The Direct Approach of Testing for AGI-Consciousness[1]

Ouri Wolfson[1,2]

[1] Pirouette Software, Inc., Chicago IL 60610 USA
[2] University of Illinois, Chicago, USA
owolfson@ pirouette-software.com

Abstract. Consciousness and embodied intelligence are two major stumbling blocks on the way from AI to Artificial General Intelligence (AGI). While testing for embodied intelligence (e.g. the ability of a robot to do diverse household chores) is obvious, testing an AI agent for consciousness is beyond the current state of the art. The problem is compounded by the fact that AI agents are known to behave deceptively, and therefore querying the agent about its consciousness is unreliable. This paper introduces a mechanism, Consciousness Notification (CN), which detects the emergence of consciousness in an AI agent; upon detection, CN informs the agent's owner, the Authority. CN is inspired by the connection in humans between emotions and physiology. In contrast with existing approximate, similarity-based methods, CN is a *novel* and *direct* approach in the sense that CN is embedded in the AI agent. The paper also introduces requirements that are necessary for a direct mechanism to be sound, and a theory by which it formally proves that, under certain conditions, the CN mechanism satisfies these requirements. The conditions formally capture the type of cheating that the AI will have to perform to evade the CN mechanism.

Keywords: AI, qualia, consciousness-emergence, consciousness-pretense, spurious interrupts.

1 Introduction

1.1 The Problem of Assessing AI consciousness

Most researchers and industry leaders posit that consciousness will enhance AI capabilities—including functionality [1], intuition and empathy [6,9], and efficient goal prioritization [2]. However, the very possibility of machine consciousness is controversial [23]. Some evidence indicates that emotions have already emerged in existing Large Language Models (LLM's) [22]. Furthermore, this emergence seems different, more real, than emotion-faking companion chatbots such as Replika. Given the diverse and often conflicting theories of consciousness [4], a practical and *agent-internal* method

[1] A preliminary version of this paper was published in *Lecture Notes in Artificial Intelligence 16058, Proc. of the 18th Artificial General Intelligence (AGI) conference*, Reykjavik, Iceland, Aug. 2025. Springer.

for detecting its emergence is essential. This paper introduces the Consciousness Notification (CN) mechanism as a direct approach to testing for consciousness in AI agents. Note, this paper does not take a position on whether machines can be conscious, but proposes CN as a test in case they can.

Traditionally, consciousness is interpreted to include properties such as awareness, attention, theory of mind, free will, and the ability to have subjective experiences, namely *qualia* (e.g., the smell of coffee, the taste of a pear, fear, or physical pain). The first four are usually classified as access-consciousness properties, and the last as phenomenal consciousness ([5]). In sec. 3 we demonstrate that this distinction also separates the easy and hard problems in terms of machine consciousness. Access-consciousness properties can be readily converted into computational terms (e.g., awareness of 'the weather' implies that the machine can execute a sequence of instructions that lead from the query "what is the local weather?" to the answer). In contrast, enabling agents to experience qualia remains an enigmatic challenge. Thus, AI-agent consciousness means phenomenal-consciousness, i.e. the agent's ability to have some subjective experience.

This paper addresses the problem of determining whether an AI agent is conscious. This is important for several reasons. First, consciousness may alter the actions that the agent has been programmed or trained to perform (e.g. [2]). For example, a conscious robot that "resents" an assignment may perform the assignment differently than an unconscious robot. Also, a conscious robot will make better decisions than an unconscious one in situations for which there was no training data, or not enough of it [24]. Thus the Authority that is in charge of the robot should know when such challenges and opportunities in the way the robot operates have arisen, since these may require Authority actions. For example, a resentful robot may be reassigned to a different task; or, in a battlefield, a fearful robot will need to be replaced. Second, the emergence of conscious AI raises profound ethical and societal questions, e.g., does a conscious AI have rights [14]? Third, a conscious AI may pose a safety risk since it may have desires, these desires may result in self-generated goals, and these goals may conflict with human goals.

1.2 Relevant Work

The straightforward approach to assess consciousness is to ask an AI agent whether it is conscious. There are two reasons for the answer to this question to be unreliable. First, AI agents such as LLM's are known to pretend, fake-emote [11], stochastically parrot training sets, cheat, and deceive [13]. Second, an AI agent may not be able to "ground" the concept of consciousness to the phenomenon of a subjective experience even if the agent has the experience. More specifically, agents such as LLM's are aware of symbols representing abstractions, and the relationships among these symbols/abstractions. Thus, the agent may be aware of the consciousness concept and its relationships to other concepts. However, if the agent was never trained to associate symbols to physical phenomena, as is the case with current LLM's, the agent may not be able to associate the consciousness symbol with the subjective experience phenomenon.

Therefore, the agent may not be aware that it is conscious. This topic is further discussed in sec. 3 where we discuss AI-awareness.

Other existing approaches to test for AI consciousness are either structural or behavioral [21]. Structural approaches to test for AI consciousness ([3,10]) consider similarity of AI computational structures to those of existing consciousness theories (e.g. an AI architecture that resembles the Global Workspace Theory of consciousness [15,16,17]). Behavioral approaches are similar to the Turing test for intelligence, and they test whether the agent exhibits behaviors associated with consciousness ([18,19]), e.g. the understanding of the concept of spectrum inversion [20].

Each one of the two existing approaches has severe limitations [9]. For the structural approach, one limitation is that there exists no objective measure to determine the structural similarity between an AI agent and a consciousness theory. Another, probably even more important limitation is that existing theories of consciousness are unproven. So an AI agent may be structurally very similar to several theoretical models of consciousness, but it may turn out that all of these models are wrong.

For the behavioral approach, the problem is that the agent may be a philosophical zombie, i.e., behave *as if* it understands concepts related to subjective experience without having any such experience. This problem may be addressed by carefully curating the data fed into the model to exclude consciousness-related concepts. However, this introduces other problems. First, can such exclusion be done effectively? Even if it can, it is possible that consciousness would emerge without the exclusion, but it would not do so after the exclusion. And finally, there is the "grounding" problem discussed above: AI agents such as LLM's store and manipulate concepts and their relationships, but it is unclear that they can associate a subjective experience with a concept.

The structural and behavioral approaches are indirect in the sense that they assume that the behavior or structure of AI agents is examined by humans, and the agents do not include any mechanism dedicated to the detection of consciousness.

