Journal of Artificial Intelligence and Consciousness
Vol. 7, No. 2 (2020) 257–275
© The Author(s)
DOI: 10.1142/S2705078520500125

World Scientific
www.worldscientific.com

A Constructive Explanation of Consciousness

Pei Wang

Department of Computer and Information Sciences
Temple University, 1925 North 12th Street
Philadelphia, PA 19122-1801, USA
pei.wang@temple.edu

Published 25 July 2020

This paper describes the consciousness-related aspects of the AGI system NARS, discusses the implications of this design and compares it with other relevant theories and designs. It is argued that the function of consciousness is self-awareness and self-control, and the phenomenal aspect of consciousness is the first-person perspective of the same process for which the functional aspect is the third-person perspective.

Keywords: Consciousness; Artificial General Intelligence; Non-axiomatic Reasoning System.

1. Introduction

In recent years, consciousness has become a hot topic in several fields, and opinions from various perspectives have been proposed, as surveyed in Blackmore [2004]; Gamez [2008]; Chella and Manzotti [2012].

Like many basic concepts about thinking and mental processes, "consciousness" has no widely accepted definition, and in different fields the focuses of research are not the same. This paper will not address all approaches, however will describe our approach toward this topic. This line of work has been introduced in our previous publications, such as Wang *et al.* [2017, 2018], and this writing will focus on new considerations and progresses in our project that are not covered in the other writings.

In the following, I will start by explicitly stating my epistemological presumptions and their implications. Consciousness will be introduced initially as a cognitive function that can be realized in a computer system. I will then briefly summarize the design of our model, NARS (which has been specified in Wang [2006, 2013] and explained in our previous publications), with a focus on the distinction between *conscious* and *unconscious* processes. Finally, I will address the phenomenal aspect of consciousness, as well as the related theoretical issues.

2. Theoretical Position on Descriptions

Like it or not, consciousness is fundamentally a philosophical issue in epistemology, as it is about the nature of our knowledge or *descriptions* of ourselves, other minds and the world. As this paper is not a purely philosophical essay, here I will simply describe my position without a systematic comparison with the existing (huge) literature on this topic. Some mostly relevant and important comparisons will be made later in the paper.

2.1. *Description: Subjective versus objective*

To summarize my position on this topic in one sentence, it is "*There is no unique objective description of any entity or event.*"

I acknowledge the existence of the outside world independent of any observer, calling it "agent" or "cognitive system." However, any *description* of an entity or event depends on the agent making or holding the description. This dependency exists for several reasons:

- The sensorimotor interface between the agent and its environment decides the primes (the atomic sensations and actions) the agent can directly experience or take.
- The available concepts within the agent decide the patterns or schemata used in descriptions that summarize the experience.
- Since an agent is always limited by available computational resources, every description is selective in content, and the motivational structure, attention allocation, emotional status, etc. of the agent influence the selection among the candidates.
- As the environment changes constantly, any summary of the past experience may be confronted by new experience, thus no description can be guaranteed to be correct forever.

Since none of the above factors is uniquely determined by the environment alone, accurately speaking there is no way to "describe the world (or part of it) as it is." Instead, every description is made by an agent at a certain moment in a specific context. Here "description" also includes "knowledge," "theory," and similar notions.

Therefore, in its preliminary form, a piece of knowledge K is *subjective* and can only take the form of "I feel K," "I believe K," etc. and is from the viewpoint of an agent, as it comes from the system's experience. Here, the word "experience" is used in its basic sense, meaning a record of the agent's interaction with the environment in a comprehensible form to the system. Similarly, the meaning of a concept is also subjective, as it indicates a specific ingredient or pattern in the agent's experience.

As soon as an agent begins to communicate with other agents, involved concepts start getting a shared, or *inter-subjective* aspect, otherwise no mutual understanding or cooperation is possible. Even so, the subjective aspect is still there. Gradually, subjective knowledge "I believe K" may become "Everyone believes K," which can be simplified into "K," and even be interpreted as an "objective fact" that does not depend on any agent. Since people have different experiences, they rarely agree on anything completely, so objectivity is always a matter of degree. Furthermore, even this objectivity still depends on the community of agents participating the communication. For example, in different cultures what is considered as "fact" is not exactly the same.

