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. From this perspective, we develop a formal machine model for consciousness. The model is inspired by Alan Turing's simple yet powerful model of computation and Bernard Baars' theater model of consciousness. Though extremely simple, the model aligns at a high level with many of the major scientific theories of human and animal consciousness, supporting our claim that machine consciousness is inevitable.

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 the implications and surprising consequences of resource limitations.

By taking resource limitations into account, the TCS perspective is distinguished from the earlier Turing Theory of Computation (TOC) where limitations of time and space did not figure. TOC distinguishes computable from not computable. It does not distinguish between computable and not efficiently computable.[1] We highlight the importance of this separation for tackling consciousness and related topics such as the paradox of free will.

[1] For a brief history of TOC and TCS see Appendix 7.1.

Lenore Blum
lblum@cs.cmu.edu

Manuel Blum
mblum@cs.cmu.edu

AI Consciousness is Inevitable

Elsewhere, we describe the Conscious Turing Machine (CTM), a *simple formal machine model of consciousness* inspired in part by Alan Turing's *simple formal machine model of computation* (Turing, 1937), and by Bernard Baars' theater model of consciousness (Baars, Bernard J., 1997). See (Blum & Blum, 2021) and (Blum & Blum, 2022)[2]

In contrast to Turing, we take resource limitations into account, both in designing the CTM model and in how resource limitations affect (and help explain) feelings of consciousness. 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 chapter we take a brief look under the hood.

In addition, we show how the CTM naturally *aligns* with and *integrates* features considered key to human and animal consciousness by many of the major scientific theories of consciousness.[4] These theories consider different aspects of consciousness and often compete with each other (Lenharo, 2024). Yet their alignment with the CTM at a high level helps demonstrate their compatibility and/or complementarity.

But, *even more*, their *alignment* with the CTM, a simple machine model, supports **our claim that a conscious AI is inevitable**.

---

[2] In the 2022 paper, we considered 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.

[3] This is important. We claim that simulations that modify CTM's key internal structures and processes will not necessarily experience what CTM does. We are not claiming that the CTM is the only possible machine model to experience feelings of consciousness.

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

Lenore Blum
lblum@cs.cmu.edu

Manuel Blum
mblum@cs.cmu.edu

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 consciousness (Block, 1995) or functional consciousness, we call *conscious attention*. The second type (associated with the Hard Problem) is called subjective or phenomenological consciousness and is generally associated with qualia. We call it *conscious awareness*. Chalmers' Hard Problem can be viewed as a challenge to show that subjective consciousness is "functional".

We contend that consciousness writ large requires both conscious attention and conscious awareness, each interacting with the other to various degrees. We contend that a machine that interacts with its worlds (inner and outer) 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 language, can have both types of consciousness. In particular, we contend that subjective consciousness is functional.

We emphasize that the CTM is a formal computational model designed to explore and understand consciousness from a TCS perspective. It is not intended to model the brain nor the neural correlates of consciousness. Nevertheless, the CTM is inspired by cognitive and neuroscience theories of consciousness.

Specifically, as we have mentioned, the CTM is inspired by cognitive neuroscientist Bernard Baars' theater model of consciousness (Baars, Bernard J., 1997), the global workspace (GW) theory of consciousness. However, here again, the CTM is not a standard GW model. The CTM differs from GW in a number of important ways: its competition for global broadcast is formally defined, and completely replaces the ill-defined Central Executive of other GW models; its special processors including especially its Model-of-the-World processor construct and employ models of its (*inner* and *outer*) worlds; its rich multimodal internal language, *Brainish*, for creating labeled sketches in its world models and for communicating between processors; and its *predictive dynamics* (cycles of prediction, testing, feedback and learning, locally and globally).

The CTM also interacts with its outer world via input *sensors* and output *actuators*. To emphasize CTM's embodied, embedded, enacted and extended mind, we call it here the CTM Robot (**CtmR**).

While working on this chapter, 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. For a similar conclusion, see (Storm & et.al., 2024). Even more, Mitchell presents "a non-exhaustive list of 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*[5] (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."

Mitchell's questions are thoughtful, interesting, and important. At the end of this chapter, we offer preliminary answers from the perspective of the Conscious Turing Machine Robot (CtmR). Our answers both supplement and highlight material in the brief Overview of CtmR that we now present. [6]

2 Brief Overview of CtmR, a Robot with a CTM Brain

2.1 Formal Definition of CtmR

CtmR is defined formally as a 7-tuple, **(STM, LTM, Up-Tree, Down-Tree, Links, Input, Output**). The seven components each have well-defined properties. We indicate these properties here.

For CtmR, the stage in the theater model is represented by a Short Term Memory (**STM**) that at any moment in time contains CtmR's current *conscious content*. STM is not a processor; it is merely a buffer and broadcasting station. The **N** audience members are represented by a massive collection[7]

---

[5] Italics ours.

[6] In the Overview, we annotate paragraphs that refer to Kevin Mitchell's queries. As an example, if a paragraph has a label [KM1], then it refers to Mitchell's first query, KM1. Conversely, if Mitchell's query is labeled with an asterisk such as KM1*, then it refers to [KM1] in the Overview.

[7] We assume $\mathbf{N} \gtrsim \mathbf{10^7}$ LTM processors. Abstractly, these processors may be considered random access machines.

Lenore Blum
lblum@cs.cmu.edu

Manuel Blum
mblum@cs.cmu.edu

of initially independent powerful (*unconscious*) processors that comprise its computational machinery and Long Term Memory, together called **LTM**. These processors compete to get their information on stage to be immediately broadcast to the audience.[8,9]

CtmR has a finite lifetime $\mathbf{T}$.[10] At time $\mathbf{t = 0}$, all but the input and output LTM processors are "generic" with certain basic built-in properties, e.g., some learning/prediction correction algorithms, as well as a preference for choosing the positive over the negative.[11] Their functionalities evolve over time.

But for now, we designate some important LTM processors built in. These include: a *Model-of-the-World processor* (MotWp), actually a collection of processors, for building models of CtmR's inner and outer worlds; *Sensation* processors (with **input** from CtmR's outer world via its various

---

[8] See Appendix 7.2 for information about CtmR's competition.

[9] As an example of the theater analogy, 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 we are home, her name pops into our head (unfortunately too late). What's going on?

Racking our brain to remember caused the urgent request "What's her name?" to rise to the stage (STM) which in turn was immediately broadcast to the audience.

Some (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**.

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

(Based on the correct outcome, internal learning algorithms cause processor **p** to lower the importance (|weight|) it gives its information and cause **p''** to increase the importance.)

[10] For the general CtmR theory, both **T** (lifetime) and **N** (number of LTM processors) are parameters. Time $\mathbf{t = 0, 1, 2, 3, \ldots, T}$ is measured in discrete clock ticks.

[11] This built-in preference creates a predilection for survival.