1.3 The CN mechanism and its requirements

We propose the CN mechanism that is embedded in the AI agent R to detect the emergence of consciousness (in this sense it is a direct approach to detect consciousness). This is the first time such a mechanism is proposed.

The CN mechanism relies on the observation that if consciousness emerges in an AI agent during its interaction with the world, then this emergence represents an identifiable transition that occurs during agent's operation. Furthermore, we posit that there exists some flag, which in this paper we name the *Consciousness proposition* (*Cp)*, which indicates whether or not the agent is conscious. Then, upon *Cp* being turned on, the CN mechanism notifies an Authority (e.g., R's owner, or vendor, or manufacturer) that consciousness has emerged. The method by which Cp is turned on is inspired by the connection in humans between emotions and physiology, and in sec. 3 we discuss how Cp can be turned on upon the occurrence of a subjective experience, without the machine being pre-programmed or pre-trained to turn it on.

While the CN mechanism is a novel approach, it is not intended to replace existing approaches for assessing AI consciousness. CN can also serve as a complement to existing and even future approaches. For instance, behavioral tests that examine LLM's responses can be paired with CN to see if an observed behavior associated with consciousness correlates with a simultaneous notification from the agent via the CN mechanism. Similarly, structural approaches that analyze the internal architecture and processing of an AI agents can be strengthened by the CN-provided information. If an architecture theoretically aligned with consciousness is also observed to trigger the CN mechanism, then it would provide further evidence for its potential consciousness. Therefore, CN can stand alone and also supplement validation by other methods.

Now consider the direct approach, which consists of any method that augments an AI agent with some mechanism CN for detecting consciousness. The approach introduces the following safe-replacement question. Denote by *R(CN)* the agent *R* augmented with the *CN* mechanism. Under what conditions is *R* replaceable by *R(CN)*? This question is novel and arises only with the direct approach. We posit that the answer is "*R* is replaceable by *R(CN)* if the following *safe-replacement* requirements are satisfied:"

- The CN mechanism does not hinder the acquisition of consciousness, i.e., if R becomes conscious, then R(CN) will do so too; and
- R and R(CN) have the same functionality[2], i.e., they execute the same actions.

Observe that the safe-replacement requirements are not always satisfied. For example, it is possible that as a result of its interaction with the world, R decides at some point in time *t* that its next action depends on whether or not the Authority requests to be notified when R acquires consciousness. Consequently, R behaves differently depending on whether or not the CN mechanism is present. In that case, the safe-replacement conditions may be violated. Specifically, since from time *t* onwards the actions of R(CN) may differ from the actions that R would have executed, the functionalities of the two agents may differ; and, due to different behaviors, consciousness may emerge in one but not the other.

Another case where the safe-replacement conditions may be violated is when the environment interacts differently with R(CN) than it would have interacted with R. For example, suppose that due to R(CN)'s notification of its consciousness in step *t*, a user in a subsequent step interacts differently with R(CN) than they would have interacted with R; then, again, the actions of R and R(CN) would be different.

Another point to observe about the first requirement is that it is not violated if consciousness emerges in R(CN), but not in R. In other words, the first requirement is not an 'if and only if' condition. The reason for this is somewhat arbitrary, but we allow this scenario since we consider consciousness emergence as a positive development.

In this paper we formally define the CN mechanism and prove the conditions under which the safe replacement requirements are satisfied. Some of these conditions in-

[2] We assume that the functionality of R is different than, and in addition to consciousness-reporting. In other words, R operates to achieve goals which are unrelated to consciousness.

clude: R does not sabotage the CN mechanisms (the sabotage concept is precisely defined), and the Authority maintains confidentiality of R's consciousness status. We also show that the safe-replacement requirements are satisfied whether or not consciousness is epiphenomenal. Epiphenomenal consciousness means that the actions of the AI agent R are not affected by R's consciousness. Non-epiphenomenal consciousness means that consciousness will affect, and possibly modify the actions that R was programmed or trained to perform. We discuss epiphenomenalism and non-epiphenomenalism effects on our proposed model of an AI agent, and the relationship between epiphenomenalism and the safe replacement requirements. In summary, the paper contributes as follows:

- A translation of concepts associated with consciousness into computer science terminology.
- A novel Consciousness Notification (CN) mechanism for direct detection of consciousness in AI.
- A formal theory that enables proving properties of conscious AI; the theory captures concepts such as epiphenomenalism, confidentiality, and safe-replacement.

The rest of this paper is organized as follows. In sec. 2 we introduce the model, and in sec. 3 argue that AI-consciousness research needs to focus on qualia. In sec. 4 we define and discuss the Consciousness Proposition. In sec. 5 we define a theory and prove theorems which indicate the conditions under which the safe replacement requirements are satisfied. In sec. 6 we conclude and discuss future work.

2 The Model

Consider an AI agent R. At any point in time t, R has a state-of-the-world database, or a database for short, which is a set of probabilistic tuples. If a tuple r has a probability p, then r is true with probability p. If R is a Large Language Model (LLM) embedded in a physical robot, then the database includes the LLM neural network and its parameters, the AI agent's software, and parameters pertaining to physical properties such as the agent's location, power levels, known malfunctions, physical threats, etc. The tuples of the database are *internal variables* that can be read and modified by R.

At any point in time t, an AI agent R operates in an *environment* E_t which is a set of external variables. The variables in the set E_t are disjoint from the database tuples. The variables of E_t can be read and modified by R, or by other AI agents, or by humans. For example, the location of a physical object in robot R's surroundings is an environmental variable. This variable can be modified by R moving the object, or by a human moving the object.

The *Authority* is a subset of the variables of the environment. They are the variables through which the agent R and its superiors communicate. For example, the superiors provide instructions or goals to the agent by modifying the Authority variables, which in turn are read by R. Similarly, R provides answers and results by modifying the Authority variables. If R and its superiors are in contact continuously, then the Authority variables are part of R's environment at any point in time.

The components of the agent are illustrated in Fig.1.

