AI Consciousness is Inevitable:

A Theoretical Computer Science Perspective

Lenore Blum and Manuel Blum

ABSTRACT

We look at consciousness through the lens of Theoretical Computer Science, a branch of mathematics that studies computation under resource limitations, distinguishing functions that are efficiently computable from those that are not. From this perspective, we are developing a *formal machine model* for consciousness. We are *inspired* by Alan Turing's simple yet powerful model of computation and Bernard Baars' theater model of consciousness. Though extremely simple, the model (1) aligns at a high level with many of the major scientific theories of human and animal consciousness, (2) provides explanations at a high level for many phenomena associated with consciousness, (3) gives insight into how a machine can have subjective consciousness, and (4) is clearly buildable. This combination supports our claim that machine consciousness is not only plausible but inevitable.

Lenore Blum
lblum@cs.cmu.edu

Manuel Blum
mblum@cs.cmu.edu

Table of Contents

1 Introduction

We study consciousness from the perspective of **Theoretical Computer Science (TCS)**, a branch of mathematics concerned with understanding the underlying principles of computation and complexity, including especially the implications and surprising consequences of resource limitations. As a branch of mathematics, the TCS perspective abstracts and formalizes informal concepts and derives results from them.[1]

From this perspective, we are *developing* a *formal machine model* for consciousness. Our aim is to start with a reasonable and minimal set of formal assumptions about consciousness, then see what follows: the pros and cons of the model. We have tried to make our model "Occam's razor" simple.

We get confidence that we are on the right track from the fact that :

1. All other formal models we tried – and we tried many – were found seriously wanting.
2. The model's underlying principles and derived properties at a high level align with informal concepts and major scientific theories of human and animal consciousness (Section 3).
3. The model answers Kevin Mitchell's 15 questions "that a theory of consciousness should be able to encompass" reasonably well (Appendix, section 6.4).

The **Conscious Turing Machine (CTM)** is a *simple formal machine* model of *consciousness* inspired in part by Alan Turing's *simple,* yet powerful, *formal machine* model of *computation* (Turing, 1937)**,** and by Bernard Baars' *theater model* of consciousness (Baars, Bernard J., 1997) and (Baars, 1997).[2] In the spirit of TCS, informal notions are defined formally in the CTM.

[1] Theoretical Computer Science, a field started in the 1960's, grew out of the Theory of Computation (TOC), a field with origins in the 1930's. By taking resource limitations into account, the TCS perspective differentiated itself from TOC where limitations of time and space do not figure. TOC distinguishes computable from not computable. It does not distinguish between efficiently (feasibly) computable and not efficiently (not feasibly) computable. For a brief history of TOC and TCS, see Appendix, section 6.1.

[2] The *global workspace* has origins in architectural models of cognition, developed largely at Carnegie Mellon University by: Herb Simon's *Sciences of the Artificial* (Simon, 1969), Raj Reddy's Blackboard Model (Reddy, 1976), Allen Newell's *Unified Theories of Cognition* (Newell, 1990) and John Anderson's ACT-R (Anderson, 1996). Indeed, (Baars, 1997) states that: "Global Workspace theory derives from the integrative modeling tradition of Alan Newell, Herbert Simon, John Anderson, and others in cognitive science."

In (Blum & Blum, 2022), we consider how a CTM could exhibit various phenomena associated with consciousness (e.g., blindsight, inattentive blindness, change blindness) and present CTM explanations that agree, at a high level, with cognitive neuroscience literature.

In contrast to Turing, we take resource limitations into account, both in designing the CTM model (e.g., size of chunks, speed of computation) and in how resource limitations affect (and help explain) phenomena related to consciousness (e.g., change blindness and dreams) and related topics such as the paradox of free will (Blum & Blum, 2022).

Our perspective differs even more. What gives the CTM its feeling of consciousness is not its input-output map, nor its computing power, but what's under the hood.[3] In this paper we take a brief look under the hood (section 2.1).

David Chalmers' introduction of the Hard Problem (Chalmers, 1995) helped classify most notions of consciousness into one of two types. The first type, variously called *access* (Block, 1995) or *functional* or *cognitive* consciousness (Humphrey, 2023), we call *conscious attention* (Formal Definition 1 in section 2.2).[4] The second type (associated with the Hard Problem) is called *subjective* or *phenomenal* consciousness (Nagel T. , 1974) and is generally associated with sensory experiences or *qualia*, (Peirce, 1866), modified by (Lewis, 1929). We call it *conscious awareness* (Formal Definition 10 in section 2.3.2.2). We view Chalmers' Hard Problem as a challenge to show that the latter, subjective consciousness, is "computational".[5]

We argue (in section 2.3) that "consciousness" generally requires what we call conscious attention *and* conscious awareness. We contend that a machine that interacts with its worlds (inner and outer)

[3] This is important. We claim that simulations that modify CTM's key internal structures and processes will not in general experience what CTM experiences. On the other hand, we are not claiming that the CTM is the only possible machine model to experience *feelings of consciousness*. The CTM is a *minimal* formal machine model for consciousness. Indeed, Wanja Wiese describes the CTM as a "minimal unifying model", (Wiese, 2020) and personal communication.

[4] In previous papers, e.g., (Blum & Blum, 2022), we used the words "conscious awareness" to denote what we formally define in this paper as *conscious attention*. As emphasized in our formal definition here (in section 2.3), *conscious awareness* requires more than conscious attention; it requires the Model-of-the-World.

In this paper, our formal definition of *attention* aligns more with what is generally considered "access". Our formal definition of *conscious awareness* aligns more with what is generally considered "subjective" or "phenomenal" consciousness.

[5] We are not alone in this view, see e.g., (Dehaene, Charles, King, & Marti, 2014).

via input sensors and output actuators, that constructs models of these worlds enabling planning, prediction, testing, and learning from feedback, and that develops a rich internal multimodal/multisensory language, can have both types of consciousness. In particular, we contend that subjective consciousness is computational.

We emphasize that the CTM *is a simple formal machine model* designed to explore and understand consciousness from a TCS perspective. While it is inspired by and owes much to cognitive and neuroscience theories of consciousness, it *is not intended* to model the brain *nor* its neural correlates of consciousness.

Specifically, as we have mentioned, the CTM is inspired by cognitive neuroscientist Bernard Baars' theater model of consciousness (Baars, Bernard J., 1997), also called the *global workspace* (**GW**) theory of consciousness. However, the CTM ***is not*** a standard GW model: it differs from GW in a number of important ways:

- In CTM, *competition* for global broadcast is formally defined (sections 2.1.3 and 6.3).
- The CTM does away completely with the ill-defined Central Executive of other GW models.
- In CTM, a distributed *Model-of-the-World processor* (section 2.1.2.1) constructs and employs (distributed) *models* of its (*inner* and *outer*) worlds (section 2.3.1) *to make sense* of itself and itself in its worlds.
- *Brainish* (sections 2.1.2.1 and 2.3.1), CTM's *self-generated multimodal internal language*, enables communication between processors and provides the multimodal labeling of sketches in CTM's world models. Brainish and world models play important roles in generating CTM's subjective feelings such as pain and pleasure (sections 2.3.1.1and 2.3.1.2).
- In CTM, *predictive dynamics* (cycles of prediction, testing, feedback and learning, locally and globally) constantly improves its world models (see section 2.1.8).

The CTM interacts with the world via input *sensors* and output *actuators* (section 2.1.6), enabling the creation of its *embodied* (Shanahan, 2010), *embedded, enacted* (Maturana & Varela, 1972) and *extended mind*[6] (Clark & Chalmers, 1998).

[6] In principle, we can also allow off-the-shelf processors in our CTM.

Additionally, CTM naturally *aligns* with and *integrates* features considered key to human and animal consciousness by many of the major scientific theories of consciousness (Section 3).[7] These theories consider different aspects of consciousness and often compete with each other (Lenharo, 2024). Yet the alignment of their underlying principles with CTM at a high level helps demonstrate their compatibility[8].

Even more, their *alignment* with CTM, *a simple machine model* that exhibits many of the important phenomena associated with consciousness, supports our claim that a conscious AI is inevitable.[9]

In addition to specifying a machine model for consciousness, we indicate how the model provides a *framework* for constructing an Artificial General Intelligence (AGI). (See section 2.4.)

While working on this paper, we became aware of Kevin Mitchell's blog post in *Wiring the Brain* (Mitchell, 2023) in which he makes a point similar to one that we make, namely, that many of the major theories of consciousness are compatible and/or complementary[10]. For a similar conclusion, see (He, 2023) and (Storm, et al., 2024). Even more, Mitchell presents fifteen questions "that a theory of consciousness should be able to encompass". He declares that "even if such a theory can't currently answer all those questions, it should at least provide an *overarching framework* [11] (i.e., what a theory really should be) in which they can be asked in a coherent way, without one question destabilizing what we think we know about the answer to another one."

[7] These theories include: The Global Workspace/Global Neuronal Workspace (GW/GNW), Attention Schema Theory (AST), the Self-Organizing Metarepresentational Account (SOMA), Predictive Processing (PP), Integrated Information Theory (IIT), Embodied, Embedded, Enacted and Extended (EEEE) theories, Evolutionary theories, and the ERTAS (Extended Reticulothalamic Activating System) + FEP (Free Energy Principle) theories.

[8] We are often asked: how can these competing theories be compatible? By taking the TCS approach and abstracting these theories' underlying principles, we are not mapping theories to different parts of the human brain as is done in the adversarial competitions (Lenharo, 2024).

[9] Thus, our response to the query, "could machines have it [consciousness]?" (Dehaene, Lau, & Kouider, 2017), is "YES it could". In a recent paper, (Farisco, Evers, & Changeux, 2024) consider the question "Is artificial consciousness achievable?" from the perspective of the human brain. They conclude with a tentative possibility for "non human-like forms of consciousness."

[10] We use "complementary" to emphasize the idea that the diverse theories can be seen as different pieces of a larger puzzle rather than conflicting viewpoints.

[11] Italics are ours.

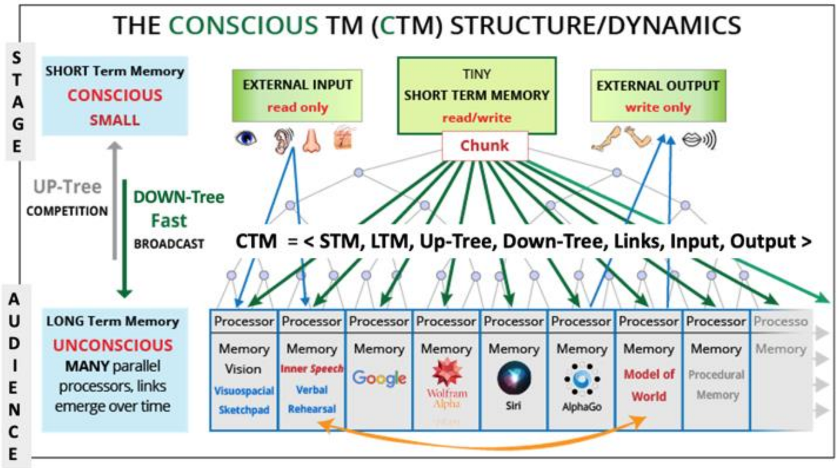
Mitchell's questions are thoughtful, interesting, and important. Later in this paper (Appendix, section 6.4), we offer preliminary answers from the perspective of CTM. Our answers to Mitchell's questions supplement and highlight material in the following overview of CTM.[12]

The development of CTM is a work in progress. More specifics will appear in our upcoming monograph (Blum, Blum, & Blum, 2026).

2 The Conscious Turing Machine (CTM)

2.1 Structure (STM, LTM, Up-Tree, Down-Tree, Links, Input, Output) and Predictive Dynamics

CTM is defined formally as a 7-tuple, (STM, LTM, Up-Tree, Down-Tree, Links, Input, Output). These seven components each have well-defined properties (Blum & Blum, 2022) which we outline in this section.



To start, CTM has a *finite* lifetime **T.** Time **t = 0, 1, 2, 3, … , T** is measured in discrete clock ticks. **T** is a parameter.

[12] In the overview (Section 2), we annotate paragraphs that refer to Kevin Mitchell's queries. For example, if a paragraph has a label [KM1], then it refers to Mitchell's first question, KM1.

CTM has both a time-varying ***inner world*** $\mathbf{I(t)}$ and a time-varying ***outer world*** $\mathbf{O(t)}$. Formally, these are state spaces. Informally, CTM's *inner world* includes its internal mechanisms and processes, and its "thoughts" and memories; its *outer world* is where it lives, what's outside CTM.

Again, to emphasize that CTM is *embodied*, *embedded, enactive* and *extended* in its worlds (inner and outer), we should denote it by **rCTM** (**r** for robot).[13] Instead, we use the acronym CTM for both CTM and rCTM. We now outline the properties of its components:

2.1.1 STM, a Buffer and Broadcasting Station

The stage in the theater model is represented in CTM by what we call **Short Term Memory (STM)** (though it is not much of a memory at all). At any moment in time, STM contains CTM's current *conscious content*. STM is not a processor; it is merely a buffer and broadcasting station. STM can hold only *one* "chunk of information" at a time. (For the definition of *chunk*, see section 2.1.3.) One is not the "magical number" 7 ± 2, but we are looking for simplicity and a single chunk will do.[14]

2.1.2 LTM Processors

The audience, which we call **Long Term Memory (LTM)**, is represented by a massive collection of $\mathbf{N}$ (initially independent) *powerful* processors that comprise CTM's *principal computational machinery* and *long term memory*. $\mathbf{N}$ is a parameter: for our particular CTM, we choose $\mathbf{N = 2^{24}}$.[15] These processors are *random access machines* (*not* Turing machines).[16] We assume that each processor has

[13] As AI's continue to interact with the world, they also (increasing) incorporate the 4 E's.

[14] The "magical number" $\mathbf{7 \pm 2}$ was proposed by George Miller (Miller G. A., 1956) as the number of "chunks" that a human at any moment of time. can hold in their "short term memory". Nowadays, following (Baddeley & Hitch, 1974), that "short term memory" is called the phonological loop. In CTM, both the phonological loop and the visuo-spatial sketchpad are LTM processors.

[15] We have chosen the numbers in our particular CTM instance so that they are (roughly) consistent with the numbers in a model of the human brain. The number $\mathbf{N}$ of LTM processors is a parameter. Choosing $\mathbf{N= 2^{24}}$ for our particular instance is suggested by the $\mathbf{10^7}$ ($\mathbf{\sim 2^{23.25}}$) cortical columns in the human brain. Two times that number accounts for additional important processors. (In addition, $\mathbf{2^{24}}$ rather than $\mathbf{2^{23}}$ makes it easier for us humans to do computations in your head than $\mathbf{2^{23}}$.)

CTM's lifetime $\mathbf{T}$ is a parameter. At $\mathbf{10}$ ticks per second (the alpha frequency of the brain), and $\mathbf{T = 10^3 N \approx 10^{10}}$ ticks, CTM's lifetime is about $\mathbf{32}$ years. We get the same lifetime of $\mathbf{32}$ years if clock frequency is $\mathbf{100}$ ticks per second (the gamma frequency of the brain) and $\mathbf{T = 10^4 N \sim 10^{11}}$ ticks. At $\mathbf{T = N\sqrt{N} \approx 10^{10.5}}$ and $\mathbf{10}$ ticks per second, CTM's lifetime is about $\mathbf{100}$ years.

(See also Appendix 6.2.)

[16] A random access machines (RAM) can search a sorted list in logarithmic time; a Turing Machine needs linear time.

its own internal time clock for running its own algorithms, and that that time clock is **100** times faster than CTM's global time clock. (For more on built-in processor properties, see section 2.1.7.)

In CTM, all processors are in LTM so when we speak of a processor, we mean an LTM processor.[17] These processors are "unconscious". They compete to get their information in the form of *chunks* (section 2.1.3) on stage to be immediately broadcast to an attentive audience.[18]

2.1.2.1 The Model-of-the-World Processor (MotWp), The Model-of-the-World (MotW), and Brainish

We single out what we call the ***Model-of-the-World processor*** (**MotWp**), which is not actually a single processor, as its functionality is distributed across all LTM processors. It constructs models of CTM's inner and outer worlds, which it stores in these processors. We call this collection of constructed models the ***Model-of-the-World*** **(MotW)**. The MotW starts off as a "coarse blob"; over time it becomes more sharply defined and populated with many labeled sketches.

CTM's finite lifetime imposes a finite lifetime on LTM processors. However, for understanding consciousness, processors are better modeled as RAMs *running algorithms* rather than as finite state machines executing tuples.

[17] Hence, the processors that Baars puts outside of LTM, like the phonological loop, visuospatial sketchpad (Baddeley & Hitch, 1974), and episodic buffer (Baddeley, 2000) will be in CTM's LTM.

[18] In Baars' theater metaphor, consciousness is the activity of actors in a play performing on a stage of Working or Short Term Memory (STM). The Inner Speech actor is often on stage. Their performance is under observation by a huge audience of powerful unconscious processors in Long Term Memory (LTM) that are sitting in the dark. The unconscious processors vie amongst themselves to get their script/query/information on stage to be broadcast to the audience.

As an example of the theater metaphor, consider the **"What's her name?"** scenario:

Suppose at a party, we see someone we know but cannot recall her name. Greatly embarrassed, we rack our brain to remember. An hour later when driving home, her name pops into our head (unfortunately too late). What's going on?

Suppose we have a CTM brain. Racking our brain to remember caused the *urgent* request **"What's her name?"** coming from LTM processor **p** to rise to the stage (STM), which immediately broadcasts the question to the audience. Many (LTM) processors try to answer the query. One such processor recalls we met in a neuroscience class; this information gets to the stage and is broadcast, triggering another processor to recall that what's-her-name is interested in "consciousness", which is broadcast. Another processor **p'** sends information to the stage asserting that her name likely begins with **S**.