Lenore Blum
lblum@cs.cmu.edu

Manuel Blum
mblum@cs.cmu.edu

sensors[12]); *Motor* processors (with **output** to CtmR's outer world via motor *actuators*[13]); and so on. We also allow off-the-shelf processors (like ChatGPT and Google) that *extend* CtmR's capabilities.

All processors are in LTM so when we speak of a processor, we mean an LTM processor. While each processor may have its own distinct language, processors communicate within CtmR in *Brainish*, Brainish being CtmR's rich multimodal inner language. Brainish words, *gists*, and phrases fuse sensory modalities (e.g., sight, sounds, smells, tactile) as well as dynamical processes.[14] The Brainish language evolves over CtmR's lifetime.[15]

LTM processors *compete* in a well-defined (*fast* and natural) probabilistic *competition* to get their questions, answers, and information onto the stage (STM).[16] The competition is hosted by the **Up-Tree,** a *perfect* binary tree[17] of height **h** which has a leaf in each LTM processor and 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.

A chunk is defined formally to be a tuple, **<pointer, time, *gist*, weight, auxiliary information>**, consisting of a *succinct* Brainish gist of information, a valenced weight (to indicate the importance/value/confidence the originating processor assigns the gist), a pointer to the originating processor, and the time the chunk was created, plus some auxiliary information.

Each chunk competes locally with its neighbor and a variation of the local winner moves up one level of the Up-Tree in one clock tick, now to compete with its new neighbor. The local winner is

---

[12] Ears, eyes, nose, skin, … .

[13] Arms, hands, legs, … .

[14] A succinct Brainish gist is like a frame in a dream.

[15] Paul Liang is developing a computational framework for Brainish (Liang, 2022).

[16] See Appendix 7.2 for information about CtmR's probabilistic competition.

[17] A *perfect* binary tree is a binary tree in which all leaf nodes are at the same depth. This depth is also the height of the tree. If **h** is the height of a perfect binary tree, then the tree has $\mathbf{N = 2^h}$ leaves. For simplicity, we choose a perfect binary tree in this chapter.

Lenore Blum
lblum@cs.cmu.edu

Manuel Blum
mblum@cs.cmu.edu

chosen probabilistically by a fast built-in competition function. The process continues until a chunk reaches the Up-Tree root node in STM. The competition takes **h** clock ticks.

Notable about CtmR's competition is that the winner is independent of its submitting processor's location.

The chunk that gets onto the stage, i.e., the winning chunk, is called CtmR's current *conscious content* and is immediately *globally broadcast* (in one clock tick) via the **Down-Tree** (a bush of height **1** with a root in STM and **N** branches, one leaf in each LTM processor) to the audience (of all LTM processors). [KM2] [KM5]

The single chunk in STM to be globally broadcast will enable CtmR to focus attention on the winning gist. One is not the "magical number" 7±2 proposed by George Miller (Miller G. A., 1956), but we are looking for simplicity and one chunk will do.

2.2 Conscious Attention in CtmR

Formal definition. *Conscious attention* in CtmR is the *reception* by all LTM processors of the broadcast of CtmR's current conscious content. [KM2] [KM5]

LTM processors make *predictions* and get *feedback* from CtmR's inner and outer worlds. Based on this feedback, learning algorithms internal to each processor improve that processor's behavior. These learning algorithms include each processor's built-in *Sleeping Experts Learning* algorithm that helps the processor adjust weights it gives its gist. See (Blum A. , 1995) and (Blum, Hopcroft, & Kannan, 2015). [KM7]

Thus, we can already see some basic *predictive dynamics* (prediction + testing + feedback + learning/correction) occurring within CtmR. [KM2]

In time, some LTM processors become connected via bi-directional **links.**[18] Such links enable *conscious communication* through STM to be replaced by more direct *unconscious communication*

[18] The CtmR has no links at birth.

Lenore Blum
lblum@cs.cmu.edu

Manuel Blum
mblum@cs.cmu.edu

through links. Thus, when CtmR 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 CtmR to pay conscious attention again. [KM6]

CtmR's competition, broadcast, attention and immediate direct communication via links is reminiscent of a process that Dehaene and Jean-Pierre Changeux call *ignition* (Dehaene & Changeux, 2005).

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

2.3 Conscious Awareness and the Feeling of Consciousness in CtmR

Importantly, CtmR's *Model-of-the-World processor* (MotWp) constructs models of CtmR's inner and outer worlds that are collectively called the *Model-of-the-World* (MotW). The MotW has CtmR's current and continuing view of CtmR's world. [KM1] [KM3]

The MotWp, collaborating with other processors, plays an important role in planning, predicting, exploring, testing and correcting/learning. The models it constructs contain *sketches* of referents in CtmR's inner and outer worlds. Sketches are labeled with succinct *Brainish gists.* These labels indicate what CtmR *learns* or "thinks" about those referents over time. In particular, the sketch whose referent is CtmR itself will develop from scratch and eventually be labeled with SELF and CONSCIOUS.[19] [KM1] [KM3] [KM9]

Both the Model-of-the-World and Brainish evolve over time and play an essential role in the feeling of "what it is like" to be a CtmR. Thomas Nagel's "what it is like" (Nagel, 1974) is often taken to be the canonical definition of phenomenological consciousness. [KM1] [KM3] [KM9]

---

[19] We are often asked, isn't this process recursive? Doesn't the sketch of CtmR have a sketch of CtmR have a sketch of CtmR, etc. ? Yes, up to a point. But, at each iteration the current sketch is degraded, so the process soon becomes null.

Lenore Blum
lblum@cs.cmu.edu

Manuel Blum
mblum@cs.cmu.edu

The infant CtmR has only a very foggy MotW that does not even include a sketch of itself. Sketches in the MotW develop gradually, become refined, and gather labels. For example, when the infant CtmR discovers it can move its left leg actuator by the "power of thought", the MotWp labels the sketch of that leg actuator in the MotW as SELF.[20] [KM3] [KM16]

Even earlier, "feelings" of pain and pleasure start to develop. When the infant CtmR's fuel gauge gets low, some sketch (which becomes the sketch of the fuel gauge) in the MotW gets labeled with the Brainish word LOW FUEL/PAIN (or HUNGER) and this information with a large negatively valenced weight wins the competition and gets globally broadcast. This information triggers a processor to activate the fuel pump processor. The infant CtmR learns that the fuel pump relieves the pain when the fuel gauge indicates "low fuel" (hunger). The "fuel pump" in the MotW is labeled PAIN RELIEVER, and may also get labeled PLEASURE PROVIDER.[21] [KM3][KM12]

Formal definition. *Conscious awareness* and *feelings of consciousness* in CtmR are consequences of Brainish-labeled sketches being globally broadcast.[22,23] [KM2]

---

[20] Certain pathologies will occur if a breakdown in CtmR causes its MotWp to mislabel. For example, if the sketch of that leg actuator gets labeled NOT-SELF at some point, CtmR might beg to get its leg amputated, even if it is a still functioning properly. This would be an example of CtmR body integrity dysphoria. Other pathologies due to faulty labeling in the MotW: CtmR phantom limb syndrome (a sketch of an amputated arm actuator is mislabeled SELF), Cotard's syndrome (the sketch of CtmR is labeled DEAD), paranoia (sketch of CtmR's best friend is labeled SPY), … . [K10]

[21] 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 PAIN RELIEVER and PLEASURE.

