Towards Understanding and Dissolving the Hard Problem of Consciousness

Jiří Wiedermann
Jan van Leeuwen

Technical Report UU-PCS-2025-01
March 2025

Philosophy of Computer Science
Department of Information and Computing Sciences
Utrecht University, Utrecht
The Netherlands

Series: UU-PCS

Department of Information and Computing Sciences
Utrecht University
Princetonplein 5
3584 CC Utrecht
The Netherlands

Towards Understanding and Dissolving the Hard Problem of Consciousness*

Jiří Wiedermann[1] and Jan van Leeuwen[2]

[1] Institute of Computer Science of Czech Academy of Sciences and Karel Čapek Center for Values in Science and Technology, Prague, Czech Republic
jiri.wiedermann@cs.cas.cz
[2] Dept of Information and Computing Sciences, Utrecht University, the Netherlands
J.vanLeeuwen1@uu.nl

Abstract. We propose a novel approach to the hard problem of consciousness, framing it as a tractable scientific inquiry within theoretical computer science. We explore how an agent might process sensory inputs, gain subjective experiences, and generate meaningful behaviors by considering a simplified, idealized model of a cognitive, embodied agent. Exploiting the semantic properties of computations, we investigate mind-like properties traditionally studied in the philosophy of mind. Our framework enables us to formulate and prove core propositions that offer partially non-constructive answers to Chalmers' questions of how and why consciousness arises. We bridge the explanatory gap between the physical and mental by endowing the agents with an anticipation ability via their learned world model. This ability connects subjective experiences, triggered by the impact of experiential qualities of sensory events on the agent's data processing, with the corresponding anticipated behavior. Our findings provide a functional explanation for the hard problem of consciousness and suggest that consciousness and subjective experience, in various forms, extend beyond biological brains. This knowledge has significant implications for understanding the nature of consciousness in a wide range of natural and artificial systems.

> *"Can you imagine that human consciousness is just one example of this huge spectrum of conscious-like experiences that can be instantiated in other systems that could be artificial or they could be organic and what we consider sort of this wondrous quality of being a human being is actually just a pedestrian example of something that can take on so many other forms?"*
>
> Brian Green, in a video podcast with David Chalmers and Anil Seth, 2024 [12]

Keywords: behavioral agents, aphantasia, events, experiential qualities, hard problem of consciousness, hard problem of matter, interaction, non-trivial properties, non-uniformity, subjective experience, world model

1 Introduction

1.1 Motivation and background

In the philosophy of mind, the famed *hard problem of consciousness* epitomizes our profound ignorance of the relationship between physical processes such as occurring in the brain and consciousness. The problem asks for an explanation of how and why physical brain activity gives rise to qualia, phenomenal consciousness, or subjective experience. Can it be understood at least from a theoretical perspective? Could it apply to AI systems?

* The research of the first author was partially supported by Grant No. CK04000150 EBAVEL of the Czech Technology Agency, programme Strategy AV21 "Breakthrough Technologies for the Future", and the Karel Čapek Center for Values in Science and Technology.

Subjective experiences are a subjective aspect of information processing within any conscious cognitive entity. Obviously, such an entity has subjective experiences in any situation, but the question is why and what it feels like for the entity. In 1974, Nagel [18] put it as follows, phrased in our terms: *what is it like to be that conscious entity?* For example, humans can experience various stimuli, be it the red color of the setting sun, the cool breeze, the smell of the sea, the sour taste of the sorrel, the glorious tones of the organ, and so on. Experiencing is not only evoked by external stimuli but also by various internal stimuli: experiencing emotions, pains, mental images of past experiences, experiencing the flow of thoughts, et cetera. The emergence of all these experiences has a physical basis – the processing of the relevant information – but why, as asked by Chalmers [7], is this processing accompanied, as it were, by these internal experiences, this mental experiencing? This is precisely the crux of Chalmers' hard problem of consciousness.

The hard problem of consciousness has since become the subject of intense research and debate among philosophers, researchers in artificial intelligence, and other thinkers. The hard problem is usually contrasted with the *'easy problems,'* roughly the remaining problems related to consciousness, like the ability to discriminate, integrate information, perform behavioral functions, et cetera. These problems are amenable to functional explanation – that is, explanations that are mechanistic or behavioral [6]. However, despite of all efforts, the hard problem still needs to be resolved [31, 42]. It is a challenging question how one could still proceed.

By its very definition, the hard problem of consciousness regards human cognition. A majority, if not all, of the efforts mentioned above to resolve the issue have been made in this framework. For a broader perspective we note that only recently animal consciousness has come into the focus of consciousness researchers. Even a bold hypothesis has lately appeared, in the form of a declaration by experts that there is *'a realistic possibility'* for elements of consciousness in reptiles, insects, and mollusks [15].

If even relatively simple organisms exhibit signs of consciousness, why not consider the possibility of (artificial) consciousness in *AI systems* whose complexity competes with such creatures? For example, with the recent appearance of large language models, the hard problem of consciousness has gained new momentum. Could it be the case that these large language models actually exhibit signs of consciousness [1, 2, 19]? The problem of conscious AI systems is generally seen as an important obstacle in achieving qualitatively advanced AI systems that mimic higher-level human mental abilities.

In line with Brian Green's quotation at the beginning, why not abandon the anthropocentric context of the mainstream investigations of the theory of mind that focus on the human brain in favor of more general, natural, and artificial cognitive systems? Why don't we look for common cognitive properties of such systems, regardless of their operating environment, architecture, technological base, substrate, simplicity, or complexity? Such studies may be vital to answering the tantalizing hard problem of consciousness. This idea is the motivation for this study.

1.2 A novel mindset

We base our methodology on an approach that radically differs from most previous attempts to model high-level mental activities. The idea is to design an abstract mathematical model, or a computational one for that matter, that focuses on key properties attributed to the mind while abstracting from low-level details. The resulting model of *embodied cognitive behavioral agents* (CBAs) should cover a large spectrum of natural and artificial cognitive systems, ranging from humans and animals to complex and straightforward cyber-physical systems equipped with sensors, effectors, and data processing abilities, all potentially performing meaningful goal-oriented activities while exhibiting essential aspects of consciousness.

To study the properties of the abstract model, we take inspiration from two sources within theoretical computer science. The first source deals with non-uniform models of

computation, whose architecture can be tuned to the concrete computational task to be solved, as studied in e.g. non-uniform computational complexity theory. In our case, we are especially interested in non-uniform aspects of embodiments of our models. In this instance, referred to as *embodiment non-uniformity* (or just non-uniformity for simplicity), the corresponding systems all process multi-modal, potentially infinite input streams of data but their design depends on the embodiment that reflects their mission. Different missions normally require different embodiments ("architectures") and different programs. Non-uniformity is at the heart of many problems related to modeling and explaining consciousness and subjective experience, in the many forms in which they occur in biological and artificial systems.

The second source of inspiration comes from studying the semantics of programs, i.e., of their behavior when run. We draw an analogy between the behavior of embodied cognitive agents, which we see as embodied Turing machines of some kind, and the semantic properties of Turing machine computations. In this context, special attention is given to so-called *non-trivial properties* of computations. A non-trivial property is neither true for every program nor false for every program. (In the theory of computation, a result known as Rice's Theorem (cf. [21]) shows that non-triviality is the essential feature of all properties that are undecidable in general form.)

For embodied cognitive agents, a semantic property is 'non-trivial' if it is a property of the behavior of some agents but not of all agents: some have the property, and some do not. Examples of properties of agents that one may be interested in from the viewpoint of non-triviality are: consciousness, subjective experience, attention, creativity, adherence to given principles, etc. Even this simple discrimination between the behaviors of agents will be sufficient for proving the non-existence of philosophical zombies.

The resulting mindset of non-uniformity and semantic properties brings a significant methodological benefit when analyzing the behavior of embodied cognitive agents. Instead of considering several high-level mental properties attributed to the human mind separately, such as consciousness, subjective experience, creativity, ethical behavior, and the like, it allows a consideration of the *intrinsic semantic properties* of agents in general. This vastly simplifies the number of specific cases of the previously considered mind-like properties. It allows their generalization to a broad range of analogous properties of both natural and artificial systems. This generalized approach offers new insights into the hard problem of consciousness and related fields of consciousness studies. It has the potential to change the landscape of the philosophy of mind substantially.

1.3 Contribution and results

We posit that consciousness is a *computational phenomenon* arising from the interaction of embodied cognitive agents with their respective environment over time and the streamlining of their behavior towards their mission under all circumstances. It is the product of specific computational acts applied to sensory data, internal states, and learned knowledge of the agents. This position is strongly aligned with traditional computationalism, but surpasses it in essential aspects. Note that traditional agent models in computationalism are mostly anthropocentric. Most are based on our understanding of the brain and its functioning (cf. [3–5, 14, 23, 24]) and focus on abstract disembodied and instantaneous information processing.

In contrast, our model of embodied cognitive behavioral agents (cf. Section 2) is primarily *explanatory* and *non-anthropocentric*, covers a large spectrum of natural and artificial systems, and incorporates the core elements of embodiment and perceptual data processing mechanisms. An explanatory model concentrates on defining, in a structured way, WHAT the model is doing rather than on the details of HOW it does it. The model does not give a recipe for constructing an embodied cognitive agent endowed with a form of consciousness, but rather is a high-level guide for understanding the goals of its work.

In our model, we characterize the behavior of embodied cognitive agents by the semantic properties that represent their *intrinsic* (defining, fundamental, inseparable) *properties* (cf. Section 3.2). We will argue that different semantic properties lead to a noticeable difference in an agent's behavior and imply the non-existence of philosophical zombies (cf. Proposition 1 and Proposition 3). The latter has been considered by some as a strong argument supporting the significance of the hard problem of consciousness (cf. [8]).

The main results of this study are in Section 4. Within the framework of embodied cognitive behavioral agents, we first define the notion of experiential quality of events and of subjective experience. Here, *events* are perceived through multiple 'sensations' over time (cf. Definition 8). Next, using that agents have an *anticipation ability* via their learned *'world model'*, we bridge the explanatory gap between the physical and the mental. The anticipation ability referred to connects subjective experiences, as triggered by the impact of the sensations of experiential qualities of events on an agent's data processing, with the corresponding anticipatory behavior represented in the agent's world model.

The effects of the respective data processing mechanisms and their concurrent interplay lead, on the one hand, to emotions, qualia, and feelings of the experiential qualities of events perceived by the agent as subjective experiences (cf. Proposition 4). On the other hand, we argue that they give rise to the development of the non-uniform, agent-dependent 'what-is-it-like' aspect of consciousness (cf. Section 4.3).

Eventually, this leads to a simple and concise general definition of consciousness covering the broad spectrum of embodied cognitive agents considered in this study:

> *Consciousness is the ability of embodied cognitive behavioral agents to perceive experiential qualities of events in their surroundings and be responsive to their subjective experience.*

This insight offers a novel understanding of consciousness as an interplay between physical interaction, internal data processing abilities, and the historical context of interactions. No resort to emergence to explain consciousness as a spontaneous, unpredictable rise of a new property is needed; no notion of complex systems needs to be invoked.

The generality of our approach allows us to apply it in answering the so-called *hard problem of matter* raised by Russell [22] and Strawson [25]: *what are the intrinsic qualities of physical phenomena, or more generally, of matter?* (cf. Section 5).

The results demonstrate that the hard problem of consciousness, traditionally seen as an impenetrable barrier between the physical and the mental realms, is not a fundamental obstacle. Within our model, we transform the problem into a series of 'easy problems' that are believed to be solvable by tractable scientific inquiry.

2 Cognitive behavioral agents

We now describe our main object of interest – embodied *cognitive behavioral agents* (CBAs or 'CBA agents'). Embodied cognitive behavioral agents intend to model all living or non-living (biological or non-biological) cognitive entities operating in the actual world or its subsets or, in the case of artificial agents, in formalized domains.