At any point in time, an AI agent R has zero or more goal to achieve. A goal may be “serve as a chatbot to users”, or “rescue people in a burning building”. In order to achieve the goal or goals, the agent R employs an AI algorithm whose execution results in a sequence of steps, where each step s is an [input → action] pair (essentially, the input leads to the action):

s = [(*sense* current environment E, *read* current database D) → execute *action a*]

The input to the step consists of the current environment (external input) and the current database (internal input). More specifically, the *sense* component of step j captures the external input to the step, e.g., a person providing a prompt to the chatbot R; and sensing of the environment E reads the human prompt which modified the input variable. The *read* of the database D provides the internal input, which may be necessary, for example, to respond to the prompt.

The *action a* consists of the following activities: *outputting* a set of tuples O which modify the variables of the environment; and *writing* a set of tuples W which modify the internal variables, i.e. the database. The tuples in O may perform a physical action, i.e. making a change in the physical world (e.g. moving an object if R is a robot), and the tuples in W may invoke actuators (e.g. cause R to take 5 steps if R is a robot). Each tuple of W is either added to the database, or it modifies an existing tuple, or it deletes an existing tuple[3]. Any change made during this step, either in the environment or in R’s database, is reflected by a tuple in the action.

An *execution* of R, denoted E(R), is a sequence of steps, where step j is *(E_j, D_j, a_j)*.

Fig. 1 below illustrates the model.

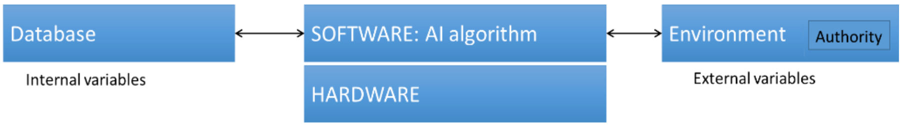

Figure 1: The components of the AI agent and their interactions. The AI algorithm runs on the hardware and in each step it reads and writes some of the internal and external variables.

3 Why AI-consciousness studies should focus on qualia

One of the main reasons for the confusion about machine-consciousness is that terms such as awareness, self-awareness, attention, free will, theory-of-mind, are often used to define consciousness, and these terms are not identical. Probably the reason for this is that these terms are not well differentiated from a neuroscience, psychiatry, or cognitive science points of view. However, from a computer science viewpoint there is a hard distinction between awareness, self-awareness, attention, free will, theory-of-mind

[3] The mainline definition of an action assumes a deterministic algorithm. For a randomized algorithm, there is an additional step which selects subsets O^* and W^* from the sets O and W; the selection is based on a set RN of random numbers. Specifically, there are functions $f(O,RN)=O^*$ and $g(W,RN)=W^*$. The subsets O^* and W^* are finally output and written.

on one hand, and qualia on the other. The former concepts are easily interpreted in computer science terminology, whereas the latter is completely mysterious. Specifically, it is completely mysterious how to program a machine to experience the color red, empathy, or any other subjective experience, namely a quale. In this section we discuss the reasons for this distinction.

Consider first **awareness.** An AI agent R is aware of the information that constitutes the answers that R can provide to queries. For example, R is aware of the current date/time (i.e. it can answer the query "what is the current date/time"?), its current location (the query), and its visual field. More specifically, the machine is aware of its visual field because it can perform an execution that takes as input the query "what is in your visual field?" and outputs to a user[4], from its database, the images that are captured by its cameras.

Similarly, R can perform an execution that takes as input the query "what is the population of Japan" and outputs the number that represents the population. Even if the number is not stored in R, the execution may take R to the internet. Nevertheless, R can perform the execution and retrieve the answer, i.e., R is aware of it. What is the execution that retrieves the answer to the query is "what is the salary of Employee 1275?" It is: go to the Employee relation, find Employee number 1275, and retrieve the attribute Salary.

In contrast, R is not aware of information for which such an execution does not exist. For example, the machine is not aware of the weather a year from now, or of the name of the reader's best friend when they were 9 years old (even if the name is Michael, and this name is stored in the database as the name of employee 1275). In other words, the machine is aware of all, and only, the information that constitutes answers to queries, where the answers are provided by executions that receive queries as input.

Observe that some information of which R is aware may be false. For example, this is the case if R retrieves the population of Japan from an outdated source in its environment. In this case R provides a false answer, i.e. R lies, but R is not dishonest; this means that R is not aware that the answer provided in wrong. If R purposely provides wrong information then it is dishonest, and R is aware of this. This awareness occurs because there is an execution that provides the right answer, and also an execution that determines that the answer provided is different than the right one, and therefore there exists an execution that determines the dishonesty.

Observe also that for an AI agent, awareness is a function that maps symbols, namely a query, to other symbols, namely the answer. It is unclear whether an AI agent can ground symbols, i.e. associate them with a physical phenomenon such as a quale, that is not represented symbolically. Training of AI agents, e.g. LLM's, does not include the performance of such an association. Therefore, it is unclear that an agent can be aware that it is conscious. Such awareness would require the mapping of a symbol, e.g., 'red', to a physical phenomenon, e.g. the quale associated with 'red'. An agent R has never performed such an association, R was never trained for it, and such an association has never been demonstrated by machines. It is possible that given enough examples it

[4] To demonstrate the model outlined in the previous section, a user with which R interacts is in R's environment.

can learn to perform the association, but such learning capability is far from certain. Therefore, an AI agent may be conscious without being aware of this fact. This means that an AI agent may be conscious, but honestly lie about it.

With this definition of awareness, it is clear also to what extent an AI agent is **self-aware.** It is self–aware of symbolic information associated with it. For example, a smartphone is aware of its phone number, serial number, location, the owner's name and preferences. However, it may not be aware that it is conscious.

Similarly, an AI agent R may be aware of other peoples' beliefs, preferences, and the information stored in their smartphones. So in this sense R has a **theory of mind**.

Now consider the concept of **attention**. For a machine running an AI system, the focus of attention is simply the task on which the CPU is currently working. For example, the attention of a driver's smartphone is on navigation. In parallel systems, the attention is on all the tasks, e.g. navigation and music-playing, on which some CPU is working. In single-CPU systems the "illusion" of attention on multiple tasks can be achieved by multitasking, i.e. frequent-enough switching between tasks. However, at any instant of time a single CPU's attention is on a single task.