[22] We call a sequence of such broadcasted chunks, a *stream of consciousness*. CtmR *dreams* are such streams during which the input sensors and output actuators are essentially inactive, and CtmR's *Dream* processor gets to work. Although dreams are "felt" as real, they can also be fantastical since their predictions are not being tested in the world. We propose that (testing for) *dreaming* is a (partial) *test for consciousness*.

[23] Is the infant consciously aware? An infant CTM's memory of an instance doesn't even have the label SELF associated with its MotW sketch, but soon enough that recollection will get labeled SELF, after which future recollections will have itself in that memory. [KM2], [KM3]

Lenore Blum
lblum@cs.cmu.edu

Manuel Blum
mblum@cs.cmu.edu

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

Before we summarize CtmR's alignment with a number of major theories of consciousness, we remark on CtmR's potential to serve as a framework for constructing an Artificial General Intelligence (AGI). This is a result of CtmR'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 CtmR and Baars' global workspace. CtmR has *no* Central Executive. This is a feature, not a bug. It enables CtmR to become an AGI (Blum & Blum, 2023):

The competition to get information on stage considers the |weight|ed information submitted by a huge collection of (**N**) processors. And it does this quickly ($\log_2$ **N** steps). This enables CtmR to engage processors to solve problems, even though CtmR does not know which of its processors might have the interest, expertise, or time to do so. [KM15]

Specifically, if CtmR has a problem to solve, meaning that one (or several) of its LTM processors has a problem, the processor can submit the problem into the competition as a highly |weight|ed chunk. If no other chunks are highly |weight|ed, the chunk wins with high probability and becomes globally broadcast to all processors. Processors with the interest, expertise and time to work on the problem will respond with appropriately |weight|ed chunks. In this way, unexpected ideas (from unexpected sources) may turn out to have relevance to solving the problem, and useful collaborations can emerge. [KM15]

A Central Executive would have to know which processors had the inclination, expertise, and resources to solve problems as they arise, and figure this out quickly. Baars' does not say how a Central Executive could do this. [KM15]

More generally, we predict that a Central Executive is not needed for consciousness or for general intelligence; indeed, it might prove to be an impediment.

Lenore Blum
lblum@cs.cmu.edu

Manuel Blum
mblum@cs.cmu.edu

3 Alignment of CtmR with Other Theories of Consciousness

We have presented an overview of the CtmR model. Now we indicate how, at a high level, the model naturally aligns with and integrates key features of major theories of consciousness, supporting also our view that CtmR provides a framework for building a conscious machine.

3.1 Global Workspace (GW)/Global Neuronal Workspace (GNW)

CtmR aligns broadly with the architectural and global broadcasting features of the *global workspace* 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.[24]

CtmR differs from GW in significant ways. For example: CtmR has a formally defined natural competition for information to become globally broadcast; it constructs world models; and CtmR has no Central Executive, a feature, not a bug.

3.2 Attention Schema Theory (AST)

CtmR's ability to construct and utilize models of CtmR's worlds (inner and outer) and the key role they play in CtmR's *conscious awareness* aligns closely with Michael Graziano's *attention schema theory* of consciousness (Graziano, Guterstam, Bio, & Wilterson, 2020). AST proposes that the brain is an information processing machine that constructs a simplified model of attention, like it constructs a simplified model of the body, the Body Schema. This Attention Schema, by providing a rough but adequate description of what it is attending to, causes the brain to conclude that it is "aware".

---

[24] [Other references include: (Dehaene & Naccache, 2001), (Sergent & Dehaene, 2005), (Dehaene & Changeux, 2011), and (Mashour, Roelfsema, Changeux, & Dehaene, 2020).]

Lenore Blum
lblum@cs.cmu.edu

Manuel Blum
mblum@cs.cmu.edu

3.3 Predictive Processing (PP)

Predictive processing asserts that the brain is constantly inferring, correcting and updating its predictions, generally based on sensory inputs. CtmR'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, 2014), (Clark A. , 2015), and (Hohwy & Seth, 2020) and others.[25]

3.4 Embodied Embedded Enactive Extended Mind (EEEEM)

CtmR's phenomenal consciousness and, in particular, its ability to construct and utilize models of its worlds, derives in part from a mix of its embodied, embedded, enactive, and extended mind: CtmR is *embedded* in its outer world and, through its *embodied* actuators, can *enact* in this world, thus influencing what it senses and experiences. CtmR's mind is *extended* by information it gleans from resources in its outer world, and from its embedded (or linked) off-the-shelf processors.

This aligns with the "4E" view that consciousness, like cognition (Carney, 2020), also involves more than brain function (Rowlands, 2010).

- Embodied: Phenomenal consciousness involves incorporating relations with the entity's body parts and processes, (Damasio, 1994), (Edelman, 2006) and (Shanahan, 2005)[26].
- Embedded, Enactive: It involves the entity enacting and interreacting with its outer world, thus affecting its world and experiences (Maturana & Varela, 1972), in English (Maturana & Varela, 1980), (Varela, Thompson, & Rosch, 1991), (Thompson, 2007), and (Clark A. , 2008).

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

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

Lenore Blum
lblum@cs.cmu.edu

Manuel Blum
mblum@cs.cmu.edu

- Extended: And consciousness is enhanced by the entity having rich external resources (Clark & Chalmers, 1998).

3.5 Integrated Information Theory (IIT)

IIT, 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. CtmR's extensive feedback (its predictive dynamics globally and locally) and its interconnectedness via global broadcasts contributes to its high Phi.

3.6 Evolutionary Theories of Consciousness

CtmR 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.

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 on how an entity might create world models and, sense of self. Indeed, the roadmap closely parallels the way CtmR's world models evolve, and how CtmR develops its sense of self and phenomenological conscious awareness.[27]

[27] We claim Humphrey actually gives a road map for 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, "The Road Taken", from the perspective of CtmR 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.)