To achieve a high level of generality and machine independence, we present our CBAs in the spirit of the epistemic approach to computation [36]. In this approach, the data processing acts of agents are seen as processes generating knowledge, in terms of behavior, over (or in) their operating domain, in the framework of a corresponding knowledge theory. The knowledge theories describe both the operating domains of the agents and the allowed operations in that domain, and thus define the *grounding* of the agents in their operating domains. Knowledge theories can be described formally or informally. For details we refer to [36].

Following the philosophy of epistemic computation, in the description of the agents we focus on WHAT an agent has to do (in terms of generated knowledge or behavior)

rather than on HOW it achieves its goals. Hence, we see embedded CBAs as black boxes whose 'calling' it is to generate a set of actions for each set of input stimuli that they perceive while preserving a certain semantic condition put on the behaviors they generate. The agents preserve a causation link between the incoming stimuli and the corresponding 'mental processes'. In turn, at the physical level, these processes lead to agent behaviors while following a specific goal guaranteed by their adherence to certain suitable semantic properties (cf. Section 3).

2.1 Informal description of the 'CBA model'

An embedded cognitive behavioral agent is (or, models) a *physical system*, i.e., an arrangement of parts or elements that exhibits behavior based on that of the individual constituents. The individual constituents make up the agent's 'body'. Agents have facilities for the interchange of information or energy (but not matter) with their environment. They satisfy the following properties:

– *Finite specification:* Each agent is to be finitely describable at any given time. This refers to the fact that we assume that the entire constitution and behavior of an agent can be specified using a finite description of some kind over a finite alphabet of symbols and that the representation of the data it gathers over time is necessarily finite at all times too.

– *External and internal sensations:* Each agent is 'grounded' in its environment via its sensors and effectors. It continuously senses its environment (i.e., operating domain) and its body. In general, an agent can possess sensors that register external and internal stimuli and the qualities of various modalities. The sensors deliver representations of the events or objects they register and information about their objectively measurable qualities (intensity, pitch, speed, number of revolutions, acceleration, direction, proximity, position, place, state, size, temperature, salinity, humidity, pressure, voltage, level of resource utilization, activation of internal mechanisms, specific pattern occurrence, and the like). Often, agents will work with sensor arrays comprising a homogeneous group of sensors, e.g. deployed in a specific geometric pattern. The advantage of using a sensor array over a single sensor is that an array adds more dimensions to the observation, helping to estimate more parameters and improving the quality of the performances.

– *Mood sensations:* In addition to objectively measurable sensations and their qualities, some agents may use context-dependent subjective sensations, delivered by the interplay of external and internal sensors and mechanisms that return specific quality signals reporting the 'moods' of various agents. Moods emerge upon the registration of particular configurations of external and internal events occurring in the context of related previous experiences, characterized by the simultaneous activation of specific mechanisms or patterns in the agent's body. They correspond to the instantaneous complex 'cognitive state' of an agent as the result of current and remembered past circumstances. The common feature of moods is their private nature, which is ineffable in a natural language, intrinsic to and inseparable from the agents. Moods are not objectively measurable. Their arousal and registration in particular situations depend on an agent's design and constitution and manifest themselves through the agent's reactions. In humans, they correspond to feelings, emotions, and qualia. Moods occur in various qualities and intensities and are inseparable parts of the qualities of events under specific circumstances.

– *Effectors:* Depending on sensations and moods, an agent activates its external and internal effectors (actuators) and acts in its environment and in its body. In this way, it generates behavior corresponding to the situation as mediated by its incoming external and internal percepts, moods and their respective qualities, possibly depending on the history of previous interactions. Effectors can have sensors sending feedback information concerning the success of their actions.

– *Interactive behavior:* A step consisting of, possibly several, 'readings' of sensory data input data and the 'writing' of a corresponding bundle of outputs (data, signals, stimuli)

to the body and the environment of an agent, is called an *interaction*. The duration of an interaction, i.e. the time that elapses between the entering of the input data and the issuing of the respective output, is always finite but need not be constant; it can depend on the time complexity of the processing of the input data and on any ongoing internal processing before the bundle of output data is generated (cf. the next item in this description). We measure the duration of interaction steps in *perception cycles.* Within each perception cycle, any agent performs just one reading of its sensors.

– *Data-processing ability:* Agents process streams of input stimuli from internal and external sources, generating actions while preserving appropriate semantic conditions on the behavior that are brought about. By the philosophy of epistemic computation, we do not make any assumptions about HOW the input data are actually processed. The underlying processes, deterministic or otherwise, may be realized by whatever computational, biological, chemical, mechanical, quantum, or other physically conceivable principles, including not yet known ones, that allow an agent to link its percepts to the corresponding actions. In particular, we allow for any data processing using the mental abilities of biological brains.

In our context, we see the body of any agent as a 'black box' whose behavior depends on its current percepts, moods, previous interactions, and on how it is processing the respective data. In Subsection 4.2.1 we will specify some more details of what is taking place 'inside' the agents, but we do not need it now. Agents are allowed to inform their environment about all ongoing ('mental', internal, private) processes inside their bodies (i.e., black boxes). We posit that the data-processing abilities of agents are substrate-independent and enable the realization of the 'easy problems' of consciousness (cf. Section 1.1).

– *Energy consumption:* The abilities that underly perception, data processing, and action of agents are assumed to be based on physical processes and thus consume energy when activated. This excludes any non-physical influencing on agents.

– *Non-uniformity of agents:* The agents are essentially nonuniform, and operating domain- and purpose-specific. This means that no uniform algorithm can be assumed that would generate all instances of agents, given their operating domain and task in that domain. Indeed, assuming that there are unboundedly many domains, nonuniform complexity theory learns that there may be uncountably many ways in which an agent can be up to its mission.

From the informal description one may observe that CBA agents are related to cyber-physical systems (CPSs) and embedded 'intelligent' agents as known in software engineering and artificial intelligence, respectively. However, the notion of CBA agents is more general since CBAs are not necessarily driven by objectively measurable internal sensations and standard computations like CPSs are and have qualities not found in common agent systems. The model of embodied cognitive behavioral agents with the properties informally described above, will be referred to as the *CBA model* in the sequel.

2.2 Discussion of the CBA model

Before we continue the study of the CBA model, we examine some of the key assumptions that were made in the list of properties in Subsection 2.1.

– The first assumption, stating that agents must be *finitely describable* at any given time, is of utmost significance. It eliminates 'infinite' agents, but does allow agents to 'grow' in size with time. For example, this option allows agents to store and retrieve data from past interactions, as long as their description remains finite.

– Next, in the description of the sensory qualities of agents, the possibility of *sensor arrays* must be included. The arrays are integral components of many biological and artificial systems and significantly influence their environmental interaction. Important examples include the complex arrays of photosensitive cells in the eye retinas of animals, which process a wide range of visual stimuli and contribute to the navigation of the animals in their environment, and the tactile arrays of mechanoreceptors in the mammalian skin, which allow

mammals to detect various tactile stimuli such as pressure, vibration, and texture. Both arrays of photosensitive cells and mechanoreceptors have their technological counterparts in artificial agents.

– By positing the existence of 'private' internal phenomena called*moods,* to be captured by special *mood sensors*, our model deviates from other approaches to conscious-like information processing in the theory of mind. Moods arise from the evaluation by an agent's internal control mechanisms of the impact of past and present external and internal events (cf. Definition 8). The evaluation occurs in the context of related past experiences, characterized by specific activation patterns within the agent's body. While standard sensors measure physical, chemical, or biological quantities and convert these into measurable outputs (typically electrical signals), mood sensors are *persistent semantic processes* that detect patterns or trends across the data from various agent mechanisms. These sensors assess attributes such as the rate, frequency, order, speed, intensity, and duration of activations – which can metaphorically be interpreted as the agent's moods.

A defining characteristic of moods is their private nature - essentially inexpressible in natural language and inherently tied to the agent. Unlike standard sensor data, moods are not directly measurable online, in real-time. Instead, they are latent within the data, necessitating longitudinal analysis of event sequences for detection. This is because events are logically linked, enabling both retrospective identification and continuous monitoring of relationships. These analytic procedures, analogous to sensory fields, reveal additional modal dimensions and trends within event sequences, as the corresponding data capture a latent spatiotemporal representation of selected event properties [20].

Processing moods requires complex, multi-modal data analysis, considering incoming external and internal inputs, their quality, intermediate results from prior interactions, and potentially unforeseen physical influences stemming from data processing patterns within the body such as feelings, emotions, qualia, and thought processes. Recognizing the presence and behavioral influence of moods is crucial for the investigation of the hard problem of consciousness.

– The data-processing ability of agents is crucial to the model. The assumption of the variable length of *interaction steps*, which may include several perception cycles, is essential for mood processing. It allows the 'observation' and temporal influence of events whose perception extends over several cycles and includes 'experiential qualities' crucial for the origination of subjective experience (cf. Section 4.2). The assumed *substrate independence* of the data processing implies that the ability of cognitive behavioral agents to satisfy their intrinsic properties (of which the mental-like properties of living creatures are a special case, as we shall see later) do not vary with the physical substrate used for the underlying data processing. To quote Tegmark [28]: *'it's only the structure of the information processing that matters, not the structure of the matter doing the information processing.'* However, the sensory data being processed and the motor data generated depend on the physical characteristics of the respective sensors and motors.

– The description of the data-processing ability so far leaves it open what the agents are actually supposed to do and remember based on the input data they perceive, in terms generated behavior and (stored, accumulated) data from ongoing and earlier interactions. Depending on context and purpose of the agents, the CBA model may be *tuned* by specifying in more detail what the data processing by an agent's body is to bring about when it is confronted with certain internal or external 'events' in the input. This allows for the wide applicability of the CBA model. In particular, it will be pursued in this study in Section 4.2 where we aim to model the mental facilities in agents and explain the emergence of subjective experiences in agentic behaviors.

Many systems are modeled by CBAs. As a straightforward example, consider the thermostat controlling a heating system. Its actions are simple: if the air temperature drops below 20° Celsius, say, it sends a signal switching on the heater. If the temperature is over

22° Celsius, the heater is off. The ability to maintain the temperature within the given bounds is the intrinsic property of the system; at the same time, it is its purpose.

In general, there are plenty of examples of systems in practice of which at least the core activities can be modeled by CBAs. The examples include humans, animals, plants, bacteria, robots, self-driving cars (cf. [39]), and various cognitive cyber-physical systems (cf. [38]). This claim can be verified case by case by inspecting whether such systems satisfy the general properties that characterizing CBAs. Software and AI systems like neural networks, large language models, or AI chatbots are also CBAs, their sensors being the input devices and their effectors being the output devices.

To give examples of agents that are not CBAs, think of physical systems that do not comply with all the properties of CBAs. For instance, systems that are not finitely describable; have a behavior that is not well defined, that are not physical systems, or that do not process information: rocks, water, rain, air, fire, crystals, etc.

2.3 Formalization of the CBA model

When it comes to formalizing the model of embodied CBAs, it makes sense to formalize only the *interfaces* between their sensors and the body and between the body and the external and internal effectors. After all, due to the black-box principle and the substrate independence of the data processing, we do not want to make any assumptions concerning how the sensory data are processed and how the instructions for the effectors are generated. Instead, we are interested in explaining the purpose of the actions of the agents, i.e., what they do, in various situations. As noted before, we only assume that the agents can solve 'easy problems of consciousness'. This restriction aligns with the explanatory nature of our model and the underpinning epistemic theory.

Let $\mathbb{A}$ be an agent, let Σ be the finite set of basic signals that can be delivered to and recognized by the external, internal and mood sensors of $\mathbb{A}$, and let Γ be the finite set of basic signals that can be sent to, received and interpreted by the effectors of $\mathbb{A}$. Assuming k input and ℓ output signals, for any $t \geq 0$ the pair $i_t = (s_t, b_t)$ consisting of an input *situation* $s_t \in \Sigma^k$ and a corresponding *behavior* $b_t \in \Gamma^\ell$ as output is called the t-th *interaction* of agent $\mathbb{A}$.