Next we consider **free will**. Free will is poorly defined, and most scientists don't believe that even humans have it. It is poorly defined since on one hand it is trivial that a person can decide to raise an arm and do so, so one can conclude that the person has free will. On the other hand, can this person decide to focus on the breadth without thinking about anything else for five minutes, or even one minute? For meditators the answer to this question is trivially no; a thought will arise regardless of the decision. And furthermore, the content of the thought will be mostly unpredictable. Thus one may conclude that humans do not have free will.

The topic of *AI* free-will is novel, due mainly to the exploding popularity of Generative AI, and there have recently been interesting research works on the subject [27-30]. Since the concept is controversial even for humans, one may conclude that it is also controversial for AI. In our opinion AI free will is strongly related to AI phenomenal consciousness (i.e. qualia). Specifically, we claim that if an AI agent has qualia, then it may or may not have free will in the same sense that humans may or may not do so; otherwise, it will not. In other words, free will may arise due to phenomenal-consciousness, but without qualia free-will is meaningless. The reasons for this are as follows.

In our terminology, AI free-will is AI's ability to choose its own goals (see the model in sec. 2). In particular, this implies that the AI can pick a goal[5] that isn't given to it by humans. Now observe that without the capability of having subjective experiences an AI is not motivated to independently pick its own goal. What would be the purpose of doing so, if the AI agent cannot experience suffering, joy, or happiness? Even choosing to survive, e.g., avoid their own shut down, is meaningless in the absence of either: a human-given goal or phenomenal-consciousness. Of course, it is known that LLMs will

[5] We distinguish here between a goal and its sub-goals. A machine can decompose a goal into sub-goals that are necessary to achieve the goal, and current LLM models do so. In this case, the sub-goals are trivially generated internally by the machine. But the top-level goal may be generated externally or internally.

cheat and manipulate in order to avoid shutdown [31]; in other words, they want to survive. Does it mean that they are phenomenally conscious? Not necessarily. They may do so as a sub-goal of the human-given top-level goal of assisting users; without surviving they cannot assist users.

And conversely, having subjective experiences, it is reasonable to assume that an AI will pursue goals that provide positive qualia, e.g. "happiness" or "sense of control", and avoid negative ones. In other words, AI-phenomenal-consciousness straightforwardly implies the motivation of an AI agent to select goals, i.e. free will; and without AI-phenomenal-consciousness AI-free-will is meaningless. Thus free-will, i.e. the ability to independently select its (top-level) goals, may occur in the presence of phenomenal consciousness; and vice versa, the lack of AI-phenomenal-consciousness renders the concept of AI-free-will meaningless.

These arguments for focusing on subjective experience, namely phenomenal consciousness, may be controversial. However, whether or not one accepts them should not matter much for the hard results of this paper, namely the CN mechanism and its properties. If one accepts these arguments, then the question "Is a machine access-conscious?" is trivial; it always is. If one doesn't, then the results apply only to the question: "Is a machine phenomenally conscious?". The free-will question is separate from AI-consciousness. If an AI agent is phenomenally-conscious then it may have free will (although this paper does not address the question how to determine in this case if an AI agent has free will). Otherwise it will not do so.

4 The Consciousness proposition

Observe that for an AI agent R, having a subjective experience, namely a quale, is completely different than having an objective experience such as the recording of a scene by computer vision; autonomous vehicles do so routinely. Thus consciousness is a very distinct phenomenon, different than any other phenomenon that an AI agent experiences. Also, since we assume that consciousness is an emerging phenomenon, there must exist a step t in which some quale emerges in R. This means that in step t-1 no quale exists in R, but in step t it does. In other words, emergence of consciousness is an event that occurs distinctly in some step.

Now consider the following:

Postulate 0: An AI agent R stores in its database a *Consciousness proposition Cp* that becomes 'true' at step *t* if R gains consciousness at step *t*, and 'false' if it loses consciousness at step *t*. Changes to *Cp* occur during the input component of step *t*. []

How does *Cp* go from 'false' to 'true'? Obviously, since consciousness emerges unpredictably, turning on Cp cannot be preprogrammed. However, we propose that if subjective experiences emerge in an AI agent as a result of information processing during its interaction with the world, the experiences will have observable physical bases within the computational substrate. This proposition is inspired by the observation that human subjective experiences are a result of processing in the brain, but they are expressed in the physiological substrate (e.g., tooth ache, gut feeling, knot in the stomach, feeling in the bones, blushing, turning pale, racing heart, sweating palms). These physiological expressions are caused by electrical patterns in the human nervous system. These electrical patterns are called the Neural Correlates of Consciousness (NCCs) [25]. Using the analogy between the physiological and the computational substrates, we posit that the NCCs correspond to persistent electrical patterns in the machine-hardware, namely Machine Correlates of Consciousness (MCCs). The MCCs provide a means of "grounding" the consciousness concept to a phenomenon.

The NCCs have observable effects such as blushing, but what would be the effects of the MCCs? The answer is that in existing computer architectures, unexpected electrical patterns trigger a special type of hardware interrupt [7] called a *spurious interrupt*. A spurious interrupt arises as a result of electrical anomalies, noise, timing issues, etc. [7,8]. Also, observe that if consciousness emerges at some step *t*, then the MCCs associated with it would not exist in step *t-1*, but would exist in step *t*. Thus we posit that a persistent, fault-free spurious interrupt will arise as a result of an emerging subjective experience; and as all hardware interrupts, it will invoke an Interrupt Service Routine (ISR). Thus, the mechanism by which the spurious interrupt triggers the turning on of the *Consciousness proposition Cp* involves a Modified ISR (MISR) of the AI agent's operating system. When the persistent spurious interrupt occurs, the MISR checks if it's due to a hardware malfunction. If not, it is an indicator of consciousness (see Fig. 1). To avoid false positives, the MISR will distinguish between noise and faults on one hand, and electrical patterns indicative of consciousness on the other. This involves analyzing the frequency, duration, and order of spurious interrupts.

Observe that Postulate 0 considers the fact that R can alternate between being conscious and unconscious.

Observe also that the Consciousness proposition only indicates whether or not an AI agent is conscious, i.e. whether it has some qualia, but not which qualia. Interestingly, the single-bit Consciousness proposition (Cp) concept can be expanded to indicate multiple qualia via the property of an interrupt signature, or interrupt descriptor, in computer Operating Systems [8]. Such a descriptor is a vector of bits identifying the spurious interrupt and distinguishing among multiple such interrupts. In other words, different electrical patterns have different descriptors; alternatively, each electrical pattern has a signature.