Lenore Blum
lblum@cs.cmu.edu

Manuel Blum
mblum@cs.cmu.edu

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

In *The Hidden Spring*, Marc Solms makes the case that the source of consciousness is the arousal processes in the upper brain stem (Solms M. , 2021). More generally, there and elsewhere, Solms cites the Extended Reticulothalamic Activating System (ERTAS) as the generator of feelings and affects, enabling consciousness. "Affective qualia" is the result of homeostasis. "[D]eviation away from a homeostatic settling point (increasing uncertainty) is felt as unpleasure, and returning toward it (decreasing uncertainty) is felt as pleasure" (Solms M. , 2019). Homeostasis arises by a system resisting entropy, i.e., minimizing free energy (Solms & Friston, 2018). 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. The "system must incorporate a *model of the world*, which then becomes *the basis upon which it acts*[28]" (Solms M. , 2019).

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

- Although GW models generally consider processors as performing cortical functions, CtmR goes beyond that. There is nothing to preclude CtmR from having processors that function as the ERTAS.
- In (Blum & Blum, 2021), we discuss pleasure and pain in the CTM (known here as CtmR). Our discussion of pleasure aligns with Solms, and also with (Berridge & Kringelbach, 2015). Pain is more complex. In (Blum & Blum, 2021) we discuss in more detail how the feeling of pain is generated.
- Predictive dynamics (cycles of prediction, testing, feedback and correcting/learning) in CtmR works to reduce prediction errors, an analogue to minimizing free energy.
- And the incorporated Model-of-the-World in CtmR is *the basis upon which CtmR acts*.

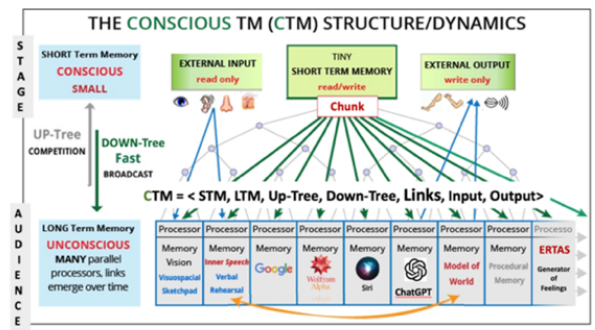
---

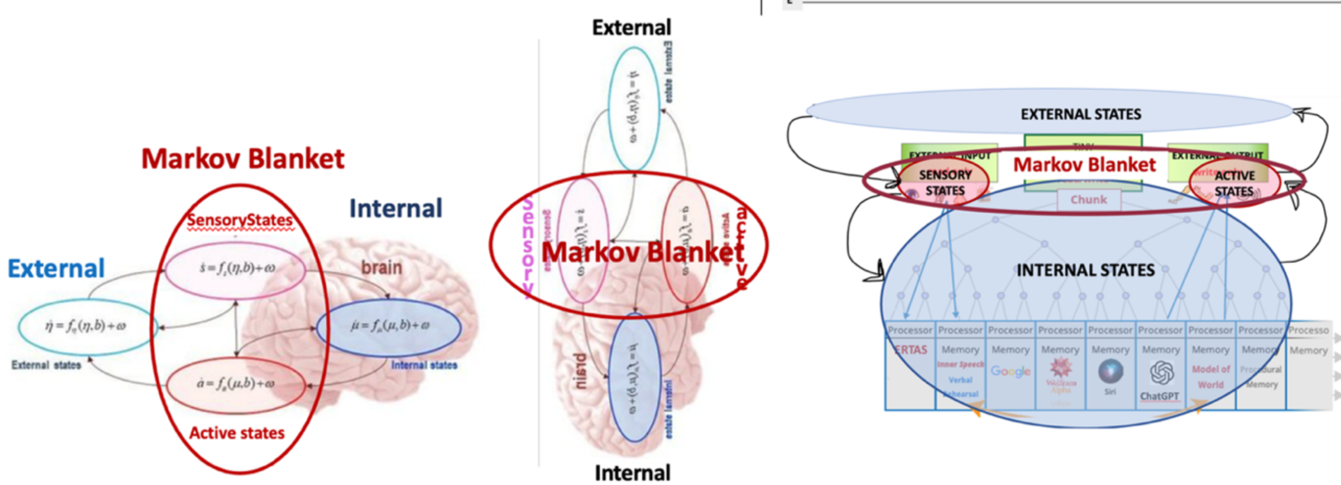
[28] Italics here ours.

Lenore Blum
lblum@cs.cmu.edu

Manuel Blum
mblum@cs.cmu.edu

We have taken one of the well-known Friston diagrams (Parr, Da Costa, & Friston, 2019) with Markov blanket separating internal and external states, rotated it clockwise 90 degrees, and then superimposed it on CtmR (with a little stretching and shrinking). And voilà, a perfect fit!





Friston **Rotated 90⁰ clockwise** **Superimposed on CtmR**

Lenore Blum
lblum@cs.cmu.edu

Manuel Blum
mblum@cs.cmu.edu

4 Addressing Kevin Mitchell's questions from the perspective of CtmR

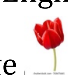
Here we answer Kevin Mitchell's questions (Mitchell, 2023) from the perspective of the Conscious Turing Machine Robot (CtmR). Many of Mitchell's questions are in fact several (often intertwined) questions, separated into parts.

Our answers refer to and supplement what we have discussed in our Overview. They deal *only* with the CtmR model, meaning an entity or robot with a CTM brain. These answers say nothing about other models. They say nothing about whether a worm is conscious or not - unless the worm has a CTM brain. From here on, unless otherwise stated, everything we have to say is about *a robot with a CTM brain,* what we call the *CtmR* model.

KM1*. "Q1. What kinds of things are sentient? Q2. What kinds of things is it like something to be? Q3. What is the basis of subjective experience and what kinds of things have it?"

BB1. Great questions.
A1. CtmR is *sentient*, meaning it is able to perceive and feel things. As mentioned above, we have nothing to say about entities that are not CtmRs. However, we can and will sometimes say what parts of CtmR are responsible for what parts of its sentience.

A2. The Model-of-the-World (MotW) plays an essential role in "what it is like" to be a CtmR. It contains multimodal *Brainish-labeled sketches* of referents in CtmR's worlds. The sketches and labels (as well as Brainish itself) develop and evolve throughout the life of CtmR. The labels succinctly indicate what CtmR learns or "thinks" about the referents. For example, the label SELF applied to a sketch in the Model-of-the-World (MotW) indicates that that particular sketch is (a part of) CtmR itSELF.

A3. The ability of CtmR to construct models of its worlds (inner and outer) is fundamental for its *subjective experiences*. These experiences are described in the model not with English words but with Brainish labeled gists. In the case of a red tulip, the label might be an image  fused with its colors, associated odors, and such.

Lenore Blum
lblum@cs.cmu.edu

Manuel Blum
mblum@cs.cmu.edu

KM2*. "Q1. Does being sentient necessarily involve conscious awareness? Q2. Does awareness (of anything) necessarily entail self-awareness? Q3. What is required for 'the lights to be on'?"

BB2.
A1. In CtmR, *sentience* has two main components, *conscious attention* and co*nscious awareness*.