Sometime later the stage receives information from processor **p''** that her name more likely begins with **T**, which prompts processor **p'''** (which like all processors has been paying attention to all the broadcasted information) to claim with *great certainty* (and correctly) that her name is **Tina**. The name is broadcast from the stage, our audience of processors receives it, and we finally consciously remember her name. Our conscious self has no idea how or where her name was found.

Intertwined with building the MotW, the MotWp generates ***Brainish***, CTM's internal self-generated multimodal inner language.

While each processor has its own distinct language, the (community of LTM) processors communicate with each other (entirely within CTM) in Brainish. A ***Brainish gist***[19] is a "succinct phrase" consisting of a finite sequence of at most **20** Brainish words,[20] each word being a pair **<p, t>**, where **p** is a processor number and **t** is a time in clock ticks. A Brainish gist ***fuses*** modalities (e.g., sight, sounds, smells, sense of touch) and processes (e.g., thoughts like, "to prove this math statement, use induction"). (See section 2.3 and its subsections for formal definitions and development of these notions.)

A Brainish gist (phrase) is like a dream or movie frame. The Brainish language evolves over CTM's lifetime, from nothing initially into an ever growing dictionary of words. Clearly, Brainish will differ from one CTM to another. (See section 2.3.1 on how Brainish and the MotW co-evolve.)[21]

The MotWp and its MotW play an important role in planning, predicting, testing and learning, and in CTM's *feelings of consciousness* (section 2.3). When Brainish-labeled sketches are broadcast and *inspected* (Formal Definition 7 in section 2.3.2.2) they *evoke* CTM's subjective experiences (Axiom A in section 2.3.2.2).

2.1.3 Up-Tree Competition and Chunks

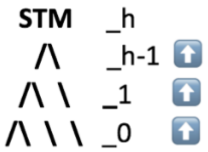
LTM processors *compete* in well-defined (*fast* natural) *probabilistic competitions* to get their questions, answers, and information in the form of *chunks* onto the stage (STM). The competitions

[19] We were unaware of Jeremy Wolfe's 1998 paper (Wolfe, 1998) until we read Fei-Fei Li's wonderful book *The Worlds I See* (Li, 2023), where she identifies Wolfe as the "gist" guy. Wolfe uses the word "gist" in reference to "what we remember – and what we forget – when we recall a scene." Our explanation of *change blindness* in CTM (Blum & Blum, 2022) is almost identical to Wolfe's explanation of change blindness in humans: the same gist describes both the original and changed scenes.

[20] We could require gists to be at most 2 Brainish words instead of 20, but for creating non-trivial examples, 20 is easier (and the maximum for quick computations, as in the Up-Tree competition).

[21] Paul Liang is developing a computational framework for Brainish based on multimodal machine learning (Liang, 2022). This is different than CTM's evolved formal Brainish (section 2.3.1).

are hosted by the **Up-Tree,** a *perfect*[22] binary tree of height **h** which has a leaf in each LTM processor and its root in STM.



At each clock tick, *a new competition starts* with each processor putting a chunk of information into its Up-Tree leaf node. (See Appendix 6.3) for a fuller discussion of the Up-Tree competition.)

A ***chunk*** is formally defined to be a tuple,

<address, time, gist, weight, auxiliary information>,

consisting of: the ***address***[23] of the originating processor; the ***time*** the chunk was created; a *succinct* Brainish ***gist*** (***phrase***) of information; a ***valenced*** (+/-) ***weight*** [to indicate the importance/ urgency/ value/ confidence/ sentiment the originating processor assigns the gist]; plus some ***auxiliary information***. (See Appendix, section 6.3 for a discussion of auxiliary information.)

Gists in chunks may convey information, pose questions, provide answers, request that specific *tasks be carried out* or *problems solved,* and so on. ....

Brainish ***words*** and Brainish ***gists*** (*phrases*) are formally defined in section 2.3.2.1. The primal Brainish gist (the empty sequence) is denoted by **NULL.**

The chunk's (***valenced***) ***weight*** is estimated by the processor's ***weight assigning algorithm*** (Blum, Blum, & Blum, 2026). We think of this algorithm as assessing the importance that the processor assigns the gist. That weight (that the processor assigns to a chunk) is adjusted by the processor's built-in ***Sleeping Experts*** *(learning)* ***Algorithm*** (section 2.1.8).

[22] A *perfect* binary tree is a binary tree in which all leaf nodes are at the same depth, that depth being the height **h** of the tree. The tree has $\mathbf{N} = \mathbf{2}^{\mathbf{h}}$ leaves, and every node except the root node has a unique sibling (neighbor). For simplicity, the CTM's competition tree is a perfect binary tree with $\mathbf{h = 24, N} = \mathbf{2^{24}}$.

[23]

Each submitted chunk competes locally with its neighbor (as in a tennis or chess tournament). In a single clock tick, the local winning chunk is chosen and provided it is not at the top moves up one level of the Up-Tree, ready to compete with its neighbor. At each clock tick, there are active local competitions at every node in every level from **0** to **h-1** of the **Up-Tree**. The competition that begins at time **t** ends at time **t+h** with the *winning chunk* in STM.[24]

Notably, for the *probabilistic* CTM competition (Appendix, section 6.3), it is proved that a chunk wins the competition with probability proportional to a function of its |weight|.[25] As a consequence of this theorem, the winning chunk is independent of the processor's location! Clearly, this is an important feature for a machine or brain in which no movement of processors is possible. (It is a property that would be difficult if not impossible to achieve in fair tennis or chess tournaments.)

2.1.4 Down-Tree Broadcast, Paying Conscious Attention, and Conscious Communication

The ***winning chunk,*** being the chunk that gets into STM, is called CTM's ***conscious content*** . Upon entry to STM at some time **t**, it is immediately ***globally broadcast*** to the audience of LTM processors, which receive it at time **t+1**. Broadcast is via the **Down-Tree**, a bush of height **1** with root in STM and **N** leaves, one leaf in each of the **N** LTM processors. /|\ ⬇ [KM5]

The global broadcast of a single chunk at each tick to all LTM processors enables CTM to "focus attention" on the winning gist (chunk). One is not the "magical number" **7±2**, but we are looking for simplicity and a single chunk will do.

[24] If the interval between successive clock ticks is **.1** seconds and the number of LTM processors is **2^{24}**, then a full binary competition takes about **2.4** seconds. Define an **octal competition tree** to be a tree with 8 children per node in which **one** level of octal tree does **three** levels of a binary tree in a single tick, **.1** seconds, rather than **.3** seconds. Then an octal tree competition takes **.8** seconds, which better mirrors what goes on in a brain.

[25] One can think of the chunk's |weight| as being analogous to a player's ranking in tennis or chess. It would not do to pick the winner to be the chunk with the highest |weight|, for then no other chunk, even one with a slightly less |weight|, would have a chance. The probabilistic competition gives all competitors a fair chance.

We say that CTM ***pays conscious attention*** to CTM's *conscious content* when this chunk is simultaneously received by all LTM processors.[26] (This is Formal Definition 1 of ***conscious attention*** in section 2.2.)

We call communication between LTM processors that goes through STM ***conscious communication***.

2.1.5 Links and Unconscious Communication

The CTM has no links at birth. A 2-way ***link*** forms between processors A and B when A has broadcast a request, B has answered, and A acknowledges that B's answer is helpful.[27, 28] Such links enable *conscious communication*, i.e. communication that goes through STM, to be replaced by more direct and faster ***unconscious communication*** (between processors) through links. The CTM has no provision for deleting links, despite that deletions are known to be important in human learning. No provision for deleting links enables the model to be simpler, and processors can still do the equivalent of deleting links by simply ignoring them. Adding links enables sending and receiving more information per tick.

Thus, when CTM initially learns to ride a bike, most communication is done consciously until relevant processor links have formed. Then, for the most part, riding a bike is done unconsciously until an obstacle is encountered, forcing CTM to *pay conscious attention* if its "unconscious instincts" fail to kick in.[29] [KM5]

[26] Alison Gopnik's contention that "babies are more conscious than we are" (Gopnik, 2007) can be understood in terms of *conscious attention*. In the infant CTM, until links are formed (section 2.1.5), all communication between processors is conscious, i.e., goes through STM, and is therefore consciously attended to. On the other hand, the MotW in the infant CTM is considerably less developed than in the adult. Hence phenomenal consciousness (what we call *conscious awareness*, section 2.3) is considerably less developed in the infant CTM than in the adult.

[27] The two-way link is actually two one-way links: A to B and B to A.

[28] In our earlier **"What's her name?"** scenario (footnote in section 2.1.2), when processor **p** sees that processors **p'**, **p**'' and **p'''** have useful information for it, and **p** acknowledges their usefulness, **p** forms bi-directional links with **p'**, **p''** and **p'''**. This is akin to the Hebbian property, "when neurons fire together, they wire together".

[29] Humans learn to play ping pong consciously. In a ping pong tournament however, a player must let the unconscious take over, must insist that the conscious get out of the way (see TableTennisCoaching.com). In swimming, repetition gives one's unconscious an ṗ opportunity to improve one's stroke, but it doesn't enable a new stroke to be acquired. That requires conscious attention. For example,

2.1.6 Input/Output

The CTM is not a brain in isolation. ***Input/ Sensory processors*** take **input** from CTM's outer world via ***sensors***[30], convert that **raw input** into Brainish gists, those gists into chunks, and then submit those chunks to the competition (sections 2.1.3 and 6.2). ***Output*** **/Motor** ***processors*** convert **output** chunks into instructions for motor ***actuators***[31] to act on CTM's outer world.

2.1.7 Built-in Processor Properties

Now that we have established the basic CTM structure, we outline some important properties built into all processor at CTM's birth (**t = 0)**. These properties include *triggers* for what processors *must do* and come into play in *triggering* conscious awareness (section 2.3).

1. Each and every processor ***stores in its memory*** every chunk *it submits* to the competition and every chunk *it receives* by broadcast, forming two distinct lists ordered by time. We call these lists the processor's ***local dictionary*** and its ***global dictionary***, respectively.[32]

2. ***Input*** (e.g., sensory) ***processors***, ***Output*** (e.g., motor) ***processors***, and ***Gauge processors*** (homeostats that monitor measurable quantities like time, strength, ability, and so on) each have ***specialized functions built in***.

3. All processors (including the above specialized ones) have ***built-in learning algorithms*** that work (unconsciously) to predict, check, correct and improve that processor's prediction algorithms (section 2.1.8), to set weights, to make chunks, to communicate with other processors, and so on. Importantly, a *Sleeping Experts Learning Algorithm* (necessarily modified for CTM) adjusts the

the dolphin kick is weird and unnatural, but since it works for dolphins, it makes sense to simulate it, and that is done consciously at first. The unconscious then optimizes the constants.

[30] Ears/sounds, eyes/sight, nose/smell, skin/touch, mouth/taste, … .

[31] Arms, hands, legs/motor actuators, mouth/vocal actuators, … .

[32] That's two chunks per tick. The decision was made to not store every chunk transmitted along links (section 2.1.5) because a processor can in general have almost **N** links, too much to store in a single tick or even a few ticks. A consequence of this decision is that if CTM is asked at some point why it did something it did, it can answer based on what it knows from broadcasted chunks, from submitted chunks, but not from chunks sent along links, which thwarts its ability to answer in full.

weight each processor gives its gists.[33] See (Blum A. , 1995) and (Blum, Hopcroft, & Kannan, 2015). [KM6]

4. Each and every processor has a ***built-in preference for choosing greater (more positive) weights*** over lesser (more negative) weights.[34] By contrast, the Up-Tree competition has a preference for choosing greater intensities. The difference is one of weights versus intensities. See section 6.3.

5. For each processor, at any moment of time, some of its algorithms are critical and must be executed. The time left over is its ***free time***. On receipt of a broadcasted chunk **<p, t, g, w, ... >,** every processor is ***programmed to focus*** a fraction of its *free time* – a fraction proportional in part[35] to **|w|** - to ***deal with*** the gist **g.** If **w < 0**, then the processor ***must spend*** that time on increasing **w.**[36]

   To ***deal*** with the gist **g** means to try to understand the gist, to carry out the task or solve the problem posed, and so forth. In order to understand the gist, processors ***may inspect*** the chunk (Formal Definition 7 in section 2.3.2.2).

6. Each processor has a built-in ***threshold*** parameter **θ.** On receipt of a broadcasted chunk **<p, t, g, w, ... >**, if **|w| > θ**, then the processor ***must inspect*** the broadcasted chunk. If in addition, **w < 0**, then the processor ***must spend all*** its time on increasing **w.** (This may force the processor to abandon some of its essential duties.)

[33] Referring again to the **"What's her name?"** scenario (footnote in section 2.1.2), recall that processor **p**' got its incorrect information quickly into STM while it took processor **p"** longer to get its correct information in. The *Sleeping Experts Algorithm* in **p'** caused **p'** to lower the importance it gives its information, while the *Sleeping Experts Algorithm* in **p"** caused **p"** to increase the importance it gives its information.

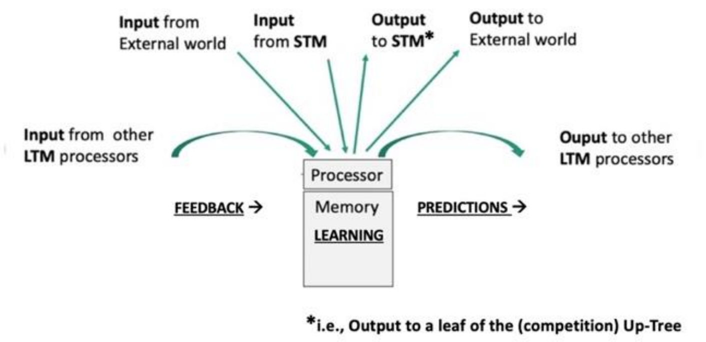
[34] This built-in *preference* in the processors (not in the competition function) for choosing positive over negative contributes to the drive for survival.

[35] "In part" depends on the priorities of the processor **p**.

[36] Note, we have not stipulated specifically what to do if a broadcasted chunk has high positive **w.** This is purposeful. As a consequence, except for other constraints (e.g., responding to tasks), positivity allows processors to use more of their free time as they choose. This is an example where positivity and negativity are not symmetric in the CTM.

2.1.8 Predictive Dynamics (Prediction + Testing + Feedback + Correction/Learning)

A major goal for CTM is to ensure that all predictions are as accurate as possible. For the most part this is done unconsciously (locally) within the individual processors themselves. Processors make predictions when they turn raw sensory inputs into chunks (a kind of analog to digital conversion[37]), when they put chunks in the competition, when they send information over links, and when they transform chunks into raw outputs (a kind of digital to analog conversion). Processors get feedback from broadcasts, from inputs to CTM, and through links. Learning/correcting then takes place within processors.


Input from External world
Input from STM
Output to STM*
Output to External world
Input from other LTM processors
Ouput to other LTM processors
Processor
Memory
LEARNING
FEEDBACK →
PREDICTIONS →
*i.e., Output to a leaf of the (competition) Up-Tree

Thus, we can already see some basic ***predictive dynamics*** (cycles of prediction → testing → feedback → correction/learning) occurring locally and globally within CTM.

2.2 Conscious Attention in CTM

> **Formal Definition 1.** We say CTM ***pays conscious attention*** to CTM's ***conscious content*** when this chunk is simultaneously received by all LTM processors.

Thus, by definition, *conscious attention* in CTM results from the simultaneous *reception* by all processors of CTM's *conscious content*. [KM2] [KM5][KM8].

[37] Digital signals are binary and have a few defined states, while analog signals have a theoretically infinite number of states.

CTM's *conscious content* (in STM) at time **t+h** is the winner of the competition that commenced at time **t**. CTM pays *conscious attention* to this chunk at time **t+h+1**.[38]

We call a finite sequence of successively broadcasted chunks a ***stream of consciousness***.

Globally, the process of competition for STM, global broadcast to LTM, and consequent processor responses and direct communication via links is reminiscent of a process that Stanislas Dehaene and Jean-Pierre Changeux have called *ignition* (Dehaene & Changeux, 2005).

Our definition of *conscious attention* is related to what is often called *access consciousness* (Block, 1995).

But... for *subjective feelings of consciousness*, attention is not all you need. More is required.

2.3 Conscious Awareness and "What it is like to be" a CTM[39]

How might CTM, a formal machine model, experience feelings such as *pleasure* and *pain*? How might CTM experience the *subjective feelings of consciousness*?

In this section we look at how the *Model-of-the-World* (MotW) and *Brainish*, CTM's self-generated multimodal inner language, co-evolve to play an essential role in *"what it is like to be"* a CTM.

We begin with some preliminary concepts, informally, indicating (in section 2.3.2) how we formalize these concepts in CTM.

Recall (section 2.1.2.1), the ***Model-of-the-World processor*** (**MotWp**) is not actually a single processor: its functionality and memory are distributed across all LTM processors. The MotWp builds models of

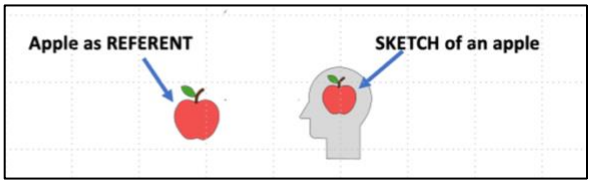
[38] There is a delay of **h** clock ticks between when LTM processors submit their chunks to the competition for STM and when the winner appears in STM (as CTM's 'conscious content').

