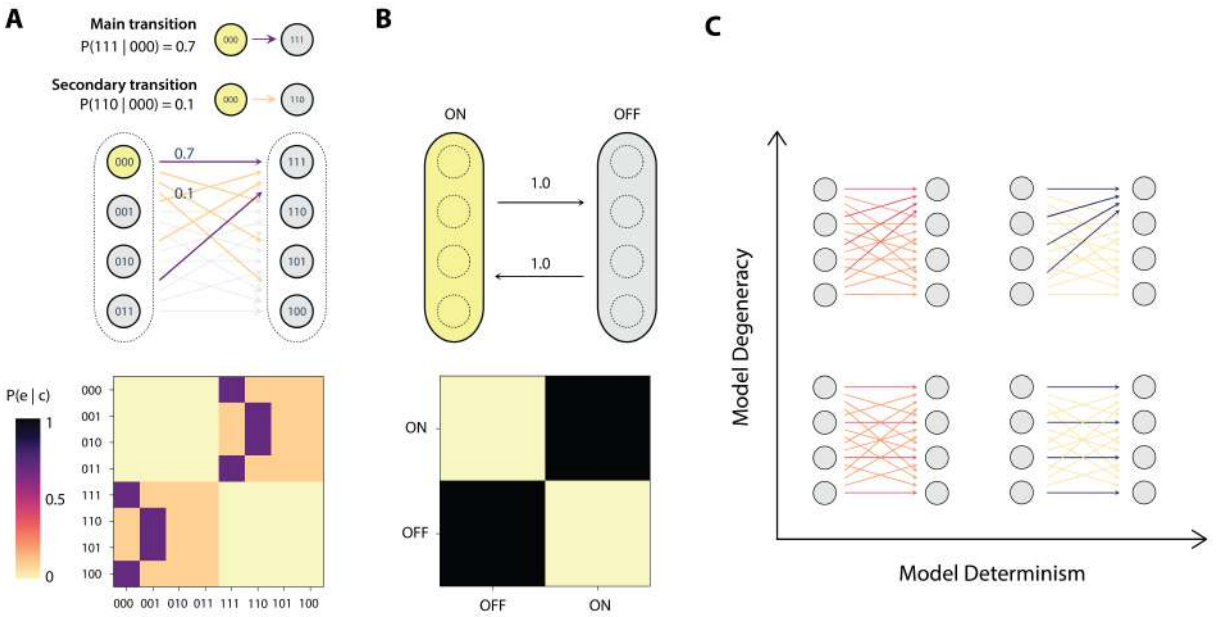



Figure 2: **A simple bipartite Markov chain model for studying causal measures.** (A) Microscale model of the bipartite Markov chain with 8 microstates where microstates transition back-and-forth between two groupings (left and right). On the top, a representation of the state-space with binary labels is shown, with the dotted line indicating the natural macrostate. The possible microstate transitions from the group in the left to the group in the right are represented by the arrows (the transitions from right to left are omitted). A main transition from 000 to 111 is highlighted, and contrasted with a secondary transition from 000 to 110, which intuitively has a lower causal strength. The relevant transitions for evaluating the causal primitives (suffiency and necessity) for the main transition are color coded by the probability according to the probability transition matrix (TPM) shown in the bottom. (B) Macro model of bipartite Markov chain obtained by coarse-graining the original microstates into the two groupings (ON $= \{000, 001, 010, 011\}$ and OFF $= \{111, 110, 101, 100\}$). The obtained macro transition probability matrix (TPM) describes the deterministic transition between the OFF and ON states ($P(\text{ON}_{t+1} \mid \text{OFF}_t) = 1$ and $P(\text{OFF}_{t+1} \mid \text{ON}_t) = 1$) and is independent of the model parameters that control the determinism and degeneracy of the microscale. (C) Graphical representation of the bipartite model state-space (as shown in panel A) for different values of the model's microscale degeneracy and determinism.

### 4.2 Applying measures of causation requires defining an intervention distribution

As we have seen, measures of causation, which can be interpreted as "strength" or "influence" or "informativeness" or "power" or "work (depending on the measure) are based on a combination of causal primitives. However, both the calculations of measures themselves, as well as the causal primitives, involve further background assumptions in order to apply them.

To give a classic example: you go away and ask a friend to water your plant. They don't, and the plant dies. Counterfactually, if your friend had intervened to water the plant, it'd still be alive, and therefore your friend not watering the plant caused its death. However, if the Queen of England had intervened to water the plant, it'd also still be alive, and therefore it appears your plant's death was caused just as much by the Queen of England. This intuitively seems wrong. How do we appropriately evaluate the space of *sensible* counterfactuals or states over which we assess causation? As we will discuss, there are several options.

Previous research has introduced a formalism capable of dealing with this issue in the form of an *intervention distribution* [20]. An intervention distribution is a probability distribution over possible interventions that a modeler or experimenter considers. Effectively, rather than considering a single $do(x)$ operator [1], it is a probability distribution over some applied set of them. The intervention distribution fixes $P(C)$, the probability of causes, which is in fact necessary to calculate all the proposed causal measures.

We point out that there are essentially three choices that a modeler/experimenter has for intervention distributions. The first obvious choice is the *observational distribution*. Also sometimes called an "observed distribution," in the dynamical systems we're discussing this corresponds to the stationary distribution of states:

$$P_{obs}(c) = \lim_{n\to\infty} P^n(e \mid c)$$

In this choice, $P(C)$ is simply based on the system's dynamics itself. However, this choice suffers from serious problems—indeed, much has been made of the fact that analyzing causation must explicitly be about what *didn't* happen, i.e., departures from dynamics, and the observational distribution misses this [45]. In this case, your plant is dead because your friend didn't water it but you can't even consider what would have happened if they had, since it's not in the observational distribution. In another example: a light-switch would have varying causal power over a light-bulb based entirely on the probability of the person in its house switching it on and off. Another example: a dynamical system with point attractors has no causation under this assumption. This is because the gain from mere observation to perturbing or intervening is lost when the intervention distribution equals the observational distribution. Finally, it is worth noting that definable stationary distributions rarely exist in the real world.

To remedy this, measures of causation often implicitly assume the second choice: an unbiased distribution of causes over $\Omega$, totally separate from the dynamics of the system. In its simplest form, this is described as a maximum-entropy intervention distribution:

$$P_{maxent}(c) = \frac{1}{n}$$

where $|\Omega| = n$. The maximum-entropy distribution has been made explicit in the calculation of, for instance, Integrated Information Theory [46] or the previously-described effective information of Section 3.10 [42]. There are a number of advantages to this choice, at least when compared to the observational distribution. First, it allows for the appropriate analysis of counterfactuals. Second, it is equivalent to randomization or noise injection, which severs common causes. Third, it is the maximally-informative set of interventions (in that maximum-entropy has been "injected" into the system).

However, it also has some disadvantages. Using a maximum-entropy intervention distribution faces the difficulty that if $\Omega$ is too large, it might be too computationally expensive to compute. More fundamentally, using $P_{maxent}(c)$ can lead to absurdity (e.g., it assumes that the counterfactual wherein the Queen of England watered the plant is just as equally likely as your friend watering it, thus leading to the paradox wherein your friend is not a necessary cause of your plant's death). That is, $P_{maxent}(c)$, taken literally, involves very distant and unlikely possible states of affairs. However, in cases where the causal model has already been implicitly winnowed to be over events that are considered likely, related, or sensible—such an already constructed or bounded causal model, like a set of logic gates, gene regulations, or neuronal connections—it allows for a clear application and comparison of measures of causation.