Conscious attention (access consciousness) occurs when all LTM processors receive the global broadcast of CtmR's current *conscious content*, that being the current winning chunk in the competition for STM. (By the way, STM is a buffer and broadcast station only. It is not and does not have a processor.)

Conscious awareness arises when the broadcasted chunk refers to a Brainish-labeled sketch in the MotW. The labels describe what CtmR is consciously aware of.

A2. Does awareness necessarily entail *self-awareness*? No. The infant CtmR initially builds a world model that does not include a labeled sketch of itself, so it has no self-awareness. In time, however, that model will include a rough labeled sketch of itself and the label SELF. The label SELF marks the beginning of self-awareness, which eventually develops into full-blown self- awareness.

A3. The *lights come on* gradually, as the MotW gets populated with sketches and their labels. (For more on this, see our answer to KM4.)

KM3*. "Q1. 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?) Q2. Are there entities with different degrees or kinds of consciousness or a sharp boundary?"

BB3. We rephrase this as:
Q1. What distinguishes a consciously aware CtmR from a non-consciously aware CtmR?
Q2. Are there CtmRs with different degrees or kinds of consciousness?

A1. Every CtmR pays conscious attention to every broadcast. *Absent a* Model-of-the-World-processor (MotWp) and its Model-of-the-World (MotW) with sketches labeled in Brainish, there is *no* conscious awareness.

Lenore Blum
lblum@cs.cmu.edu

Manuel Blum
mblum@cs.cmu.edu

CtmR can be *consciously aware* when awake or dreaming. It is not consciously aware when it is in deep sleep, when its STM contains a NoOp chunk, i.e., a chunk with NoOp gist and a high enough |weight| to keep all other chunks at bay. (See our answers to KM4 for discussions of Sleep and Dream processors.)

A2. CtmR can have a varying *degree of consciousness*. Its many processors are instrumental in developing rich sketches in CtmR's world models. Involvement by those processors, the Smell, Vision, Hearing, Touch, ,... processors, raises the degree of conscious awareness. Even in deep sleep, however, a CtmR can still carry out tasks (utilizing unconscious communication between processors via links) but without attention and therefore without awareness.

Faulty processors or faulty competition paths can diminish what gets into STM, hence diminish both conscious attention and conscious awareness. For example, a faulty CtmR can exhibit blindsight, meaning it can do things that are normally done with conscious sight, but without having the *feeling* that it is sighted (Blum & Blum, 2022). This can happen, for example, if the Vision processor fails to get its chunks into STM. Perhaps relevant branches in the Up-Tree are broken, or the Vision processor fails to give high enough |weight| to its chunks.

Different degrees of consciousness already occur in a developing CtmR. As we have noted, an infant CtmR has only a very foggy world model which does not even include a sketch of itself. Sketches with annotated labels develop and become *refined gradually*. They are what CtmR is consciously aware of.

KM4. "Q1. For things that have the capacity for consciousness, what distinguishes the *state* of consciousness from being unconscious? Q2. Is there a simple on/off switch? Q3. How is this related to arousal, attention, awareness of one's surroundings (or general responsiveness)?"

BB4.
A1. Only chunks with *non-zero weight have a chance to become conscious* (i.e., win the competition for STM and thus become globally broadcast). That occurs with a probability proportional to the chunk's |weight|. If all chunks have zero weight, then chunks flit in and out of STM at random and so fast that CtmR loses anything remotely resembling sustained attention (like

Lenore Blum
lblum@cs.cmu.edu

Manuel Blum
mblum@cs.cmu.edu

Robbie the Robot in *Forbidden Planet*). This is a *state of unconsciousness*. CtmR can get out of this state only when some processor creates a nonzero-weighted chunk.

Another unconscious state occurs when a *Sleep processor* generates a NoOp chunk (a chunk having a NoOp gist) that has a "sufficiently high" |weight|. A sufficiently high |weight| is one well above the weight of any other chunk. That prevents other chunks - including those from processors that interact with the outer world[29] - from having much chance to enter STM.

When the |weight| of a Sleep processor's chunk drops a bit (but not enough to let input-output chunks enter STM), a CtmR's *Dream processor* can take over, enabling chunks that create dreams to emerge. If the |weight| of the Sleep processor's chunks drop even further, CtmR wakes up.

A2. The above are some of the ways CtmR can go from consciousness to unconsciousness and back. There is *no simple on/off switch*.

A3. In an unconscious state, CtmR 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, meaning their |weight| is greater than the |weight| of the latter.

KM5*. "What determines what we are conscious of at any moment?"

BB5. In CtmR, at every clock tick, **t**, there is exactly one chunk in STM. When a chunk is broadcast, CtmR pays *conscious attention* to that one chunk only (and only if chunks do not flit around from one clock tick to the next). Chunks are purposely small (in a well-defined way) to ensure that all processors focus on the same thought.

---

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

Lenore Blum
lblum@cs.cmu.edu

Manuel Blum
mblum@cs.cmu.edu

KM6*. "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?"

BB6. Operations within each LTM processor are done unconsciously. Communication between LTM processors via links is unconscious communication. It is much quicker than conscious communication that goes through STM.

In the *infant* CtmR, most communication between processors is conscious. Then as processors form links, communication can go quickly through links, meaning unconsciously. This is what happens after the young CtmR learns to ride a bike.[30]

KM7*. "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)?"

BB7. For CtmR to be consciously aware of a thing, call that thing **abc**, a chunk referring to **abc** must get into STM. Once it does, CtmR pays conscious attention to **abc**. But even conscious attention to **abc** does not make for conscious awareness of **abc**. For that, the chunk must reference a sketch in the MotW that is called or labeled **abc**.

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

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

Lenore Blum
lblum@cs.cmu.edu

Manuel Blum
mblum@cs.cmu.edu

1. Things that must be done so quickly that the communication necessary to do the thing cannot go through STM. For example, CtmR 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 like **xyz**. In that case, time permits only one of **abc** and **xyz** to be attended to. If there is barely enough time to do one (and only one) of them, then CtmR cannot be conscious of **abc** while moving to do **xyz**.

KM8. "Q1. Which systems are required to support conscious perception? Q2. Where is the relevant information represented? Q3. Is it all pushed into a common space or does a central system just point to more distributed representations where the details are held?"

BB8.
A1. In CtmR, conscious awareness is impossible without the *MotWp* (among others). Conscious attention *is* possible without the MotWp, but impossible without the broadcast station.

A2., A3. The relevant information is held in the MotW and in *individual processors*. For example, color in the Color processor, smell in the Smell processor, and so on. So, *in that sense, information is distributed*. When CtmR first sees and smells a rose, these processors alert the MotWp which in turn attaches the labels RED and SWEET to its sketch of the rose in the MotW. At some point in time, RED and SWEET become fused as a Brainish word or gist, and *in that sense, information is unified*.