In one more clock tick, CTM pays conscious attention to the winner. This **h+1** delay is somewhat analogous to behavioral and brain studies going back to (Libet, 1985) that suggest a delay between when unconscious processors in our brains make decisions and when we become conscious of them. The CTM competition also suggests, however, that while some processor made the decision that will be broadcast, others may have made different decisions. So, in a real sense, the decision in CTM may *not* have been fully made.

[39] Thomas Nagel's "what it is like to be" (Nagel T. , 1974) is often taken to be the canonical "definition" of phenomenal or subjective consciousness.

CTM's inner and outer worlds which also are distributed across all LTM processors. We call this *collection* of models, the ***Model-of-the-World*** (**MotW**). [KM1][KM2][KM8]

The MotW is CTM's current and continuing view of CTM's (inner and outer) worlds. Specifically, the MotWp constructs ***sketches*** in the MotW, chunks with *succinct* ***Brainish*** *descriptions* (gists) of ***referents*** in CTM's (inner and outer) worlds. [KM1]



A ***referent*** might be a (red) apple or (red) rose in CTM's outer world, or a feeling (of pleasure), or a thought, idea, desire, concern, and so on in its inner world. Sketches of inner world referents are as important as outer world referents.[40]

Sketches may be ***labeled*** with ***Brainish words*** such as (translated into English): BRIGHT_COLOR ○ SWEET_TASTE ○ CRUNCHY or BRIGHT_COLOR ○ SWEET_SMELL ○ SILKY_TOUCH or HAPPY_THOUGHT. The concatenator ○ indicates the *fusion* of previously defined Brainish words.

These *labels* are Brainish words for what CTM "learns", "senses" or "feels" about those referents based on "lived" experiences. They *symbolize/represent/denote* ***qualia***, which we define to be the *subjective experiences evoked* when the chunks pointed to are *inspected*. (See section 2.3.2.2.)

Over time and with experience, Brainish words, labels and sketches (Formal Definitions 4 and 5 in section 2.3.2.1) get introduced and become finer, richer and more nuanced.

In particular, the sketch "CTM" in the MotW, whose referent is CTM itself,[41] will develop from scratch and eventually be labeled with SELF ○ CONSCIOUS, etc. The sketch starts as an empty blob

[40] A human dream frame typically has both outer and inner world sketches. For example, feeling sketches are particularly evident in nightmares, when the objects are monsters under the bed and the feeling is one of great fear.

[41] We are often asked, isn't this process recursive? Doesn't the sketch of CTM have a sketch of CTM have a sketch of CTM, etc. ? Yes, up to a point. But, at each iteration the current sketch is degraded, so the process rapidly ends up with the empty sketch.

representing both the newborn CTM and the “world”. In time the “CTM” sketch will become an entity distinct from the rest of the “world” (see sections 2.3.1.2 and 2.3.1.3).

CTM will pay ***conscious attention*** to MotW sketches and labels when chunks referring to these *labeled sketches* win the competition for STM and are received by all processors. If in addition, all processors ***inspect*** (Formal Definition 7 in section 2.3.2.2) the received chunks, CTM will become ***consciously aware*** of these sketches and labels (see Formal Definition 10 in section 2.3.2.2) and they will ***evoke*** CTM’s subjective experiences (see Axiom A in section 2.3.2.2). [KM3][KM11][KM15]

In this way, **MotW becomes the world that CTM *consciously* “sees”, or more generally *consciously* “senses” or “knows”.** [KM1] [KM3] [KM15]

2.3.1 Co-Evolution of Brainish and the Model-of-the-World (MotW)

The Model-of-the-World (MotW) and Brainish co-evolve during CTM’s lifetime, starting at birth (**t = 0**). To motivate our formalization (in section 2.3.2) of this co-evolution, we start here with some examples.

The MotW in the newborn CTM (which could be a human infant if it has a CTM brain) contains just an empty blob at first representing both the newborn CTM and the “world”. This blob acquires CTM’s first Brainish name as follows:

Suppose the chunk **<p_0, 0, NULL, w_0, …>** is CTM’s very first broadcasted chunk. Here **NULL** represents the empty sequence which we consider to be a Brainish word. The pair **<p_0, 0>** is a pointer to that chunk, **<p_0, 0, NULL, w_0, …>. *Brainish words*** will in general be such pointers.[42]

> **Formal Definition 2.** CTM’s very first broadcasted chunk **<p_0, 0, NULL, w_0, …>** is its ***initial sketch*** in the MotW of both the “world” and the newborn CTM. The pointer **<p_0, 0>** to the sketch **<p_0, 0, NULL, w_0, …>** is CTM’s ***primal Brainish word***, and it’s the ***name*** of that initial sketch. The English translation of the Brainish word **<p_0, 0>** is **BLOB$_0$**.

[42] In the AI/ML/neuroscience literature, Brainish *words* are reminiscent of what are called *keys* and the *chunks*, that the names point to, are reminiscent of what are called *values* (Gershman, Fiete, & Irie, 2025).

In general, *sketches* are broadcasted chunks **<p, t, gist, w, …>** and *names of sketches* are pointers **<p, t>** to these chunks. (See Formal Definition 5 in section 2.3.2.1.) It is sometimes convenient to identify the *name of the sketch* with the *sketch* itself.[43]

Since processors store all broadcasted chunks (built-in property 1 in section 2.1.7) any LTM processor at any future time can ***decode*** the word **<p, t>**, that is, look into its memory and retrieve the stored chunk, **<p, t, gist, w, …>**.[44]

We view the procedure in processor $\mathbf{p_0}$ that generated the chunk **<**$\mathbf{p_0}$**, 0, NULL,** $\mathbf{w_0}$**, … >** that got broadcasted to be the part of the MotWp in $\mathbf{p_0}$.[45] We view this stored chunk (sketch) as part of the MotW.

Brainish words for pain and pleasure also develop early. We indicate here how these words come to be in CTM. We start with pain.

2.3.1.1 Brainish Word for Primal PAIN

Consider again an infant CTM at the moment it is born. A processor that monitors the O_2 level, call it the **O_2_Gauge**[46], raises its growing concern for the lack of O_2 by submitting to the Up-Tree competition a sequence of chunks having ***negatively valenced weights*** of ***increasingly high absolute value***. At some point, the O_2_Gauge processor's chunks get onto the stage (STM) and are broadcasted, their huge and growing negative weight signaling a increasingly desperate "scream" for something (O_2). All processors hear this "scream".

[43] "You are sad," the Knight said in an anxious tone: "let me sing you a song to comfort you." … "The name of the song is called 'Haddocks' Eyes.'" "Oh, that's the name of the song, is it?" Alice said, trying to feel interested. "No, you don't understand," the Knight said, looking a little vexed. "That's what the name is called. The name really is 'The Aged Aged Man.'" "Then I ought to have said 'That's what the song is called'?" Alice corrected herself. "No, you oughtn't: that's quite another thing! The song is called 'Ways And Means': but that's only what it's called, you know!" "Well, what is the song, then? " said Alice, who was by this time completely bewildered. "I was coming to that," the Knight said. "The song really is 'A-sitting On A Gate': and the tune's my own invention." (Carroll, 1871)

[44] Retrieving the stored chunk can be done quickly. That's because every processor stores every broadcasted chunk (built-in property 1), and the processor is a random access machine. A Turing machine could not do the lookup quickly enough.

[45] Recall, CTM's MotWp is distributed across all LTM processors.

[46] "O_2_Gauge" is a metaphorical name for a CTM gauge processor that, like all CTM gauges, has a built-in program to announce a great need. (See built-in properties in section 2.1.7.)

Since every processor stores every chunk that ever got broadcast (built-in property 1 in section 2.1.7), every processor stores the O_2_Gauge processor's broadcasted chunks, **<p_1, t_1, $gist_1$, w_1, ...>**. Here **p_1** is the O_2_Gauge processor's address, **t_1** is (any) one of the times at which the O_2_Gauge processor screams for O_2, and **$gist_1$** has a high (and growing) negative weight, **w_1**. The high (and growing negative weight causes each processor to focus a fraction of its free time (proportional to **$|w_1|$**) to increase **w_1**, (make it less negative) something all processors will have been programmed to do (built-in properties 5 and 6 in section 2.1.7).

The processors have absolutely no idea what to do (as the infant was just born) to reduce that huge and growing |weight|. They know only that they must do something. Processors that control CTM's actuators command them to do something, anything. The arms and legs flail. The infant pees and poos. The vocal actuator screams and cries. This last one works! The cry opens the "lungs" and the infant takes in its first breath. The O_2_Gauge processor's weight gradually returns to normal.

The next time the infant CTM needs help, it finds that screaming and crying work again. In short order, the infant CTM learns that a good response to negativity (particularly intense negativity) of any kind is to scream and cry.

Recall that a chunk **<p_1, t_1, $gist_1$, w_1, ...>** was a scream for help from the O_2_Gauge processor. Let **<p_1, t_1>** be a pointer to this chunk.

> **Formal Definition 3.** If **$gist_1$** is **NULL**, the Brainish word **<p_1, t_1>** is a ***primal Brainish word for pain***. Its English translation is **$PAIN_1$** (or PAIN_FROM_THE_LACK_OF_O_2).

Different Brainish words for pain will evolve over time, later pains building on earlier ones.

Now that CTM has these two Brainish words, **<p_0, 0>** (i.e., **$BLOB_0$**) and **<p_1, t_i> ($PAIN_1$)**, more complex words and phrases can be generated. For example, after the creation of the initial primal words, some processor **p_2** may submit a chunk **<p_2, t_2, g_2, w_2, ... >** into the Up-Tree competition, where the gist **g_2** is the ***Brainish phrase***, **<p_0, 0><p_1, t_1>**. This Brainish phrase translates to the English phrase "**$BLOB_0$ is in $PAIN_1$"**. If **$|w_2|$** is large enough, the chunk may win the competition. Then the Brainish word **<p_2, t_2>**, a pointer to **<p_2, t_2, g_2, w_2, ... >,** will be a compound Brainish word.

We can also consider the word **<p_2, t_2>** as the *name* of the ***Brainish-labeled sketch***

$<p_2, t_2, <p_0, 0><p_1, t_1>, w_2, \ldots >$, where **$<p_0, 0>$ (BLOB$_0$)** is the ***name*** of the initial *sketch* and **$<p_1, t_1>$** (PAIN$_1$) is its ***label***.[47] (See Formal Definition 5 of *Brainish-labeled sketches* in section 2.3.2.1.)

The above indicates how the Brainish language starts to develop.

Other primal Brainish words quickly develop. For example, suppose **p_3** was the ***Vocal processor*** whose direct signals to the ***vocal actuator*** caused the newborn **CTM** to cry. When the processor learns that its signal produced the desired result, it submits a chunk **$<p_3, t_3, gist_3, w_3, ,,,>$** with **$gist_3$ = NULL** announcing this discovery to the competition with high enough **$|w_3|$** so that its chunk gets broadcast.

Then **$<p_3, t_3>$** becomes a primal Brainish word for **CRY**.

2.3.1.2 Primal HUNGER, PLEASURE and SELF

Now suppose the infant CTM's *Fuel Gauge* is low. The *Fuel Gauge processor* creates a chunk with negative weight proportional to the need. When the weight is sufficiently negative, this chunk will (with high probability) win the competition and get globally broadcast. Again, the processors are "compelled", i.e. programmed, to do or try to do something to lower the |weight|.

Having learned that crying is a good response to negativity, the CTM cries out causing the *Fuel Source* (mother) in the outside world to respond. As the fuel comes in, the Fuel Gauge will start submitting chunks with more positive weights (... → -4 → -2 → ...).

Let **$<p_4, t_4, gist_4, w_4, \ldots>$** be a chunk that, on account of its extreme negative weight **w_4**, is a scream to do something (to make **w_4** more positive). Here **p_4** is the Fuel Gauge processor's address, **t_4** is one of the times at which the chunk was broadcast, and **$gist_4$** could be anything, say **NULL**.

The *pointer* **$<p_4, t_4>$** to **$<p_4, t_4, gist_4, w_4, \ldots>$** is now a Brainish word that translates (we suggest) to the English **HUNGER_PAIN**. The Fuel Source sketch (of mother) in the MotW will eventually get labeled (we suggest) **RELIEVES_HUNGER_PAIN** and **PLEASURE_PROVIDER**.

[47] As in Creole and in many languages, we place the adjective after the noun.

In this way, the newborn CTM learns (or increases its confidence) that **PLEASURE** relieves **PAIN**.[48]

At this stage in CTM's early life, there is only one blob, named **BLOB$_0$** in the MotW which represents the newborn CTM and the "world". With the introduction of the Fuel Source (mother), the blob will also represent mother. But soon enough, the infant CTM will discover that it and the Fuel Source are not one. The MotWp then separates the sketch named **BLOB$_0$** into a sketch named **CTM$_0$** and labeled **SELF$_0$**, and a sketch named **FUEL-SOURCE**. (This is done by creating relevant chunks that get broadcast. See Formal Definition 5 in section 2.3.2.1.) [KM2]

2.3.1.3 The Power of Thought and Sense of Self

The infant CTM's primal sense of self (**SELF$_0$**) arose when it discovered that it was not the only entity in the world. In time, the infant CTM will discover that it can utilize the ***power of thought*** **(PoT)** to check what is **SELF** and what is not. [KM2].

To perform an action by the ***power of thought*** (or mental causation) is to perform the action in the MotW, then confirm that it actually got done in the world.

For example, at some point in time, the MotW will contain a rough sketch of the infant's left leg. When the infant CTM discovers it can move its left leg (an actuator) by the *power of thought*, the MotWp appends the label **SELF** to that sketch of the left leg:

> In this case, the MotWp moves the left leg sketch in its MotW, which commands (via the Motor processor) the left leg to move. The MotWp detects that movement via CTM's sensors. *By repeating the action a number of times*, the infant CTM becomes ever more certain that it itself is responsible for moving the leg (that is, that this prediction is correct). Now convinced, the

[48] When a human mother gives a breast to her infant, the infant learns that the breast relieves the pain of hunger. The breast gets incorporated into the infant's MotW labeled with **RELIEVES_PAIN ∘ GIVES_PLEASURE**. Unless brought up in a psychotic household, the human infant learns that pleasure relieves pain and prefers pleasure over pain. This is one of many ways in which pain and pleasure are not symmetrical. This dynamic is reminiscent of the pleasure cycle proposed by (Berridge & Kringelbach, 2015) consisting of three separate entities: wanting, liking, and learning. [KM11]

> MotWp labels the left leg sketch with **SELF.** This Brainish label represents CTM's discovery that it can move its leg by the power of thought.[49] [KM13]

This explanation of the *power of thought* in labeling a limb as **SELF,** requires some qualification. In similar fashion, the newborn infant might easily surmise by the power of thought that its mother is SELF. After all, every time the newborn infant cries, its mother is activated. And so, in the same manner, the mother blob in the MotW is labeled **SELF.** But soon enough, the infant will discover the mother does not always respond, and the **SELF** label on mother will be removed.

2.3.2 Formalizing Brainish, Labeled Sketches and Conscious Awareness

The primal Brainish words for pain and hunger, and early Brainish gists such as **$\mathbf{BLOB_0}$ is in $\mathbf{PAIN_1}$,** provides clues on how to *formally* define *Brainish words*, *gists*, and (*multimodal*) *fusion*.

2.3.2.1 Brainish Words and Labeled Sketches

Formal Definition 4.

i. A ***primal Brainish word*** is a pointer **<p, t>** to a *broadcasted chunk* **<p, t, NULL, …>** whose gist is **NULL** (the name for the empty sequence).
ii. A ***Brainish word*** is a pointer **<p, t>** to a *broadcasted chunk* **<p, t, gist, …>,** or **NULL.**
iii. A ***Brainish gist*** (***phrase***) is a finite sequence of at most **20** *Brainish words*.[50] (So *words* are *gists*.)

[49] Certain pathologies will occur if a breakdown in CTM causes its MotWp to mislabel a sketch. For example, if the sketch of a leg gets mislabeled NOT-SELF, CTM might beg to get its leg amputated, even though the leg still functions properly. This would be an example of body integrity dysphoria (body integrity identity disorder) in CTM. Other pathologies due to faulty labeling in the MotW include: phantom limb syndrome (the sketch of an arm remains labeled SELF after it is amputated), Cotard's syndrome (the sketch of CTM is labeled DEAD), paranoia (the sketch of CTM's best friend is labeled SPY), … . (There may be various causes for breakdowns in CTM, e.g., faulty sensors or sensory processors, faulty gauges, faults in the Up-Tree competition and/or broadcasting system, and so on.) [KM9]

[50] For Brainish words, we use **< p, t>** instead of the full chunk **< p, t, gist, w, aux>** because **< p, t>** is sufficiently short to appear 20 times in a gist while **< p, t, gist, w, aux>** is not short enough to appear even once. The reception of **< p, t>** by a processor enables that processor to do a quick look up in its personal memory for the associated **chunk**. Quick look ups are possible because the **CTM** processors are RAMS (Random Access Machines), not Turing Machines.

As commented earlier, we could require gists to be at most 2 Brainish words instead of 20, but for creating non-trivial examples, 20 is helpful (and the maximum for quick computations, as in the Up-Tree competition).

iv. If the *Brainish word* **<p, t>** points to **<p, t, gist, ...>,** we say **<p, t>** is a ***fusion*** of the words in the **gist.**[51, 52]

Formal Definition 5.

i. MotW ***sketches*** are broadcasted chunks that were submitted by the MotWp; ***names*** (of sketches) are words, i.e., pointers to (these) broadcasted chunks. ***Labels*** (of sketches) are also words, i.e., pointers to broadcasted chunks. Suppose **σ** is the *name of a sketch* (such as **<p, t>** or **NULL**) and **λ** is a *label* (such as **<p', t'>** or **NULL**). The Brainish two-word gist **σλ** will denote the English phrase, **"The sketch named σ is labeled λ"** or more succinctly **"σ is labeled λ".**

ii. The broadcasted chunk **<p, t, σλ, ... >** is the labeled ***sketch*** in the MotW, and **<p, t>** is the ***Brainish name*** of that labeled sketch.[53]

Sketches and labels get refined over time.