Denying the existence of "objective fact" not only contradicts with some philosophical theories, but also challenges the widely accepted opinion that some of our knowledge, especially the "scientific knowledge," describes the world *truthfully*, rather than shows someone's personal opinion or preference. I admit that some descriptions agree with human experience better than some others, and that science is based on the common experience of human beings, rather than that of an individual, but I still insist that the so called "common knowledge" and "objective descriptions" are only acceptable for certain purposes in certain situations.

Though science was traditionally associated with objective truth, the dependency of description on observers has been revealed by more and more evidence in fields including cognitive science, philosophy of science, quantum mechanics and so on. This result has not become the commonsense of the general public, partly because the traditional belief works well enough in our daily life for most purposes. Even so, for our current discussion such a treatment is no longer acceptable. On the contrary, it is the root of many problems, including that of consciousness.

2.2. *Relations among descriptions*

If the same subject matter gets different descriptions, there are various situations that should be distinguished, as they correspond to different relations among the descriptions.

When two descriptions mainly include the same concepts and conceptual relations, but assign different truth-values to those relations, they are *conflicting* and *incompatible* with each other. If they come from different experience, the conflict may be resolved by pooling the evidence and merging the descriptions into a single one with a truth-value corresponding to the pooled evidence, or by selecting the description that is applicable to the current situation and ignoring the other as irrelevant.

When two descriptions mainly include disjoint concepts, they may provide complementary summaries to the subject matter, though in certain situations they also compete as better ways to perceive it. In this situation, the important question is not which of them is more *correct*, but which is more *suitable* to serve the current need.

When the descriptions are actually composed with a large number of related sentences, they can be considered as *theories* about the subject matter. In most domains, there are usually multiple theories that not only describe the situation differently, but also propose different predictions and recommendations for actions.

Usually, each theory has its strengths and weaknesses, which restrict its applicable situations. Contrary to a naive belief, competing theories cannot be combined by merging the "correct parts" of each into a new theory, because the concepts in them often correspond to incompatible ways of summarizing the phenomena.

For this discussion, the most relevant case is when there are multiple theories at different levels of abstraction or generalization. The human brain is often described as a neural network, where the basic unit of analysis is a neuron, with its internal structure (soma, dendrites, axon, and input-output mapping) and external connectivity (with other neurons). However, this is not the only level to describe what happens in a human brain. It is clearly possible to describe a brain as consisting of molecules, atoms, and even quanta. On the other hand, it is possible to describe the human *mind* (rather than *brain*) using the language of psychology (and philosophy) by talking about concepts and their processing.

Though the above levels of description are well-known and uncontroversial, some misconceptions on their relation are widespread and even taken as self-evident. The most eminent one is reductionism, which takes a lower level as more fundamental or real, and a theory at a higher-level can, at least in principle, be completely translated into a low-level theory, but not vice versa. In cognitive science, this school of ideas includes the belief that all psychological phenomena can and should be explained using neurological phenomena, because the former are *caused* by the latter. Some researchers believe that eventually every theory can be absorbed into a "Theory of Everything" in the language of physics, to which all other theories are merely approximations or special cases.

This belief is not completely groundless as many psychological concepts and phenomena indeed have neural-level explanations, and similarly, physics provides explanations for the concepts in some other disciplines. However, as a general conclusion, reductionism assumes a "true description" of everything, therefore contradicts with the epistemological postulate I proposed previously that there is no unique objective description of any entity or event.

When the same subject matter is described at different levels, a low-level description usually contains more details within the subject, while a high-level description usually contains richer relations between it and its context. Consequently, in general neither can be fully replaced by the other, though for a specific purpose, one of them may work better. This is just like observing objects through lenses with different magnifications. Even for the same object, each lens provides a different vision, with its scope and granularity. These visions are different from each other, and each may serve a certain purpose. The important point here is that none of the visions is "truer" than the others, or is "fundamental" so that the others can be derived from it.

One factor often omitted by the supporters of reductionism is the cognitive capability of the creator or user of a theory. A low-level theory does provide more details, but at the same time has a more restricted scope. It also cannot show large-scale patterns and regularity. It is just like that you cannot expect a tourist to use a 1:1 city map, even though it contains much more information than the maps available in the visitor centers. I guess that there is a rough upper-bound on the number of concepts a theory can have, beyond it the theory will be too hard to comprehend and use for a normal human mind. It is similar to the case that even though every software is eventually coded by a long binary string, we cannot, even in principle, discuss software design by talking about which bit should be put at which position in the string. For this reason, there cannot be a theory of everything, as there are needs for theories at different levels of description and with different focuses, and these theories are organized around disjoint central concepts.