The core of an agent's activities is presented in the following definition. It defines the next activity of an agent, on a discrete time scale, as a function of the current situation and the history of all previous interactions.

Definition 1 (Next move). *For all $t \geq 0$, interaction i_{t+1} is the pair (s_{t+1}, b_{t+1}) where s_{t+1} is the (representation of the) environment at time $t+1$ as read by $\mathbb{A}$'s sensors and b_{t+1} is the corresponding (representation of the) behavior of $\mathbb{A}$ in the environment at time $t+1$ defined as*

$$b_{t+1} = F(s_{t+1}, i_0, i_1, \ldots, i_t) \qquad (*)$$

where F is the 'next-move function' realized by $\mathbb{A}$.

It is important to realize that Definition 1 is the basis for all of the agent's functions related to its presumed consciousness later, because it represents a relationship between perception, cognition, and action. Iterating the interactions, agent $\mathbb{A}$ works like a *transducer* producing unbounded sequences (*streams*) of consecutive interactions $(i_0, i_1, \ldots)$ over time [38].

Definition 2 (Interactive run). *Any sequence of consecutive interactions $(i_0, i_1, \ldots)$, generated by $\mathbb{A}$ over time in response to a sequence of input situations $(s_0, s_1, \ldots)$. is called an* interactive run *of* $\mathbb{A}$.

Definition 1 is related to the paradigm of *predictive processing* considered in cognitive neuroscience (cf. [10, 24]). Namely, we can see b_{t+1} as a *prediction,* or *anticipation*, of what agent $\mathbb{A}$ should do when being in situation s_{t+1} and having processed interactions

$i_0, i_1, \ldots, i_t$. Predicting b_{t+1} requires a form of looking ahead – thinking of and deciding about the future. Second, a high probability of the 'correctness' of prediction b_{t+1} assumes a high degree of the agent's *understanding* of its previous actions. Third, simultaneously, the realization of the respective anticipation is an expression of the agent's elementary *intentionality* under the given conditions. Last but not least, since mood sensations are involved, we see 'germs' of subjective experience and the formation of subjective points of view (cf. [11]). All these are essential constituents of self-awareness.

These observations already indicate the potential of the CBA model for modeling the mental faculties of AI systems. However, we will not follow the individual mental traits mentioned since our arguments toward the dissolution of the 'hard problem of consciousness' will be based on the general *semantic properties* of such traits rather than their particular properties (cf the next section and beyond).

Expression $(*)$ in Definition 1 shows that 'the next move' of an agent is dependent on its current situation and its complete history of previous interactions. This dependence is necessary to enable the reaction of the current behavior b_{t+1} to events occurring at arbitrary times in the past. From the data-processing point of view, the evaluation of $(*)$ calls for storing the entire history of agents' interactions, causing a potentially unbounded increase in time and space *complexity* of the underlying processes. In practice (as in LLMs, say), an agent will use heuristics to identify and store only the 'important' past events to mitigate this increase. On the one hand, this leads to space savings, but on the other hand, it leads to cases of catastrophic 'forgetting' in which incorrect reactions of agents are invoked.

Note that the concept of interactions does not refer to a current 'state' of the agent (whatever it could be) – interactions are just pairs of tuples of input and output signals linked together by the agent's operation. This aligns with our strategy to see the innerness of agents as substrate-independent black boxes with an unknown internal structure of which we can only observe the input/output behavior.

A problem with the view of agents proceeding in discrete steps might arise when the agents work continuously, with individual interactions not separated. In such a case, we will consider the discretization of the respective runs into a series of short-time steps. The "granularity" of such steps must be chosen so that only a fixed constant number of input and output changes happen during each step (cf. [33]).

Definition 3 (Agent behavior). *Let $\mathcal{L}_{\mathbb{A}}$ denote the set of all interactive runs of agent $\mathbb{A}$, generated by $\mathbb{A}$ in response to all possible sequences of inputs situations that $\mathbb{A}$ can encounter. $\mathcal{L}_{\mathbb{A}}$ is called the* behavior language *of (or, generated by) $\mathbb{A}$.*

$\mathcal{L}_{\mathbb{A}}$ represents the *behavior* that any instantiation of $\mathbb{A}$ can generate, under any sequence of circumstances that it can face.

2.4 Families of CBA agents

Instead of focusing solely on individual agents, we will often want to consider sets or specific collections of agents that are in some way related.

Definition 4 (Family of CBA agents). *A* family of CBA agents *is any set of cognitive behavioral agents operating in the same environment, sharing identical alphabets of input and output signals and processing the same streams of situational inputs.*

There may be many reasons for considering agents collectively. For example, agents may form a family because they have architectural or behavioral similarities (or both). It is not uncommon to assume that families of agents are *closed* under certain kinds of modification, construction or 'creation' of agents.

Definition 5 (Behavioral equivalence). *Two agents $\mathbb{A}$ and $\mathbb{B}$ from the same family are called* behaviorally equivalent *if and only if $\mathcal{L}_{\mathbb{A}} = \mathcal{L}_{\mathbb{B}}$.*

Thus, two agents from a same family are behaviorally equivalent if and only if they have the same behavior under all sequences of circumstances.

3 Agents, properties, and the problem of experience

In the general CBA model, an agent's behavior is characterized by its *semantic properties*. By definition, semantic properties are properties of interactive runs of an agent. The properties cannot be recognized syntactically from inspecting an agent's specification, since the behavior of an agent will normally vary with time and depend on its input and output signals and on the history of its previous interactions. In general, semantic properties can only be guaranteed by *design*.

In this section we explore the role semantic properties play in the CBA model and what the properties may tell us about the agents in a given family. We also consider the possible implications for our study of the hard problem of consciousness which, as argued by Chalmers [6], is essentially the problem of explaining what it is like for agents to *experience* (events, emotions and so on). In the sequel, when we speak of properties of agents we will always mean semantic properties of their behavior.

3.1 Intrinsic properties

An interactive run $(i_0, i_1, \ldots, \ldots)$ of agent $\mathbb{A}$ is said to satisfy propery P if and only if P holds for every initial segment of the run, i.e., if $P(i_0, i_1, \ldots, i_t)$ is satisfied for any $t \geq 0$. We are especially interested in properties that hold for *all* interactive runs of an agent.

Definition 6 (Intrinsic property). *We say that P is an* intrinsic property *of agent $\mathbb{A}$, or that $\mathbb{A}$* satisfies *or* can satisfy *P, if and only if P is satisfied for (every initial segment of)* every *interactive run of $\mathbb{A}$.*

Examples of intrinsic properties of human-like agents are, for instance, consciousness, attention, creativity, courtesy, understanding, adherence to given principles, and the like. For driver-less cars, the ability to apply the brakes or to respect the driving rules in any situation. For a large language model, the ability to react to any prompt.

An intrinsic property P of agent $\mathbb{A}$ is an *invariant* property of the agent's behavior that must be fulfilled for any given run of $\mathbb{A}$. It cannot happen that an agent satisfies P and "decides not to exercise it". Such a behavior would violate P. The adjective *'intrinsic'* stresses that condition P belongs to $\mathbb{A}$ by its very nature – $\mathbb{A}$'s interactions satisfy P under all possible circumstances. An agent that does not satisfy P under all circumstances cannot be equal to $\mathbb{A}$ (cf. Proposition 1).

If P is an intrinsic property of agent $\mathbb{A}$ then, for any interactive run $(i_0, i_1, \ldots)$ of $\mathbb{A}$, $P(i_0, i_1, \ldots, i_t)$ must hold for all $t \geq 0$. The dependence of P on all initial segments, thus on all previous interactions at any moment during the run, reflects $\mathbb{A}$'s ability to react to its history of interactions (which, primarily, reflects its learning capabilities). In this way, it displays elements of *self-awareness.* Under this point of view, intrinsic properties represent the *world model* in which the agent operates. This model drives the agent's behavior by capturing and understanding the relationships between current and past actions.

Note the crucial role of moods in shaping the world model. Namely, at any time $t \geq 0$, the moods are functions of $(s_{t+1}, i_0, i_1, \ldots, i_t)$. Hence, they are implicitly part of the next move function of an agent (cf. Definition 1). Therefore, moods must be implied by some intrinsic property P of the agent, capturing semantic properties of it.

An intrinsic property P depends on the agent's mission, design, and data processing abilities. It can be defined in various ways. For example, P may be specified as a quantified predicate over the sets of inputs from Σ^k and outputs from Γ^ℓ. P may also be defined by means of a ω-Turing machine (cf. [30]) that recognizes precisely those streams that satisfy it, or by any other suitable formalism reflecting the underlying data-processing mechanism. It is natural to assume that P has a finite description.

In the case of sophisticated agents operating in complex worlds, the definition of a property P may be very extensive and complicated since it must model the real world and the

agent's behavior under all circumstances as they may have occurred during the agent's mission. For example, consider large language models. Practically, the entire agent's 'body' is realized by a very large neural network that represents the aforementioned world model, enabling the realization of the next move function satisfying P. By Definition 1, this network and the respective inference mechanisms instruct the agent what to do in any situation it can encounter while respecting its previous interactions.

A single agent can satisfy several intrinsic properties at the same time. For instance, an agent can be curious but obedient to his master. This ability is a consequence of the compositionality of the quantified predicates that describe each intrinsic property. We use this ability of agents when explaining a 'what is it like to be' argument in Section 4.

3.2 Non-trivial properties

An important question is what properties might reveal of the 'internal processes' of a given CBA agent. In order to find out, we focus on families of agents. Clearly, different agents of the same family that process the same streams of situational inputs may exhibit very different behaviors. This is precisely what we are interested in: families populated by CBA agents with different behaviors even on the same inputs. We consider what their semantic properties might tell us.

Definition 7 (Non-triviality). *Let $\mathfrak{F}$ be a family of agents. Property P is said to be a* non-trivial property *for $\mathfrak{F}$ if and only if it there are agents $\mathbb{A}, \mathbb{B} \in \mathfrak{F}$ such that P is an intrinsic property of $\mathbb{A}$ but not an intrinsic property of $\mathbb{B}$.*

A property P will be called *trivial* for $\mathfrak{F}$ if and only if P is not non-trivial for $\mathfrak{F}$. As a consequence, P is trivial for $\mathfrak{F}$ if and only if either P is an intrinsic property of every agent $\mathbb{A} \in \mathfrak{F}$ or P is not an intrinsic property of any agent $\mathbb{A} \in \mathfrak{F}$.

Non-triviality can be a very practical notion. For example, consider a family of self-driving cars where some obey property PCR defined as *'prefer the cars arriving from the right',* and some do not. Property PCR is a non-trivial property of this family of cars. In this family, some cars always prefer the cars arriving from the right and some that do not always prefer the cars arriving from the right. The following general observation can be made for families of agents that satisfy a non-trivial property.

Proposition 1. *Let $\mathfrak{F}$ be a family of agents and P a non-trivial property for $\mathfrak{F}$. Let $\mathbb{X}$ and $\mathbb{Y}$ be two agents from $\mathfrak{F}$.*

(i) If P is an intrinsic property of one of the agents but not of the other, then $\mathbb{X}$ and $\mathbb{Y}$ are not behaviorally equivalent.

(ii) The converse statement does not always hold, i.e., if $\mathbb{X}$ and $\mathbb{Y}$ are not behaviorally equivalent, then it does not necessarily follow that P is an intrinsic property of one of the agents but not of the other.