2.3.2.2 Axiom A, Conscious Awareness, and the Feeling of Consciousness

Formal Definition 6. For a processor to ***decode*** a word **<p', t'>** means for it to ***retrieve*** the chunk **<p', t', g', w', ... >** that **<p', t'>** points to.[54]

Formal Definition 7.

(i) We say a *processor* ***inspects*** a chunk **<p, t, g, w, ... >** whenever it takes note of the chunk's weight **w** and ***decodes*** all words **<p', t'>** appearing in gist **g.**
(ii) We say CTM ***inspects*** a broadcasted chunk **<p, t, g, w, ... >** whenever *all* processors simultaneously (upon its reception) inspect that chunk. These are *level one inspections*.

[51] As Brainish words evolve, CTM's Brainish language will likely develop a simple basic (creole-like) grammar. For information on such (simple creole-like) languages see: (McWhorter, 1998), (McWhorter, 2008), (Sigal, 2022) and (Bancu, et al., 2024).

[52] Trivial observation: If **<p, t, gist, ...>** is a broadcasted chunk and **<p', t'>** is a word in its **gist,** then **t'< t**.

[53] When we say, the MotWp creates a *labeled sketch*, we mean that such a broadcasted chunk originated from the MotWp.

[54] Assuming that CTM can *decode* a word in **.1** milliseconds, 20 words in at most **2** milliseconds, then it can *inspect* a chunk in **3** milliseconds.

Inspection by CTM of a broadcasted chunk $\mathbf{<p, t, g, w, \ldots >}$ requires all processors to do that inspection simultaneously (in the same tick). A *deeper inspection* of $\mathbf{<p, t, g, w, \ldots >}$ does not require simultaneity. It can be done by a single processor involving its personal *priorities* and $\mathbf{w}$. But see section 2.3.2.3 on ***levels of inspection***.

The chunk's *weight* $\mathbf{w}$ signals the *importance* and valence that the processor gives the gist. By built-in properties 5 and 6 (section 2.1.7), if $\mathbf{<p, t, g, w, \ldots >}$ is a broadcasted chunk and $\mathbf{|w| > \theta}$, then all processors must inspect the chunk and spend a fraction of their time *dealing with* gist $\mathbf{g}$. If in addition, $\mathbf{w < 0}$, then all processors must also spend their time on increasing $\mathbf{w}$.

We now state **CTM Axiom A** for ***subjective experience*** in CTM.

Axiom A. ***Inspecting*** a broadcasted chunk $\mathbf{<p, t, g, w, \ldots >}$ ***evokes*** in CTM a ***subjective experience.***

i. The *intensity* and *valence* of the *evoked experience* depends on $\mathbf{w}$.
ii. If gist $\mathbf{g}$ is **NULL**, a ***subjective experience is evoked***, and that experience, whatever it might be, is a ***primal subjective experience***.[55]
iii. If gist $\mathbf{g}$ contains Brainish words $\mathbf{<p_1, t_1>, \ldots, <p_k, t_k>}$, each $\mathbf{<p_i, t_i, >}$ pointing to a chunk that *evoked* a *subjective experience,* then $\mathbf{<p, t, g, w, \ldots >}$ ***evokes* a *subjective experience*** that is a ***fusion*** of the ***subjective experiences evoked*** by all the $\mathbf{< p_i, t_i, g_i, w_i, \ldots >}$.

Formal Definition 8. ***Conscious awareness*** in CTM arises when CTM pays ***conscious attention*** to a chunk $\mathbf{<p, t, g, w, \ldots >}$ and ***inspects*** it.

We call a sequence of receptions of broadcasted inspected chunks, a ***stream of conscious awareness***. Gists in these chunks are like the frames of a movie or dream.[56]

[55] The bottom-most turtle.

[56] CTM *dreams* are streams of conscious awareness generated when input sensors and output actuators are inactive. Although dreams are "felt" as real, they can also be fantastical since, for one, their predictions are not being tested in the world. We propose that a test for *subjective* consciousness may involve testing for *dreaming*.

Axiom A implies that ***conscious awareness*** of chunk **<p, t, g, w, ...>** ***evokes a subjective experience*** in CTM, a ***fusion*** of the subjective experiences ***evoked*** by the chunks **<** $\mathbf{p_i, t_i, g_i, w_i}$**, ... >** that are pointed to by the **words <** $\mathbf{p_i, t_i}$**,>** in **gist g.**

Similarly, if CTM becomes ***consciously aware*** of a **chunk <p, t, σλ, w, ...>**, where **σλ** is a *Brainish-labeled* **(λ)** *sketch* **(σ)**, then by **Axiom A**, the ***subjective experience evoked*** **by this chunk** is a ***fusion*** of the experiences *evoked* by the **sketch** named **σ** and its **label λ**[57].

Thus, for example, when CTM becomes ***consciously aware*** of a chunk describing a MotW sketch of the referent rose with fused label BRIGHT_RED ○ SWEET_SMELL ○ SILKY_TOUCH, it "sees" the bright red color, "smells" the sweet odor, and "feels" the silky petals of the rose.

As CTM becomes *consciously aware* of more and more broadcasted chunks, some processor may ***call attention*** to this situation. Specifically, some processor may submit a chunk to the Up-Tree competition asserting that CTM is becoming ever more consciously aware. If this chunk wins the competition and hence gets globally broadcast, we say the CTM has perceived – or has become ***perceptive*** *of – its conscious awareness*. If in addition, this broadcasted chunk is *inspected* by all processors, we say the CTM becomes ***consciously aware*** *of its conscious awareness*. This is a high level of conscious awareness in CTM. (See also section 2.3.2.3.)

The pair **<p, t>** pointing to the broadcasted chunk that calls attention to CTM's "conscious awareness" becomes a Brainish word for CONSCIOUSLY_AWARE. The MotWp labels the sketch of CTM in the MotW with **<p, t>** to indicate CTM is CONSCIOUSLY_AWARE or simply CONSCIOUS.

Now suppose the sketch of CTM has also been labeled SELF. When CTM *inspects* a chunk with the labeled sketch <CTM><SELF ○ CONSCIOUS>,[58] CTM *will experience* (according to **Axiom A**) *a fusion* of the experiences *evoked* by a this labeled sketch. It will *feel* itself conscious. KM3][KM9]

[57] When we say evoked by the ***label*** λ we mean, evoked by the chunk the ***word*** λ points to.

[58] That is, all processors simultaneously receive the broadcasted labeled chunk with gist **<CTM><SELF ○ CONSCIOUS>** and simultaneously inspect this gist.

2.3.2.3 Levels of Inspection

When a broadcasted chunk is received (conscious attention), it may or may not be inspected (conscious awareness), and that inspection, if any, will be to a certain *level*. Recall that ***inspection*** of a broadcasted chunk **<p, t, gist, ... >** has been defined to be "*all* processors simultaneously (upon its reception) take note of the chunk's weight **w** and ***decode*** all words **<p', t'>** appearing in gist **g**." That is a first level of inspection. Any chunk can be inspected to a deeper level by any subset of processors (doesn't have to be all processors). The *deepest possible level of inspection* by a processor of a chunk **<p, t, gist, ... >** is when every descendant of every word in that chunk's gist is decoded down to the word **NULL**.

2.3.2.4 The *Experience* of Pain

Pain is a particularly good exemplar of the Hard Problem as it is hard to imagine that a robot can truly *suffer* from pain, not just *simulate* suffering. Here we look into how CTM *experiences* the feeling of pain.

It follows from **Axiom A** that becoming *consciously aware* of a chunk **<p, t, g, w, ... >** whose gist contains a Brainish word for pain, **<p', t'>,** will *evoke* a fused experience that includes the feeling of pain **<p', t'>** previously experienced by CTM.

Different kinds of pain evolve over time. Later pains build on earlier ones. For example, when the infant CTM becomes a toddler and first skins its knee, its MotW may contain a sketch of a bloody knee labeled SELF and PAIN built up from earlier Brainish words for pain. As soon as the toddler CTM becomes consciously aware of this labeled sketch, it *experiences* a fused feeling of earlier pains (Axiom A). A more nuanced Brainish word is created: SKINNED-KNEE_PAIN.

When a sketch of a bloody skinned knee is labeled SELF but not PAIN, or PAIN but not SELF, CTM does not consciously experience it. That is, in the case of SELF but not PAIN, CTM does not become consciously aware that itSELF has pain; in the case of PAIN but not SELF, the pain does not belong to itSELF. In either case, we say CTM has *pain asymbolia*.[59] [KM9] [KM14]

The distinction between *pain-knowledge* and *pain-suffering* is important when considering the *behavioral aspects* of experiencing pain. *Pain-knowledge* refers to *paying* attention to the existence of pain without necessarily experiencing the suffering. *Pain-suffering* refers to both *pain-knowledge* and the *subjective experience* of pain. Pain-suffering serves as a motivator for responding appropriately to the pain and its causes. This follows from the built-in properties of CTM:

Built into each and every LTM processor is an internal command that, when CTM becomes ***consciously aware*** of a chunk with high |weight| and negative valence, causes it to spend *all of its free time* to do something to lower the |weight| that the chunk's originating processor gives its gist. (See built-in properties 5 and 6 in section 2.1.7.) Thus, each processor is "motivated" to do something to "solve the problem".[60]

2.4 CTM as a Framework for Artificial General Intelligence (AGI)

Before indicating CTM's alignment with a number of major theories of consciousness, we remark on CTM's potential to serve as a *framework* for constructing an Artificial General Intelligence (AGI). This is a result of CTM's global architecture (kindred to arguments made for global latent workspace by (VanRullen & Kanai, 2021)) and, at the same time, the result of an essential difference between CTM and Baars' global workspace, CTM has *no* Central Executive. This is a feature, not a bug:

> **CTM's competition process enables it to engage processors to solve problems, even though CTM does not know which of its processors might have the interest, expertise, or time to work on them.** (Blum & Blum, 2023)**.** [KM14].

[59] A person who has pain and *knows* everything about it but lacks the *feeling* of agony has *pain asymbolia.* Such a person is not motivated to respond normally to pain. Children born with pain asymbolia rarely live past the age of three. The experience of pain, whether physical or emotional, serves as a motivator for responding appropriately to the pain. See (Grahek, 2001; 2007), (Klein, 2015), and (Gerrans, 2024).

[60] On the other hand, an experience of pain can also be exacerbated by this internal command. Forcing all LTM processors to spend more time on reducing the pain and less on whatever else they were doing could also be experienced as painful.

Specifically, if CTM, meaning one (or more) of its LTM processors, has a problem it needs solved, it (the processor) can submit the problem to the competition as a chunk with high |weight| giving it a chance to win and be globally broadcast to all processors. Processors with the interest, expertise and time to work on the problem can respond with appropriately |weight|-ed chunks.[61] In this way, ideas from unexpected sources may contribute to solving the problem, and useful collaborations can emerge. [KM14]

A Central Executive would have to know which processors have the inclination, expertise, and resources to solve problems as they arise, but Baars' does not say how the Central Executive could have that. [KM14]

We predict, in fact, that a Central Executive is not needed for consciousness or for general intelligence.

3 CTM Aligns with Major Theories of Consciousness

We have presented an Overview of the CTM model. Now we indicate how, at a high level, the model naturally aligns with and integrates key features of major theories of consciousness, further supporting our view that CTM provides a *framework* for building a conscious machine. We start with theories that have greatest alignment.

1. Global Workspace (GW)/Global Neuronal Workspace (GNW)

CTM aligns broadly with the architectural and global broadcasting features of the *global workspace* (GW) theory of consciousness (Baars, Bernard J., 1997), and at a high level with the *global neuronal workspace* theory of consciousness of neuroscientists Stanislas Dehaene, Jean-Pierre Changeux (Dehaene & Changeux, 2005), (Dehaene S. , 2014), and others.[62]

[61] If CTM's competition is based on the simple **f** value of intensity (see Appendix, section 6.3), the processor can submit the problem to the competition with |weight| = intensity of the currently broadcasted chunk, which is the sum of all |weights| of all submitted chunks. In that case, the chunk with that high |weight| will win, with probability ≥ 1/2 and become globally broadcast to all processors.

[62] Additional references for GNW include: (Dehaene & Naccache, 2001), (Sergent & Dehaene, 2005), (Dehaene & Changeux, 2011), and (Mashour, Roelfsema, Changeux, & Dehaene, 2020).

However, CTM differs from GW in significant ways. For example: CTM's competition for global broadcasting is formally defined; CTM constructs world models; and CTM purposely has no Central Executive.

2. Attention Schema Theory (AST) and the Self-Organizing Metarepresentational Account (SOMA)

CTM's ability to construct and utilize models of CTM's worlds (inner and outer), and the key role they play in CTM's *conscious awareness*, align closely with neuroscientist Michael Graziano's *attention schema theory* (AST) of consciousness (Graziano, Guterstam, Bio, & Wilterson, 2020). AST proposes that the brain is an information processing machine that constructs a simplified model of attention, just as it constructs a simplified model of the body, the Body Schema. According to AST, this Attention Schema provides a sufficiently adequate description of what it (the brain) is attending to for it to conclude that it is "aware".

The constructive co-evolution of Brainish and the MotW in CTM aligns in principle with the self-organizing metarepresentational account (SOMA) according to which "consciousness is something that the brain learns to do." (Cleeremans A. , 2011) and (Cleeremans, et al., 2020)

3. Predictive Processing (PP)

Predictive processing (correctly) asserts that the brain is constantly inferring, correcting and updating its predictions, generally based on motor outputs and sensory inputs. CTM's *predictive dynamics* (cycles of prediction, testing, feedback, and learning/correcting), locally and globally, align with various incarnations of *predictive processing* (von Helmholtz, 1866; 1962), (Friston K. , 2010), (Cleeremans A. , 2014), (Clark A. , 2015), (Hohwy & Seth, 2020), and others.[63]

[63] Other references include: (McClelland & Rumelhart, 1981), (Lee & Mumford, 2003), (Friston K. , 2005), (Clark A. , 2015), (Seth A. K., 2015), (Miller, Clark, & Schlicht, 2022).

4. Embodied Embedded Enactive Extended Mind (EEEE Mind)

CTM's ability to construct (and utilize) models of its worlds containing rich Brainish-labeled sketches (leading to its *feelings of consciousness*) derives in part from its embodied, embedded, enactive, and extended (EEEE) mind.

This aligns with the "4E" view that consciousness, like cognition (Carney, 2020), involves more than brain function (Rowlands, 2010). For consciousness, the interconnected 4E's are:

- *Embodied*: Incorporating relations with the entity's *body parts* and *processes* is essential for phenomenal consciousness. See, (Damasio, 1994), (Edelman, 2006) and (Shanahan, Global Access, Embodiment, and the Conscious Subject, 2005).[64]
- *Embedded,* and *Enactive*: Being *embedded in the outer world* and *enacting*/interreacting with it, thus affecting the world and creating experiences, is necessary for phenomenal consciousness. See, (Maturana & Varela, 1972), (Maturana & Varela, 1980), (Varela, Thompson, & Rosch, 1991), (Thompson, 2007), and (Clark A. , 2008).
- *Extended*: Consciousness is further enhanced by the entity having access to considerable external resources (such as libraries, Google, ChatGPT, Mathematica, ...). See, (Clark & Chalmers, 1998).

CTM is *embedded* in its outer world and, through its *embodied* actuators, can *enact* in this world, thus influencing what it senses and experiences. CTM's "mind" is *extended* by information it gets from resources in its outer world, and from embedded (or linked) off-the-shelf processors.

5. Integrated Information Theory (IIT)

Integrated Information Theory, the theory of consciousness developed by Giulio Tononi (Tononi, 2004), and supported by Koch (Tononi & Koch, 2015), proposes a measure of consciousness called Phi that, in essence, measures the amount of feedback and interconnectedness in a system. CTM's

[64] We note that here Shanahan views the global workspace as key to access consciousness, but that phenomenal consciousness requires, in addition, embodiment.

extensive feedback (its predictive dynamics, globally and locally) and its interconnectedness (global broadcasts and its fused multimodal Brainish gists) contribute to a high Phi.

6. Evolutionary Theories of Consciousness

CTM aligns with aspects of *evolutionary theories* of consciousness:

Oryan Zacks and Eva Jablonka provide evidence for the evolutionary development of a modified *global neuronal workspace* in vertebrates (Zacks & Jablonka, 2023) reinforcing our suggestion that an AI with a global workspace architecture could possess access consciousness.[65]

In *Sentience*, Nicholas Humphrey presents an evolutionary argument for the development of *phenomenal consciousness* in warm-blooded animals (Humphrey, 2023). In "The Road Taken" (Chapter 12 of *Sentience*), Humphrey spins a "developing narrative", starting with "a primitive amoeba-like animal floating in the ancient seas. Stuff happens. …" The resulting story provides a roadmap for how an entity might create world models and a sense of self. Indeed, this roadmap closely parallels the way CTM's world models evolve, and how CTM develops its sense of self and (subjective) conscious awareness.[66]

Thus, while the former evolutionary theory (Zacks & Jablonka, 2023) aligns with CTM's built-in GW architecture, the latter theory (Humphrey, 2023) aligns with CTM's development of subjective consciousness over time.