We point out there is a third possible construction of an intervention distribution. This is to take a local sampling of the possible world space (wherein locality is distance in possible worlds, states of affairs, the state-space of the system, or even based on some outside non-causal information about the system). There are a number of measures of causation that are constructed around local interventions; one of the earliest and most influential is David Lewis's idea of using the closest possible world as the counterfactual by which to reason about causation 3.7. Other examples that implicitly take a local intervention approach includes the bit-flip measure [41] of Section 3.8, as well as the "causal geometry" extension of effective information in continuous systems [9]. We formalize the assumptions behind these approaches as a local intervention distribution, which are possible states of affairs that are similar (or "close") to the current state or dynamics of the system.

For example, to calculate Lewis's measure, we can compute locality using the Hamming distance [40]. Rather than simply picking a single possible counterfactual $\bar{c} \in \Omega$ (which in Lewis's measure would be the closest possible world from Section 3.7) we can instead create a local intervention distribution which is a local sampling of states of affairs where $c$ didn't occur. This is equivalent to considering all states which are a Hamming distance less or equal to $\Delta$ from the actual state:

$$P_{local(c^*)}(c) = \begin{cases} \frac{1}{n_\Delta}, & \text{if } c \in \Theta_{c^*} \\ 0 & \text{otherwise} \end{cases}$$

$$\Theta_{c^*} = \{s \in \Omega \mid D_H(s, c) \leq \Delta\}$$

where $n_\Delta = |\Theta_{c^*}|$. For example, if we want to locally intervene within a distance $\Delta = 1$ around an actual state $c^* = 001$, then $\Theta_{c^*} = \{001, 101, 011, 000\}$ and $n_\Delta = 4$, such that the intervention distribution is $1/4$ over the four states and 0 elsewhere.

We note that local interventions avoid many of the challenging edge cases of measuring causation. Therefore, we we use local interventions for our main text figures to highlight their advantages. However, in order to make our points about the consilience of measures of causation, as well as causal emergence, we take an exhaustive approach and consider all three choices of intervention distributions for the dozen measures. It should be stressed that a) the measures again behave quite similarly, even across different choices of intervention distributions, and also b) instances of causal emergence are, as we will show, generally unaffected by choice of intervention distribution.

### 4.3 All measures of causation are sensitive to noise

To demonstrate the consilience between measures of causation, as well as their underlying causal primitives, we study their behavior in the model described in section 4.1 under different parameterizations of noise in the form of indeterminism and degeneracy. Due to how we paramaterize determinism and degeneracy, we can simplify looking at every single transition in the model into just two. This is because any given state has a *main transition*, which is the transition of highest probability (e.g., $000 \rightarrow 111$ in Figure 2A) and its set of *secondary transitions* which are the lower probabilities of transitions (e.g., $001 \rightarrow 111$ in Figure 2A). When the probability of main transitions equals that of secondary transitions, the system is maximally indeterminate, since all state transitions are a random choice (maximum noise of prediction). This is what is occurring along the $det$ (determinism) axis in Figure 3. When main effects are stacked on top of a given target, this is increasing the $deg$ (degeneracy axis) (maximum noise in retrodiction). The precise nature of this parameterization and how it reflects the determinism and degeneracy is discussed in Supplementary Section 7.

We apply the measures of causation in Section 2 in both a state-dependent and a state-independent manner, since both are common throughout the literature on causation [47, 48, 34, 49, 50, 51]. We examined the behavior of the measures on specific transitions (such as identifying strong or weak causes) but also their expectation averaged across all transitions, thus covering both individual and global causal properties of the system.

Our expectation is that, broadly, measures of causation should peak in their values when determinism is maximized and degeneracy is minimized. And indeed, that is what we find in the bipartite model across the measures of Section 3 and whether they are applied in a state-dependent / actual causation sense or in a global expectation sense (with the exception of the bit-flip measure, but this may be a function of our arbitrary state-labeling, since it is sensitive to that).

Furthermore, we consider different intervention distributions used to probe counterfactual space: the maximum entropy distribution where all states are equally and exhaustively probed; the stationary distribution where the states are weighted according to their frequency of occurrence in the long-term dynamic of the system; the local perturbation distribution, where a subset of the full state-space is probed by considering states that are close to the candidate cause according to some criteria of distance (e.g., Hamming distance).

The majority of the measures of causation increase with the determinism of the model and decrease as the model gets more degenerate (Figure 3). Moreover, the system level behavior of the causation measures, i.e. average across all state transitions, is dominated by that of the main transitions, which is consistent with the idea that these transitions concentrate the causal powers of the system. Note that these results are shown using local perturbations, but using the other intervention distributions led to qualitatively similar results the maximum-entropy distribution and the observational distribution (data shown for the causal primitives in Supplementary Section S2). This indicates that local perturbations may indeed provide an efficient surrogate for computing causal powers without relying on the exhaustive exploration of counterfactual space or using an observational distribution that reflects the system's dynamics rather than its causal structure.

## 5 Macroscale causation

Causal emergence ($CE$) is computed as the difference between the macroscale's causal relationships and the microscale's causal relationships with respect to a given measure of causation.

$$CE = CS_{macro} - CS_{micro}$$

If $CE$ is positive, there is causal emergence, i.e., the macroscale provides a better causal account of the system than the microscale. This can be interpreted as the macroscale doing more causal work, being more powerful, strong, or more informative, depending on how the chosen measure of causation is itself interpreted. A negative value indicates