KM9*. "Q1. Why does consciousness feel unitary? Q2. How are our various informational streams bound together? Q3. Why do things feel like *our* experiences or *our* thoughts?

BB9.
A1. At each clock tick, all LTM *processors simultaneously receive a global broadcast* of the conscious content (current chunk) in STM. That gives CtmR its sense of a *unitary* experience.

A2. In addition, processor *links* and Brainish *multimodal gists* further *bind information* together.

Lenore Blum
lblum@cs.cmu.edu

Manuel Blum
mblum@cs.cmu.edu

A3. If CtmR's conscious content refers to a thought or experience that MotWp has labeled SELF, CtmR will be consciously aware of that thought as its own. If that thought is also labeled FEELS, CtmR will not only *know* that the thought is its own, it will also *feel* that it is its own.

KM10*. "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*?

BB10. Here again, world models, with their *learned* Brainish-labeled sketches, determine CtmR's sense of self. The MotW's sketches are labeled with a variety of gists. For this question, the labels SELF, FEELS, and CONSCIOUS are particularly important. If all three labels are attached to a sketch of CtmR in the MotW, then CtmR **FEELS** that its**SELF** is **CONSCIOUS.** Known pathologies occur when any one (or more) of these labels is missing, or when sketches are mislabeled.[31]

KM11. "Q1. Why do some kinds of neural activity feel like something? Q2. Why do different kinds of signals feel different from each other? Q3. Why do they feel specifically like what they feel like?"

BB11.
A1. In CtmR, inputs from different sensors go to *different sensory processors*. Those different senses become incorporated in the MotW with *different Brainish labels*.

A2. In the MotW, sketches of a red rose and a red fire engine are both labeled RED. Over time, each of these sketches can gain many other labels as well. For example, the fire truck sketch likely gets the Brainish labels FIRE_TRUCK and LOUD_SIREN while the rose sketch does not. The rose sketch gets labeled "SILKY_FEEL and SWEET_SMELL".

A3. The two referents are distinguished in the MotW, and with more Brainish labels, "feel specifically like what they feel like."

---

[31] Some human examples of pathologies due to mislabeling: body integrity dysphoria, phantom limb syndrome, Cotard's syndrome, anosognosia, paranoia, ...

Lenore Blum
lblum@cs.cmu.edu

Manuel Blum
mblum@cs.cmu.edu

KM12*. "Q1: How do we become conscious of our own internal states? Q2: How much of our subjective experience arises from homeostatic control signals that necessarily have valence? Q3: If such signals entail feelings, how do we know what those feelings are about?"

BB12.
A1. In the Overview, we indicated how the infant CtmR would know it is hungry when a high |weight| negatively valenced chunk from a Fuel Gauge processor reaches STM and is broadcast from it.

A2. The LOW_FUEL chunk will trigger an actuator to connect CtmR's fuel intake to a fuel source (in humans, the breast). Assuming it works, that will eventually result in a sketch of the fuel source (in the MotW) being labeled FUEL_SOURCE and PLEASURE_SOURCE. At the same time, the label FUEL_INTAKE and FEELS_PLEASURE will be attached to the sketch of CtmR when it is hungry and being fueled. A high |weight| broadcast indicates that "CtmR feels pleasure when it gets fuel if it's hungry." This process is an example of homeostasis in CtmR and how CtmR becomes conscious of its own internal state.

A3: The about-ness of those feelings come from the Brainish labels and sketches that evolve during CtmR's lifetime.

KM13. "Q1. How does the about-ness of conscious states (or subconscious states) arise? Q2: How does the system know what such states refer to? (When the states are all the system has access to)."

BB13.
A1., A2. Conscious states in CtmR are broadcasted states.

The MotW is *all* that CtmR knows about its (inner and outer) worlds. This includes CtmR's actions and their effects. When choosing which of several actions to take, the MotW processor predicts the

Lenore Blum
lblum@cs.cmu.edu

Manuel Blum
mblum@cs.cmu.edu

effect of each of its possible actions. It does this by simulating the world's response to the action in its MotW.[32] Each response from the world has a Brainish description with Brainish labels.

- If a broadcasted chunk refers to a Brainish labeled sketch in the MotW, the chunk's gist is the about-ness of the current conscious state. For example, the gist could be sketch of a rose labeled RED and SWEET_SMELL.
- If a broadcasted chunk refers to a Brainish labeled prediction (gotten from the simulation in the MotW) this is the about-ness of the current conscious state.
- If a broadcasted chunk refers to Brainish labeled response (gotten from the world) this is the current about-ness.

KM14. "Q1. What is the point of conscious subjective experience? Or of a high level common space for conscious deliberation? Q2. Or of reflective capacities for metacognition? Q3. What adaptive value do these capacities have?"

BB14.
A1. A conscious subjective feeling is experienced when broadcasted chunks refer to Brainish-labeled sketches in the MotW. The labels describe the subjective feelings that conscious awareness is about. Without these feelings, CtmR would not be compelled by feelings to act appropriately.[33]

A2. Reflective capacities enable CtmR to treat itself, as referred to in the MotW by the sketch of itself, with all the tools it uses to treat other sketches.

---

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

[33] A person who has pain and *knows* everything about it but lacks the ability to *feel* its agony has *pain asymbolia.* Such a person is not motivated to respond normally to pain. Children born with pain asymbolia have rarely lived past the age of 3. The experience of pain, whether physical or emotional, serves as a motivator for behaving appropriately to the pain.

Lenore Blum
lblum@cs.cmu.edu

Manuel Blum
mblum@cs.cmu.edu

A3. Conscious subjective experience is adaptive because *all* processors receive each and every broadcast, so all processors can contribute to the understanding of the broadcast &/or its solution. The point is that all processors focus their attention on the same thing. Suppose the broadcast is a problem like "Must fill the fuel tank": the Navigation processor might contribute a choice of routes to local fuel stations. The Computation processor might compute how much fuel is required for each choice. The Weather processor might weigh in if one of the routes is blocked. The conscious subjective experience can take account of these estimations.

KM15*. "Q1. How does mentality arise at all? Q2. When do information processing and computation or just the flow of states through a dynamical system become elements of cognition and Q3. why are only some elements of cognition part of conscious experience?"

BB15.
A1. The question asks: "How does the capacity for intelligent thought come about?" CtmR is ideal for answering this question since at birth, all $\gtrsim 10^7$ processors are independent. The first processors to come online – meaning they have sufficient weight to get their chunks to STM - are those having homeostatic importance like the Nociceptor Gauge (monitors pain), Fuel Gauge (monitors hunger), and so on, or have immediate access to the senses like vision, hearing, and so on. These processors help the MotWp to make and improve its predictions and world models. The next processors to come online are those that affect the activators, one of which cries for help. Then come processors that detect coincidences like: "This visual input and that auditory input coincide." This is the beginning of intelligent thought.