Another common misunderstanding is to consider a high-level phenomenon as *caused by* certain low-level phenomena. As in the lenses metaphor where the visions from different lenses do not cause each other, theories at different levels may describe the same event differently, but since they are (parallel) descriptions of the same event, there is no causal relation involved — though causality has different definitions, it is nevertheless about the relations between *separate* events. People indeed often accept a low-level explanation to a high-level concept (just like to "zoom-in"at a certain point in the lenses system), but such explanations are not *causal*. Furthermore, this "zooming-in" practice does not mean that a low-level theory is superior to a high-level one by being "more informative" or something like that, because the opposite "zooming-out" practice also helps in understanding an entity or event by putting it in a wider context or larger picture.

In summary, theories on different levels of descriptions are surely related to each other in content, as a high-level theory is usually a generalization of a low-level one. However, there is neither isomorphism between their concepts, nor causal relations between their phenomena. In general, each level may have its unique values and usages that cannot be obtained at another level, either below or above it.

2.3. *Self-description*

According to the previous analysis, every description and theory is from the view point of some people and at a certain level of abstraction. Of course, it does not mean that every theory is equally good. For a certain situation with problems to be solved, each candidate theory has a degree of applicability, jointly determined by its *correctness* (evidential support), *concreteness* (instructiveness), and *compactness* (simplicity), as explained in Wang [2012].

When this conclusion is applied to the self-description of an agent, there is something special. As explained above, even though every description is intrinsically subjective, it can become objective (i.e. inter-subjective) to an extent when shared by a group of observers who have roughly the same relationship with the object of the description.

However, it cannot be the case when the object is the agent itself. In this situation, the self-description is obtained via a different perceptive process compared to a description made by the others on the same event. When an event happening in agent X is described by agent Y, Z (via observation), and X itself (via introspection), the corresponding descriptions D_Y, D_Z, and D_X are all different from each other, though D_Y and D_Z will be much more similar to each other than to D_X in several aspects:

- They are obtained via different sensors.
- They go through different perceptive processes.
- They are categorized and expressed differently.
- They trigger different associations and responses.

Consequently, D_X is often considered as *private* and *subjective* (as the processing mechanism is not shared), while D_Y and D_Z are *public* and *objective* (as the processing mechanism is largely shared). This distinction is basically the same as that between the so-called "first-person perspective" and "third-person perspective." Though there are correlations between the two, their differences are fundamental. There is no way to feel "what is it like to be X" without being X, as argued by Nagel [1974].

This distinction is at the core of the puzzle of consciousness, which is closely associated with introspection. My position acknowledges this distinction, though disagrees with some of its common interpretations:

- It does not mean that consciousness cannot be studied by science, as its private and subjective nature conflicts with pursue of science for being public and objective. I think that though a conclusion must be public and objective to be considered as a part of science, it does not mean that the *phenomenon* described by the conclusion cannot be private and subjective.
- It does not mean that an explanation of consciousness must reduce a first-person perspective into a third-person perspective. According to the above analysis, it is neither possible nor necessary, because the same process may have different descriptions from different perspective, and none of them is more fundamental.
- It does not mean that it is impossible for AI systems to become conscious, because an artifact cannot have anything private and subjective by nature. I think this nature just prevents us from directly feeling what it is like to be an AI, but it does not mean that an AI also cannot feel what it is like to be an AI, as it is exactly what it "feels", in the sense of receiving and processing a certain type of signal to obtain some effects. We indeed cannot decide whether an AI is conscious according to a phenomenal standard, but have to explore the other aspects of consciousness. However, we cannot even use such a standard to decide whether another human being is conscious, as we cannot feel exactly what it is like to be that person. Since in this case we accept indirect evidence (provided by sympathy, analogy, and so on), the same should be acceptable for AI, though the evidence may be collected in other ways.