Proof. (i) Assume without loss of generality that P is an intrinsic property of $\mathbb{X}$ but not of $\mathbb{Y}$. This means that P is satisfied by all interactive runs of $\mathbb{X}$, but not by all interactive runs of $\mathbb{Y}$. It follows that $\mathcal{L}_{\mathbb{X}}$ and $\mathcal{L}_{\mathbb{Y}}$ cannot be equal and thus, that $\mathbb{X}$ and $\mathbb{Y}$ are not behaviorally equivalent (cf. Definition 5).

(ii) If $\mathbb{X}$ and $\mathbb{Y}$ are not behaviorally equivalent, it can still happen that the behavior of both $\mathbb{X}$ and $\mathbb{Y}$ satisfies P or that the behavior of both does not satisfy P. This is easily argued e.g. in the example of the family of self-driving cars given above. It follows that P is not necessarily an intrinsic property of one of the agents but not of the other. □

Note that Proposition 1 *(i)* does not claim that the behavioral inequivalence of agents $\mathbb{X}$ and $\mathbb{Y}$ will be observable along *every* identical sequence of situational inputs. Rather, the opposite will be true – in many identical situations, $\mathbb{X}$ and $\mathbb{Y}$ could well exhibit the same behavior. Behavioral inequivalence only means that there must *exist* identical sequences

of situational inputs that will lead, in finite time, to different behaviors of the agents, as they must satisfy the complementary pair of properties (always) P and *'not always'-P* respectively.

Proposition 1 leads to a simple, but significant principle for families of agents with a non-trivial property. Namely, if P is a non-trivial property for a family of agents $\mathfrak{F}$, the proposition asserts the existence of behavioral differences between at least some pairs of agents in $\mathfrak{F}$, implying a *physical discrimination* between the agents that always satisfy P and those that do not always satisfy P. The reason is that the behavioral differences between agents $\mathbb{X}$ and $\mathbb{Y}$ in Proposition 1 *(i)* must be caused by the different ways in which they process their input and output streams, since one of them has to satisfy P all the time, while the other agent can have runs not satisfying P all the time. However, note that Proposition 1 *(i)* is *non-constructive* – it only states the *existence* of runs with different properties and thus a different behavior, but not say how to find such runs.

The considerations show that notion of semantic properties is essential for defining and understanding the behavior of CBA agents. However, deciding whether an agent satisfies a given property intrinsically or not can be computationally difficult. It is worth mentioning here that, in general, deciding whether a given agent satisfies a given non-trivial property is an *undecidable* problem, i.e., there is no universal algorithm to fulfill such a task in finitely many steps.

Proposition 2 ([39]). *There exist families of CBA agents $\mathfrak{F}$ and non-trivial properties P for $\mathfrak{F}$ such that the problem of deciding whether a given agent $\mathbb{A} \in \mathfrak{F}$ always satisfies P is algorithmically unsolvable. In particular, the problem is algorithmically unsolvable by another agent in $\mathfrak{F}$.*

Reference [39] proves the result for (infinite) families of very general agents and for all non-trivial properties P that are in a well-defined sense 'natural' for the CBA model. It makes Proposition 2 similar to *Rice's Theorem* in computability theory (cf. [21]), now adjusted to the premises of the CBA model.

Proposition 2 is essential from a 'practical' viewpoint – it shows that there is no universal agent that can always determine from another agent's description ('program') whether this other agent satisfies a given intrinsic property. In practical terms, this result means that the boundary between a 'genuine' agent, satisfying a non-trivial intrinsic property (such as adherence to human values), and a simulating agent, not always satisfying this property, may be hard to establish.

Another important lesson from this finding is that no 'universal' empirical benchmark would allow deciding whether an arbitrary agent (or system) possesses consciousness, contrary to the beliefs of some thinkers. For an extensive discussion of this issue, cf. [9]. Especially if emotions, qualia, or feelings are, for some agents, their intrinsic properties, then the distinction between the 'genuine' emotional experience and its sophisticated simulation ('conscious-seeming AI') becomes blurred.

Note that Proposition 2 does not preclude the existence of specialized benchmarks for testing the presence of well-defined consciousness or, in general, of intrinsic conditions in specific entities.

3.3 On the problem of experience

Proposition 1 proves useful in our further considerations of the hard problem of consciousness. Chalmers [6] argued that the essence of the hard problem (for human agents) is to explain what it is that makes agents *experience* (thoughts, emotions and so on). While agents are processing information, performing all sorts of functionalities internally and externally that can usually be functionally defined and explained, why and how does this also give rise to subjective experiences ([6], p. 6)?

In reviewing various possibilities, Chalmers [6] argues that purely physical accounts of experience as a phenomenon are not very likely. He even claims that ([6], p. 12):

> *"The facts about experience cannot be an automatic consequence of any physical account, as it is conceptually coherent that any given process could exist without experience. Experience may* arise *from the physical, but it is not* entailed *by the physical."*

However, this conclusion seems to go too far. We will show that Chalmers' claim is *contradicted* by the observations underlying Proposition 1 which point to a marked difference between agents that satisfy opposing intrinsic properties.

For a general argument we consider any family of CBA agents satisfying a non-trivial property P (with 'experience' being a special case) whose satisfaction is externally perceptible through behavioral reactions of the agents. An example of such a property is the sensation of damage that can occur both in biological and artificial agents, with some agents reacting to it and some not. A different example might be the behavior of a missile defense system that reacts differently to hostile missiles heading towards their targets and those going astray. We identify any such property P with the agent's experience of satisfying P. The following proposition contradicts Chalmers' claim explicitly.

Proposition 3. *Let $\mathfrak{F}$ be a family of CBA agents as described above, and assume that 'experience' as a property is non-trivial for $\mathfrak{F}$. Then, at least for some agents, experience is* entailed *by the physical.*

Justification. We will view agents as processes, to be compatible with Chalmers' parlance. Assume that, as claimed, any given process satisfying a non-trivial property P could also exist without satisfying property P. Let $\mathbb{X}$ be any process satisfying P and let $\mathbb{Y}$ be any process not satisfying P. ($\mathbb{X}$ and $\mathbb{Y}$ exist by non-triviality of P.) By Proposition 1 *(i)*, $\mathbb{X}$ and $\mathbb{Y}$ are not behaviorally equivalent. Hence, $\mathbb{X}$ and $\mathbb{Y}$ must be different processes. Therefore, it is not conceptually coherent that process $\mathbb{X}$, satisfying P, could also exist as a process $\mathbb{Y}$ not satisfying P. This contradicts the assumption. Hence there must be processes $\mathbb{X} \in \mathfrak{F}$ satisfying P that cannot exist without satisfying P. For these processes, P is an automatic consequence of $\mathbb{X}$: satisfaction of P is entailed by the physical. □

Proposition 3 has severe consequences for understanding the hard problem of consciousness. Namely, one of the arguments defending the hard problem relies on the existence of *philosophical zombies*. A philosophical zombie is a being in a thought experiment in the philosophy of mind that is physically identical to a normal human being but does not have conscious experience [8]. Proposition 3 states that the existence of philosophical zombies is logically incoherent: if they existed, their behavior could not faithfully mimic the behavior of humans under all circumstances.

In the next section, we will argue that experience is an essential component of consciousness.

4 Events and their experiential qualities

We will now (finally!) elaborate on our framework of CBA agents, for the specific purpose of gaining further insights into the hard problem of consciousness. Our ultimate aim is to find answers to the key questions at the heart of the problem as identified by Chalmers: *'why and how does subjective experience arise'* (cf. [6], p. 3), and *'why is the performance of cognitive functions accompanied by experience'* (cf. [6], p. 5). We will describe what mechanisms in the agents could potentially account for it.

Our considerations will be speculative and hypothetical because we will attempt, within our agential model, to bridge the so-called *explanatory gap* using an uncharted mindset. The explanatory gap is the core of the 'hard problem' as we have been describing it above (cf. [6], p. 6), namely the challenge of explaining how and why the physical properties of processes realized by agents can give rise to corresponding subjective experiences. Until now, no adequate physicalist solution to this problem has been described.

To resolve it in our context, we extend the CBA model with a novel framework for coping with the *experiential qualities* of *events* and propose *mechanisms* that translate these qualities to *subjective experiences*. Then, we will investigate to what extent the solution we find for the model is plausible and corresponds to the insights and ideas we have for the case of human consciousness.

This lengthy section is organized as follows. First, in Section 4.1, we define several new notions that are necessary to explain our approach: events, experiential quality, subjective experience, and longitudinal perception, and we give some examples. Next, in Section 4.2, we take a closer look at what is happening internally in CBA agents and explain the mechanisms that give rise to subjective experience in them. In Section 4.3 we discuss the meaning of our findings, e.g. for explaining the *'what-it-is-like'* aspect and the *inseparability* of subjective experience. Finally, in Section 4.4 we return to Chalmers' questions above and show how they are answered by our theory.

4.1 Concepts

Before tackling the problem of subjective experience, we will describe the various key concepts we will be using in detail. First, we devise the following definition of *events* that reflects their spatial, temporal, and structural aspects as perceived by the agents.

Definition 8 (Events). *Events are perception records of internal, external, and mood sensations and their qualities during bounded time periods. Events come in two forms:* elementary *and* composed *ones.*

(i) *An* elementary event *is a discrete time-bounded period of sensations lasting over several cycles received from the agent's body or environment and perceived by the agent, that cannot be further decomposed into simpler events.*

(ii) *A* composed event *is a well-distinguishable, goal-oriented, time-bounded sequence of sensations which the agent receives from its body or environment. A composed event is characterized by its purpose, grounded in the agent's goals, onset conditions, announcing the event's arrival, progression pattern, characterizing the event's time and space evolution, and termination condition. The purpose of a composed event is formally described as an intrinsic property R that the agent has to satisfy during the event at hand. Property R is implied by the overall intrinsic property P satisfied by the agent.*

(iii) *Events can be composed of other events to form a more complex structure.*

Thus, an event is characterized by sensory information of 'what' happens in the agent's environment (the onset aspect), 'where' it happens (spatial aspect), 'when' it happens (temporal aspect), 'how' it happens (qualitative and structural aspects), and the purpose 'why' it happens.

Examples of elementary events are: typical human qualia in the philosophy of mind like the redness of an evening sky, the scent of a lemon, the triumphal sound of a trumpet, and the like. For artificial systems, like mobile phones: the reception of a strong GPS signal, of an incoming call, or acceleration changes. For missile defense systems: the interception of an adversarial missile, for heating systems: an abrupt temperature fall, etc. Perception of all those events lasts for several interaction cycles.

For humans, an example of a composed event is a long jump in athletics, consisting of four phases: approach, takeoff, flight, and landing. Each phase is a well-recognizable, simpler event which, however, can be further decomposed into elementary events. For self-driving cars: emergency braking. For mobile phones: call by voice.

Continuing the exposition, we consider the *experiential qualities* of an event. The notion depends on the given event and on the time moment at which an agent registers the event.

Definition 9 (Experiential quality). *The* experiential quality *of an event $\mathcal{E}$ registered by agent $\mathbb{A}$ at time $t \geq 0$, i.e. after processing interactions $i_0, i_1, \ldots, i_{t-1}$, is the complete record integrating $\mathbb{A}$'s external, internal, and mood sensations at time t.*

Experiential qualities of events are the analogs of word-to-vector encodings of words in large language models (cf. [43]). Their encodings represent multi-modal high-dimensional vectors of sensory and motor signals as registered by an agent at a given time.

Now, we are ready to define the notion of *subjective experience*.

Definition 10 (Subjective experience). *The* subjective experience *of agent $\mathbb{A}$ of event $\mathcal{E}$ at time $t \geq 0$ refers to the impact of $\mathbb{A}$'s processing of the experiential qualities of $\mathcal{E}$ as registered by $\mathbb{A}$ on its internal structures at time t and, indirectly, on $\mathbb{A}$'s future doings. It encompasses the sensory, cognitive, behavioral, and structural responses that $\mathcal{E}$ evokes when realizing its progression pattern, fulfilling the agent's intrinsic property R (cf. Definition 8).*

In other words, subjective experience is a *process*, or algorithm, pertinent to the registration of $\mathcal{E}$ and its influence on the current and future behavior of the agent. Processing subjective experiences related to experiential qualities of perceived events is at the heart of many applications – in fact, of any agent's activities. For example, subjective experiences are important for face recognition, where they allow to discriminate between known and unknown faces and eventually identify the person at hand or recognize a person's mood. Another example is the classification of obstacles by autonomous vehicles.