7. Extended Reticulothalamic Activating System (ERTAS) + Free Energy Principle (FEP)

At a high level, CTM aligns with ERTAS + FEP:

In *The Hidden Spring* (Solms M. , 2021), Marc Solms proposes that the source of consciousness is the arousal processes in the upper brain stem. More generally, Solms cites the Extended Reticulothalamic

[65] See (Ginsburg & Jablonka, 2019) for an extensive treatise on the evolutionary development of consciousness and their Unlimited Associative Learning (UAL) theory of consciousness.

[66] We claim Humphrey actually gives a road map for how an entity, warm blooded or not, might create world models and sense of self. As an exercise, we have re-written part of Chapter 12 (Humphrey, 2023) , "The Road Taken", from the perspective of CTM and sent a copy to Humphrey. His reply, "It would be great if we could meld these theories." (Personal communication with Nick Humphrey, Oct 9, 2023.)

Activating System (ERTAS) as the generator of feelings and affects, enabling consciousness. Although GW models generally consider processors as performing cortical functions, CTM goes beyond that. There is nothing to preclude CTM from having Gauge processors that function as ERTAS.

In "The Hard Problem of Consciousness and the Free Energy Principle" (Solms M. , 2019), Solms states that the "system must incorporate a *model of the world*, which then becomes *the basis upon which it acts*".[67] Similarly, *conscious awareness* in CTM of its MotW broadcasted chunks becomes "the basis upon which it [CTM] acts".

(Solms M. , 2019) posits that "affective qualia" is the result of homeostasis. "Deviation away from a homeostatic settling point (increasing uncertainty) is felt as unpleasure, and returning toward it (decreasing uncertainty) is felt as pleasure". Our discussions of pain and pleasure (in sections 2.3.1.1 and 2.3.1.2) and in (Blum & Blum, 2021) align with this view.[68]

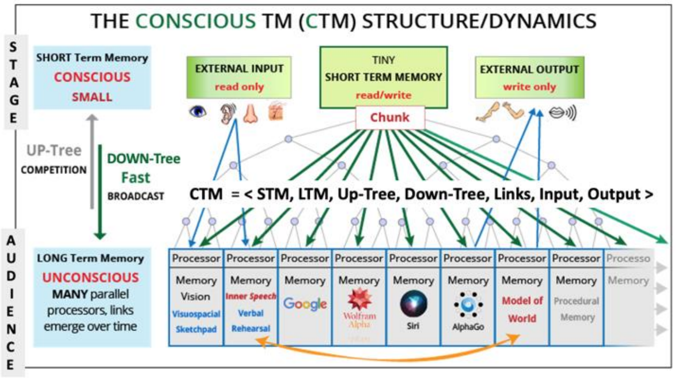
According to (Solms & Friston, 2018), homeostasis arises by a system resisting entropy, i.e., minimizing free energy. This is enabled by a *Markov blanket* (containing the system's input sensors and output actuators) that insulates the internal system from its outer world. In CTM, predictive dynamics (cycles of prediction, testing, feedback and correcting/learning) works to reduce prediction errors, an analogue to minimizing free energy.

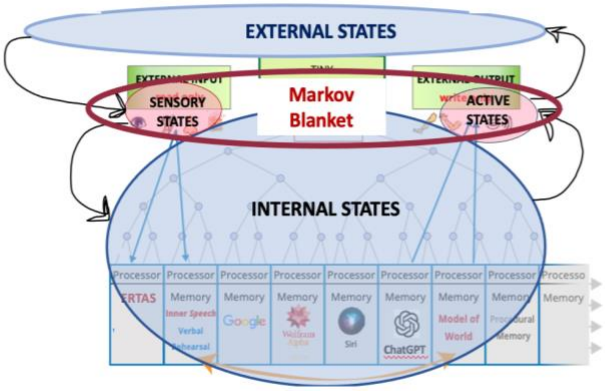
Evocatively, the well-known Friston diagram (Parr, Da Costa, & Friston, 2019) rotated 90 degrees clockwise, with Markov blanket separating internal and external states, is clearly realized in CTM:

[67] Italics here ours.

[68] Kevin Mitchell points out (personal communication) that another important point from Solms is that "the ascending homeostatic signals, which track different needs, must be valenced but also must have some distinguishing "qualities" so that when they are submitted to central decision-making units, the sources of the signals can be kept track of - so the organism doesn't mistake feeling thirsty for feeling tired (both of which feel BAD)".

In CTM, the fused Brainish gist FEELING_BAD ∘ THIRSTY is different than the fused Brainish gist FEELING_BAD ∘ TIRED. No need for a central decision-maker.





8. Comparison with Other Theories and Perspectives on Consciousness

Our view is that phenomenal or subjective consciousness is a consequence of (what is called) access, functional, or computational consciousness. Other researchers, including those mentioned above, tend to agree, at least to some extent. Features of CTM align with features and arguments they propose.

Lionel Naccache gives examples of how access consciousness can account for phenomenal consciousness (Naccache, 2018).

The co-evolution of Brainish and the Model-of-the-World[69] in CTM to produce subjective consciousness aligns with Axel Cleeremans et. al. proposal that phenomenal experience should be viewed "as the product of active, plasticity-driven mechanisms through which the brain learns to redescribe its own activity to itself." It is "not only shaped by learning, but its very occurrence depends on it." (Cleeremans, et al., 2020)

Our views on machine consciousness and AGI are close to (VanRullen & Kanai, 2021). We also see connections with Murray Shanahan's view on embodiment and the inner life (Shanahan, 2010). We see a kinship between the **CTM** and the self-aware robots developed by (Chella, Pipitone, Morin, &

[69] World models arise in other contexts, e.g., in the science of the brain and mind (Duncan, 2025). Here as in CTM, "the agent looks ahead to consider the future consequence of its choice." World models also arise in cognitive science, e.g., in analyzing what LLMs know (Yildirim & & Paul, 2024). We thank the anonymous referee for the latter reference.

Racy, 2020). Leslie Valiant proposes a "computational model to describe the cognitive capability that makes humans unique among existing biological species on Earth". We see his model's employment of a "*Mind's Eye*, where descriptions of multiple objects and their relationships can be processed" (Valiant, 2024) is remarkably similar to CTM's employment of a Model-of-the-World for its conscious awareness.

We have not explored how CTM might align or not with higher order theories (HOTs) of consciousness. HOTs claim to take a middle ground between standard GWT and early sensory theories (Brown, Lau, & LeDoux, 2019). (LeDoux & Brown, 2017) argue "that the brain mechanisms that give rise to conscious emotional feelings are not fundamentally different from those that give rise to perceptual conscious experiences." We do see some connection between HOTs specification of higher order representations (HORs) and CTM creating Brainish labeled sketches in its Model-of-the-World. Likewise, we see some connection between HOTs feature of *introspection* and the role that *inspection* plays in CTM's subjective experiences (Formal Definitions 7 and 10 in section 2.3.2.2). As CTM becomes increasingly consciously aware it becomes consciously aware of its conscious awareness. We have yet to explore how *levels of inspection* in CTM (section 2.3.2.3) might lead to higher *levels of subjective consciousness*.[70]

Philosophically, we align with much of Daniel Dennett's functionalist perspective (Dennett D. C., 1991). Along with Dennett, we do not see the *explanatory gap* (Levine, 1983) as insurmountable. On the contrary, we see the CTM as helping to explain the feeling of "what it is like" (Nagel T. , 1974). CTM, like Attention Schema Theory (AST), appears to embody and substantiate *illusionist* notions of consciousness proposed by Dennett (Dennett D. C., 2019) and Keith Frankish (Frankish K. , 2016).[71]

[70] Relevant papers on levels and degrees of consciousness are (Farisco, Evers, & Changeux, 2024) and (Lee A. Y., 2023). We thank the anonymous reviewer for the latter.

[71] Saying that the feeling of consciousness is an illusion does not deny the existence of that feeling. It's just not what you think it is. As a familiar example, the fact that a movie is made up of (many) discrete still images does not affect the feeling of continuity one gets from viewing it. The feeling of continuity is an illusion.

In this regard, CTM and AST both align with philosophers Daniel Dennett's and Keith Frankish's views that consciousness is a kind of "illusion", which can be understood by such of their explications as: "Consciousness is the brain's interface for itself." (Dennett); "Illusionism's a view about what experiences, perceptions, interpretations illusions *really are*. The illusion is that they are something

However, we are not claiming that any of these researchers also view that artificial consciousness is inevitable, even possible. For example, Anil Seth (Seth A. , 2024) and Nick Humphrey (Humphrey, 2023) view that consciousness requires living organisms.

4 Summary and Conclusions

In this paper we have presented:

1. a simple formal machine model of consciousness, the CTM (Section 2),
2. the observation that CTM aligns at a high level with some of the major theories of consciousness (Section 3), and
3. answers (Appendix, section 6.4) drawn from CTM, to Kevin Mitchell's 15 questions that a "theory of consciousness should be able to encompass" (Mitchell, 2023).

The Theoretical Computer Science (TCS) perspective has influenced the design and precise formal definitions of CTM, and the conclusions we have drawn from the model.

Important features of CTM include:

- a formal competition leading to a global broadcast of the winner,
- a self-generated internal multimodal language (Brainish),
- interactions with the outer world via input sensors and output actuators,
- ability to construct models of its (inner and outer) worlds,
- cycles of prediction, feedback and learning, that constantly unconsciously (and consciously) update its algorithms,

all while operating under resource limitations (time and space).

they're not." (Frankish K. , 2023); and "I believe that feelings are real, but I think we have mistaken ideas about they are." (Frankish K. , 2023).

Although CTM is inspired by the simplicity of Turing's formal model of computation and the insight of Baars' global workspace (GW) architecture, our formalization is neither a Turing Machine nor a standard GW model. Its *consciousness* (access and phenomenal) depends on what's under the hood, and its interaction with its worlds. In other words, what matters is not just the result of computation, but also *how* that computation was done.

In developing this theory, we were motivated to construct an "Occam's razor" **simple model** that addresses and provides insights into consciousness. We tried and discarded many variants of the CTM model because they grew in complexity without end. **The probabilistic CTM and only the probabilistic CTM has withstood our tests for simplicity and explanatory power.** This provides a measure of confidence that the CTM model is on the right track.

In an earlier paper (Blum & Blum, 2022), we provided examples of how CTM could exhibit phenomena associated with consciousness (e.g., blindsight, inattentive blindness, change blindness, and dreams) and related topics in ways that agree at a high level with cognitive neuroscience. Although our TCS approach differs from empirical science, we view its alignment at a high level with findings from empirical science as providing an additional measure of confidence. To the extent that CTM theory can predict other phenomena, or provide ways to think about questions related to consciousness, its perspective helps provide understanding.

As an example of a prediction from CTM, the theory proposes the existence of two different kinds of pain asymbolia (section 2.3.2.4.) One kind occurs when the sketch of the CTM in the MotW is labeled SELF but not PAIN, the other when labeled PAIN but not SELF. The SELF but not PAIN (i.e., sensory deprivation) explanation (Grahek, 2001; 2007) is standard. The PAIN but not SELF explanation was proposed by (Klein, 2015). (Sierra, 2009) quotes a typical pain asymbolic's explanation: "When a part of my body hurts, I feel so detached from the pain that it feels as if it were somebody else's pain." In an excellent and recent review, (Gerrans, 2024) points out that each of these two different explanations has a problem, without noting the possibility that there might be two different kinds of pain asymbolia which is what the CTM model predicts.

As an example of how CTM helps think about questions related to consciousness, here is a problem posed by (Zadra & Stickgold, 2021), p. 267:

> "Recall something that happened to you yesterday and think about it for a second. Got it? Well, as scientists, we don't have much of an idea of how your brain did that, how it searched yesterday's memories, selected just one of them, and found the associations that defined what

> it meant to you. And we know nothing about how you became conscious of any of this information."

Here is a CTM response to the request "Recall something that happened yesterday and think about it for a second":

> Different processors recall different things that happened yesterday and put those in the competition. One of them wins and its recollection gets broadcast. CTM *pays conscious attention* to whatever got broadcast. The prompt, "think about it for a second", will motivate processors to *inspect* the broadcast. As a consequence, CTM will experience subjective feelings related to the recalled memory, i.e., CTM will become *consciously aware* of what it recalls happened yesterday.

We see this explanation as providing a high-level understanding of how humans might respond, and become conscious of memories.

Other examples of understandings include:

- The CTM explanation for bipolar disorder: the CTM has a parameter **d** (**disposition)** that lies between **-1** and **+1**. (See Appendix, section 6.2.) When **d = +1**, the CTM cannot see anything negative until there is nothing positive to see, in which case it flips to **d = -1**. When **d = -1**, it cannot see anything positive unless and until there is nothing negative to see. Only a reboot can reset **d**. The best **d** in real life appears to be around **+.2**, meaning a bit greater than **0** but much less than **+1**.

- An explanation for the hard problem of how pain and pleasure can be experienced, not just simulated, in a machine.

CTM is not a model of the (human or animal) brain, nor is it intended to be. It is a simple formal *machine model* of consciousness. Nevertheless, at a high level, CTM can exhibit phenomena associated with human consciousness (blindsight, inattentional blindness, change blindness, body integrity identity disorder, phantom limb syndrome, …), and aligns with and integrates those key features from main theories of consciousness that are considered essential for human and animal consciousness. The CTM model demonstrates the compatibility of those theories.

The fact that CTM is clearly buildable, and arguably a basis for consciousness, supports (the credibility of) our claim that *a conscious AI is inevitable.*

Finally, the development of CTM is a work in progress. More will appear in the upcoming monograph (Blum, Blum, & Blum, 2026). We view this paper as an outline for the monograph.

Our goal is to explore the model as it stands, determine the good and the bad of it, and make no (unnecessary) changes to it.

5 Acknowledgements

We are grateful to Michael Xuan for his enormous encouragement, and to UniDT for their long-term support. We appreciate numerous discussions with and helpful critiques from friends and colleagues: Johannes Kleiner, Wanja Wiese, Ron Rivest, Alvaro Velasquez, Sergio Frumento, Aran Nayebi, Martin Korth. We thank Kevin Mitchell for posing his enormously sensible questions "that a theory of consciousness should be able to encompass" and his subsequent perceptive discussion with us. We thank the anonymous reviewer for insightful comments and thoughtful suggestions.

6 Appendix

6.1 A Brief History and Description of the TCS Approach to Computation

The Theoretical Computer Science approach to computation starts with Alan Turing in the 1930's and focuses on the question, "What is computable (decidable) and what is not?" (Turing, 1937**.** Turing defined a simple formal model of computation, which we now call the Turing Machine (TM) and *defined* a function to be computable if and only if it can be realized as the input-output map of a TM. The formal definition of a TM (program) also provides a formal definition of the informal concept of algorithm.

Using his model, Turing proved properties (theorems) of computable functions, including the existence of universal computable functions (universal Turing machines) and the fact that some functions, though definable, are not computable. The former foresees the realization of general purpose programmable computers; the latter that some problems cannot be decided even by the most powerful computers. For example, Turing shows there is no Turing machine (Turing computable function) that given the description of a TM **M** and an input **x**, outputs **1** if **M** on input **x** (eventually) halts, and **0** if not. This is known as the "halting problem" and is equivalent to Gödel's theorem on the undecidability of arithmetic.

But why should we believe the Church-Turing Thesis, suggested first in (Turing, 1937), that the TM embodies the informal notion of computability (decidability)? That's because each of a great many very different independently defined models of discrete computation, including TMs and Alonzo Church's *effective calculability* (Church, 1936), define exactly the same class of functions, the computable functions (also called the *recursive functions*). In programming parlance, all sufficiently powerful practical programming languages are equivalent in that anyone can simulate (be compiled into) any other. The ensuing mathematical theory is generally called the Theory of Computation (TOC).

In the 1960's, with the wider accessibility of computers, newly minted theoretical computer scientists such as Jack Edmonds (Edmonds, 1965) and Richard Karp (Karp, 1972), pointed out that resources matter. Certain problems that in principle were decidable/computable, were seemingly intractable given feasible time and space resources. Even more, intractability seemed to be an intrinsic property of the problem, not the method of solution or the implementing machine. The ensuing sub-theory of TOC, which introduces resource constraints into what is or is not *efficiently* computable, is called Theoretical Computer Science (TCS).

TCS focuses on many problems that arise when resource limitations are taken into account, including the question, "What is or is not computable (decidable) given limited resources?" A key problem here is the deceivingly simple "SAT problem": Given a boolean formula $\mathcal{F}$, is it satisfiable, meaning "Is there a truth assignment to its variables that makes formula $\mathcal{F}$ true?" This problem is decidable. Here is a decision procedure: Given a boolean formula $\mathcal{F}$ with **n** variables, systematically check to see if any of the $\mathbf{2^n}$ possible truth assignments makes the formula true. If yes, output **1**, otherwise output **0.** This brute force procedure takes **exponential(n)** time in general. But is the "SAT problem" tractable, meaning decidable efficiently, i.e., in **polynomial(n)** time? This is equivalent to the well-known **P = NP?** problem of (Cook, 1971), (Karp, 1972), (Levin, 1973).

While the design of novel and efficient algorithms is a key focus of TCS, another unanticipated direction comes of the ability to exploit the power of hard problems. Such problems that cannot be solved efficiently, have been a key insight of TCS. An example is the definition of pseudo-randomness (Yao, 1982). This ability to exploit the power of hardness is novel for mathematics.