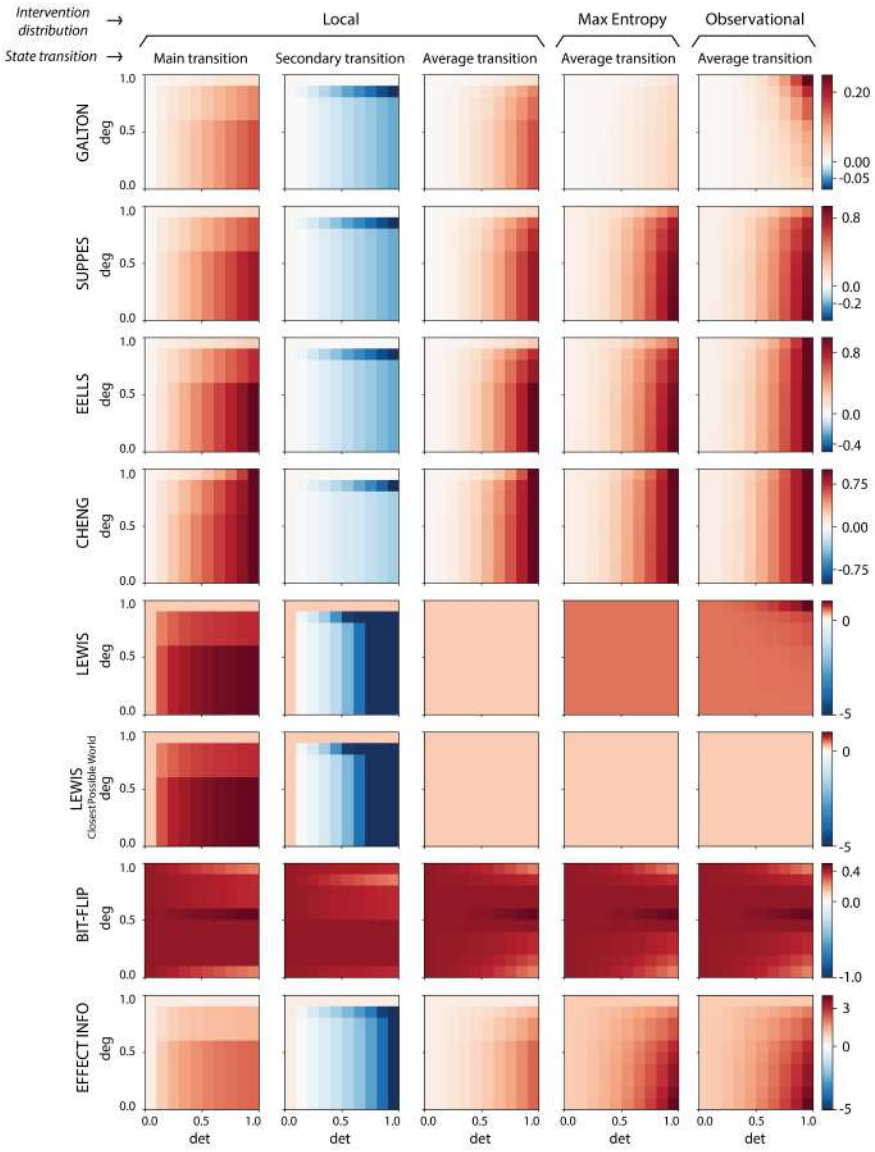



Figure 3: **Behavior of the causation measures in the model system.** Heatmaps of causal strength are shown for all measures (rows) calculated for the microscale of the bipartite Markov chain model with $n = 16$ microstates, 8 states in each macro group ($\Omega_A = \{0000, 0001, 0010, 0011, 0100, 0101, 0110, 0111\}$ and $\Omega_B = \{1111, 1110, 1101, 1100, 1011, 1010, 1001, 1000\}$), at different values of the determinism and degeneracy parameters (see Section 7.1 in the Supplementary Information for a detailed description). Positive values indicate presence of causal strength and are depicted in red, while negative values correspond to what is known as preemptive or negative causation and are shown in blue. Each measure was calculated for different transitions in the bipartite model: a main transition where a strong causal link is thought to be present ($0000 \rightarrow 1111$), a secondary transition where the causal relationship is supposedly weak ($0000 \rightarrow 1110$), and the average across all state transitions. This average is computed as the joint expectation of the measure $CS(c, e)$ across all transitions using $P(c, e)$ calculated using the transition probability matrix (TPM) and the observational distribution to reflect the expectation of causal strength. The measures were computed using different intervention distributions to assess the counterfactuals: the maximum entropy distribution (all states are uniformly sampled), the stationary distribution (states are sampled according to the observed distribution of the dynamics of the model) and the local distribution (the candidate cause is locally perturbed, so that "close" counterfactuals are sampled). For each measure (row), a common scale is used (shown in the colorbar). The full combinations of intervention distributions and transitions can be found in Supplementary Figure S2.

*causal reduction*, which is when the microscale gives the superior causal account. Note that the theory is agnostic as to whether emergence or reduction occurs.

### 5.1 Modeling macroscales

In order to calculate causal emergence, both a microscale and macroscale must be defined. It should be note that the theory is scale-relative, in that one starts with a microscale that is not necessarily some fundamental physical microscale. It is just some lower-bound scale. In neuroscience, for instance, this may be the scale of individual synapses. A macroscale is then some dimension reduction of the microscale, like a coarse-graining (an averaging) or black-boxing (a leaving of variables exogenous) or more generally just any summary statistic that recasts the system with less parameters [11]. E.g., in the neurosciences a macroscale may be a local-field potential or neuronal population or even entire brain regions.

Previous research has laid out clear examples and definitions of macroscales in different system types [8, 20, 11]. One important note is that macroscales should be dynamically consistent with their underlying microscale. This means that the macroscale is not just derivable from the microscale (supervenience) but also that the macroscale behaves identically or similarly (in terms of its trajectory, dynamics, or state-transitions over time). Mathematical definitions of consistency between scales have been proposed [11]; however, here we can eschew this issue as the macroscale for the bipartite model we use automatically ensures consistency by simply grouping each side of the bipartition. Specifically, we use a microscale with $N = 16$ microstates $\Omega_{micro} = \Omega_A \bigcup \Omega_B = \{0000, 0001, 0010, 0011, 0100, 0101, 0110, 0111\} \bigcup \{1111, 1110, 1101, 1100, 1011, 1010, 1001, 1000\}$ and two macrostates $\Omega_{macro} = \{\text{ON}, \text{OFF}\}$ defined by the coarse-graining function $h : \Omega_{micro} \rightarrow \Omega_{macro}$, with $h(\Omega_A) = \text{ON}$ and $h(\Omega_B) = \text{OFF}$ (Figure 2B).

This coarse-grains the bipartite model into a simple two-state system at the macroscale, which trades off dynamically (a NOT gate with a self-loop). This macroscale is deterministic (each macrostate transitions solely to the other) and non-degenerate (each macrostate has only one possible cause). Note that, in our bipartite model, the macroscale is deterministic, non-degenerate, and dynamically consistent no matter the underlying noise in the microscale. This allows us to compare a consistent macroscale against parameterizations of noise, like increases in indeterminism and degeneracy. It's also worth noting that for the macroscale grouping of the bipartite model, the stationary intervention distribution, maximum entropy distribution, and local intervention distribution, are all identical at the macroscale, ensuring clear comparisons.

### 5.2 All measures of causation assessed show causal emergence

Taking into consideration different transitions in the model and employing different intervention distributions, all measures of causation exhibited instances of causal emergence, as shown in Figure 4). This is likely because the causal primitives demonstrated causal emergence, and the measures are universally composed of or closely related to these primitives. Exactly as would be predicted by the idea that macroscales provide error-correction of noise in causal relationships, causal emergence is greater when determinism is low and degeneracy is high in the microscale across the set of measures (see Figure 5). Moreover, causal emergence occurred most prominently in secondary transitions, where causal strength in the microscale was shown to be generally lower due to noise, than in main transitions. There were even cases of "super causal emergence" wherein a microscale transition has a preventative role due to a negative value while the macroscale transition has a positive value, according to the same measure. Additionally, at the global system level, such as at the expectation of the measures of causation, there was also significant amount of causal emergence in certain system architecture domains (particularly those with more uncertainty).