Subjective experiences are essential in determining an agent's *next move* as a function of $\mathcal{E}$ and all previous interactions (cf. Definition 1). In the discussion after Definition 2 we have stressed that, philosophically speaking, following a next-move function amounts to an educated guess concerning the immediate future development of an agent's environment. Based on the agent's experience and current situation, this guess anticipates the agent's forthcoming action.

This effect of anticipation and its exploitation in the definition of subjective experience has several significant consequences for the theory of mind. For one, it means that subjective experiences directly relate to the concept of consciousness and its numerous particular cases mentioned following Definition 2. Note that Definition 10 also points to the relationship of consciousness to the anticipation of future states. This relationship is the game-changer for philosophical zombies. It explains the observable difference in behavior between agents that are zombies and agents that are not, as was explained right after Proposition 3.

In the context of the hard problem of consciousness, the central message of Definition 10 lies in the fact that it bounds the subjective experiences of agents to their world model. Metaphorically, to paraphrase Stephen Hawking, the latter fact *'breathes the fire'* into the agents and makes their world model the place to behave. (The metaphor refers to Hawking's famous question:*'What is it that breathes fire into the equations and makes a universe for them to describe?* which he posed in [13].)

This insight provides a strong argument in favor of Definition 10. Namely, it represents a bridge over the *explanatory gap* in the philosophy of the mind: the physical properties of the impact of events give rise to the subjective experiences of the agents that correspond to the evolution over time of their behavior as learned in their world model.

The problem with the definitions mentioned above is their generality, which stems from our effort to cover a large class of agents. The definitions focus on *subjective perception*, which differs from agent to agent not only among agents of diverse construction but also among agents of the same type. This is due to variations in the technical parameters of their sensory-motor equipment.

The situation is better for artificial systems since we can shift our perspective from subjective human experiences to objective, measurable qualities like speed, acceleration, frequency, or energy consumption. Only with a further specification of an agent's properties, like embodiment, sensory-motor equipment and their technological idiosyncrasies, data-processing abilities, environment, and purpose, can we give a more specific definition of what may be meant by the experiential quality of an event and its impact as registered

by the agent. The situation here is similar to the definition of intrinsic properties: subjective experiences are non-uniform properties of an agent's embodiment.

Finally, we define is the notion of a *longitudinal perception* by an agent.

Definition 11 (Longitudinal perception). *A* longitudinal perception *is the repeated perception of the same sequence of percepts during the repeated occurrences of an event.*

The term 'longitudinal perception' refers to the situation when the same perceptions or sequence of different perceptions (a cause) are processed throughout the instances of the respective elementary or composed events. If this occurs, longitudinal perception manifests itself as a subjective, composed perception of the event's impact (i.e., subjective experience) on the agent. It includes sensory, cognitive, behavioral, and structural responses (the effects) evoked by the event and realized by the agent's internal mechanisms. The *external* evidence of such causation is the observed behavior of the agent, where the cause is responsible for the effect, and the effect is dependent on the cause. The *internal* evidence is the arousal of the subjective experiences in the form of sensory, cognitive, structural, and behavioral responses that the event evokes.

4.1.1 Examples To illustrate the concepts of experiential qualities and subjective experience, Table 1 provides examples of natural and artificial agents with nonhuman sensory capabilities. The agents use their sensors to sense their environment and detect the experiential qualities of events.

Type	Agent	Nonhuman sensor	Experiential quality	Subjective experience	A possible form of consciousness
Biological	Bats	Sonar Echolocation	Spatial map of environmental echoes	*"insects flying close in front of me"*	3D environment perceiving through echoes
	Bees	Ultraviolet vision	UV images of flowers	*"approaching a honey-bearing flower"*	Consciousness in the color spectrum invisible to humans
	Sharks	Electroception (EM field sensing)	Electrical pulses from other animals	*"a big shark on my left"*	The world as a symphony of electromagnetic signatures
Plants	Sunflower	Photoreceptors Chemoreceptors Mechanoreceptors Thermoreceptors	Sunlight as a source of heat and direction of orientation. Gravity as a directing force for growth. Water as nourishment and relief.	The sensation of thirst. The setting of the flower against the sun. Perception of day length and climatic changes. The "joy" of pollination.	Perception of the environment through subjective changes in experiential qualities mediated by receptors at various time scales.
Artificial	Autonomous vehicles	Lidar, radar, high-resolution cameras, GPS, ultrasonic sensors	Object distance, position measurements speed, acceleration, etc.	*"safe distance from the previous vehicle"*	Circular 2D perception of the world through the position, distance and movements of objects
	Industrial AI systems	Acoustic emission sensors	Subtle structural changes	*"no deviation from normal"*	Perception of stress, fatigue, and material failure
	Weather forecast	Various meteorological sensors	Changes in weather parameters	*"storm front approaching"*	3D weather trends map
	Environmental monitoring	Gas concentration and vibration sensors	Air quality and vibration	*"dangerous dust particle concentration"*	Sensing chemical changes and seismic activity
	Universal human agent	Human-like sensors with extended ranges and resolution, and other nonhuman sensors	Superhuman multi-sensory model of the world	Similar to humans, but with much higher quality	Superhuman multi-sensory consciousness lacking mental imagination

Table 1. Natural and artificial agents with non-human sensing

The subjective experiences of these systems reflect the experiential qualities of events, regardless of whether they possess the mental imagery capability inherent in most humans. Agents interpret and respond to events in ways that indicate the presence of consciousness despite having 'only' nonhuman human perception.

In the table, the description of experiential and subjective qualities and possible forms of consciousness is only *metaphorical* – natural language has neither the words nor the semantics, let alone the pictorial imagination, to express these concepts in a way reflected by the given agents. In fact, these descriptions describe the agents' actions as if they were conscious in the human sense, which, of course, they are not.

4.2 Emergence of subjective experience

We now elucidate the mechanism that gives rise to the subjective experience of CBA agents. To do so, we adhere to the philosophy of epistemic computation [36]. It means that, when describing the impact of events, we concentrate on WHAT is happening rather than on HOW things happen. Our findings will be formulated in Proposition 4 below .

The proposition presents an informed *hypothesis* supported by empirical introspection, evidence from neuro-imaging, and analogy with the functioning of current large language models (LLMs). The details are facilitated by the features of embodied CBA agents, the anticipation ability based on their world models, and the mechanisms generating experiential qualities of events.

In Section 4.2.1 we first give details of what goes on inside our CBA agents. In Section 4.2.2 we present and justify the Proposition 4 that describes how subjective experiences emerge. For simplicity, we will not distinguish between elementary and composed events in what follows. We will see the former events as specific cases of the latter, with no internal longitudinal structure. The purpose of elementary events is to draw attention to the ongoing phenomenon.

4.2.1 Closer look at the internal working of CBA agents Before describing the mechanism underlying their qualitative experiencing, we sketch the general scheme of the internal operation of a CBA agent.

We see CBA agents as entities processing streams of elementary events. The basic assumption is that an agent observes but a finite number of such events, each having only a finite number of parameters associated with it. All this observed data is kept and memorized in some way by the agent in a searchable *catalog* in its internal black box structure. The finiteness assumption is based on the fact that every agent possesses only a finite number of perceptual modalities, each being represented by a finite construct of finite tuples over a finite set of signals (cf. Section 2.3).

The catalog also stores additional information pertinent to each event. Upon perceiving and registering an event, an agent finds it in its catalog and proceeds according to the executive information attached to it. As far as the mechanisms realizing the corresponding operations are concerned, we assume that their implementation is part of the 'easy problems' of consciousness and is functionally realized. Due to the non-uniformity of the agents, the implementations may differ from case to case and will be hidden within the black-box architecture of the agents.

The information associated with events in the catalog is of two kinds:

- First, routing information connects the given 'entry' of an event with its possible successor events. Such information is helpful in the case of composed events. In general, 'chaining' of events allows an agent to predict and anticipate the expected circumstance and choose an optimal way of reacting to a given event.
- Second, there is additional, auxiliary information characterizing each event. Such information comprises any temporal, spatial, and qualitative data pertaining to the event (cf. Definition 8).

The interconnection of all events gives rise to a giant finite *network* of events with auxiliary information attached to each event. One can see such a network as a specific model of the world in which a given agent operates. Of course, the respective model will generally be extensive but, in any case, finite.

There is an analogy between the processing of words of a natural language in LLMs and the response of CBAs to ongoing event processing. Events are like the words of a natural 'event language' through which the environment 'talks' to the agents, and the agents 'reply' through their internal and external actions. This language consists of elementary events that connect to higher-level composed events. The agents learn the semantics of the event language, which they translate into their behavior. The analogy works because our agent's event language is finite, just like the learned fragment of natural language in an LLM.

At this level of description, we will not go into further details. By the explanatory nature of our model, we will not describe how the entire catalog of events is established except to note that *learning* plays a crucial role in this process, similar to the case of LLMs. We assume that the problems involved belong to the class of 'easy problems' (in Chalmers' parlance) and can be functionally specified.

4.2.2 Subjective experience and CBA agents We are now able to describe the mechanisms giving rise to the subjective experience of CBA agents. We formulate the effect as a proposition (Proposition 4) and justify the proposition by giving supporting evidence for it. (A proof would require some kind of formal or informal background theory.)

The context of the proposition is that of the repeated perception of an event $\mathcal{E}$. Recall that the presumed impact of event $\mathcal{E}$ is correlated with an intrinsic property, say R, that the agent is to satisfy during the event (cf. Definition 8). The proposition describes a *causation link* between registered physical stimuli and the corresponding 'mental processes' of the agent at hand.

Proposition 4 (Emergence of subjective experience in CBA agents). *Let $\mathbb{X}$ be an agent. Consider its longitudinal perception of an event $\mathcal{E}$. Let R be the intrinsic property of $\mathbb{X}$ to be satisfied upon the occurrence of $\mathcal{E}$. Assume that $\mathbb{X}$ is constructed such that the sequence of percepts pertinent to the longitudinal perception of $\mathcal{E}$ gets memorized in the agent's event catalog at prior occasions of the event.*

As the onset of next occurrences of $\mathcal{E}$ gets recognized (i.e. not for the first time), $\mathbb{X}$ is repeatedly experiencing the known event, as follows:

(i) The remembered (stored) progression pattern of the longitudinal perception whose task is to satisfy R during $\mathcal{E}$ is accompanied by the ongoing progression of the actual *event observed by $\mathbb{X}$ in the real-time interaction, thereby generating the subjective experience accompanying the impact of $\mathcal{E}$ on the agent, encompassing sensory, cognitive, behavioral, and structural responses.*

(ii) Depending on the agent's design, the subjective experience invoked by $\mathcal{E}$ take the form of qualia, emotions, feelings, or internal changes that might lead to
- *the modification of perceptual qualities of the event,*
- *qualitative cognitive changes recorded in the internal mechanism of $\mathbb{X}$, and*
- *behavioral reactions implied by the satisfaction of R.*

The respective actions of $\mathbb{X}$ occur during the duration of event $\mathcal{E}$.

(iii) Thanks to the construction of its event catalog, $\mathbb{X}$ can predict the progression of $\mathcal{E}$ and perform the reactions accompanying the subjective experience generated in (ii).

(iv) Simultaneously, the ongoing persisting longitudinal sensory perception of $\mathcal{E}$ gives rise to repeated arousal of the respective sensation mechanisms that keep refreshing the impact of $\mathcal{E}$. In a substrate-independent way, the sensations of this impact give rise to a unique subjective experience of $\mathcal{E}$.