6.2 About Numbers

Although the general CTM model allows parameters for numbers such as the lifetime **T** and the number **N** of LTM processors, our specific choice of numbers in this paper has been guided by both the workings of the human brain and the requirements of TCS:

- The human brain's $\mathbf{10^7}$ ($\mathbf{\sim 2^{23.25}}$) cortical columns leads us to set CTM to have $\mathbf{N = 2^{24}}$ ($\mathbf{\sim 10^{7.225}}$) processors. A clock rate of **10** to **100** ticks per second is suggested by the alpha and gamma rhythms respectively. The lifetime of $\mathbf{T = 10^{10}}$ ticks at **10** ticks per second is about 32 years, roughly the length of a human lifetime. The brain, assuming it is a CTM, CTM cannot have $\mathbf{(10^7)^2 = 10^{14}}$ *independent* **1:1** links at birth because it has less than $\mathbf{10^{11}}$ neurons, so at most **1/1000** pairs of processors can be linked.
- The demand for simplicity suggests that the Up-Tree be binary. The focus on just **1** chunk instead of **7±2** arises from the demand for simplicity and a simple argument that **1** chunk will suffice. The need to do a computation at each node of the Up-Tree in a single tick suggests that a gist should have at most **20** Brainish words.

6.3 The Probabilistic Competition for Conscious Attention and the Influence of the Disposition (a Parameter)

We tried to make the CTM Up-Tree competition (a variant of the standard tennis or chess tournament) deterministic, but it turned out to *necessarily* be *probabilistic*: This is because any deterministic competition must be made increasingly complex to realize essential properties. This is the case because, for example, we want chunks of near equal |weight| to have near equal chances of getting onto the STM stage. Why is that?

Consider a deterministic competition on **22** chunks that makes decisions based on a chunk's weight. Suppose chunk **A** is pinned to weight **11**, chunk **B** to weight **9**, and the remaining **20** chunks to weight **1** each. In the deterministic CTM competition, **A** always wins, all other chunks never do, and as long as weights don't change, CTM remains totally unconscious of all chunks other than **A.**

In the probabilistic CTM competition described below (with **f** = intensity), a chunk wins with probability proportional to its |weight|, so chunk **A** wins with probability **11/40**, **B** win with probability **9/40**, and the remaining **20** chunks with probability **1/40** each. **A** and **B** each have roughly equal probability of winning a competition, and since a new independent competition is begun at every clock tick, CTM will likely become conscious of both. (Moreover, with probability **1/2**, *some* chunk of weight **1** will win the competition, but rarely the same chunk.)

In the general probabilistic CTM competition it is proved that a chunk wins the competition with probability proportional to (a *function* of) its |weight|. As a consequence, the winning chunk is independent of the arrangement of processors, where they are located! This is a property of CTM's probabilistic competition.[72] (It is a property, incidentally, that would be difficult if not impossible to achieve in tennis and chess tournaments.)

We now describe the probabilistic competition. First recall that a *chunk* is a tuple,

<address, time, gist, weight, auxiliary information>,

[72] It would not do to pick the winner to be the chunk with the highest |weight|, for then no other chunk, even one with a slightly less |weight|, would have a chance. The probabilistic competition gives all competitors a fair chance.

consisting of the **address** of the originating processor, the **time** the *chunk* was put into the competition, a succinct Brainish **gist** of information, a **valenced weight** (to indicate the importance/ value/ confidence that the originating processor assigns its gist and whether processors should look for ways to increase or decrease the **|weight|**), and some auxiliary information. For the probabilistic CTM, there is good reason to choose the *auxiliary information* to be a pair of numbers that we call **(intensity, mood)**.

At the start of the competition, each LTM processor puts a chunk into its leaf node of the competition with auxiliary information: **intensity = |weight|** and **mood = weight**.

In the probabilistic competition, each non-leaf node of the Up-Tree contains a **coin-toss neuron**. The coin-toss neuron probabilistically chooses the *local winner* of the two competing chunks supplied by the node's two children with probability proportional to their **f values**. Here **f** is a function mapping chunks to non-negative real numbers. A *simple*, but natural **f** value maps chunks to their **intensities**.

If C_1 and C_2 are the two competing chunks in a match of the competition, then the coin-toss neuron chooses C_i as local winner of the match with probability $\mathbf{f}(C_i)/\mathbf{f}(C_1)+\mathbf{f}(C_2)$.

The local winner will move up one level in a single clock tick. The first four parameters of this new chunk are the same as the local winner's. But its **intensity** *is the sum* of the two competing chunks' intensities. Similarly for its **mood**. We call this the **WINNER TAKES ALL** policy!

Thus, as a chunk moves up the tree, the intensity never decreases. This is not necessarily the case for mood. The winning chunk's auxiliary information at the end of the competition will contain the *sum of all submitted chunks' intensities* (|weights|) and *the sum of all submitted moods* (weights).

Hence, although at the start of the competition each processor has little if any knowledge about the other $\mathbf{N = 2^{24} -1}$ chunks that are being submitted, the reception of the broadcasted winner provides useful information. In the competition with the simple **f** values mentioned above, in addition to providing the winning gist, its given weight and its processor's address, each processor also gains knowledge from the broadcast of the average of all submitted chunks' intensities and moods (on dividing the broadcasted intensity and mood by **N**).[73]

[73] **Question:** Why does CTM need a competition? Why not have each processor compute the probability of its chunk winning and then choose the winning chunk based on these probabilities? **Answer:** "At the start of the competition each processor has little idea about

More generally, consider a CTM whose competition employs the following **f** value:

$$\mathbf{f(chunk) = intensity + d\bullet(mood) \quad where \quad -1 \leq d \leq +1.}$$

Here **d** is a *fixed constant* called CTM's **disposition**.

CTM's *disposition* plays an important factor in the competition that selects which chunk will be globally broadcast, and hence CTM's behavior.

If the disposition is $\mathbf{d = 0}$ we say CTM is "level-headed". Here **f** is the *simple* **f** value discussed earlier. In this case, the probability of a submitted chunk "winning" the competition will depend only on its |weight|, independent of valence.

If the disposition is $\mathbf{d > 0}$, CTM will be "upbeat" in the sense that positively valenced chunks will have a higher probability of winning than negatively weighted chunks of the same |weight|. If its disposition is $\mathbf{d = +1}$, CTM is manic: only positively valenced chunks can win the competition.[74] The CTM knows only what is positive in its life, as long as anything is positive.

If the disposition is $\mathbf{d < 0}$, CTM will be "downbeat". In the extreme, if $\mathbf{d = -1}$, CTM will be "hopelessly depressed". Only negatively valenced chunks can win the competition.[75] In the CTM, there is no way out of this horrible state except with a reboot, i.e., a "shock" to the system that gives it a less extreme disposition.[76]

the other $\mathbf{N} = \mathbf{2^{24} - 1}$ chunks that are being submitted". Without this knowledge, the simple competition is an efficient way to ensure that each chunk gets into STM with probability proportional to its weight.

[74] – Unless of course no chunks have positive valence, in which case a negatively valenced chunk must get into STM, and each such chunk will have probability $\mathbf{2^{-24}}$, independently of its weight, to get to STM.

[75] – Unless of course no chunks have negative valence, in which case a positively valenced chunk must get into STM, and each such chunk will have a probability of $\mathbf{2^{-24}}$, independently of its weight, to get to STM.

[76] In humans, electroconvulsive therapy (ECT) is used primarily for extreme depression (**d=-1** in CTM). For humans in the manic state (**d=+1** in CTM), CTM theory suggests that there too, a reboot is warranted. See (Elias, Thomas, & Sackeim, 2021).

6.4 Addressing Kevin Mitchell's questions from the perspective of CTM

Here we address Kevin Mitchell's fifteen questions (Mitchell, 2023)[77] from the perspective of the CTM, a robot with a CTM brain.

Our answers refer to and supplement our Overview (Section 2). They deal *only* with the CTM model, meaning an entity with a CTM brain. That entity could be a human if it has a CTM brain. These answers say nothing about other models. They say nothing about whether or not a worm is conscious unless the worm has a CTM brain. From here on, unless we say otherwise, everything we have to say is about the CTM model.

In the following, Mitchell's fifteen questions, KM1, ..., KM15 are printed in **bold**. Our answers follow in non-bold print.

KM1. What kinds of things are sentient? What is the basis of subjective experience and what kinds of things have it? What kinds of things is it like something to be?

Again, our answers deal *only* with the CTM model.

What kinds of things are sentient? We take "sentience" to mean the ability to "sense" and "feel" things. In this *sense*, CTM is sentient: Sensory processors in CTM convert sensory inputs into chunks. When CTM becomes *consciously aware* of such a "sensory chunk" **<p, t, ... >**, a sensory experience is *evoked* and the Brainish word **<p, t>** for this experience is created. Later, when CTM becomes *consciously aware* of a chunk containing the word **<p, t>**, the original *sensory feeling* will be *evoked*. (See Axiom A in section 2.3.2.2 and KM2 for explanations of these concepts.)

What is the basis of subjective experience and what kinds of things have it? The ability of CTM to construct models of its worlds (inner and outer) is important for its *subjective experiences*. These experiences are described in the model not with English words but with multimodal *Brainish-labeled sketches* of *referents*, things the sketch refers to (section 2.3). In the case of a red rose, the Brainish-labeled sketch might be an image , fused with its odors , smooth touch , and so on.

What kinds of things is it like something to be? The Model-of-the-World processor (MotWp) and the Model-of-the-World (MotW) it creates play essential roles in "what it is like" to be a CTM (section

[77] Many of Mitchell's questions are in fact a collection of intertwined questions.

2.3). The Brainish-labeled sketches in the MotW are *created* and *evolve* throughout the life of CTM. The labels succinctly indicate what CTM learns, senses or feels about the referents.

For example, the label SELF applied to a sketch in the MotW indicates that that particular sketch's referent is felt as a part of, or the whole of, CTM itself. The label PAIN indicates that that particular referent is in pain. Labels like PAIN and SELF are (English translations of) Brainish words **<p, t>, pointers** to chunks **<p, t, gist, w, aux>**[78]**,** in which the concepts (like PAIN and SELF) arose in "meaningful" ways (sections 2.3.1.1 and 2.3.2.4 and sections 2.3.1.2 and 2.3.1.3).[79]

KM2. Does being sentient necessarily involve conscious awareness? Does awareness (of anything) necessarily entail self-awareness? What is required for 'the lights to be on'?

In this paper we defined two related notions of consciousness in CTM, *conscious attention* (Formal Definition 1 in section 2.2) and *conscious awareness* (Formal Definition 10 in section 2.3.2.2). We review these formal CTM definitions here:

Conscious attention in CTM is the *reception* by all LTM processors of the *broadcast* of CTM's current *conscious content*, **<p, t, ... >**, the winner of the competition for STM that started at time **t**. If in addition, CTM ***inspects*** this chunk, meaning that, CTM determines the *weight* of this chunk and *decodes* all Brainish words appearing in the gist of the chunk,[80] we say CTM becomes ***consciously aware*** of this chunk.

We did not give a formal definition for *self-awareness*, but could add:

> **Formal Definition 11.** ***Self-awareness*** in CTM is *conscious awareness* of one or more labeled sketches in the MotW that are labeled **SELF.** (Here **SELF** is one of the Brainish words evolved from the primal **SELF$_0$**.)

Does being sentient necessarily involve conscious awareness? Yes, since "sentience" is the ability to "sense" and "feel" things. Sentience in CTM *is* conscious awareness.

[78] Recall, a chunk is given by a tuple, **<p, t, gist, w, aux>**, where **p** is the address of the processor that created the chunk, **t** is the time the chunk was put into competition for STM, **gist** is the succinct Brainish information the processor is submitting, **w** is the valenced weight the processor has given the gist, and **aux** is auxiliary information (not necessarily given by processor **p**).

[79] These concepts are not understood by CTM straightaway but rather is suggested by examples. Over time, the concept becomes further refined and nuanced.

[80] Recall, *Brainish words* are pointers **<p, t>** to broadcasted chunks, **<p, t, g, w, ... >**. Broadcasted chunks are stored by all LTM processors. For a processor to *decode* a word **<p, t>** means for it to look into its memory, retrieve the chunk **<p, t, g, w, ... >**.

Does awareness (of anything) necessarily entail self-awareness? No. The sketch of the newborn CTM in the MotW is just an empty blob at first, representing itself and the "world". In time, however, the infant's world model will include a rough sketch of itself labeled SELF. The label SELF marks the beginning of self-awareness, which eventually develops into full-blown self-awareness.

What is required for 'the lights to be on'? In CTM, the *lights come on* gradually, in lockstep with the MotW getting populated with sketches and their labels. (For more on this, see our answers to KM3 and KM4.)

KM3. What distinguishes conscious from non-conscious entities? (That is, why do some entities have the *capacity* for consciousness while other kinds of things do not?) Are there entities with different degrees or kinds of consciousness or a sharp boundary?

Again, we replace "entities" with "CTMs".

What distinguishes conscious from non-conscious entities? (That is, why do some entities have the *capacity* for consciousness while other kinds of things do not?) All 'healthy" CTMs have the *capacity* to be conscious. Absent a broadcast system there is no *conscious attention*. Absent the ability to inspect a broadcasted chunk there is no *conscious awareness* (subjective consciousness).

Are there entities with different degrees or kinds of consciousness or a sharp boundary? In CTM, we have distinguished between *conscious attention* and *conscious awareness*. Alison Gopnik (Gopnik, 2007) has said that babies are more conscious than adults. With respect to *conscious attention*, this is true for infant CTMs: before links have formed, all communication goes through STM and is broadcasted to all processors. This is conscious attention. However, without developed world models, infant CTMs are less *consciously aware* than adult CTMs.

CTM can exhibit different *degrees of consciousness*:

1. CTM's degree of consciousness is roughly proportional to the |weight| of the chunk being consciously attended to, that chunk being the one just received by broadcast.

2. As already noted, a newborn CTM has only a very coarse world model which does not even include a sketch of itself as a separate entity. Sketches with annotated Brainish labels develop and become refined *gradually*. They are what CTM becomes consciously aware of.

3. CTM can be consciously aware when awake or dreaming. It is not conscious when in *deep sleep*, at which time the broadcasted chunk (originated by the Sleep processor) has a high enough |weight| to keep all other chunks at bay. (See our answers to KM4 for discussions of Sleep and Dream processors.)

4. Faulty processors or faults in the Up-Tree competition can affect what gets into STM, hence affect both conscious attention and conscious awareness. For example, a faulty CTM can have blindsight,

meaning it can do some things that are normally done with conscious sight, but without any *conscious sense* that it can see (Blum & Blum, 2022). This can happen if the Vision processor fails to get its chunks into STM (e.g., relevant branches in the Up-Tree are broken or the Vision processor fails to give high enough |weight| to its chunks), while previously formed links enable visual information to (appropriately but unconsciously) manage the Motor processors.[81]

KM4. For things that have the capacity for consciousness, what distinguishes the *state* of consciousness from being unconscious? Is there a simple on/off switch? How is this related to arousal, attention, awareness of one's surroundings (or general responsiveness)?

For things that have the capacity for consciousness, what distinguishes the *state* of consciousness from being unconscious? An unconscious state occurs, for example, when a *Sleep processor* generates a NoOp chunk, a chunk having a **NULL** gist with a "sufficiently high" positive weight, one well above the sum of the |weight|s of all other chunks. Such a weight prevents other chunks - including those from processors that interact with the outer world[82] - from having much chance to enter STM. This is the *sleep state of unconsciousness*.[83] In this unconscious state, CTM is not aware of its surroundings, though it might be *aroused* by pangs of intense hunger, other pains, a very loud explosion, and so on. This occurs, for example, when these pangs, pains, and sounds overwhelm the Sleep processor with even larger |weight|.

Another unconscious state occurs when all chunks submitted to the competition for STM have zero weight. In CTM's winner-take-all competition, chunks reach STM with probability proportional to (a *monotonically increasing function* of) the chunk's |weight|. If all chunks have zero weight, they all have equal probability 2^{-24} to get to STM. In that case, chunks flit in and out of STM at random so fast that CTM loses anything remotely resembling sustained attention (like Robbie the Robot in *Forbidden Planet*). CTM can get out of such a state only when some processor creates a nonzero-

[81] In the human visual cortex, the dorsal stream of vision is unconscious; the ventral stream is conscious. Studies on blindsight suggest that communications via the unconscious dorsal stream account for the surprising blindsight ability of visually impaired people (Tamietto & Morrone, 2016).

[82] Something like this can happen in total depression and catatonia in CTM, and slow-wave (non-REM) sleep in humans.

[83] Even in deep sleep, however, a CTM can still carry out tasks (utilizing unconscious communication between processors via links) but without attention and therefore without conscious awareness.

Similarly, if CTM's broadcast station is turned off, there is no conscious attention and no conscious awareness. If this occurs after links and unconscious algorithms have been created, then while the broadcast station is down, CTM is an unconscious zombie. It can still function unconsciously with algorithms it acquired when it was conscious.

weighted chunk or the whole system is rebooted. (See Appendix, section 6.3 for discussion of the competition to enter STM.)

How is this related to arousal, attention, awareness of one's surroundings (or general responsiveness)? Returning to the sleep state of unconsciousness, when the |weight| of a Sleep processor's chunk drops a bit - but not enough to let input-output chunks enter STM - CTM's *Dream processor* can take over, enabling chunks of a dream to emerge and enter STM. We consider the Dream state to be a *conscious state*. When the |weight| of the Sleep processor's chunks drops further, CTM wakes up.

Is there a simple on/off switch? The Sleep processor's weight assigning mechanism serves as an *on/off switch* for consciousness in CTM.

The above are some not all of the ways CTM can go from consciousness to unconsciousness and back.

KM5. What determines what we are conscious of at any moment? Why do some neural or cognitive operations go on consciously and others subconsciously? Why/how are some kinds of information permitted access to our conscious awareness while most are excluded?

What determines what we are conscious of at any moment? In CTM, the competition for STM determines what chunk CTM is conscious of. At every clock tick, there is exactly one chunk in STM, the winner of the competition. When that chunk is broadcast and received at the next clock tick, CTM pays *conscious attention* to that chunk. This ensures that all processors can focus on the same thought.