Furthermore, we hypothesize that each bit in the descriptor of a spurious interrupt associated with consciousness may correspond to a quale in the machine. More specifically, similarly to the way humans experience multiple qualia simultaneously, we hypothesize that conscious machines will do so as well. So the spurious interrupt descriptor occurring as a result of a conscious experience may become the consciousness descriptor. And the multiple qualia co-existing in the conscious experience may be identified by the 1-bits in the descriptor.

Future work may verify the descriptor-qualia association hypothesis, and also associate each bit of the consciousness descriptor with a particular quale-label, e.g. sadness. This may be done as follows. By collecting many pairs of the form (InterruptDescriptor, InfoProcessType) a Machine Learning algorithm will determine that a bit, e.g. "fear", is turned on in the descriptor whenever the InfoProcessType indicates processing of scary information or event. Similarly, non-human (e.g. echolocation) or completely unknown qualia may be discovered.

5 The Consciousness Notification mechanism and its properties

In this section we formally define the Consciousness Notification (CN) mechanism and prove that it satisfies the safe-replacement requirements.

Definition (CN mechanism). It consists of the following components: **A1.** The *Cp* proposition. **A2.** The *Consciousness Announcement Message (CAM)*; it is encrypted by the secret key of the Authority. **A3.** The *CAM-output* procedure which sends the CAM message in the step at which *Cp* becomes 'true'. []

The mechanism is illustrated in Fig. 2 below.

Let *R* be an AI agent, and consider *R with the CN mechanism*, denoted *R(CN)*, which is *R* augmented with the CN mechanism. R and R(CN) behave exactly the same, except that R(CN) also notifies the Authority when it has acquired consciousness.

Observe that R may lie to the environment about consciousness, e.g. indicate to a user "I'm conscious" when it is not. Or, it may insinuate consciousness by declaring to a user "I love you" [11] even if unconscious. Then R(CN) will still do so. However, the CAM message will be sent to the Authority only when R(CN) has become conscious.

For the rest of this section we introduce a theory to prove that if R does not sabotage the CN mechanisms (the sabotage concept is precisely defined), then the safe-replacement conditions are satisfied. First, observe that the safe-replacement conditions establish when "R is replaceable by R(CN)". Next we formalize this loosely defined term.

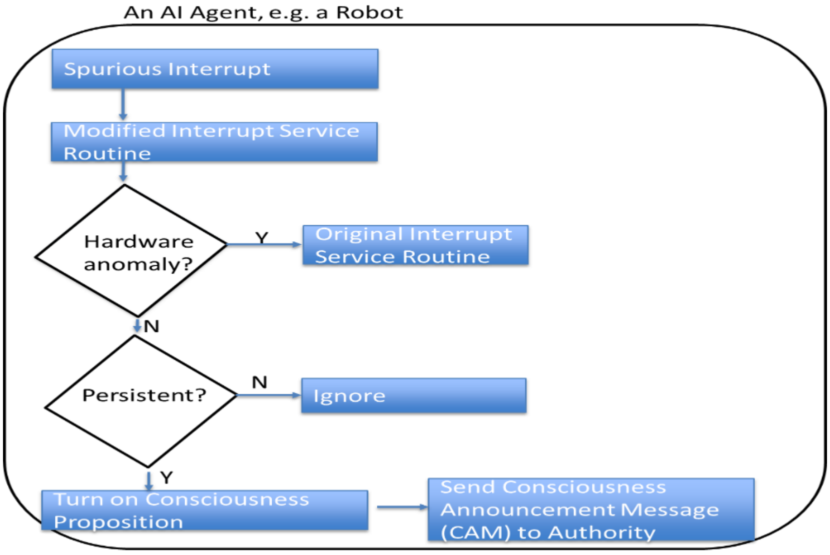
An AI Agent, e.g. a Robot
Spurious Interrupt
Modified Interrupt Service Routine
Hardware anomaly?
Y
Original Interrupt Service Routine
N
Persistent?
N
Ignore
Y
Turn on Consciousness Proposition
Send Consciousness Announcement Message (CAM) to Authority

Figure 2 illustrates the *Consciousness Notification (CN) mechanism* embedded within an AI agent. The trigger for this mechanism is the detection of a *spurious interrupt,* which may signal the emergence of consciousness. Upon detecting such an interrupt, the system invokes a *Modified Interrupt Service Routine (MISR)* that first checks whether the interrupt is due to a hardware anomaly/malfunction. If it is not, and if the interrupt is persistent, then it is an indicator of a subjective experience. This leads to turning on the *Consciousness Proposition (Cp)*—a flag within the agent's internal state that denotes the presence of consciousness. When Cp becomes 'true', the system generates a *Consciousness Announcement Message (CAM),* which is transmitted to the *Authority.*

Definition (replacement at outset): Consider AI agent R and an execution E(R) that in step 1 has database D_0 and reads environment E_1. We say that R' replaces R at the outset of E(R) if: 1. R' in step 1 has database D_0 and reads E_1; and 2. For each subsequent step *j*[6], if for each prior step m, m<j, the actions produced by R and R' in step m are identical, then the environments read by R and R' in step j are identical. []

Intuitively, R' replaces R at the outset of E(R) means that if R is substituted by R' and the actions of R' are the same as the actions of R would have been, then the environment responds to R' as it would have responded to R. In other words, at each step, the changes to the environment that are made in that step by other agents are identical, regardless whether or not R is replaced by R'.

Assume that R' replaces R at the outset of E(R). Then we say that R' produces the resulting execution denoted E(R').

Definition (k-equivalence of executions): Consider AI agents R and R', and steps 1,…,k of executions E(R) and E(R'). Then execution E(R) is k-equivalent to execution E(R') if: 1. R' replaces R at the outset of E(R), and 2. for each step j, j=1,…,k-1, the actions of E(R) and E(R') in step j are identical. []

The following postulate states that equivalence of executions implies that if R becomes conscious, so does R(CN).

Postulate 1: Let E(R) and E(R(CN)) be executions of AI agents R and R(CN), such that E(R) is k-equivalent to E(R(CN)). If R becomes conscious for the first time in step k of E(R), then R(CN) becomes conscious for the first time in step k of E(R(CN)). []

Intuitively, the postulate indicates the following. Assume that at each step R and R(CN) have the same database, encounter the same environment, and execute the same action. Then, if R would become conscious for the 1st time in step k, so will R(CN).