(v) Depending on the agent and its purpose, this emerging subjective experience takes various forms that may be substrate-dependent, with varying degrees of vividness and intensity.

Justification. (i) From a physical point of view, the ability of $\mathbb{X}$ to satisfy R upon the occurrence of $\mathcal{E}$ as described here is a consequence of the construction of $\mathbb{X}$ – the agent is designed in this way (cf. Section 4.2.1). The primary cause for subjective experience invocation is the agent's qualitative response triggered by the onset of event $\mathcal{E}$.

The processing and the actions realizing the longitudinal perception of $\mathcal{E}$ are not for free – they consume a certain amount of *energy*. This fact qualifies the ability of $\mathbb{X}$ to satisfy R as a physical process. This process occurs whenever the onset of $\mathcal{E}$ is recognized, and R is satisfied along the prefix of the respective run until the present interaction. It continues during all activities of $\mathbb{X}$ throughout event $\mathcal{E}$.

(ii) The description of the actions of $\mathbb{X}$ invoked by its qualitative response at the occasion is based on empirical evidence and introspection. It is also supported by the recent findings in functional neuroimaging research concerning the neural underpinning of 'human' experience [17, 26].

Unfortunately, and understandably, the non-uniformity of embodied agents makes it generally impossible to specify concrete details of the impact of experiential qualities of events like $\mathcal{E}$ on an agent's sensory, cognitive, and behavioral reactions. From the description of the causation mechanism, it is clear that its qualitative effect depends on the properties of the agent and its mission.

Note that it is here where the *explanatory gap* (cf. Section 4.1) between the 'physical' and the 'mental' is crossed: experiential qualities of events, as determined by the various sensors of an agent ('physical phenomena'), give rise to the agent's subjective experience ('mental phenomena') that manifest themselves through the agent's subsequent internal and/or external actions.

(iii) This claim follows from the organization of the event catalog, allowing a prediction of the expected progression of $\mathcal{E}$ and the expectation of the respective series of sensations. Here, an additional mechanism must be invoked to select the most promising event that will succeed the current event. The working of this mechanism depends on the agent.

(iv) The emergence of a unique subjective experience and the description of the underlying mechanism as stated here is a bold hypothesis. We give two plausible arguments supporting this hypothesis.

- First, we note several reasonings due to Tegmark. In [27], he writes: *'I believe that consciousness is the way information feels when being processed.'* In our model, consciousness is the impact of an agent's data processing, pertinent to determining its next move. Tegmark's arguments supporting this statement refer to the human brain and stem from a combination of insights from physics, neuroscience, and information theory.
  In [29], Tegmark claims: *'If consciousness is the way that information feels when it is processed in certain ways, then it must be substrate-independent; it is only the structure of the information processing that matters, not the structure of the matter doing the information processing. In other words, consciousness is substrate-independent twice over!'* The claim supports the substrate-independency of the respective phenomenon.
- A different line of argument may be based on 'reasoning by analogy'. Recall that, by Proposition 3, the emergence of subjective experience in 'many' CBA agents is entailed by a physical process (e.g. realized by electrochemical signaling in the brain). If one considers this across all agents, one sees that 'by analogy' subjective experience can be a special case of the physical data processing in an agent's black boxes.

(v) Note the cautious formulation here. Namely, one cannot claim that the subjective experience accompanying the (processing of) experiential qualities of events will always take a clear and telling form. An example is given by the experiential qualities of the relatively little known phenomenon of *aphantasia*.

Aphantasia refers to the reduced ability, or complete lack of it, to voluntarily generate mental imagery in the human mind ([44]). It occurs in varying degrees of vividness and extends across multiple senses, including visual, auditory, gustatory, olfactory, tactile, and

motor imagery. In our general framework, we cannot exclude the existence of this phenomenon in any properly designed agents.

Aphantasia is not a disorder – it is a different experiencing of the world that still suits its purpose. This conclusion aligns with the findings in neuroscience (cf. [44]). If people's experiencing of the world varies, the same can be expected from subjective experiences in whatever (non-uniform) families of CBA agents. □

Concerning mental imagery, aphantasia represents a challenge to the notion of qualia and, in general, subjective experience. Nevertheless, our algorithmic description of the emergence of unique subjective experiences and the ensuing processes shows that even in cases of a complete loss of imagery in some agents, the ensuing processes still work undisturbed. Subjective experience still 'does its job' of providing a link between experiential qualities of sensation, cognition, and action as a core of consciousness.

4.3 Significance of Proposition 4

Proposition 4 describes how CBA agents handle their streams of 'events' and what causes them to have 'subjective experiences'. In this way Proposition 4 contributes essentially to our understanding of the 'hard problem of consciousness'. We discuss some further aspects and implications, from a philosophical viewpoint

In Section 4.3.1 we first give some general comments on the various components of Proposition 4. In Section 4.3.2 we argue that the proposition provides interesting support for Nagel's *'there is something it is like'* argument for conscious organisms, here CBA agents. Finally, in Section 4.3.3 we present a proof that subjective experience is *inseparable* for an agent.

4.3.1 General remarks The chief contribution of Proposition 4 is its elucidation of the circumstances and mechanisms invoking experiential qualities of events and their impact on CBAs: the recognized onset of 'familiar' events $\mathcal{E}$, triggering an agent's subjective qualitative reaction, altering its properties, and the production of an adequate behavior.

Some further observations can be made when we consider each of the items *(i)* through *(v)* of Proposition 4 separately.

– Ad *(i)*: The justification of *(i)* dismantles the seemingly intractable philosophical mystery around the origin of subjective experience or, why the performance of mind-like functions (or, the induction of non-trivial intrinsic properties) is accompanied by experience, as Chalmers has put it. The primary cause is the agent's qualitative response triggered by the onset of event $\mathcal{E}$.

– Ad *(ii)*: The statement of *(ii)* is not entirely satisfactory since, in the formulation, we concentrated on WHAT the qualitative feedback causes rather than HOW it is to be realized. However, this is the best we can do, due to the non-uniformity of agents.

Note, as consciousness occurs in many different forms, so do experiential qualities of events, due to the same non-uniformity. See also the remarks ad *(v)* below.

– Ad *(iii)*: The ability to predict the progression of events, stated in *(iii)*, implicitly allows agents to monitor this progression, anticipate the consequences of possible actions, plan effective sequences of actions accordingly, and possibly adjust their behavior. This ability of CBA agents is of good use e.g. in biological systems supporting rational behavior. The same applies to artificial systems. For instance, for a defense system under attack by an adversary, it is crucial to monitor and predict the trajectories of incoming missiles and focus on those targeting inhabited areas, while disregarding missiles aimed at uninhabited areas. See also the comments on 'understanding and thinking' in Section 5.

– Ad *(iv)*: The concluding effect of experiencing events, stated in *(iv)* for CBA agents, reminds of Nagel's argument that *'no matter how the form may vary, the fact that an organism has conscious experience* at all *means, basically, that there is something it is like to* be *that organism'* ([18], p. 436). Applied to CBA agents, this 'something it is like to

be that organism' is given by an agent's own specific internalization of the mechanisms in Proposition 4. We will discuss this in more detail in Section 4.3.2 below.

– Ad *(v)*: The nature of an agent's qualitative response to an event depends on the agent's embodiment and properties. Examples of the qualitative response to elementary events mentioned earlier (cf. Section 4.1) include, for instance, more intensive heating in case of an abrupt temperature fall or the retuning to a different provider during an unexpected signal loss for a mobile phone.

In the case of human agents, the kind of response is not easy to determine. For instance, it can take the form of a positive reinforcement of the attention paid to the ongoing events, making the subjective experiences more 'vivid' or telling than without this reinforcement. Other alterations may be included as well. For instance, the emotional intensity of repeated recalls of distressing or sad events tends to diminish with time. Examples of such responses are grief, envy, regret or euphoria. Their temporal variability indicates that qualia and emotions are not fixed properties of agents. Qualitative responses occur in real time and include possible effects of not explicitly known mechanisms influencing the data processing inside the agent's black boxes via mood sensors.

On the other hand, there are simple agents, like automatic door opening systems, that can operate entirely without experiential qualities. There are many different types of agents in between these two extreme examples, all realizing a qualitative response that serves their purposes in a specific, hence non-uniform way. Therefore, the formulation of item *(ii)* must remain in its present less specific form, to cover a broad spectrum of possible cognitive agents.

4.3.2 'There is something it is like' Further to the remarks ad *(iv)* above, we observe that Proposition 4 supports Nagel's argument [18] to the effect that *'an organism has conscious mental states if and only if there is something that it is like to* be *that organism – something it is like* for *the organism'* ([18], p. 436). This follows once we see that 'there is something it is like' for a CBA agent, i.e. for that agent, to experience, as implied by its own specific internalization and realization of the mechanisms in Proposition 4.

Namely, one can view the entire lifespan of an agent as the elaboration of a *complex composed event* that must satisfy a host of distinct non-trivial inherent properties characterizing the components of consciousness, like wakefulness, self-awareness and awareness of the environment, attention, curiosity, creativity, understanding, adherence to certain principles, etcetera. Each component event generates its own specific experiential quality. All properties are intrinsic and are continuously and concurrently to be satisfied by an agent, under any circumstance that occurs throughout the agent's lifespan.

In each interaction, the respective multi-sensory perceptions of an agent 'pour together' as a never-ending stream of simultaneous elementary events that are continuously processed by the agent. The ongoing concurrent satisfaction of the corresponding longitudinal events is manifested via qualitative responses as complex causations, giving rise to subjective physical phenomena that invoke an interplay between the respective experiential qualities and the corresponding 'something-it is-like-to-be-that-agent' aspect of the agent. At the same time, thanks to the predictive abilities and other aspects of the mechanisms in Proposition 4, the underlying agentic processes allow the agent to exert ongoing long-term control over its doings.

4.3.3 Inseparability of subjective experience A further observation can be made which basically continues the given argument. In the philosophy of mind it is widely accepted that (consciousness and) subjective experiences are inherently linked to the individual experiencing them and that these cannot be 'separated' from that individual's perspective. We argue that this holds for individual CBA agents as well.

The argument is essentially implicit in Section 4.3.2 where we concluded that, when it comes to the conscious experience of a CBA agent, 'there is something it is like to be that

agent' to have that particular experience. However, using Proposition 4, the *inseparability* of subjective experiences for CBA agents can also be argued more intuitively.

Proposition 5 (Inseparability of subjective experience.). *Let $\mathbb{X}$ be an agent, $\mathcal{E}$ an event, and R the intrinsic property of $\mathbb{X}$ to be satisfied upon the occurrence of $\mathcal{E}$. Then the ability of $\mathbb{X}$ to satisfy R, generating the respective subjective experience pertinent to $\mathcal{E}$, is inseparable from $\mathbb{X}$.*

Proof. According to the items of Proposition 4, the ability of $\mathbb{X}$ to satisfy R during event $\mathcal{E}$ under all circumstances manifests itself as a substrate-independent physical process with behavioral qualitative effects on $\mathbb{X}$. Consequently, $\mathbb{X}$ changes its relationship to its environment and innerness as this happens. The effects of these changes become manifest in the next interaction. Thus, in any interaction, under all circumstances, $\mathbb{X}$ acts as a 'producer' of the physical processes related to the *'satisfaction of R'* (i.e., of the longitudinal perception), to become in the very next interaction a 'consumer' of the effects of these processes (i.e., the beneficiary of the longitudinal perception). This cyclic causality means that the phenomenon related to the *'satisfaction of R'* cannot become evident without this particular agent's simultaneous presence; hence, this phenomenon is inseparable. □

Proposition 5 offers a convincing argument for the inseparability of subjective experience from the agent that it concerns, caused by the impossibility of separating its 'production' from its 'consumption'. At the same time, the proposition confirms that the ability of agent $\mathbb{X}$ to satisfy R, and agent $\mathbb{X}$ as such are markedly different kinds of things. While the ability is manifested as a process, the agent is material, and both cannot be separated. By the way, this statement is a strong argument against 'dualism' in the philosophy of mind.