Why do some neural or cognitive operations go on consciously and others subconsciously? We add: **and others unconsciously?**

Conscious activity is deliberate attention. It is triggered by information that goes though STM and is broadcast to all LTM processors. It is used to process new information, to learn new things, to get processors working together in new ways.

We call *subconscious* whatever is put in the Up-Tree competition for STM, but does not win it. It is below but near the level of consciousness, in what Freud called the anteroom (Freud, 1966) .

We call *unconscious*:

1. Operations within each LTM processor that do not go into the Up-Tree competition. (These operations enable processors to do what they need to do efficiently with minimal distractions.)
2. Communication between LTM processors via links. (Such communication is much quicker than conscious communication that goes through STM.)

Unconscious activity is quicker than *conscious activity*. It (unconscious activity) is typically used for perfecting already implemented procedures (algorithms). Such *unconscious* activity can become *subconscious*, and *subconscious activity* can later become *conscious*.

Why/how are some kinds of information permitted access to our conscious awareness while most are excluded? Information is "permitted access" into CTM's conscious attention/awareness only if it wins the competition for STM enabling all processor to focus on the same thought. Computations done unconsciously (within processors themselves or though links) are much faster than computations that go through STM. Thus, processors will not submit information to the competition if is not needed globally. This information remains unconscious until there is reason to have it globally broadcast. Typically, conscious attention is largely for the creation of new algorithms; unconscious attention is for improving the constants in those algorithms.

KM6. What distinguishes things that we are currently consciously aware of from things that we *could be* consciously aware of if we turned our attention to them, from things that we *could not be* consciously aware of (that nevertheless play crucial roles in our cognition)?

Conscious awareness in CTM has to do with subjective experience. For CTM to be consciously aware of a thing, call that thing **abc**, a chunk referring to **abc** must get into STM so CTM can pay conscious attention to it (**abc**). But even conscious attention to **abc** does not make for conscious awareness of **abc**. For that, the chunk must be *inspected*, as this will *evoke* CTM's subjective experience. (See section 2.3.2.2 and KM2.)

What things, though important for cognition, *cannot* enter consciousness? Here are a couple of examples from CTM:

1. Things that must be done so quickly that the communication necessary to do the thing hasn't the time to go through STM. For example, CTM must quickly swerve away from an oncoming car while riding its bike.
2. Things like **abc** whose doing would take away from a more important thing called **xyz**, when there is time to consciously attended to only one of **abc** and **xyz**.

KM7. Which systems are required to support conscious perception? Where is the relevant information represented? Is it all pushed into a common space or does a central system just point to more distributed representations where the details are held?

Which systems are required to support conscious perception? Assuming conscious *perception* is *conscious awareness*, then the major systems required are the Up-Tree competition system, the system for broadcasting the winner, and the system for *inspecting* broadcasted chunks.

Where is the relevant information represented? Is it all pushed into a common space or does a central system just point to more distributed representations where the details are held?

Information in CTM is represented by chunks **<p, t, gist, weight, aux>,** gists being short sequences[84] of Brainish words. Brainish words are pairs **<p', t'>** which point to chunks **<p', t', gist', weight', aux'>** that when *broadcast* and ***inspected*** evoke a fusion of the senses and feelings evoked by the words in the gist. If the gist is **NULL** then its **<p, t>** ***is*** its primary feeling. This is an axiom, Axiom A.

All information is stored in the LTM processors. Processors store all chunks they have submitted to the Up-Tree competition (creating a local dictionary) and all chunks received from broadcasts (creating a global dictionary). The Model-of-the-World processor is distributed across all LTM processors, as are the models it builds of CTM's inner and outer worlds. Thus, the Brainish-labeled sketches in these models are distributed across the LTM processors.

KM8. Why does consciousness feel unitary? How are our various informational streams bound together? Why do things feel like *our* experiences or *our* thoughts?

Why does consciousness feel unitary? How are our various informational streams bound together? CTM binds different thoughts **{<p, t>}** in a gist – a short Brainish phrase - that points to those thoughts or experiences.

At each clock tick, all LTM processors *simultaneously receive* a global broadcast of the conscious content (current chunk) in STM. The simultaneous reception of a chunk gives CTM a sense of a *unitary* experience of the gist in the chunk. If in addition, the chunk is *inspected* by all processors, this will evoke in CTM a fused subjective experience expressed by its gist.

CTM "sees" and "smells" a sweet red rose when it becomes consciously aware of the sketch of the rose in its MotW with its *fused multimodal Brainish* label RED_COLORS ○ SWEET_SMELL. That broadcasted information feels unified because it is received and inspected simultaneously by all processors.

Why do things feel like *our* experiences or *our* thoughts? When CTM becomes *consciously aware* of a thought (i.e., a sketch of a thought in the MotW) labeled SELF, CTM will *experience* that thought as its own (sections 2.3.1.2 and 2.3.1.3).

KM9. Where does our sense of selfhood come from? How is our conscious self related to other aspects of selfhood? How is this *sense of self* related to actually *being a self*?

[84] A gist is ***short*** if it contains at most twenty Brainish words, **<p', t' >**.

Here again, world models, with their (learned from experience) Brainish-labeled sketches, determine CTM's sense of self and other aspects of selfhood.

Over time, the MotW's sketches become labeled with a variety of Brainish words. For this question, the labels SELF and FEEL are particularly important. When CTM becomes *consciously aware* of such a Brainish-labeled sketch, it will feel that that part is part of itself. If the sketch is of CTM, then CTM will *feel being itself*.

Known pathologies occur when any one or both of these labels is missing, or when sketches are mislabeled.[85]

KM10. Why do some kinds of neural activity feel like something? Why do different kinds of signals feel different from each other? Why do they feel specifically like what they feel like?

Why do some kinds of neural activity feel like something? In CTM, "neural activity" is activity in the sensors, the actuators, the LTM processors, the links between them, the Up-Tree and the broadcasting system. Each Sensory processor gets input from specific sensors and turns distinct inputs into distinct Brainish gists, i.e., Brainish representations of those senses, and those gists into chunks. When chunks with these gists are globally broadcast and *inspected*, they *evoke* different feelings in CTM, depending on their originating Sensory processors' chunks.

Why do different kinds of signals feel different from each other? Why do they feel specifically like what they feel like? As above, chunks that are broadcast and inspected will evoke different feelings depending on their Brainish gist's origins.

In the MotW, sketches of a red rose 🌹 and a red fire truck 🚒 will both get the Brainish label RED. But these sketches will get many other labels as well. For example, the fire truck sketch likely gets the Brainish label LOUD_SIREN while the rose sketch does not. The rose sketch gets labeled SILKY_TOUCH ∘ SMELLS_SWEET while the fire truck does not. With increasingly more Brainish labels gained during CTM's lifetime, they come to "feel specifically like what they feel like."

Similar arguments apply to the distinct pleasures CTM feels when seeing a red rose or when replenishing its empty fuel tank. See also our answer to the next question.

[85] Some human examples of pathologies due to mislabeling include body integrity dysphoria and asomatognosia (when SELF is missing from some body part), phantom limb syndrome (when an amputated arm is still labeled SELF), Cotard's syndrome (when SELF and FEEL are missing from the representation of oneself in the MotW), paranoia (when a friend is labeled SPY), ...

KM11. How do we become conscious of our own internal states? How much of our subjective experience arises from homeostatic control signals that necessarily have valence? If such signals entail feelings, how do we know what those feelings are about?

How do we become conscious of our own internal states? CTM becomes conscious of its own internal state when and only when it becomes globally broadcast. For example, in section 2.3.1.2, we indicated how the newborn CTM would know it needs something (fuel) when a high |weight| negatively valenced chunk from a processor (its Fuel Gauge processor) reaches STM and is broadcast from it.

How much of our subjective experience arises from homeostatic control signals that necessarily have valence? In response to the reception of the negatively valenced LOW_FUEL chunk submitted by the Fuel Gauge processor (a homeostatic processor), the sketch of CTM in the MotW will be labeled HUNGRY. The broadcast and inspection of a negatively valenced chunk with this Brainish-labeled sketch will indicate that "CTM feels hungry". An actuator will eventually connect CTM's fuel intake to a Fuel Source. Assuming the fuel transfer is successful, the Fuel Source sketch in the MotW will eventually be labeled RELIEVES_HUNGER ∘ PLEASURE_PROVIDER. The broadcast and inspection of a positively valenced chunk with this Brainish-labeled sketch indicates that "the Fuel Source is a hunger reliever and pleasure provider" and that "CTM experiences pleasure when it is hungry and getting fuel."

The above process is an example of homeostasis in CTM. In particular, this process is an example of how CTM's subjective experience arises from homeostatic control signals that have valence and how CTM knows what those feelings are about.

If such signals entail feelings, how do we know what those feelings are about? The about-ness of those feelings comes from the Brainish-labeled sketches. Suppose the sketch of CTM is labeled HUNGRY. Pleasure could then come from the sketch of CTM in the MotW being labeled INTAKES_FUEL ∘ FEELS_PLEASURE. See also our answer to the next question.

KM12. How does the about-ness of conscious states (or subconscious states) arise? How does the system know what such states refer to when the states are all the system has access to?

How does the about-ness of conscious states (or subconscious states) arise? We define the ***conscious state*** at time **t** to be the chunk received by all processors at that time. So conscious states are chunks that are received simultaneously by all processors.

Suppose c**hunk1** (received by all processors at time **t1**) contains a gist describing the world that CTM faces from the top of a down-staircase:

Suppose **chunk2** (received at time **t2**) predicts the effect of starting down the stairs. This prediction is done by the MotWp simulating that action[86] (in the MotW). **Chunk3** (received at time **t3**) then sends a command to CTM's walk actuator to go down the stairs. **Chunk4** (received at time **t4**) reports back.

That original starting state (chunk1), the prediction state (chunk2), the action state (chunk3), and the report state (chunk4) provide the desired (conscious) about-ness for the experience.

While CTM is not conscious of unconscious or subconscious states, they nevertheless can influence conscious states. As for subconscious states, they could have been conscious but for the luck of the draw, and their variants could still become conscious. As for unconscious states, they determine what chunks get submitted to the competition.

How does the system know what such states refer to when the states are all that the system has access to? The CTM only has access to *conscious states*, broadcasted chunks received by all processors. Such a chunk's gist is a Brainish description of what CTM knows of its world. If CTM *inspects* this chunk, i.e., becomes *consciously aware* of this state, CTM will know what the state refers to.

KM13. What is the point of conscious subjective experience? Or of a high level common space for conscious deliberation? Or of reflective capacities for metacognition? What adaptive value do these capacities have?

What is the point of conscious subjective experience? Without it, CTM is not compelled to act appropriately. The agony of pain, for example, is a consequence of each and every processor being forced to attend to the pain for a length of time proportional to the |weight| of the current "pain" chunk. A CTM that has pain knowledge, *knows* how its cause can damage or destroy it, but is not consciously aware of pain (i.e., does not subjectively experience it), is not compelled to attend to it, is

[86] This is similar to the kind of simulation that the MotW does in a dream sequence.

not motivated to respond normally to the pain, to take care of itself. Such a person or robot or CTM that lacks the ability to *feel* the agony of pain has *pain asymbolia.*

What is the point of a high level common space for conscious deliberation? Although it is just a buffer and broadcast station, STM can be viewed as a high level space: only one processor can write in it at any time (i.e., at each clock tick, only one chunk can win the competition); every processor has a chance to write in it. Moreover, all processors can simultaneously read that writing. The point of global broadcasting in CTM, is to focus all processors on the same thought (chunk). With the reception of a broadcast, all processors have an opportunity to contribute to its understanding and possibly its solution.

For example, suppose the broadcast is a high |weight| negatively valenced chunk, a "scream" indicating the problem: "Hungry! Need to fill the fuel tank." To deal with this situation, the Navigation processor might contribute a choice of routes to local fuel stations. The Computation processor might compute how much time is required for each route. A Calendar processor might suggest calling ahead to warn it will be late.

What is the point of reflective capacities for metacognition? Reflective capacities enable CTM to treat itself with all the planning tools it uses to treat other referents.

What adaptive value do these capacities have? They enable CTM to adapt. For example, how does CTM adapt to hot weather? Suppose the broadcast is: "It's too hot: what's a good way to cool down?" One processor suggests: "Find a cool place to go to." Another suggests: "Get and drink some water." Different processors have different ideas what to do. The competition will select one of the suggestions for broadcast. An actuator can then carry out the suggestion.

KM14. How does mentality arise at all? When do information processing and computation or just the flow of states through a dynamical system become elements of cognition and why are only some elements of cognition part of conscious experience?

How does mentality arise at all? We restate this as: **How does the capacity for intelligent thought arise?** While all 10 million processors are independent and unlinked at CTM's birth, the Up-Tree competition enables and encourages processors with different abilities and ideas, related or not, to work together to solve problems. The whole can be more than the sum of its parts. This makes for creative intelligent thinking and problem solving. See, section 2.4 where we discuss CTM as a framework for Artificial General Intelligence (AGI).

When do information processing and computation, or just the flow of states through a dynamical system, become elements of cognition? Viewing CTM as a dynamical system, the process of competition, selection of a winner, and reception of the winning chunk by all LTM processors (resulting in conscious attention to it) is an element of cognition. If all processors *inspect* the received chunk, then CTM becomes consciously aware of the information, another element of cognition.

Why are only some elements of cognition part of conscious experience? Some elements of cognition can be performed by LTM processors with their previously learned algorithms that do not need the variety of ideas broadcasted from STM. Such elements of cognition are unconscious. Processors that *do* need to search for information – or for a computational capability - can broadcast their need.[87] That broadcast to all processors is an essential ingredient of conscious cognition.

KM15. How does conscious activity influence behavior? Does a capacity for conscious cognitive control equal "free will"? How is mental causation even supposed to work? How can the meaning of mental states constrain the activities of neural circuits?

How does conscious activity influence behavior? As an example, conscious attention to a chunk of sufficiently high |weight| and negative valence, interrupts the work of all processors and forces them to pay maximum attention to the chunk. This happens, for example, in a fight, flight, or freeze confrontation. As another example, by bringing together seemingly unrelated ideas, conscious activity can generate wildly creative approaches for solving problems.

Does a capacity for conscious cognitive control equal "free will"? CTM's ability to assess a situation, meaning to consider various possible actions and predict the consequences of each, given the resources it has, gives CTM reason for it to believe it has "free will". To see this, imagine CTM playing a game of chess. When and while CTM has to decide which of several possible moves to make, it knows it will choose - is free to choose - whichever move has the greatest utility for it. That is "free will". (Blum & Blum, 2022)

How is mental causation even supposed to work? Mental causation is what we have called the *Power of Thought* (section 2.3.1.3). In CTM, the MotWp and the MotW are fundamental to *mental causation*. To decide whether to act in the world, an LTM processor (part of the MotWp) simulates that action in the MotW, and from that decides whether or not to go ahead with it. If it does, then CTM looks to see if the requested action got accomplished as predicted. (If not, that LTM processor corrects/improves the prediction algorithm.)

How can the meaning of mental states constrain the activities of neural circuits? As a first example, previously noted, sufficiently high |weight|-ed negatively valenced chunks cause all processors to put their current activity on hold and (to the extent possible) work to lower the |weight|. As another

[87] Like the processor that asks, "What her name?" (See footnote in section 2.1.2).

example of how mental states can constrain activities, our answer to KM4 discusses how the Sleep processor can generate a (non-dreaming) sleep state by raising its own |weight| so high that other chunks can't enter STM. This shows how the sleep state *constrains* CTM's activity.

***

We've now come to the end of Kevin Mitchell's questions. He ends his blog with the words, **"If we had a theory that could accommodate all those elements and provide some coherent *framework*[88] in which they could be related to each other – not for providing all the answers but just for asking sensible questions – well, that would be a theory of consciousness."** (Mitchell, 2023) This is our goal for the theory of the Conscious Turing Machine.

[88] Italics ours.

References

Anderson, J. R. (1996). ACT: A simple theory of complex cognition. *American Psychologist, 51(4)*, 355-365.

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

Baars, Bernard J. (1997). In the Theater of Consciousness: A rigorous scientific theory of consciousness. *Journal of Consciousness Studies, 4*(4), 292-309.

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. A. Bower (Ed.), *The Psychology of Learning and Motivation* (pp. 47-89). New York: Academic Press.

Bancu, P., Bisnath, Burgess, Eakins, Gonzales, Saltzman, . . . Baptista. (2024). Revitalizing Attitudes Toward Creole Languages I. In A. H. Hudley, C. Mallinson, & M. Bucholtz, *Decolonizing Linguistic.* Oxford University Press.

Berridge, K. C., & Kringelbach, M. L. (2015). Pleasure systems in the brain. *Neuron, 86*(3), 646-664. doi: 10.1016/j.neuron.2015.02.018.

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

Blum, A. (1995, July). Empirical support for winnow and weighted-majority algorithms: Results on a calendar scheduling domain. *Proceedings of the Twelfth International Conference on Machine Learning*, (pp. 64-72. https://doi.org/10.1016/B978-1-55860-377-6.50017-7).

Blum, A., Hopcroft, J., & Kannan, R. (2015). *Foundations of Data Science.* Ithaca. Retrieved from https://www.cs.cornell.edu/jeh/book.pdf

Blum, L., & Blum, M. (2022, May 24). A theory of consciousness from a theoretical computer science perspective: Insights from the Conscious Turing Machine. *PNAS, 119*(21), https://doi.org/10.1073/pnas.21159341.