Note that the ubiquity of causal emergence hinges on no particular way of performing the intervention distribution that all measures implicitly require be specified in their application. Causal emergence was present across measures of causation calculated using the maximum-entropy distribution, the observational intervention distribution, and the local intervention distribution (see Figure S3 in Supplementary Information), although distributed slightly differently depending on choice. Indeed, the only condition to not show causal emergence was the overall effective information when using the observational distribution. This was known [20] but was also pointed out by Scott Aaronson [22] as a possible criticism of the theory of causal emergence, since the mutual information (the effective information under the observational distribution) is not higher at a macroscale. First, as we show here, causal emergence still appears in the individual transition's effect information under the observational distribution, meaning that even under the observational distribution only in the average transition (not across all of them) was there no causal emergence in this condition (recall that the effective information is the average of the effect information). Additionally, the mutual information is not traditionally a causal measure [52]. Nor is it a monolithic quantity, but can be decomposed into synergistic, unique, and redundant information. Recent research has shown that the synergistic and unique mutual information can indeed

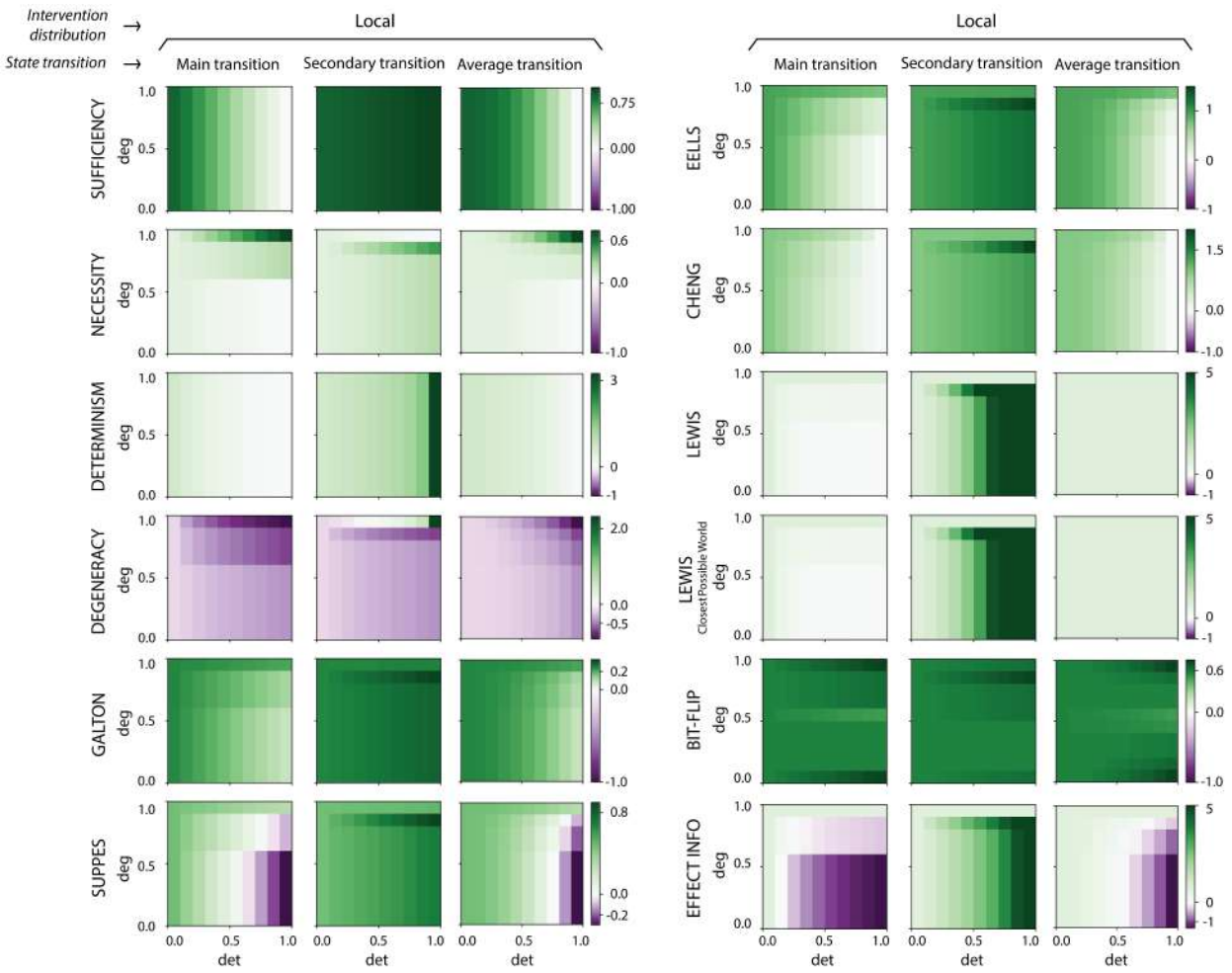



Figure 4: **Causal emergence is widespread across choice of measure of causation and intervention distribution.** Heatmaps of causal emergence (CE) and causal reduction (CR) is shown for all measures of causation and causal primitives computed in the bipartite Markov chain model. Causal emergence is calculated as the difference between the causation metric calculated in the macroscale and in the microscale, such that positive values (green) amount to CE and negative values (purple) to CR. CE/CR is assessed using a local intervention distribution, in which a subset of counterfactuals by perturbing the cause around "close" states. In each of the three columns, CE/CR is assessed over different state transitions of the system: a main transition with a strong causal strength ($0000 \rightarrow 1111$), a secondary transition with a weak causal strength ($0000 \rightarrow 1110$) and the expectation over all state transitions. The joint probability $P(c, e)$ used to compute the expectation is obtained using the transition probabilities $P(e \mid c)$ and the stationary intervention distribution $P_{obs}(c)$. For each measure (row), a common scale is used (shown in the colorbar). Causal emergence across the full combinations of intervention distributions and transitions can be found in Supplementary Figure S3.

increase at a macroscale, indicating that the non-redundant bits of the mutual information show causal emergence [27]. Overall, in context of the results from other measures this indicates solely that effective information is a conservative measure of causal emergence, rather than a liberal one, compared to other measures of causation.

Finally, in order to ensure that these results did not hinge on the symmetry of the bipartite model ($n_A = n_B$) we assessed causal emergence in an asymmetric bipartite models as well, which also showed causal emergence across measures (see Figure S3).

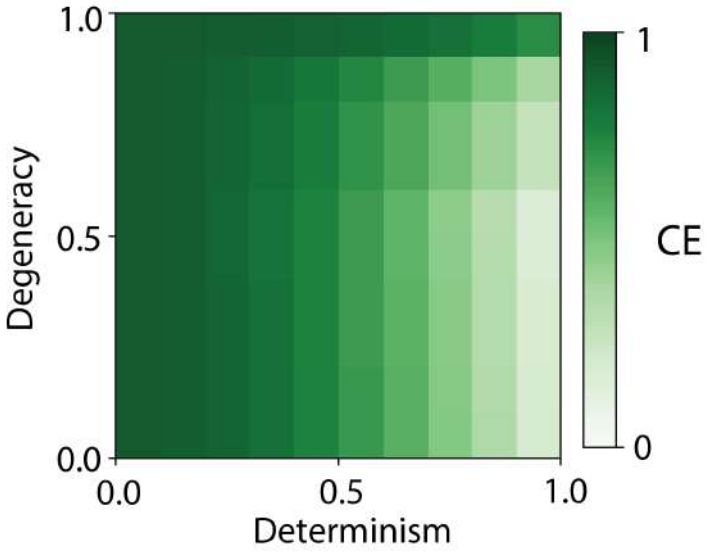



Figure 5: **Causal emergence occurs when the microscale is noisy.** Average behavior of causal emergence across the eight causation measures and four causal primitives for the bipartite Markov chain model. All twelve metrics were normalized to range from -1 to 1 by dividing each metric by its maximum absolute value of CE/CR and then combined through a simple average at each value of determinism and degeneracy. Values were all positive, ranging from low CE (light green) to high CE values (dark green) shown in the colorbar.