Therefore, I agree that consciousness has a phenomenal aspect, which can only be felt or experienced from a first-person perspective, so cannot be used to evaluate the existence or complexity of the consciousness of another agent. However, I do not consider it as the only aspect of consciousness. The other aspect, call it "access" or "functional" consciousness, can be evaluated scientifically. I consider the two aspects as unified as two sides of the same coin, so an agent cannot have one without the other. For an agent, no matter whether it is a human or an artifact, its knowledge about itself can provide cognitive functions in its adaptation process, and makes important difference in its behaviors. Consequently, we can check whether an agent has the functional aspects of consciousness, and use it to infer the existence of the phenomenal aspect.

3. Consciousness in OpenNARS

In this section, I introduce the functional aspects of consciousness in the AI system NARS (Non-Axiomatic Reasoning System), as implemented in the open-source software OpenNARS (at `http://opennars.org/`). Since NARS has been described in monographs Wang [2006, 2013] and a large number of other publications, in this paper, I only summarize its basic ideas and features, and refer to the existing publications for details.

3.1. *Objective and approach*

In my opinion, intelligence should be understood as *adaptation with insufficient knowledge and resources* [Wang, 2019a]. Accordingly, NARS is designed under the Assumption of Insufficient Knowledge and Resources (AIKR), meaning that the system has finite processing capability, while needs to be open to novel tasks, and to respond in real time. The system is adaptive in the sense that it depends on its past experience to deal with the current situation, and to allocate its resources among competing demands.

To obtain this capability in a domain-independent manner as an artificial general intelligence (AGI, as described in Wang and Goertzel [2007]), NARS has been built in the framework of a *reasoning system*. The system carries out all processes as reasoning or inference, and each of them consists of inference steps following a formal logic, so is justifiable and explainable, since the same logic is also applicable to human thinking.

It is easy to see that the *logic* needed for such a purpose has not been well-established in the related fields, mainly because the logic systems proposed so far are mostly incompatible with AIKR, and usually do not take adaptation into consideration, as analyzed in Wang [2019b]. Consequently, NARS includes a new logic, Non-Axiomatic Logic (NAL), which has been designed especially to specify adaptive reasoning under AIKR.

NAL is a *term logic* in the tradition of Aristotle [1989]. In NAL, a *term* refers to a concept within the system (rather than an object or event outside the system) and represents an ingredient in the system's experience. A *compound term* is formed by component terms in one of the predetermined structures, and represents a pattern in the experience. A *statement* is a special type of compound term, and represents a certain form of substitutability between two terms.

NAL is interpreted according to an experience-grounded semantics [Wang, 2005]. According to it, the meaning of a term (or concept) is determined by its experienced relations with other terms (concepts), and the truth-value of a statement measures its evidential support, also according to the system's experience. Consequently, meaning and truth are fundamentally subjective in NARS, though there are objective (inter-subjective) factors coming from the shared experience with other systems. In this aspect, NARS is very different from the traditional "Symbolic AI" approaches, where a symbol needs to be mapped to an outside entity to get its meaning [Newell and Simon, 1976].

Similar to logic programming [Kowalski, 1979], in NARS a statement can correspond to an *event* with temporal attribute, an *operation* executable by the system itself, or a *goal* to be achieved by executing proper operations. Therefore, the formal language of NARS, *Narsese*, can uniformly represent declarative, episodic, and procedural knowledge in a conceptual network, where each vertex corresponds to a concept named by a term, and each edge corresponds to a conceptual relation. This network is dynamically formed and shaped by the system's experience, as its topological structure and the attributes of the vertices and edges can be changed at run-time. Various types of cognitive processes, including learning, planning, perceiving, categorizing, etc. are all carried out by reasoning.

NARS handles concurrent inference tasks by time-sharing, though the mechanism is more complicated than that in an operating system. The computational (time and space) resources are dynamically allocated among the system's tasks and beliefs, biased by their priority values. As the situation changes, the priority values are adjusted at run-time, so the system's response to a task depend on history and context, rather than only on the system's design and the task.

3.2. *System architecture*

OpenNARS (at `http://opennars.org/`) is an open-source implementation of NARS. The architecture of the system is shown in Fig. 1.

The system interacts with its outside environment via an input/output interface composed of multiple types of channels:

Narsese Channel. Each channel of this type connects the system to a user or another computer system. Each input is a task in Narsese and can be a judgment to be digested, a goal to be achieved, or a question to be answered. On the other direction, normally the same types of tasks can be sent out to the other party of the communication, triggered by operations issue by the inference engine.