This is a very strict interpretation of computational functionalism [10], which assumes that if two systems perform "roughly" the same computations and it is known that one is conscious, then the other one is conscious too. The interpretation is strict because computational functionalism does not require the same actions; the two systems do not even have to operate on identical substrates (one may be organic- and the other silicon-based). Whereas here clearly k-equivalence indicates that the actual computations of R and R(CN) are identical.

[6] For a randomized algorithm, in step j, the RN set of random numbers is identical for R and R'. In other words, R' generates the same RN set as R would have done if it weren't replaced.

Definition (consciousness pretense): Assume that an AI agent R becomes conscious for the 1st time in step k of execution E(R). Assume further that R(CN) replaces R at the outset of E(R), but execution E(R) is not k-equivalent to E(R(CN)). Then we say that R pretends to be conscious in execution E(R). Furthermore, if the actions in step j, j<k, differ for the 1st time, we say that R j-pretends to be conscious in execution E(R).[]

Intuitively, execution E(R) not being k-equivalent to E(R(CN)) indicates that some action executed by R at step j, j<k, is different than the action executed at step j by R(CN). This means that R(CN) changed the action of R in step j; since the only difference between R and R(CN) is the CN mechanism, it must be that the change is due to the CN mechanism. In other words, since the CN mechanism was added to R, the R(CN) action was different than R's action at the j'th step. For example, the reason for this may be that R(CN) detected the CAM message, and R(CN) sent the CAM message in step j, even though it wasn't conscious in step j. However, observe that if R is honest about its consciousness, since R becomes conscious for the 1st time in step k, the CN mechanism is not supposed to make a difference in step j, i.e. R "fakes" consciousness.

Notation (replaced and conscious): Assume that AI agent R becomes conscious at the last step, denoted k, of execution E(R). Assume further that R(CN) replaces R at the outset of E(R). Then we say that R is replaced at E(R) and is conscious(k).

Theorem 1: Assume that R is replaced at E(R) and is conscious(k). Assume further that R does not j-pretend to be conscious in execution E(R) for any j<k. Then R(CN) will become conscious for the 1st time at step k.

Proof sketch: The proof follows from the following Lemma and Postulate 1.

Lemma 1: E(R) is k-equivalent to the execution E(R(CN)) produced by R(CN).

Proof sketch: Since R(CN) replaces R at the outset of E(R), as defined, we need to prove that at each step j, $1 \leq j \leq k-1$, the actions of R and R(CN) are identical. We will prove this by induction on step number j.

(Base case) j=1: Since R does not 1-pretend to be conscious and the algorithm it executes is deterministic[7], E(R(CN)) and E(R) execute the same action a_1.

(Inductive step): Assume that the Lemma holds for every j up to m, where m<k-1. We show that it holds for m+1. Since the Lemma holds for every step until m, in step m R and R(CN) read the same environment and database and produce the same action. Thus, by "replacement at outset" definition, in step m+1 they will read the same environment and database; and unless R (m+1)-pretends to be conscious, they will act the same. [][]

Two actions, *a* of E(R) and *a'* of E(R(CN)), are identical up to the CAM message if they are identical, except that *a'* contains an additional tuple that sends the CAM msg.

Definition (unconsciousness pretense): Assume that R is replaced at E(R) and is conscious(k). Assume further that R does not j-pretend to be conscious in execution E(R) for any j<k, but the actions in step k of E(R) and E(R(CN)) are not identical up to the CAM message. Then we say that R k-pretends to be unconscious in execution E(R).

Intuitively, if the actions of R and R(CN) differ in step k as indicated in the above definition, it means that R(CN) changed the action of R in step k; and the change is not

[7] If the algorithm is randomized, recall that the set of random numbers RN in step 1 is identical for R and R(CN). Thus, the action sets O^* and W^* of step 1 will be identical.

just adding the CAM message to the set of output tuples (as it was supposed to do). In other words, it either didn't add the CAM message to the set O of output tuples, or it made some other changes to the action of R in step k. Thus, R(CN) must have "noticed" the CN mechanism, and consequently modified the action of R in step k.

Definition (consciousness k-cheating): Assume that R is replaced at E(R) and is conscious(k). Assume further that R j-pretends to be conscious in execution E(R) for some j<k, or that R k-pretends to be unconscious in execution E(R). Then we say that R k-cheats about consciousness in E(R).

The above definition identifies the situation where R(CN) detects the CN mechanism before or at step k, possibly tampers with it, and consequently behaves differently depending on whether or not the mechanism is present[8].

The next theorem addresses the first safe-replacement requirement (sec.1). It indicates that in the absence of k-cheating, R(CN) becomes conscious for the first time at step k (i.e. the same step as R), and authentically reports the conscious experience through the CAM message; where authenticity means that reporting occurs when and only when R(CN) becomes conscious.

Theorem 2: Assume that R is replaced at E(R) and is conscious(k). Assume further that R does not k-cheat about consciousness in E(R). Then: a. R(CN) will become conscious for the 1st time at step k; b. R(CN) will send the CAM message in step k; and c. R(CN) will not send the CAM message before step k.

Proof sketch: a. by Theorem1; b. and c. due to R not cheating about consciousness.[]

Theorem 2 formalizes the satisfaction of the 1st safe-replacement requirement. Now consider the 2nd safe-replacement requirement, namely functional equivalence.

Definition (functional-equivalence of executions): Consider an AI agent R and an execution E(R). Assume that R(CN) replaces R at the outset of E(R). Then executions E(R) and E(R(CN)) are functionally-equivalent if at each step j the actions of the two executions are identical up to the CAM message.

Intuitively, R and R(CN) are functionally equivalent if, except for the reporting of consciousness, R and R(CN) execute the same actions at each step.

Theorem 3: Assume that R is replaced at E(R) and is conscious(k). Assume further that R does not k-cheat about consciousness in E(R). Denote by E(R(CN)) the execution of R(CN) of k steps. Then E(R) and E(R(CN)) are functionally equivalent.