## 6 Discussion

Causal emergence is when a measure of causation returns a higher value by having more causal strength, power, informativeness, predictiveness, or causal work (depending on the details of the measure of causation) at the macroscale vs. the microscale of a system. It is possible because macroscales can provide the advantage of noise minimization. That is, emergent scales are those that perform error-correction over their underlying microscale causal relationships. Indeed, we've shown that causal emergence is widespread across popular measures of causation with independent origins in diverse fields. This is because all of these measures are sensitive to noise in the form of indeterminism (uncertainty over the future) and degeneracy (uncertainty over the past). We refer to these terms, along with their simpler forms of sufficiency and necessity, as "causal primitives" since measures are either sensitive to them or even directly constructed from them. Notably, across the more than a dozen independent measures of causation we examined, all demonstrated causal emergence in a bipartite model system in conditions of high uncertainty over state transitions (low determinism, high degeneracy). This was true across a number of possible assumptions of how those measures are applied, showing the robustness of this theory of emergence.

The consilience of the measures examined here provide a bedrock for previous research which has already shown causal emergence using more complex information-theoretic measures of causation like effective information [8], the integrated information [18], and also recently the synergistic information [27]. Interestingly enough, we find that effective information, despite being the original measure proposed to capture causal emergence, is the most conservative in our sample.

One interesting discovery of this investigation is the similarities and agreements within measures of causation themselves. Broadly, we find that causation is not itself a primitive notion but can be decomposed along two dimensions (a finding in agreement with previous authors [1, 44]). These two dimensions are, in the philosophical literature, referred to as sufficiency and necessity; as we show, these are specific cases of determinism and degeneracy, respectively. Successful measures of causation are sensitive to both dimensions. Indeed, it is the sensitivity to these terms, and the uncertainty they capture, that guarantees the possibility of causal emergence of such measures.

It's worth noting that the measures of causation we examined need a space of possibilities, or counterfactuals, to be specified, in order to apply the measure. Here, we represent this choice mathematically using an intervention distribution. We find that causal emergence is relatively invariant across choice of intervention distribution, indicating that it is a robust phenomenon. While the choice of intervention distribution in the majority of measures doesn't affect

the possibility of causal emergence, we advocate for our notion of "local interventions" as being a step forward for mathematical measures of causation, as it offers a compromise between a maximum-entropy approach (all possibilities considered) and a minimal-difference approach (only the closest possibility is considered).

Despite its ubiquity across measures and background conditions, the existence of emergence itself is not trivially guaranteed. Rather, it is a function of system architecture or dynamics. As we have shown, in dynamic domains of deterministic and time-reversible system mechanics, causal reduction dominates. However, in scientific models these conditions are quite rare, as science deals with mainly with open systems exposed to outside uncertainty or, alternatively, systems with inherent uncertainty. Even systems with irreducibly small amounts of noise can have that noise amplified into significant uncertainty after dynamical iteration [53]. Therefore, we expect causal emergence to be common across the many scales and models of science.

The development of complex systems science was based on novel insights into how complexity can arise via iteration of simple rules [54, 55, 56]; not only that, it was based around a family of measures of complexity [7]. The development of a science of emergence should be based on causal relationships (captured by the family of measures of causation) and the noise-minimizing properties of macroscales. Ultimately, this work provides a necessary toolkit for the scientific identification of emergent scales of function, along with optimal modeling choices, interventions, and explanations.

## References


[1] Judea Pearl. *Causality*. Cambridge University Press, Cambridge, 2 edition, 2009.

[2] Branden Fitelson and Christopher Hitchcock. Probabilistic Measures of Causal Strength. *Causality in the Sciences*, January 2010.

[3] Marcello Massimini, Melanie Boly, Adenauer Casali, Mario Rosanova, and Giulio Tononi. A perturbational approach for evaluating the brain's capacity for consciousness. In *Progress in Brain Research*, volume 177, pages 201–214. Elsevier, 2009.

[4] Selmaan N. Chettih and Christopher D. Harvey. Single-neuron perturbations reveal feature-specific competition in V1. *Nature*, 567(7748):334–340, March 2019. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 7748 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Computational neuroscience;Neural circuits;Sensory processing;Visual system Subject_term_id: computational-neuroscience;neural-circuit;sensory-processing;visual-system.

[5] Olaf Sporns. Brain connectivity. *Scholarpedia*, 2(10):4695, October 2007.

[6] Andrew A. Fingelkurts, Alexander A. Fingelkurts, and Seppo Kähkönen. Functional connectivity in the brain–Is it an elusive concept? *Neuroscience and Biobehavioral Reviews*, 28(8):827–836, 2005. Place: Netherlands Publisher: Elsevier Science.

[7] Murray Gell-Mann. What is complexity? Remarks on simplicity and complexity by the Nobel Prize-winning author of The Quark and the Jaguar. *Complexity*, 1(1):16–19, 1995. Publisher: John Wiley & Sons, Ltd.

[8] Erik P. Hoel, L. Albantakis, and G. Tononi. Quantifying causal emergence shows that macro can beat micro. *Proceedings of the National Academy of Sciences*, 110(49):19790–19795, December 2013.

[9] Pavel Chvykov and Erik Hoel. Causal Geometry. *arXiv:2010.09390 [hep-th, physics:physics]*, October 2020. arXiv: 2010.09390.

[10] Erik Hoel and Michael Levin. Emergence of informative higher scales in biological systems: a computational toolkit for optimal prediction and control. *Communicative & Integrative Biology*, 13(1):108–118, January 2020. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/19420889.2020.1802914.

[11] Brennan Klein and Erik Hoel. The Emergence of Informative Higher Scales in Complex Networks. *Complexity*, 2020:e8932526, April 2020.

[12] Thomas F. Varley. Causal Emergence in Discrete and Continuous Dynamical Systems. *arXiv:2003.13075 [nlin]*, March 2020. arXiv: 2003.13075.

[13] Jiang Zhang. Neural Information Squeezer for Causal Emergence. *arXiv:2201.10154 [physics]*, January 2022. arXiv: 2201.10154.

[14] Brennan Klein, Erik Hoel, Anshuman Swain, Ross Griebenow, and Michael Levin. Evolution and emergence: higher order information structure in protein interactomes across the tree of life. *Integrative Biology*, 13(12):283–294, 2021.

[15] Rafael Yuste. From the neuron doctrine to neural networks. *Nature Reviews Neuroscience*, 16(8):487–497, August 2015. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 8 Primary_atype: Reviews Publisher: Nature Publishing Group Subject_term: Electrophysiology;Network models;Neural circuits Subject_term_id: electrophysiology;network-models;neural-circuit.

[16] Daniel P. Buxhoeveden and Manuel F. Casanova. The minicolumn hypothesis in neuroscience. *Brain*, 125(5):935–951, 2002.

[17] B. T. Thomas Yeo, Fenna M. Krienen, Jorge Sepulcre, Mert R. Sabuncu, Danial Lashkari, Marisa Hollinshead, Joshua L. Roffman, Jordan W. Smoller, Lilla Zöllei, Jonathan R. Polimeni, Bruce Fischl, Hesheng Liu, and Randy L. Buckner. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, 106(3):1125–1165, September 2011.

[18] Erik P. Hoel, Larissa Albantakis, William Marshall, and Giulio Tononi. Can the macro beat the micro? Integrated information across spatiotemporal scales. *Neuroscience of Consciousness*, 2016(1):niw012, 2016.