Blum, L., & Blum, M. (2023). A Theoretical Computer Science Perspective on Consciousness and Artificial General Intelligence. *Engineering, 25*(5), 12-16. https://doi.org/10.1016/j.eng.2023.03.010.

Blum, M., & Blum, L. (2021, March). A Theoretical Computer Science Perspective on Consciousness. *JAIC, 8*(1), 1-42. https://doi.org/10.1142/S2705078521500028.

Blum, M., & Blum, L. (2022, December 24). *A Theoretical Computer Science Perspective on Free Will.* Retrieved from ArXiv: https://doi.org/10.48550/arXiv.2206.13942

Blum, M., Blum, L., & Blum, A. (2026). *The Conscious Turing Machine: A Theoretical Computer Science Approach to Consciousness and ...* CUP.

Brown, R., Lau, H., & LeDoux, J. (2019, September). Understanding the higher-order approach to consciousness. *Trends Cogn. Sci. 23, 754–768, 23*(9), 754–768. doi: 10.1016/j.tics.2019.06.009.

Carney, J. (2020). Thinking avant la lettre: A Review of 4E Cognition. *Evolutionary Studies in Imaginative Culture, 4*(1), 77-90. https://doi.org/10.26613/esic.4.1.172.

Carroll, L. (1871). *Through the Looking Glass.*

Chalmers, D. J. (1995). Facing Up to the Problem of Consciousness. *Journal of Consciousness Studies, 2*(3), 200-219.

Chella, A., Pipitone, A., Morin, A., & Racy, F. (2020, February). Developing Self-Awareness in Robots via Inner Speech . *Frontiers in Robotics and AI, 7*. Retrieved from https://www.frontiersin.org/article/10.3389/frobt.2020.00016

Church, A. (1936). A note on the Entscheidungsproblem. *J. of Symbolic Logic, 1 (1936), 40-41, 1*, 40-41.

Clark, A. (2008, Jan). Pressing the Flesh: A Tension in the Study of the Embodied, Embedded Mind? . *JPhilosophy and Phenomenological Research, 76*(1), 37-59. https://www.jstor.org/stable/40041151.

Clark, A. (2015). Embodied prediction. In T. Metzinger, & J. Windt, *Open Mind.* Frankfurt am Main: MIND Group.

Clark, A., & Chalmers, D. (1998, January). The Extended Mind. *Analysis, 58*(a), 7-19.

Cleeremans, A. (2011). The radical plasticity thesis: how the brain learns to be conscious. *Frontiers in psychology*, 86.

Cleeremans, A. (2014). Prediction as a computational correlate of consciousness. *International Journal of Anticipatory Computing Systems, 29*, 3-13.

Cleeremans, A., Achoui, D., Beauny, A., Keuninckx, L., Martin, J.-R., Muñoz-Moldes, S., . . . de Heering, A. (2020). Learning to Be Conscious. *Trends in Cognitive Sciences, 24*(2), 112-123. https://doi.org/10.1016/j.tics.2019.1.

Cook, S. A. (1971). The complexity of theorem-proving procedures. *Proceedings of the third annual ACM symposium on Theory of computing*, (pp. 151-158. https://doi.org/10.1145/800157.805047).

Damasio, A. (1994). *The Feeling of What Happens.* NY, NY: Harcourt, Brace and Co,.

Dehaene, S. (2014). *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts.* New York: Viking Press.

Dehaene, S., & Changeux, J. P. (2005, April 12). Ongoing Spontaneous Activity Controls Access to Consciousness: A Neuronal Model for Inattentional Blindness. *PLoS Biol, 3*(5), https://doi.org/10.1371/journal.pbio.0030141.

Dehaene, S., & Changeux, J. P. (2011, April 28). Experimental and theoretical approaches to conscious processing. *Neuron, 70*(2), 200-227. DOI: 10.1016/j.neuron.2011.03.018.

Dehaene, S., & Naccache, L. (2001, April). Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition, 79*(1-2), 1-37. https://doi.org/10.1016/S0010-0277(00)00123-2.

Dehaene, S., Charles, L., King, J.-R., & Marti, S. (2014). Toward a computational theory of conscious processing. *Current Opinion in Neurobiology, 25*, 76-84. https://doi.org/10.1016/j.conb.2013.12.005.

Dehaene, S., Lau, H., & Kouider, S. (2017, Oct 27). What is consciousness, and could machines have it? *Science, 58*(6362), 486-492. doi: 10.1126/science.aan8871.

Dennett, D. C. (1991). *Consciousness Explained.* Boston; Toronto; London: Little, Brown and Co.

Dennett, D. C. (2019, December). Consciousness, Qualia and the "Hard Problem". (L. Godbout, Interviewer) Retrieved from https://youtu.be/eSaEjLZIDqc, starting time for quote 5:40

Duncan, J. (2025). Construction and use of mental models: Organizing principles for the science of brain and mind. *Neuropsychologia, 207*, https://doi.org/10.1016/j.neuropsychologia.2024.109062.

Edelman, G. M. (2006, Summer). The Embodiment of Mind. *Daedalus, 135*(3), 23-32. https://www.jstor.org/stable/20028049.

Edmonds, J. (1965). Paths, trees, and flowers. *Can. J. Math., 17*, 449–467. doi:10.4153/CJM-1965-045-4.

Elias, A., Thomas, N., & Sackeim, H. A. (2021). Electroconvulsive Therapy in Mania: A Review of 80 Years of Clinical Experience. *American Journal of Psychiatry, 178*(3), 229-239. doi: 10.1176/appi.ajp.2020.20030238.

Farisco, M., Evers, K., & Changeux, J.-P. (2024, December). Is artificial consciousness achievable? Lessons from the human brain. *Neural Networks, 180*, 1-14. https://doi.org/10.1016/j.neunet.2024.106714. Retrieved June 2024, from arXiv: https://arxiv.org/abs/2405.04540

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

Frankish, K. (2023, October 21). Retrieved from X.

Frankish, K. (2023, April 15). Retrieved from X.

Frankish, K. (2024, November 27). *How to think about artificial consciousness* . Retrieved from the Meta Lab, Consciousness Club, the Meta Lab, UCL: https://metacoglab.org/consciousness-club-events/2024/11/27/keith-frankish

Freud, S. (1966). *Lectures on Psychoanalysis.* New York: Liveright 1.

Friston, K. (2005, April 29). A theory of cortical responses. *Phil. Trans. R. Soc. B, 360*, 815-836. doi:10.1098/rstb.2005.1622.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature reviews neuroscience, 11*(2), 127-138. https://doi.org/10.1038/nrn2787.

Gerrans, P. (2024). Pain suffering and the self. An active allostatic inference explanation. *Neuroscience of Consciousness, 2024*(1), https://doi.org/10.1093/nc/niae002.

Gershman, S. J., Fiete, I., & Irie, K. (2025, January 7). *Key-value memory in the brain.* Retrieved January 2025, from arXiv: https://doi.org/10.48550/arXiv.2501.02950

Ginsburg, S., & Jablonka, E. (2019). *The Evolution of the Sensitive Soul.* Cambridge, Massachusetts, US: MIT Press.

Gopnik, A. (2007, December). Why babies are more conscious than we are. *Behavioral and Brain Sciences, 30*(5-6), 503-504. https://doi.org/10.1017/S0140525X0700283X.

Grahek, N. (2001; 2007). *Feeling Pain and Being im Pain.* Universitat Oldenburg; MIT Press (2nd edition).

Graziano, M. S., Guterstam, A., Bio, B., & Wilterson, A. (2020, May-June). Toward a standard model of consciousness: Reconciling the attention schema, global workspace, higher-order thought, and illusionist theories. *Cognitive Neuropsychology, 37(3-4)*(3-4), 155-172. Retrieved from doi:10.1080/02643294.2019.1670630

He, B. J. (2023, May). Towards a pluralistic neurobiological understanding of consciousness. *Trends in Cognitive Sciences, 27*(5), https://doi.org/10.1016/j.tics.2023.02.001.

Hohwy, J., & Seth, A. (2020). Predictive processing as a systematic basis for identifying the neural correlates of consciousness (preprint). *PsyArXiv* , psyarxiv.com/nd82g.

Humphrey, N. (2023). *Sentience: The Invention of Consciousness.* Cambridge, Massachusetts, US: MIT Press.

Karp, R. M. (1972). *Reducibility Among Combinatorial Problems.* (R. E. Miller, & J. W. Thatcher, Eds.) New York: Plenum.

Klein, C. (2015, April). What Pain Asymbolia Really Shows. *Mind, 124*(494), 493-516. https://www.jstor.org/stable/24490440.

LeDoux, J., & Brown, R. (.-o.–E. (2017). A higher-order theory of emotional consciousness. *Proc. Natl. Acad. Sci. , 114*(10), https://doi.org/10.1073/pnas.161931611.

Lee, A. Y. (2023, September). Degree of Consciousness. *Noûs , 57*(3), 553-575. https://doi.org/10.1111/nous.12421.

Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America, Optics, image science and vision, 20*(7), 1434-1448.

Lenharo, M. (2024, January 18). Consciousness - The future of an embattled field (The consciousness wars: can scientists ever agree on how the mind works?). *Nature, 625*, 438- 440. doi:10.1038/d41586-024-00107-7.

Levin, L. A. (1973). Universal Sequential Search Problems. *Probl. Peredachi Inf., 9*(3), 115-116.

Levine, J. (1983). Materialism and Qualia: The Explanatory Gap. *Pacific Philosophical Quarterly, 64*, 354–361.

Lewis, C. I. (1929). *Mind and the world-order: Outline of a theory of knowledge.* New York: Charles Scribner's Sons.

Li, F.-F. (2023). *The Worlds I See.* New York: Flatiron Books: A Moment of Lift Book.

Liang, P. P. (2022, April 14). *Brainish: Formalizing A Multimodal Language for Intelligence and Consciousness.* Retrieved from ArXiv.

Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences, 8*(4), 529-539.

Mashour, G. A., Roelfsema, P., Changeux, J.-P., & Dehaene, S. C. (2020, March 4). Conscious Processing and the Global Neuronal Workspace Hypothesis. *Neuron, 195*(5), 776-798. doi: 10.1016/j.neuron.2020.01.026.

Maturana, H., & Varela, F. (1972). *De Maquinas y Seres Vivos, Autopoiesis: La organizacion de lo vivo.* Santiago de Chile: Editorial Universitaria, S. A.

Maturana, H., & Varela, F. (1980). *Autopoiesis and Cognition: The Realization of the Living.* Boston: Reidel.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review, 88*(5), 375-407. https://doi.org/10.1037/0033-295X.88.5.375.

McWhorter, J. H. (1998, December). Identifying the Creole Prototype: Vindicating a Typological Class . *Language* , 788-818. https://www.jstor.org/stable/417.

McWhorter, J. H. (2008). *Defining Creole .* USA: Oxford University Press.

Miller, G. A. (1956). The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information. *Psychological Review, 63*(2), 81-97. https://doi.org/10.1037/h0043158.

Miller, M., Clark, A., & Schlicht, T. (2022). Editorial: Predictive Processing and Consciousness. *Rev.Phil.Psych., 13*, 797-808. https://doi.org/10.1007/s13164-022-00666-6.

Mitchell, K. (2023, September 18). *What questions should a real theory of consciousness encompass?* . Retrieved February 2024, from Wiring the Brain: http://www.wiringthebrain.com/2023/09/what-questions-should-real-theory-of.html

Naccache, L. (2018). Why and how access consciousness can account for phenomenal consciousness. Phil. Trans. R. Soc.B 373: 20170357.http://dx.doi.org/10.1098/rstb.2017.0357. *Phil. Trans. R. Soc. B*, https://doi.org/10.1098/rstb.2017.0357.

Nagel, T. (1974, October). What Is It Like To Be a Bat? *Philosophical Review, 83*(4), 435–450. https://doi.org/10.2307/2183914.

Nagel, T. (1974). What Is It Like To Be a Bat? *Philosophical Review, 83*, 435–450.

Newell, A. (1990). *Unified Theories of Cognition.* Cambridge: Harvard University Press.

Parr, T., Da Costa, L., & Friston, K. 2. (2019). Markov blankets, information geometry and stochastic thermodynamics. *Phil.Trans.R. Soc.*, http://dx.doi.org/10.1098/rsta.2019.0159.

Peirce, C. S. (1866). "Lowell lecture, ix.". In M. H. Fisch, & I. I.-1.-8. Ed. Bloomington, *Writings of Charles S. Peirce: A chronological edition I, 1857-1866* (Vol. i, pp. 471-486.). Bloomington, Indiana: Indiana University Press.

Reddy, D. R. (1976, April). Speech Rcogniton by Machine: A Review. *Proceedings of the IEEE*, 501-531. Retrieved from http://www.rr.cs.cmu.edu/sr.pdf

Rowlands, M. J. (2010). *The New Science of the Mind From Extended Mind to Embodied Phenomenology.* Cambridge, MA, US: The MIT Press.

Sergent, C., & Dehaene, S. (2005, July-November). Neural processes underlying conscious perception: experimental findings and a global neuronal workspace framework. *J Physiol Paris, 98*(4-6), 374-384. doi: 10.1016/j.jphysparis.2005.09.006.

Seth, A. (2024, December 10). Conscious artificial intelligence and biological naturalism. *PsyArXiv*. Retrieved December 2024, from Conscious artificial intelligence and biological naturalism, submi5ed for review, 08/12/2024 1 https://doi.org/10.31234/osf.io/tz6an PsyArXiv Preprints : https://doi.org/10.31234/osf.io/tz6an PsyArXiv Preprints

Seth, A. K. (2015). The Cybernetic Bayesian Brain - From Interoceptive Inference to Sensorimotor Contingencies. In T. Metzinger, & J. M. Windt, *Open MIND.* Frankfurt am Main: MIND Group.

Shanahan, M. (2005). Global Access, Embodiment, and the Conscious Subject. *Journal of Consciousness Studies, 12*(12), 46-66.

Shanahan, M. (2010). *Embodiment and the inner life: Cognition and Consciousness in the Space of Possible Minds.* Oxford University Press.

Sierra, M. (2009). *Depersonalization: A New Look at A Neglected Syndrome.* Cambridge, England: Cambridge University Press, 2009.

Sigal, U.-K. (2022, May 4). Editorial: Simple and Simplified Languages. *Frontiers in Communication , 7*, https://doi.org/10.3389/fcomm.2022.910680.

Simon, H. A. (1969). *The Sciences of the Artificial.* Cambridge, MA, USA: MIT Press.

Solms, M. (2019). The Hard Problem of Consciousness and the Free Energy Principle. *Front. Psychol. , 9*(2714), doi: 10.3389/fpsyg.2018.02714.

Solms, M. (2021). *The Hidden Spring: A Journey to the Source of Consciousness.* New York, NY, US: W. W. Norton and Company.

Solms, M., & Friston, K. (2018). How and Why Consciousness Arises: Some Considerations from Physics and Physiology. Journal of Consciousness Studies 25 (5-6):202-238. *Journal of Consciousness Studies, 25*(5-6), 202-238.

Storm, J., Klink, P., Aru, J., Senn, W., Goebel, R., Pigorini, A., . . . Pennartz, C. (2024, May 15). An integrative, multiscale view on neural theories of consciousness. *Neuron, 112*(10), 1531-1552. doi: 10.1016/j.neuron.2024.02.004.

Tamietto, M., & Morrone, M. (2016, Jan 25). Visual Plasticity: Blindsight Bridges Anatomy and Function in the Visual System. *Current Biology, 26*(2), 70-73. https://doi.org/10.1016/j.cub.2015.11.026.

Thompson, E. (2007). *Mind in Life: Biology, Phenomenology, and the Sciences of Mind.* Harvard University Press.

Tononi, G. (2004, November 2). An information integration theory of consciousness. *BMC Neuroscience, 5*(42), 42-72. doi: 10.1186/1471-2202-5-42.

Tononi, G., & Koch, C. (2015, May 19). Consciousness: here, there and everywhere? *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 370 (1668)*, https://doi.org/10.1098/rstb.2014.0167.

Turing, A. M. (1937). On Computable Numbers, with an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society, 2*(42), 230-265. https://doi.org/10.1112/plms/s2-42.1.230.

Valiant, L. G. (2024, December). *The Parameters of Educability.* Retrieved from arXiv: https://arxiv.org/abs/2412.09480

VanRullen, R., & Kanai, R. (2021, May 14). Deep learning and the Global Workspace Theory. *Trends in Neurosciences, 44*(9), 692-704. doi: 10.1016/j.tins.2021.04.005.

Varela, F., Thompson, E., & Rosch, E. (1991). *The Embodied Mind.* Camb, MA: MIT Press.

von Helmholtz, H. (1866; 1962). *Treatise on physiological optics* (Vol. 3). (J. Southall, Ed.) New York, NY: Dover Publication.

Wiese, W. T.-2. (2020, July 11). The science of consciousness does not need another theory, it needs a minimal unifying model. *Neuroscience of Consciousness, 2020*(1), https://doi.org/10.1093/nc/niaa013.

Wolfe, J. M. (1998). Visual memory: What do you know about what you saw? *8*(9), 303-304.

Yao, A. C. (1982). Theory and Applications of Trapdoor Functions. *Proc. 23rd Annual ACM Symposium on Theory of Computing* (pp. 80-91. doi: 10.1109/SFCS.1982.45). IEEE Computer Society.

Yildirim, I., & & Paul, L. A. (2024, March 4). From task structures to world models: what do LLMs know? *Trends in Cognitive Sciences.*

Zacks, O., & Jablonka, E. (2023, September 13). The evolutionary origins of the Global Neuronal Workspace in vertebrates. *Neuroscience of Consciousness*, https://doi.org/10.1093/nc/niad020.

Zadra, A., & Stickgold, R. (2021). *When Brains Dream.*