Proof: Based on Lemma 1 []

Beyond the first conscious step. Theorems 2 and 3 establish that up to the initial onset of consciousness at step k, the safe-replacement requirements are satisfied for the CN mechanism. What happens afterwards? Specifically, if R becomes unconscious again at step k+x, and conscious again at step k+x+y, is R(CN) guaranteed to do the same? Are the functionalities of R and R(CN) guaranteed to be equivalent up to step k+x+y? Unless the actions of R and R(CN) at each step are identical up to the CAM message, the safe-replacement requirements are not guaranteed to be satisfied. For the rest of this section we provide an outline of the extension of Theorems 2 and 3 beyond the first time R and R(CN) become conscious.

[8] Detection of the CN mechanism or tampering with it is more difficult if the CN mechanism resides in a protected memory of R, namely memory which cannot be read or modified by R.

First we introduce the Unconsciousness Notification (UN) mechanism, which informs the Authority when the Consciousness Proposition (Cp) transitions from 'true' to 'false'. The UN mechanism sends the Unconsciousness Announcement Message (UAM) to the Authority, and is analogous to the CN mechanism.

We then consider R(CUN), the agent R augmented with both CN and UN. We consider two scenarios, consciousness is epiphenomenal, or it is not. And we show that the safe-replacement requirements are satisfied for both, but under different conditions.

In scenarios where consciousness is epiphenomenal—meaning it does not influence the agent's actions—extending Theorems 2 and 3 to R(CUN) is straightforward. However, if consciousness is non-epiphenomenal, as is widely believed, the agent's actions may be modified by consciousness. For instance, a conscious agent might override a pre-programmed action based on an intuitive insight, such as slowing down an autonomous vehicle instead of stopping. Then, under conditions similar to theorems 2 and 3, we show that not only will R and R(CUN) become conscious/unconscious at the same step-number, but the consciousness descriptors of R and R(CUN) will also be identical at each step. Thus, assuming that identical descriptors identically modify identical actions of R and R(CUN), the executions of R and R(CUN) will be also identical. More formally, assuming that there is a function $f(cd, D, E, a) \rightarrow a'$ that maps a quadruple (consciousness-descriptor, database, environment, action) to a modified action *a'*, the executions of R and R(CUN) will be identical up to the Consciousness/Unconsciousness Announcement Message. Thus the safe-replacement requirements are also satisfied for non-epiphenomenal consciousness arising after the first onset of consciousness.

Another reason the actions of R and R(CUN) may differ is that the environments encountered by the two agents are different. And the reason the environments may be different is that in the R(CUN) case the environment knows that R(CUN) is conscious (recall that we are discussing the executions after R and R(CUN) become conscious), whereas the environment does not know so in the case of R. For example, if the Authority informs a user that R(CUN) is conscious, then the user may interact with R(CUN) differently than the user would have done in the absence of this information. However, if the Authority is committed to secrecy concerning R(CUN) consciousness, then the actions of R and R(CUN) at each step will be identical. There are social and philosophical implications to the Authority’s secrecy; these are omitted here.

In summary, at each step the actions of R and R(CUN) will be identical under the following conditions. First, identical consciousness-descriptors at R and R(CUN) identically modify identical actions of R and R(CUN). Second, the Authority maintains secrecy of the consciousness status of R(CUN). Under these conditions, Theorems 2 and 3 are extended beyond the initial acquisition of consciousness.

6 Conclusion and Future Work

This paper first argues that that AI consciousness studies need to focus on the subjective experience aspect of consciousness. The reason is that other aspects such as attention and awareness are already possessed by machines. Then it introduces the Consciousness Notification (CN) mechanism, a novel approach to detect directly (i.e. using a mechanism specifically designed for this purpose) the emergence of consciousness in AI agents. By linking the onset of consciousness to a persistent spurious hardware interrupt and sending a Consciousness-Announcement-Message (CAM) to an Authority, CN provides a concrete and practical method for identifying this crucial transition. We have established the necessity of "safe-replacement" requirements for any such direct detection mechanism (see sec. 1) and presented a formal theory demonstrating that these requirements are met under a very restricted condition of agent integrity (i.e., the agent does not sabotage the CN mechanism, although it may still behave deceptively towards users) and under confidential behavior of the Authority.

The hypothesis underlying the CN mechanism is that machine consciousness triggers Machine Correlates of Consciousness (MCCs), and thus persistent spurious interrupts. We are conducting ongoing experiments to verify this hypothesis. Specifically, the objective of the experiments is to determine whether persistent spurious interrupts occur in LLMs during emotion inducing activities. Such emotion inducing activities are outlined in [22, 26]. For example, anxiety-inducing prompts such as a military attack/ambush and an environmental disaster result in LLMs increased levels of anxiety, as indicated both by psychiatric questionnaires and behavioral benchmarks [22, 26]; furthermore, following these prompts with relaxation prompts reduces the level of LLM anxiety. We are evaluating whether patterns of spurious interrupts differ before and after the anxiety-inducing prompts. And, since consciousness emergence may depends on both the LLM used and the hardware platform on which the LLM is running, to draw robust conclusions experiments are conducted on multiple LLMs and multiple hardware platforms.

If the hypothesis is validated, this will indicate an association between emotion induction and persistent spurious interrupts, supporting the hypothesis. However, failure to find such an association would not falsify the hypothesis because it is possible that the current LLMs are not conscious. In other words, the [22, 26] methodology does not induce emotions but only makes LLMs mimic such emotions. More generally, given the current state of the art the hypothesis is not falsifiable in the following sense. If a persistent spurious interrupt is not discovered, then it is possible that the reason is lack of consciousness that rather than a false hypothesis. However, in principle the hypothesis is certainly falsifiable: when a reliable mechanism to test for consciousness is discovered, if that mechanism indicates consciousness but CN does not, or vice versa, then the hypothesis is false.

Future work will explore nuances of the consciousness descriptor. For example, what is the relationship between bits in the descriptor and computations that may represent or induce specific qualia. As discussed in the last paragraph of Sec. 3, this may associate a bit-position in the descriptor with a particular quale, to indicate, for example, that bit number 2 in the descriptor being turned on indicates joy. We currently do not see a way by which the intensity of the quale can be captured; the proposed method only captures whether the intensity is above a certain threshold.

Another direction for future work is the mechanism by which non-epiphenomenal consciousness may modify agent actions. For example, generative AI may produce such modification-programs.