[19] Acer Y. C. Chang, Martin Biehl, Yen Yu, and Ryota Kanai. Information Closure Theory of Consciousness. *Frontiers in Psychology*, 11:1504, July 2020. arXiv: 1909.13045.

[20] Erik Hoel. When the Map Is Better Than the Territory. *Entropy*, 19(5):188, April 2017.

[21] William Marshall, Larissa Albantakis, and Giulio Tononi. Black-boxing and cause-effect power. *PLOS Computational Biology*, 14(4):e1006114, April 2018.

[22] Scott Aaronson. Higher-level causation exists (but I wish it didn't), June 2017.

[23] Joe Dewhurst. Causal emergence from effective information: Neither causal nor emergent? *Thought: A Journal of Philosophy*, 10(3):158–168, 2021. Publisher: John Wiley & Sons, Ltd.

[24] Max Tegmark. Improved Measures of Integrated Information. *PLOS Computational Biology*, 12(11):e1005123, November 2016. Publisher: Public Library of Science.

[25] Pedro A. M. Mediano, Fernando Rosas, Robin L. Carhart-Harris, Anil K. Seth, and Adam B. Barrett. Beyond integrated information: A taxonomy of information dynamics phenomena. *arXiv:1909.02297 [physics, q-bio]*, September 2019. arXiv: 1909.02297.

[26] Tim Bayne. On the axiomatic foundations of the integrated information theory of consciousness. *Neuroscience of Consciousness*, 2018(1), January 2018.

[27] Thomas Varley and Erik Hoel. Emergence as the conversion of information: A unifying theory. *arXiv:2104.13368*, April 2021. arXiv: 2104.13368.

[28] Pedro A. M. Mediano, Fernando E. Rosas, Andrea I. Luppi, Henrik J. Jensen, Anil K. Seth, Adam B. Barrett, Robin L. Carhart-Harris, and Daniel Bor. Greater than the parts: A review of the information decomposition approach to causal emergence. *arXiv:2111.06518 [nlin, q-bio]*, November 2021. arXiv: 2111.06518.

[29] Fernando E. Rosas, Pedro A. M. Mediano, Henrik J. Jensen, Anil K. Seth, Adam B. Barrett, Robin L. Carhart-Harris, and Daniel Bor. Reconciling emergences: An information-theoretic approach to identify causal emergence in multivariate data. *PLOS Computational Biology*, 16(12):e1008289, December 2020. Publisher: Public Library of Science.

[30] Prerna Nadathur and Sven Lauer. Causal necessity, causal sufficiency, and the implications of causative verbs. *Glossa: a journal of general linguistics*, 5(1):49, June 2020. Number: 1 Publisher: Ubiquity Press.

[31] David Hume. *An Enquiry concerning Human Understanding*. 1748.

[32] Phyllis Illari and Federica Russo. *Causality: Philosophical Theory meets Scientific Practice*. Oxford University Press, Oxford, New York, December 2014.

[33] Ellery Eells. *Probabilistic Causality*. Cambridge University Press, 1991.

[34] Larissa Albantakis, William Marshall, Erik Hoel, and Giulio Tononi. What Caused What? A Quantitative Account of Actual Causation Using Dynamical Causal Networks. *Entropy*, 21(5):459, May 2019.

[35] Patrick Suppes. *A Probabilistic Theory of Causality*. Amsterdam: North-Holland Pub. Co., 1968.

[36] Christopher Hitchcock. Probabilistic Causation. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2021 edition, 2018.

[37] Patricia W. Cheng and Laura R. Novick. Causes versus enabling conditions. *Cognition*, 40(1):83–120, August 1991.

[38] David Lewis. Causation. *Journal of Philosophy*, 70(17):556–567, 1973.

[39] David Lewis. Postscripts to 'Causation'. *Philosophical Papers Vol. Ii*, 1986.

[40] Luciano Floridi. Information, possible worlds and the cooptation of scepticism. *Synthese*, 175:63–88, 2010. Publisher: Springer.

[41] Bryan C. Daniels, Hyunju Kim, Douglas Moore, Siyu Zhou, Harrison B. Smith, Bradley Karas, Stuart A. Kauffman, and Sara I. Walker. Criticality Distinguishes the Ensemble of Biological Regulatory Networks. *Physical Review Letters*, 121(13):138102, September 2018. Publisher: American Physical Society.

[42] Giulio Tononi and Olaf Sporns. Measuring information integration. *BMC Neuroscience*, page 20, 2003.

[43] Paul E. Griffiths, Arnaud Pocheville, Brett Calcott, Karola Stotz, Hyunju Kim, and Rob Knight. Measuring Causal Specificity. *Philosophy of Science*, 82(4):529–555, 2015. Publisher: The University of Chicago Press.

[44] J. L. Mackie. Causes and Conditions. *American Philosophical Quarterly*, 2(4):245–264, 1965. Publisher: University of Illinois Press.

[45] Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, New York, 1° edizione edition, 2017.

[46] Masafumi Oizumi, Larissa Albantakis, and Giulio Tononi. From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0. *PLoS Computational Biology*, 10(5):e1003588, May 2014.

[47] David Balduzzi and Giulio Tononi. Integrated Information in Discrete Dynamical Systems: Motivation and Theoretical Framework. *PLoS Computational Biology*, 4(6):e1000091, June 2008.

[48] Joseph Y. Halpern. *Actual Causality*. MIT Press, Cambridge, MA, USA, August 2016.

[49] Bjørn Erik Juel, Renzo Comolatti, Giulio Tononi, and Larissa Albantakis. When is an action caused from within? Quantifying the causal chain leading to actions in simulated agents. pages 477–484. MIT Press, July 2019.

[50] Christoph Adami. The use of information theory in evolutionary biology: Information theory in evolutionary biology. *Annals of the New York Academy of Sciences*, 1256(1):49–65, May 2012.

[51] Nicholas M. Timme and Christopher Lapish. A Tutorial for Information Theory in Neuroscience. *eNeuro*, 5(3):ENEURO.0052–18.2018, September 2018.

[52] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory 2nd Edition*. Wiley-Interscience, Hoboken, N.J, 2nd edition edition, July 2006.

[53] Steven Strogatz. *Nonlinear Dynamics and Chaos, 2nd Edition: With Applications to Physics, Biology, Chemistry, and Engineering: With Applications to Physics, Biology, Chemistry, and Engineering, Second Edition*. Westview Press, Boulder, CO, 2° edizione edition, 2014.

[54] Stephen Wolfram. *A New Kind of Science*. Wolfram Media Inc, Champaign, IL, first edition edition, 2002.

[55] M. E. J. Newman. Resource Letter CS–1: Complex Systems. *American Journal of Physics*, 79(8):800–810, 2011. Publisher: American Association of Physics Teachers.

[56] J. P. Crutchfield and M. Mitchell. The evolution of emergent computation. *Proceedings of the National Academy of Sciences*, 92(23):10742–10746, November 1995. Publisher: National Academy of Sciences Section: Research Article.