Proposition 5 has a significant further consequence. Namely, the agents exhibiting a more complicated rather than straightforward reactive behavior (such as automatic door opening systems) necessarily contain elements of qualitative experience. Such agents actively adjust responses to the perceived flow of events by analyzing their experiential qualities. Principally, they cannot act otherwise, as zombies. Exploitation of experiential quality of events becomes their inherent property. The behavior of conscious cognitive agents always entails experiential qualities.

4.4 Chalmers' questions

We have elaborated on the framework of CBA agents, in order to gain insight into the sources of the 'hard problem of consciousness'. Our efforts eventually culminated in the formulation of Proposition 4, which asserts a mechanism that can account for the subjective experience of CBA agents. The proposition and underlying assumptions for the CBA model are speculative, but we argued that they can well be justified by empirical evidence from various backgrounds.

The next step is to consider where we stand. In particular, can we now answer Chalmers' questions about (subjective) experience as formulated at the beginning of this section? It turns out that for the case of CBA agents we *can*. All answers are a consequence of Proposition 4 and the model assumptions we made, as we show below.

– *Why do such agents have subjective experiences?* Embodied cognitive behavioral agents have subjective experiences because of their essential role as mediators (i.e., causal links) between physical and mental phenomena and mental phenomena and physical agential responses. Without exploitation of experiential qualities of events, the agents could not develop subjective experiences and react in a rich, nuanced way to ongoing streams of events reflecting their past experiences in similar situations. They could not predict and prepare for the possible continuation of initiated actions.

– *How do such agents have subjective experiences?* This is also explained in Proposition 4. The agents are constructed (or evolutionary developed) in a way that supports the

above-mentioned two-phase mediator role of such experiences. In the first phase, the experiential qualities of the observed physical phenomenon (an event) invokes a link between the current subjective sensory percepts and their subjective impact on the agent's internal representation of this event. In the second phase, the current impact is evaluated in the context of the agent's previous experiences to generate corresponding sensory, cognitive, structural, and behavioral responses.

– *Why does subjective experience accompany the performance of cognitive functions?* This is thanks to the timing of the respective processes realizing the subjective experience of an agents. If the duration of the observed event exceeds the duration of the agent's internal reactions to this event, then these reactions are accompanied by a repeated perception of the event's subjective impact. The repeated impact of such incoming percepts produces the impression of a phenomenon we may call the 'subjective experiencing' of an event.

We reiterate that the answers only hold for the CBA model 'with Proposition 4'. The general case remains open, but the answers in the CBA case may well be indicative for the kind of ingredients that play a role.

With this in mind, our findings lead to the following simple and concise general definition of consciousness (cf. Section 1.3) covering the broad spectrum of embodied cognitive behavioral agents considered in this study:

> *Consciousness is the ability of embodied cognitive behavioral agents to perceive experiential qualities of events in their surroundings and be responsive to their subjective experience.*

This concise definition highlights experiential qualities and subjective experience as inherent aspects of consciousness. It aligns with the principle of consciousness outlined in Section 4 and emphasizes two core properties: the experiential nature of events and the subjective experiences of conscious agents. These properties define both the content and form of consciousness, as demonstrated by the formalization of the CBA model in Section 2.3 and Proposition 4. This new, pragmatic definition of consciousness, based on a functional concept of subjective experience, provides a unified view of consciousness in biological and artificial systems, encompassing many classical anthropocentric explanatory descriptions of consciousness within the philosophical theory of mind.

5 Discussion

The model of embodied CBA agents proved to be valuable for studying and hypothesizing about the difficult problems of consciousness in the philosophy of mind. In this section we make some further remarks about the model and the concepts we developed in this study. We discuss the following topics: the many faces of consciousness, from experience to understanding and thinking, a definition of thinking, the hard problem of matter, dissolving the hard problem, and the power of mathematical modeling.

– **The many faces of consciousness.** Our model of embodied cognitive behavioral agents covers a broad spectrum of natural and artificial systems. The model is especially useful in defining and studying mental properties. This power stems from the fact that, unlike all previously considered computational models of consciousness, our model leverages the semantic properties of computations. Mind-like properties, modeled by non-trivial inherent (or, intrinsic) properties, occur in many forms. The defining conditions of such properties look inconspicuous: the properties must be total, and be active in any circumstance that an agent can face. The properties must set an agent apart from any other agent (or system) that does not satisfy the original defining conditions. From the viewpoint of the embodiment of concrete conscious agents, these general conditions are non-uniform. In concrete applications, the defining properties must satisfy an additional, critical condition that is to streamline all activities of an agent toward one goal: *purposefulness*. Each agent must serve

its purpose, determined by and inferred from its world model. To this end agents must take on specific embodiments and have specific mind-like properties at diverse temporal and spatial scales.

Mind-like properties are indispensable for many, if not all, effective CBA agents in practice. To illustrate the diversity of CBA agents that need them, consider the example of missile defense systems. The systems have facilities for the detection, tracking, interception, and destruction of attacking missiles. For the effective deployment of these facilities (except the last one), the systems essentially need a form of consciousness, including subjective experience. This enables the systems, for example, to prediction and track the trajectories of enemy missiles. The geographically distributed form of the physically disconnected embodiment of missile defense systems also illustrates the variability of CBAs.

If adequate properties of artificial consciousness are lacking, CBAs may be seriously impaired. An example is given by the 2019-2021 groundings of the Boeing 737 MAX aircraft, caused by their flight control failures. A properly designed system of artificial consciousness could have prevented the loss of hundreds of human lives (cf. [38]).

– **From experience to understanding and thinking.** An interesting aspect of our explanation of the experiential qualities of events in Section 4 is the following. Can't we see the internal actions of CBA agents, viewed as an uninterrupted flow of reactions to the ongoing events perceived or realized by the agents (cf. Section 4.2.1 and 4.3.2), as (the effectuation of) a 'thinking process'? This idea could make sense, for the following reasons. First, the events occur all together in the natural order in which they occur in the environment or as consequences of the decisions and actions of the agents, thus presenting a significant 'longitudinal composed event'. Second, the semantics or purpose of this composed event is composed of the semantics of the online emerging component events (cf. Definition 8). Third, the resulting semantics of the composed event offers a description of the purpose of that event in terms of the 'internal machine language' of the agents, whatever it is. Could not this description be seen as an explanation of the doings of the agents? Of their thinking? Could not here be the source of the agents' understanding of their doings? Namely, on the onset of each event that is not new to it, an agent can 'predict' the progression of the event (cf. Proposition 4, item *(iii)*) and possibly choose any follow-up reactions. Here, a hypothesis is emerging about the mechanisms of understanding and thinking that offers CBA agents the competence to act rationally without linguistic understanding. These mechanisms radically differ from those considered in the ongoing debates on understanding by large language models (cf. [16]).

– **A definition of thinking.** Could the same considerations lead us to a definition of 'thinking' for CBA agents? Human thinking, as understood in the philosophy of mind or in the cognitive and neurosciences, is not likely a good starting point for this definition. To cover the broad spectrum of CBAs, we need a more general definition of thinking, of which human thinking is a particular case.

Within CBAs, in the broadest sense, we can view any mental processing of events as a form of thinking. It is a complex phenomenon that involves sensory, cognitive, behavioral, structural, experiential, and predictive processes. We have described these processes in Proposition 4. Due to the immense range of possible biological and artificial agents that is spanned by the CBA model and the model's non-uniformity, we can hardly expect to be more specific in defining 'thinking'. However, following the ideas of the epistemic approach to computing, we may try to base the general definition of thinking in CBAs on the *purpose* of this process?

Following this approach, we offer a specific answer based on seeing CBAs as tools for acquiring and generating artificial wisdom, enabling them to make wiser decisions and behave more intelligently [40]. Under this view, *wisdom* is the correct application of an agent's knowledge through effective behavior, which is the combined effect of cognition and action toward creating values significant for the agent. This fact is precisely what properly designed CBAs can do. Knowledge is 'stored' in their 'black boxes' (in their respective

world model), and mechanisms described in Proposition 4 generate responses to ongoing series of events in the form of 'wisdom. In this way, *thinking is the process of wisdom generation in the agent's operational domain.*

– **The hard problem of matter.** The insights into the nature of consciousness and the experiential quality of events described in this study have application also in other metaphysical inquiries. As a case in point, we consider *the hard problem of matter*.

Almost a hundred years ago, in 1927, in attempting to define consciousness, Russell [22] stated that in his view

> *'. . . we do not know enough of the intrinsic character of events outside us to say whether it does or does not differ from that of mental "events" whose nature we do know'* (cf. [22], p. 222)

More recently, Strawson ([25], p. 97) called it the *'problem of matter'* (or, 'the mystery of the nature of matter'): *what are the intrinsic qualities of physical phenomena, or more generally, of matter?* Of course, these qualities must include the 'mental "events" whose nature we do know', i.e., qualia, phenomenal experience, or consciousness.

We note that Proposition 4, which holds for biological as well as artificial agents that fit our model, directly speaks of the intrinsic character of events and thus answers the questions of the 'hard problem' for CBA agents. In the case of artificial (read: non-living) agents, events also generate experiential qualities supporting an agent's mission. Due to the non-uniformity of agents, it is hard to compare these qualities with those of humans. Nevertheless, the general conclusion is that

> *the ability to perceive experiential qualities of events in the form of subjective experience is the intrinsic property of a suitably organized matter*.

Of course, in diverse agents, these properties occur in a spectrum of varying quantity and quality, depending on the architecture of the respective agents, their cognitive abilities and mechanisms, and the purpose of why those agents have evolved or been developed. In any case, processing their cognitive information is based on similar principles described in this study.

– **Dissolving the hard problem.** Have we dissolved the hard problem of consciousness? Using our model, we have explained the relationship between physical and intrinsic phenomena, including consciousness and experiential properties of events. With the help of Proposition 4 we could answer the questions explicitly asked by Chalmers: how and why we have qualia.

However, there has been a catch: due to the non-uniformity of the agents and the unavailability of their more detailed specifications, our answers could not respect their peculiarities and dissimilarities. As a result, our answers were, in a sense, 'non-constructive' while being too general. This fact may disappoint those expecting a definition of consciousness and related phenomena; nevertheless, as explained all along (e.g. in the justification of Proposition 4), this has been an unrealistic expectation due to the non-uniformity of the agents' embodiment and the current state of the science of consciousness. Our solution works within the model of CBA agents, which is a plausible, simplistic, high-level, and idealized model underlying, thanks to its generality, also human cognition and intelligence. Within the model, we explain how and why subjective experience arises from the experiential qualities of events. Explanation of genuine experiential qualia in the human mind may involve further subtleties and cases that our model of cognitive embodied agents cannot capture. Thus, we have solved the hard problem of human consciousness only to the extent in which our model covers the basics of human mental phenomena.

– **The power of mathematical modeling.** The CBA model is essentially an abstract mathematical model for a large class of 'information processing agents'. The model proved to be well suited for a study of the concepts and hypotheses related to the hard problem of consciousness and other issues in the philosophy of mind.

It is interesting, although not too surprising, that the most potent concepts of present-day information processing play a central role in the modeling of CBA agents: *interactive, non-uniform,* and *potentially non-terminating computations.* Theoretically, such computations exceed the capabilities of classical Turing machines [32]. In practical terms it means that, without further constraints, these computations cannot be realized by classical computers – they can solve a larger class of problems than classical computers can.

As CBA agents always are finitely constraint (cf. Section 2.2), their powers as information processing entities remain limited to finite functions. However, as shown in [34], interactive, non-uniform, and potentially non-terminating computations can be realized by *lineages* ('evolving sequences') of non-uniform finite-state computational devices, like Boolean circuits, neural nets, or finite-state transducers. From an abstract point of view, such devices are precisely instances of embodied CBAs studied here.