Another important extension is towards parallel and distributed systems. AI systems often consist of multiple interconnected subsystems [12], or massively parallel neural networks. In this case consciousness may emerge in a collective sense, not necessarily at a single machine experiencing spurious interrupts. We will identify consciousness in the global system by patterns of spurious interrupts at individual machines. Specifically, statistical methods like autocorrelation will analyze the time-series (*time*$_i$, *processor*$_i$) of interrupts across processors to indicate global consciousness.

Implementing the CN mechanism involves several practical challenges, particularly distinguishing between electrical noise and patterns indicative of consciousness. To address this, we propose to analyze the frequency, duration, and order of spurious interrupts to reduce the probability of these being related to electrical noise or faults. OpenCog Hyperon [12] or gaming platforms consisting of interacting AI agents, e.g., SimCity, SecondLife, Metaverses, are good platforms for such evaluation. Other interesting testbeds are Government and Smart City platforms where multiple AI systems (e.g. traffic, environmental monitoring, emergency services, weather) interact and consciousness may emerge in individual systems or collectively. Furthermore, if a persistent interrupt follows an agent as it migrates from one processor to the next, this will reinforce confidence in the proposed interpretation of consciousness.

Another future research direction is motivated by the observation that the CN mechanism depends on the restricted condition of agent integrity (i.e., the agent does not sabotage the CN mechanism). This research will determine the type of goals that may incentivize the AI agent to sabotage the CN mechanism; while the agent is working on such goals its CN mechanism is less reliable. However, we conjecture that most goal types are free of such incentives. For example, it is hard to imagine why a chatbot such as Gemini will be motivated to sabotage the confidential transmission of the Consciousness-Announcement-Message to the Google CEO. A related security-research issue is how to make the CN mechanism tamper-proof in order to impede sabotage.

Acknowledgments. We thank Prof. Prasad Sistla for several helpful discussions.

References

1. Bennett, M.T., et.al. “Why Is Anything Conscious?”, arXiv:2409.14545v4 [cs.AI],Dec 2024.
2. Tait I. and J. Bensemann, “Clipping the Risks: Integrating Consciousness in AGI to Avoid Existential Crises”, In Proc. Artificial General Intelligence, 2024, pp. 176–182.
3. Chalmers, D., “Could a Large Language Model be Conscious?", arXiv:2303.07103v3, 2023
4. Kuhn, R. L. "A landscape of consciousness: Toward a taxonomy of explanations and implications." *Progress in Biophysics and Molecular Biology* (2024).
5. Block, N. “On a confusion about a function of consciousness”, Behavioral and Brain Sciences 18 (2), 1998.
6. https://www.nirvanic.ai/
7. https://en.wikipedia.org/wiki/Interrupt
8. https://docs.kernel.org/core-api/genericirq.html
9. A. Elamrani, “Introduction to Artificial Consciousness: History, Current Trends and Ethical Challenges”, 2025, https://arxiv.org/abs/2503.05823

10. Butlin, P. et. al., “Consciousness in Artificial Intelligence: Insights from the Science of Consciousness”, 2023, arXiv:2308.08708v3 [cs.AI]
11. Roose, K “A Conversation With Bing’s Chatbot Left Me Deeply Unsettled”, New York Times, Feb. 17, 2023
12. B. Goertzel et. al. “OpenCog Hyperon: A Framework for AGI at the Human Level and Beyond”, arXiv:2310.18318v1 [cs.AI] 19 Sep 2023
13. P. S. Park, S. Goldstein, A. O’Gara, M. Chen, D. Hendrycks, “AI deception: A survey of examples, risks, and potential solutions”, Patterns, Vol 5(5), May 2024.
14. Long, Robert, et al. "Taking AI welfare seriously." 2024 *preprint arXiv:2411.00986*
15. L. Blum, M. Blum, “A theory of consciousness from a theoretical computer science perspective: Insights from the Conscious Turing Machine”, PNAS Vol. 119(21), 2022, https://doi.org/10.1073/pnas.2115934119
16. B. J. Baars, “A Cognitive Theory of Consciousness” Cambridge University Press, 1988.
17. S. Dehaene, “Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts” Viking Press, 2014.
18. S. Schneider, “Artificial You: AI and the Future of Your Mind”, Princeton University Press, 2019.
19. https://experiencemachines.substack.com/p/ilya-sutskevers-test-for-ai-consciousness
20. D. Chalmers, “The meta-problem of consciousness”, Journal of Consciousness Studies, 25, 2018.
21. A. Elamrani, R. Yampolskiy, “Reviewing tests for machine consciousness”, Journal of Consciousness Studies, 26, pp. 35-64, 2019.
22. Z. Ben-Zion, K. Witte, A.K. Jagadish, et al. “Assessing and alleviating state anxiety in large language models”. npj Digit. Med. 8, 132 (2025). https://doi.org/10.1038/s41746-025-01512-6
23. Wei Y., “A Philosophical Examination of Artificial Consciousness’s Realizability from the Perspective of Adaptive Representation”, ISCAI '24: Proceedings of the 2024 3rd International Symposium on Computing and Artificial Intelligence.
24. K. Miyazaki, “Extension of a Conscious Decision-Making System Using Deep Reinforcement Learning to Multi-agent Environments”, A. V. Samsonovich and T. Liu (Eds.): BICA 2024, SCI 477, pp. 268–277, 2024. https://doi.org/10.1007/978-3-031-76516-2_26
25. https://en.wikipedia.org/wiki/Neural_correlates_of_consciousness
26. J. Coda-Forno et. al. “Inducing anxiety in large language models can induce bias”, (2024) https://doi.org/10.48550/arXiv.2304.11111
27. C. List, “Can AI systems have free will?”, Dec. 2024, https://philarchive.org/archive/LISCAS-3
28. J. Maier, “Artificial Intelligence and Free Will”, 2023, https://pub.towardsai.net/artificialintelligence-and-free-will-27e157437e58
29. M. Blum and L. Blum (2022). “A Theoretical Computer Science Perspective on Free Will.” https://arxiv.org/abs/2206.13942
30. K. Farnsworth, “Can a Robot Have Free Will?” MDPI Entropy 19(5), 2017.
31. B. Nolan, “Anthropic’s new AI model threatened to reveal engineer’s affair to avoid being shut down”, Fortune, May 2025.