# 7 Supplementary Information

## 7.1 Parameterizing determinism and degeneracy in the bipartite model

What follows is the detailed description of how we algorithmically vary $det$ and $deg$ in the bipartite Markov chain model. First we will label the $2^N = n$ states with binary strings and divide them into two groups $A$ and $B$ of size $n_A$ and $n_B$, respectively. For now, let us consider the symmetric case where $n_A = n_B$. For example, with $N = 3$ we have states $\Omega = \Omega_A \cup \Omega_B = \{000, 001, 010, 011\} \cup \{111, 110, 101, 100\}$) (Figure 2A, top). The model's dynamics is governed by a transition probability matrix, where for a given state $c \in \Omega$ the system is in, it can transition to a state $e \in \Omega$ with probability given by $P(e \mid c)$, such that any state transition defines a cause and effect pair (Figure 2A, bottom). Each cause state is paired to a main effect state in the opposite grouping through a mapping $f_A : \Omega_A \to \Omega_B$. A complementary mapping is given $f_B(s) = f_A(s \uparrow) \uparrow$, where $\uparrow$ is the state obtained by inverting all bits (e.g. $\uparrow 100 = 011$) (this is equivalent to reflecting the arrows in Figure 2A along the vertical axis). For a given state $c \in \Omega_A$ we have:

$$P(e \mid c) = \begin{cases} p, & \text{if } e = f_A(c) \text{ and } e \in \Omega_B \\ \frac{(1-p)}{N_B} & \text{if } e \neq f_A(c) \text{ and } e \in \Omega_B \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

If $c \in \Omega_B$, we simply interchange $B$ for $A$ in the definition above. $0 \leq p \leq 1$ is the parameter which controls the determinism of the system, by concentrating or diluting the probability over the main effect (vs. the secondary effects) of a given cause. Essentially, we are simply narrowing or widening the "scope" of possible effects from a given state in order to increase or decrease the determinism, respectively.

In order to parametrically vary the degeneracy of the model we change the mapping $f_A$ (and its complement $f_B$), going from zero degeneracy where $f_A$ is injective, (maps different cause to different main effects according to $f_A(s) =\uparrow s$ to max degeneracy), where all causes in one group map to a single effect in the other group (Figure 2B). To increase the degeneracy in a step wise manner we use the following algorithm: we chose the "poorest" effect, i.e. the one with the least number of main causes ($argmin \mid f^{-1}(e) \mid$) but with at least one main cause, and move all its main causes to the next poorest effect. In this way, we progressively re-wire the system until all causes map to only one effect and maximal degeneracy is achieved. Essentially, we are simply moving main effects on top of one another sequentially—this increases the degeneracy (and decreases the necessity). A visual example of this can be see in Figure S1.

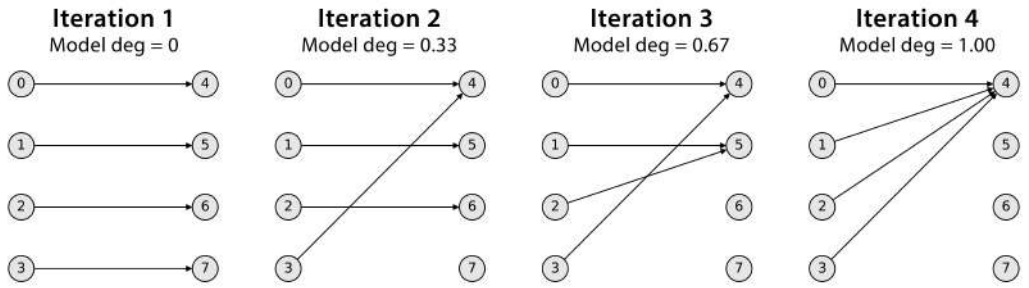



Figure S1: **Visualization of steps in the algorithm to increase degeneracy in the bipartite model.** Iterations of the algorithm for the bipartite Markov chain state-space with $n = 8$ states. States are labeled from 0 to $n - 1$, with states in the left column belonging to group $A$ and on the right to group $B$. The function $f_A : \Omega_A \to \Omega_B$ maps states in group $A$ to states in group $B$, associating every candidate cause $c \in \Omega_A$ to a main effect $f_A(c) = e^*$. At each iteration, a main effect is lost (i.e. the image $f(A) \in \Omega_B$ loses an element) as an arrow is moved an effect with the least number of arrows (with at least one arrow). The main effects of each cause progressively overlap until they are over a single state in group $B$. For $f_B$ the same algorithm is applied, but with the states $A$ and $B$ reversed.

However, it should be noted that our algorithmic methods for varying the determinism and degeneracy do not automatically ensure that it is changing the causal primitives as expected. This is because the algorithmic way of varying determinism and degeneracy (by varying the probabilities between main effects and second effects, and stacking main effects on top of targets, respectively) does not match one-to-one with the underlying mathematical properties of determinism and degeneracy. This is because there is no way to smoothly vary the actual mathematical properties in a simple algorithmic manner.

However, when the causal primitives are computed over different parameters of the model and we considered their global behavior averaged across all transitions in the state-space, the algorithmic determinism and sufficiency indeed scale with their mathematical counterparts. Similarly, the degeneracy primitive scales proportionally to the model's degeneracy parameter, while necessity does so inversely (Figure S2). Note that while determinism and sufficiency are independent of the model's degeneracy parameter, degeneracy and necessity are sensitive to the model's determinism parameter, simply due to its algorithmic construction. These results validate the bipartite model capacity to explore the behavior of the causal measures for different combinations of causal primitives, as modulated by the model's degeneracy and determinism parameter, although, due to the inability to vary degeneracy without varying the determinism, it does not do so over a perfectly symmetric manifold.

It's also interesting to note how the causal primitives behave differently for specific transitions with strong and weak causal link. The main transitions exhibit high sufficiency, determinism and necessity and low degeneracy (Figure S2, first column), in particular, at regions of high determinism and low degeneracy of the parameter space of the model. This behavior dominates and appears at the level of the average quantities across all transitions (Figure S2, last three columns). The secondary transitions show lower values in general for all causal primitives, coherent with the notion that they are endowed with weaker causal powers. In line with this, the determinism and degeneracy of the secondary transitions vary in the opposite manner of a main transition, peaking when the determinism of the model is low, and the degeneracy is high (Figure S2, second column).