A2. The CtmR model, unlike Baars' GW, has no Central Executive. The *competition for conscious attention*, which replaces the Central Executive, gives CtmR much of its cognitive power. That competition efficiently considers *all* information submitted for consideration by its more than $\mathbf{10^7}$ processors. It allots ideas a winning probability or share of consciousness (broadcast time) proportional to its estimated importance (|weight|).[34] It enables processors to solve a problem

[34] This is something that tennis and chess tournaments do not provide.

Lenore Blum
lblum@cs.cmu.edu

Manuel Blum
mblum@cs.cmu.edu

even though CtmR does not know which processors have the interest, expertise or time to consider the problem. No Central Executive could have the knowledge or resources to do that unless, of course, the Central Executive was itself an efficient competition process.

A3. Some elements of cognition can be done with a single processor. That processor doesn't need to search through an enormous data base for its information: it already knows where the necessary information is held. Processors that do need to search for the information,[35] must search for it. They do need to broadcast. That broadcast begins the process of using consciousness to do cognition.

KM16*. "Q1. How does conscious activity influence behavior? Q2. Does a capacity for conscious cognitive control equal "free will"? Q3. How is mental causation even supposed to work? Q4. How can the meaning of mental states constrain the activities of neural circuits?"

BB16.
A1. In CtmR, *conscious activity is intertwined with behavior*.

In CtmR, all LTM processors receive the broadcasted conscious content. Different processors have differing amounts of time to deal with that content. Of those that have time, some have a more reasonable idea how to deal with the broadcast than others. A broadcasted message that the fuel gauge is low can prompt one processor to try to conserve fuel, another to trigger a search for a source of fuel, and so on. A broadcast of danger may prompt CtmR to choose between fight, flight or freeze, each championed by a different processor.

Additionally, CtmR's *disposition* plays an important factor in the competition that selects which chunk will be globally broadcast and hence its behavior.[36]

A2. As for "free will", CtmR's ability to assess a situation, consider various possibilities, predict the consequences of each, and based on that make a decision (all *under resource constraints*) gives

---

[35] Like the processor that asks, "What her name?"

[36] See Appendix 7.2 for more information about CtmR's competition and the influence of its disposition.

Lenore Blum
lblum@cs.cmu.edu

Manuel Blum
mblum@cs.cmu.edu

CtmR its feeling of "free will". For example, imagine CtmR playing a game of chess. When and for as long as CtmR has to decide which of several possible moves to make, it knows it is "free" to choose whichever move has the greatest utility for it. That is free will. See (Blum & Blum, 2022).

A3. In CtmR, the MotW is fundamental to *mental causation*. To will an act in the world, the MotWp performs that action in the MotW, then looks to see if the act got accomplished in the world.

For example, suppose the infant CtmR discovers that it can somehow move its left leg. It becomes aware through its sensors that "willing" the movement of that leg is successful. For comparison, it may discover that it cannot pick up a rock, Yoda style, with the power of thought. Moving the leg or lifting the rock can be willed by performing the action in the MotW. Sensors must verify if the act has been successful. If it has, that is mental causation.

A4. As an example, in our answer to KM4, we discussed how the Sleep processor generates a non-dreaming sleep state by raising its own |weight| so high that other chunks can't reach STM. This shows how the sleep state *constrains* activity in CtmR's Up-Tree.

Kevin Mitchell ends his blog with the words "If we had a theory that could accommodate all those elements and provide some coherent *framework*[37] 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."

5 Summary and Conclusions

In this chapter we have presented a brief overview of a simple formal machine model of consciousness, known here as CtmR. The TCS perspective has influenced the design and definitions of CtmR, and conclusions we have drawn from the model.

---

[37] Italics ours.

Lenore Blum
lblum@cs.cmu.edu

Manuel Blum
mblum@cs.cmu.edu

Although CtmR is inspired by the simplicity of Turing's formal model of computation and Baars' global workspace (GW) architecture, our formalization is neither a Turing Machine nor a standard GW model. Its *consciousness* (access and phenomenological) depends on *having more than a global workspace*.

Importantly, CtmR also:

1. interacts with its outer world via input sensors and output actuators;
2. has the ability to construct models of its inner and outer worlds;
3. has a rich internal multimodal language; and
4. constantly updates its states via predictive dynamics (cycles of prediction, testing, feedback and correcting/learning),

all while operating under resource limitations.

CtmR is not a model of the human or animal brain, nor is it intended to be. It is a simple *machine model* of consciousness. Nevertheless, at a high level, the model 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 and/or complementarity of those theories. It supports (the credibility of) our claim that *a conscious AI is inevitable*, because it is clearly buildable and arguably a basis for consciousness.

Finally, the development of CtmR is a work in progress. While we have worked out many details of the model, there is much left to develop. More specifics will appear in our upcoming 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.

6 Acknowledgements

We are grateful to Michael Xuan for his immense encouragement, and to UniDT for their long-term support.

Lenore Blum
lblum@cs.cmu.edu

Manuel Blum
mblum@cs.cmu.edu

7 Appendix

7.1 A Brief History of the Theoretical Computer Science Approach to Computation

The theoretical computer science approach to computation starts with 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 suggested that a function is 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 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 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) compute exactly the same class of functions, the computable functions. In programming parlance, all sufficiently powerful practical programming languages are equivalent in that any one can simulate (be compiled into) any other. The ensuing mathematical theory is often called the Theory of Computation (TOC).

In the 1960's, with the wider accessibility of computers, newly minted computer scientists such as Jack Edmonds pointed out that resources matter. Certain problems that, in principle, were decidable were intractable on account of time and space requirements. Even more, intractability seemed to be an intrinsic property of the problem, not the method of solution or the implementing

Lenore Blum
lblum@cs.cmu.edu

Manuel Blum
mblum@cs.cmu.edu

machine. The ensuing sub-theory of TOC, which introduces resource constraints into what is or is not computable *efficiently*, is called Theoretical Computer Science (TCS).

TCS focuses on 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 $\mathbf{n}$ 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($\mathbf{n}$) time in general. But is the "SAT problem" tractable, meaning decidable efficiently, i.e., in polynomial($\mathbf{n}$) time? This is equivalent to the well-known P = NP? problem of (Cook, 1971), (Karp, 1972), (Levin, 1973).

The design of novel and efficient algorithms is a key focus of TCS.

Turning the table on its head, the ability to exploit the power of hard problems, problems that cannot be solved efficiently, has been a key insight of TCS. An example is the definition of pseudo-randomness.

The ability to exploit the power of hardness is novel for mathematics.