Our results are supporting confidence for the power of computational modeling and the respective mindset. Following the Internet [35] and large language models [41], conscious CBAs and their lineages appear to be examples of the most efficient known ways of information processing that are on the verge of the possibilities of classical physics.

6 Conclusion

We have presented an explanatory, high-level abstract model of embodied cognitive behavioral agents (CBAs) amenable to mathematical treatment and applied it to tackle one of the most enduring problems in the philosophy of mind – the hard problem of consciousness. The design of the model was not abstracted from the emulation of biological cognitive systems. Instead, it is a non-anthropocentric, non-uniform general model guided by the principles of epistemic computation, viewing information processing from the perspective of purposeful knowledge generation. Using concepts and modelings from theoretical computer science, we could overcome the explanatory gap between the physical and mental properties of agents. We proved that consciousness and subjective experience in distinct forms extend beyond the brain, and that physical matter gives rise to various forms of conscious experience, in accordance with the quotation of Brian Green at the beginning of this study.

References

1. B. Agüera y Arcas, Artificial neural networks are making strides towards consciousness, *The Economist* (June 9, 2022), https://www.economist.com/by-invitation/2022/09/02/artificial-neural-networks-are-making-strides-towards-consciousness-according-to-blaise-aguera-y-arcas
2. B. Agüera y Arcas, P. Norvig, Artificial General Intelligence Is Already Here.,*NOĒMA*, October 10, 2023, https://www.noemamag.com/artificial-general-intelligence-is-already-here/
3. L. Albantakis, L. Barbosa, G. Findlay, M. Grasso, A.M Haun, W. Marshall, W.G.P. Mayner, A. Zaeemzadeh, M. Boly, B.E. Juel, S. Sasai, K. Fujii, I. David, J. Hendren, J.P. Lang, G. Tononi, Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms, *PLoS Comput Biol.* 19(10) (2023) e1011465, https://doi.org/10.1371/journal.pcbi.1011465
4. B.J. Baars, The Global Workspace Theory of Consciousness: Predictions and Results, in: S. Schneider, M. Velmans (eds.), *The Blackwell Companion to Consciousness* (2nd ed.), Wiley-Blackwell, New York, 2017, pp. 227-242, https://doi.org/10.1002/9781119132363.ch16
5. L. Blum, M. Blum, A theory of consciousness from a theoretical computer science perspective: Insights from the Conscious Turing Machine, *Proceedings of the National Academy of Sciences (PNAS)*, 119(21), e2115934119, 20 May 2022,https://doi.org/10.1073/pnas.2115934119
6. D.J. Chalmers, Facing Up to The Problem of Consciousness, *Journal of Consciousness Studies* 2 (3) (1995) 200-219
7. D.J. Chalmers, *The Conscious Mind,* Oxford University Press, New York, 1996
8. D.J. Chalmers, Zombies on the Web, http://consc.net/zombies-on-the-web/, 2017

9. D.J. Chalmers, Could a Large Language Model be Conscious?, *Boston Review*, August 9, 2023, https://www.bostonreview.net/articles/could-a-large-language-model-be-conscious/. Also: *ArXiv*: 2303.07103, 2023, https://arxiv.org/abs/2303.07103
10. P.R. Corlett, A. Mohanty, A.W. MacDonald, What we think about when we think about predictive processing, *Journal of Abnormal Psychology* 129(6) (2020) 529-533, https://doi.org/10.1037/abn0000632
11. P. Godfrey-Smith, *Living on Earth: Forests, Corals, Consciousness, and the Making of the World,* Farrar, Straus and Giroux, New York, 2024
12. B. Green (moderator), What creates consciousness, in: *World Science Festival*, 19 July 2024, video, https://youtu.be/06-iq-0yJNM?si=sxYN86bI1sfux8Gt
13. S.W. Hawking, *A Brief History of Time - From the Big Bang to Black Holes*, Bantam Books, New York, 1988
14. J.C. Hawkins, *A Thousand Brains - A New Theory of Intelligence*, Basic Books, New York, 2021.
15. M. Lenharo, Do insects have an inner life? Animal consciousness needs a rethink, *Nature* 629, pp. 14-15 (2024), https://www.nature.com/articles/d41586-024-01144-y
16. M. Mitchell, D.C. Krakauer, The debate over understanding in AI's large language models, *Proceedings of the National Academy of Sciences (PNAS)* 120 (13) e2215907120, 21 March 2023, https://doi.org/10.1073/pnas.2215907120
17. M. Naddaf, How your brain detects patterns in the everyday: without conscious thought, *Nature* 634, p. 20 (2024), https://doi.org/10.1038/d41586-024-03116-8
18. T. Nagel, What Is It Like to Be a Bat?, *The Philosophical Review* 83(4) (1974) 435-450, https://doi.org/10.2307/2183914
19. M. Overgaard, A. Kirkeby-Hinrup, A clarification of the conditions under which Large Language Models could be conscious, *Humanities & Social Sciences Comm.* 11, Article nr: 1031 (2024), https://doi.org/10.1057/s41599-024-03553-w
20. T. Rees, Why AI Is A Philosophical Rupture - The synthesis of humans and technology portends a new AIxial age, *NOĒMA*, February 4, 2025, https://www.noemamag.com/why-ai-is-a-philosophical-rupture/
21. H. Rogers Jr., *Theory of Recursive Functions and Effective Computability*, McGraw-Hill, New York, 1967
22. B. Russell, *An Outline of Philosophy,* Routledge, London, 1927
23. A. Seth, The hard problem of consciousness is already beginning to dissolve, *New Scientist,* 1 Sept. 2021, https://www.newscientist.com/article/mg25133501-500-the-hard-problem-of-consciousness-is-already-beginning-to-dissolve/
24. A. Seth, *Being You: A New Science of Consciousness*, Dutton / Penguin Publishing Group, London, 2021
25. G. Strawson, Consciousness Never Left, in: K. Almqvist & A. Haag (eds) *The Return of Consciousness*, Stockholm: Axel and Margaret Ax:son Johnson Foundation, 2017, pp. 89-103
26. P. Tacikowski, G. Kalender, D. Ciliberti, I. Fried, Human hippocampal and entorhinal neurons encode the temporal structure of experience, *Nature* 635, pp. 160-167 (2024), https://doi.org/10.1038/s41586-024-07973-1
27. M. Tegmark, *Our Mathematical Universe - My Quest for the Ultimate Nature of Reality,* New York: Knopf / Penguin Random House, 2014
28. M. Tegmark, Substrate-Independence, in: J. Brockman (Ed.), 2017: What Scientific Term or Concept Ought to be More Widely Known?, *Edge,* https://www.edge.org/response-detail/27126
29. M. Tegmark, *Life 3.0 - Being Human in the Age of Artificial Intelligence.* Knopf / Penguin Random House, New York, 2017
30. W. Thomas, Automata on Infinite Objects, in: J. van Leeuwen (ed.), *Handbook of Theoretical Computer Science*, volume B: *Formal Models and Semantics,* Ch. 4, pp. 133-191, Elsevier Science Publ., Amsterdam, 1990
31. R. Van Gulick, Consciousness, in: E.N. Zalta, U. Nodelman (eds.), *Stanford Encyclopedia of Philosophy*, Winter 2022, https://plato.stanford.edu/archives/win2022/entries/consciousness/
32. J. van Leeuwen, J. Wiedermann, The Turing Machine Paradigm in Contemporary Computing, in: B. Engquist, W. Schmid (eds), *Mathematics Unlimited - 2001 and Beyond*, Springer, Berlin / Heidelberg, 2001, pp. 1139-1155, https://doi.org/10.1007/978-3-642-56478-9_59
33. J. van Leeuwen, J. Wiedermann, Understanding Computation: A General Theory of Computational Processes, *Technical Report* UU-CS-2019-012, December 2019, Department of Information and Computing Sciences, Utrecht University, Utrecht, The Netherlands, 2019

34. P. Verbaan, J. van Leeuwen, J. Wiedermann, The Complexity of Evolving Interactive Systems, in: J. Karhumäki, H. Maurer, G. Păun, G. Rozenberg (eds), *Theory Is Forever*, Lecture Notes in Computer Science, vol 3113, Springer, Berlin / Heidelberg, 2004, pp. 268-281, https://doi.org/10.1007/978-3-540-27812-2_24
35. J. Wiedermann, J. van Leeuwen, How We Think of Computing Today, in: A. Beckmann, C. Dimitracopoulos, B. Löwe (eds), *Logic and Theory of Algorithms*, CiE 2008, Lecture Notes in Computer Science, vol 5028, Springer, Berlin / Heidelberg, 2008, pp. 579-593, https://doi.org/10.1007/978-3-540-69407-6_61
36. J. Wiedermann, J. van Leeuwen, Epistemic Computation and Artificial Intelligence, in: V.C. Müller (ed), *Philosophy and Theory of Artificial Intelligence 2017* (PT-AI 2017), Studies in Applied Philosophy, Epistemology and Rational Ethics, vol. 44, Springer, Cham, 2018, pp. 215-244, https://doi.org/10.1007/978-3-319-96448-5_22
37. J. Wiedermann, J. van Leeuwen, Finite State Machines with Feedback: An Architecture Supporting Minimal Machine Consciousness, in: F. Manea, B. Martin, D. Paulusma, G. Primiero (eds), *Computability in Europe: Computing with Foresight and Industry* (CiE 2019), Lecture Notes in Computer Science, Vol. 11558, Springer, 2019, pp. 286-297, 2019. https://doi.org/10.1007/978-3-030-22996-2_25
38. J. Wiedermann, J. van Leeuwen, Towards Minimally Conscious Finite-state Controlled Cyber-physical Systems: A Manifesto, in: T. Bureš *et al.* (Eds.), *SOFSEM 2021: Theory and Practice of Computer Science*, Proc. 47th Int. Conference on Current Trends in Theory and Practice of Computer Science, Lecture Notes in Computer Science, Vol. 12607, Cham: Springer, 2021, pp. 43-55, https://doi.org/10.1007/978-3-030-67731-2_4
39. J. Wiedermann, J. van Leeuwen, Validating Non-trivial Semantic Properties of Autonomous Robots, in: V.C. Müller (ed), *Philosophy and Theory of Artificial Intelligence 2021* (PTAI 2021), Studies in Applied Philosophy, Epistemology and Rational Ethics, vol. 63, Springer, Cham, https://doi.org/10.1007/978-3-031-09153-7_8
40. J. Wiedermann, J. van Leeuwen, From Knowledge to Wisdom: The Power of Large Language Models in AI, *Technical Report* UU-PCS-2023-01, Department of Information and Computng Sciences, Utrecht University, Utrecht, The Netherlands, https://webspace.science.uu.nl/ leeuw112/techreps/UU-PCS-2023-01.pdf
41. J. Wiedermann, J. van Leeuwen, Large Language Models and the Extended Church-Turing Thesis, in: F. Manea, G. Pighizinni (eds), 14th *Int. Workshop on Non-Classical Models of Automata and Applications* (NCMA 2024), Electronic Proceedings in Theoretical Computer Science 407, 2024, pp. 198-213. Also: *ArXiv*, https://doi.org/10.4204/EPTCS.407.14
42. J. Weisberg, The Hard Problem of Consciousness, in: *Internet Encyclopedia of Philosophy*, https://iep.utm.edu//hard-problem-of-conciousness/, August 22, 2024.
43. S. Wolfram, *What Is ChatGPT Doing... and Why Does It Work?*, Wolfram Media, February 14, 2023, https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/
44. A. Zeman, M. Dewar, S. Della Sala, Lives without imagery - Congenital aphantasia, *Cortex*, 2015 Dec;73:378-80, https://doi.org/ 10.1016/j.cortex.2015.05.019