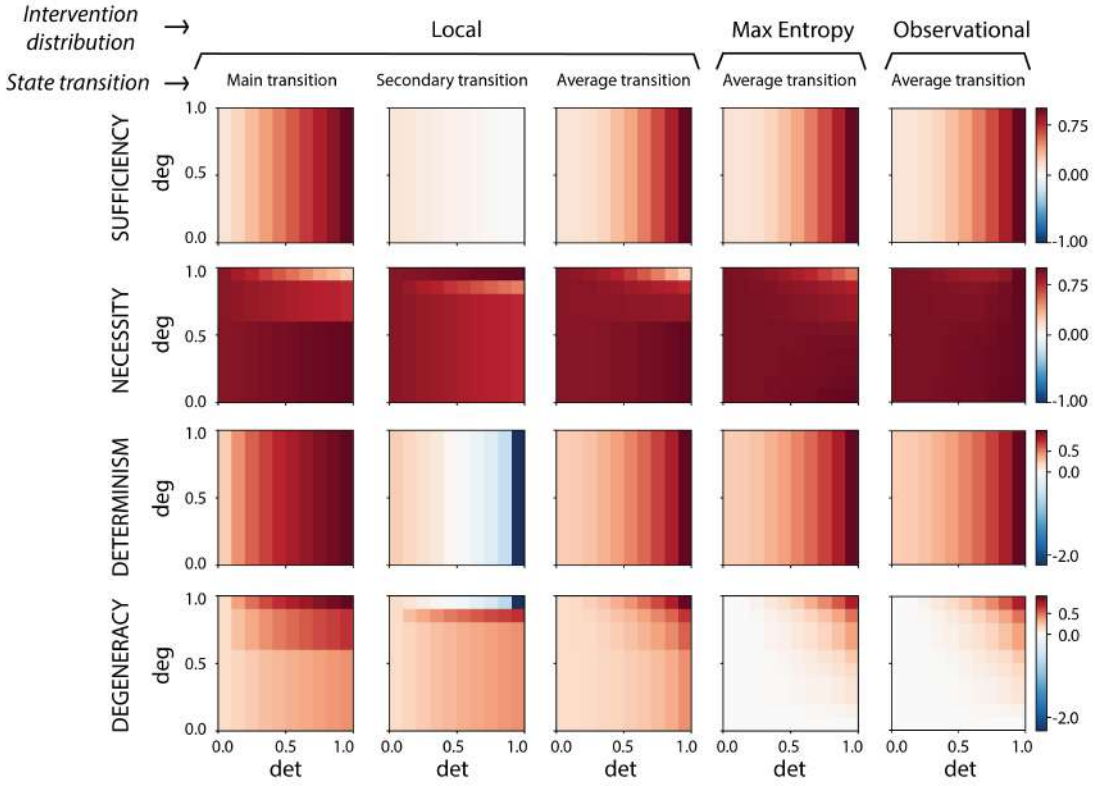



Figure S2: **Behavior of the causal primitives in the model system.** Shown along the rows are the heatmaps of the causal primitives, i.e. suffiency, necessity, determinism and degeneracy, for different values of the model's degeneracy and determinism parameters. In the first three columns, the max entropy distribution is used to calculate the causal primitives and average the across transitions. In the first two columns, the causal primitives assessed for single state transition between a cause and an effect: first, between a cause and its main effect ($0000 \rightarrow 1111$), generally a strong causal link, and second, between a cause and a non principal effect, which generally we would expect to have a weaker causal strength ($0001 \rightarrow 1111$). In the third column, the simple average of the causal primitives across all the state transitions. In the fourth column the average of the causal primitives is shown, but using the stationary distribution to estimate the primitives and also compute the average. In the last column, the causal primitives average across all transitions computed using local perturbations.

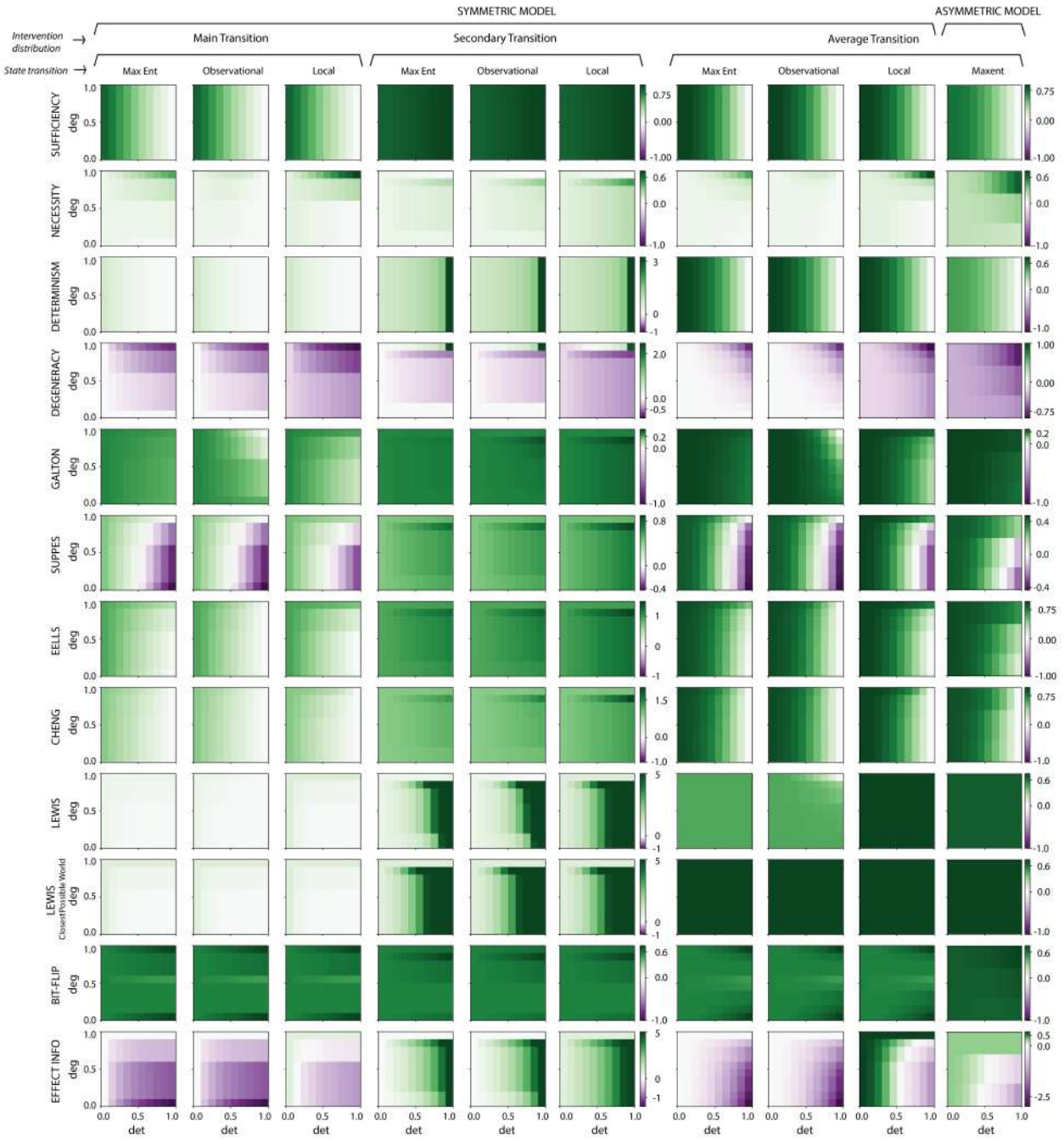



Figure S3: **Causal emergence is generally invariant to intervention distribution choice, as well as symmetry breaking.** An expanded version of Figure 4 to include three different choices of intervention distributions, as well as what happens when the bipartite model is not perfectly symmetric. Heatmaps of causal emergence (CE) and causal reduction (CR) are shown for all measures of causation and causal primitives computed in the bipartite Markov chain model. Causal emergence is calculated as the difference between the causation metric calculated in the macroscale and in the microscale, such that positive values (green) amount to CE and negative values (purple) to CR. CE/CR is assessed using a maximum entropy distribution, a local intervention distribution, and the observational distribution. CE/CR is also assessed over different state transitions of the system: a main transition with a strong causal strength ($0000 \rightarrow 1111$), a secondary transition with a weak causal strength ($0000 \rightarrow 1110$) and the expectation over all state transitions. The joint probability $P(c, e)$ used to compute the expectation using the observational intervention distribution $P_{obs}(C)$. However, the expectation of causal emergence is relatively invariant across even this choice as well (data not shown). For each measure (row), a common scale is used (shown in the colorbar). In the last column, the model was calculated using an asymmetric version of the bipartite model with $n_A = 13$ and $n_B = 3$ states on each macro group, instead of $n_A = n_B = 8$ used in the rest of the paper.