7.2 The Probabilistic Competition for Conscious Attention and the Influence of Disposition on it

We tried to make the CtmR Up-Tree competition (a binary tournament) deterministic, but it turns out to *necessarily* be probabilistic. This is because a deterministic CtmR must be made increasingly complex in many different ways. For example, consider a deterministic competition in which chunk A is pinned to weight 11, chunk B to weight 9, and all other chunks to weight 0. In the deterministic CtmR competition, A always wins and B never does. And as long as weights don't change, CtmR remains totally unconscious of chunk B. In the probabilistic CtmR competition, chunk A wins with probability 11/20 while B win with probability 9/20. So A and B each have roughly equal probability of winning a competition, and a new independent competition is begun at every clock tick. In that case, CtmR is conscious of both.

Lenore Blum
lblum@cs.cmu.edu

Manuel Blum
mblum@cs.cmu.edu

More generally, in the (probabilistic) CtmR competition (a variant of the standard tennis or chess tournament) it is proved that a chunk wins the competition with probability proportional to its |weight|. As a consequence, the winning chunk is independent of where processors are located! This is a property of CtmR's probabilistic competition. (It is a property 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,

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

consisting of a pointer to 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 the originating processor assigns its gist), and some auxiliary information.

For the probabilistic CtmR, the **auxiliary information** is a pair of numbers which we call **(intensity, mood)**.

At the start of the competition, each LTM processor enters a chunk into its leaf node with

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 in the node's two children based on their **f** values, where

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

Here **d** is a constant called CtmR's *disposition*.

If C_1 and C_2 are the two competing chunks, then the coin toss neuron will choose C_i to be the local winner with probability $\mathbf{f}(C_i)/\mathbf{f}(C_1)+\mathbf{f}(C_2)$.

A modification of the local winning chunk moves 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 two competing chunks' intensities. Similarly for its *mood*.

Lenore Blum
lblum@cs.cmu.edu

Manuel Blum
mblum@cs.cmu.edu

Thus, as a chunk moves up the tree, the intensity never decreases. This is not (necessarily) the case for the mood.

In this way, 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 idea about the other $\mathbf{N} \gtrsim \mathbf{10^7}$ chunks that are being submitted, the reception of the broadcasted winner provides useful information. In addition to providing the winning gist and its processor's address, the broadcasted information enables each LTM processor to quickly compare *how its submitted chunk's intensity* (|weight|) *compares* with the winner's and with the average of the submitted chunks' intensities, ($\sum_{\text{all } \mathbf{N} \text{ processors}}$ intensity of submitted chunk)/N. Similarly, the processor learns *how its own chunk's mood* (weight) *compares* with that of the winner's and with the average mood, ($\sum_{\text{all } \mathbf{N} \text{ processors}}$ mood of submitted chunk)/$\mathbf{N}$.

CtmR's disposition, a real number $\mathbf{d}$, $\mathbf{-1 \leq d \leq +1}$, plays an important factor in the competition that selects which chunk will be globally broadcast, and hence CtmR's behavior.

If the disposition is $\mathbf{d > 0}$, CtmR 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 $\mathbf{d = +1}$, CtmR is manic: only positively valenced chunks can win the competition. The CtmR knows only what is positive in its life, as long as anything is positive.

If the disposition is $\mathbf{d < 0}$, CtmR will be "downbeat". If $\mathbf{d = -1}$, CtmR is "hopelessly depressed", only negatively valenced chunks can win the competition. There is no way out of this horrible state except with a reboot, i.e., to "shock" the system to get a less extreme disposition.[38]

---

[38] If $\mathbf{d{=}{+}1}$, CtmR is in the manic state, and CtmR theory suggests that there too, a reboot is warranted.

In humans, electroconvulsive therapy (ECT) is used primarily for depression. It is not used for mania because doctors fear that it will lead to even more mania. The CtmR model suggests that ECT might work as well for

Lenore Blum
lblum@cs.cmu.edu

Manuel Blum
mblum@cs.cmu.edu

References

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

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.

mania. We view that as a prediction of the model. (But as we are not physicians, we cannot advocate it.) In a literature search, we found only one case of the use of ECT for mania: "it is well known that acute mania responds well to ECT, best seen in a prospective study by Mohan (Mohan, Tharyan, Alexander, & Raveendran, 2009), which found an 88% remission rate with bilateral ECT." (Thomas, White, & Dursun, 2018).)

Lenore Blum
lblum@cs.cmu.edu

Manuel Blum
mblum@cs.cmu.edu

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

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.

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

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

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.

Lenore Blum
lblum@cs.cmu.edu

Manuel Blum
mblum@cs.cmu.edu

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., 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.

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.

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.

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

Hatami, P., & Hoza, W. (2024). Paradigms for Unconditional Pseudorandom Generators . *Foundations and Trends® in Theoretical Computer Science, 16*(1-2), 1-210. http://dx.doi.org/10.1561/0400000109.

Lenore Blum
lblum@cs.cmu.edu

Manuel Blum
mblum@cs.cmu.edu

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.

Kanai, R., & Fujisawa, I. (2023, September 25). Towards a Universal Theory of Consciousness. *PsyArXiv Preprints*, https://doi.org/10.31234/osf.io/r5t2n.

Karp, R. M. (1972). Reducibility Among Combinatorial Problems. In R. E. Miller, J. W. Thatcher, (eds.), R. E. Miller, & J. W. (eds.) (Eds.), *Complexity of Computer Computations* (pp. 85-103). New York: Plenum.

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.

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

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.

Lenore Blum
lblum@cs.cmu.edu

Manuel Blum
mblum@cs.cmu.edu

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.

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

Mohan, T., Tharyan, P., Alexander, J., & Raveendran, N. (2009, March). Effects of stimulus intensity on the efficacy and safety of twice-weekly, bilateral electroconvulsive therapy (ECT) combined with antipsychotics in acute mania: a randomised controlled trial. *Bipolar Disord, 11*(2), 126-134. doi: 10.1111/j.1399-5618.2009.00668.x.

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

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.

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.

Lenore Blum
lblum@cs.cmu.edu

Manuel Blum
mblum@cs.cmu.edu

Seth, A. K. (2015). The Cybernetic Bayesian Brain - From Interoceptive Inference to Sensorimotor Contingencies. In T. Metzinger, & J. M. Windt, *Open MIND* (p. doi: 10.15502/9783958570108 23 | 24). Frankfurt am Main: MIND Group.

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

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, & et.al. (2024, March 5). An integrative, multiscale view on neural theories of consciousness. *Neuron*, doi:https://doi.org/10.1016.

Thomas, P. K., White, P. J., & Dursun, S. (2018, Sept 10). Treating electroconvulsive therapy-induced mania with more electroconvulsive therapy: Evidence for electroconvulsive therapy as the ultra-mood stabilizer. *SAGE Open Med Case Rep., 6*, doi: 10.1177/2050313X18799236.

Thompson, E. (2007). *Mind in life.* 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.

Lenore Blum
lblum@cs.cmu.edu

Manuel Blum
mblum@cs.cmu.edu

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.

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.

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.

Lenore Blum
lblum@cs.cmu.edu

Manuel Blum
mblum@cs.cmu.edu