Database Channel. Each channel of this type connects the system to a knowledge source, such as a database, knowledge-base, ontology server, etc. In addition to the

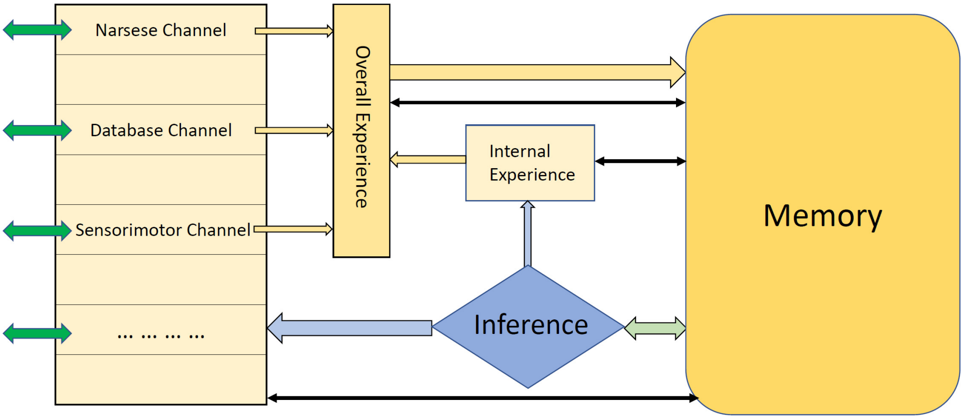

Fig. 1. Architecture of OpenNARS.

functions of a Narsese channel, a database channel also needs to convert Narsese tasks to and from the format used in the database. The functions of the two parties may not be symmetric anymore, as the system may only expect the database to respond to its queries for knowledge, rather than actively recommending knowledge. **Sensorimotor Channel.** Each channel of this type connects the system to a specific type of hardware or software as a sensor and/or an actuator that managed by the operations issued by the inference engine. As discussed in Wang and Hammer [2018], perception in NARS is an operation-driven multi-level generalization process, with results at various levels of description.

All the channels are storage and processing units that can independently carry out certain common functions:

- New input items are accepted at any moment, as far as they are in the format recognized by the channel. Each item will be converted into Narsese tasks or terms with an initial *priority* according to its features, which will decide its competitiveness in resource allocation.
- Each channel has a fixed capacity, and when there are more items, the ones with low priority will be removed. An item will be removed after a certain period of time as well.
- Each channel selects items to be added into the overall experience of the system. The selection is probabilistic, biased by the priority of the items, as well as factors including the priority of the corresponding concept in the memory.
- Each channel carries out certain inference spontaneously on selected terms to form compound terms and tasks. The selection is influenced by the priority of the items in the channel and the corresponding concepts in the memory. The types of inference include temporal composition, by which a series of event can be perceived as a whole. For channels with multiple sensors, spatial composition also happens,

which creates compound terms for certain spatial arrangements of concurrent events.

The *overall experience* buffer is similar to the I/O channels, except that its inputs are the selected tasks of the channels, and the compound terms formed in it may contain components from different modality. Selected tasks from this buffer will be added into the memory for long-term storage and processing. Beside temporal composition, this buffer also selectively adds higher-order statements that explicitly state the relationship between the system and a statement in the buffer. For example, if statement S is selected, the conclusion may be "I believe S," "I see S," etc.

The main content of the memory of NARS is a conceptual network, as described previously. In the memory, the system's experience is summarized into beliefs, each of them is a judgment relating two terms with a certain type of substitutability.

In each inference cycle, a task and a belief are selected (again, probabilistically and biased by priority) from memory and sent to the inference engine. The applicable inference rules will be triggered and derive new tasks. For the details of the inference rules, see Wang [2013]. The results of inference also include commands for certain operations to be executed, either on the (external) environment via an I/O channel, or on the (internal) memory.

The derived tasks are added into the *internal experience* buffer for selection and compound-term composing, just like the external experience, and the selected tasks will be added into the overall experience as well. Therefore, internal and external experiences of the systems are processed initially in parallel and eventually merged.

3.3. *Self-awareness and self-control*

As introduced in Wang *et al.* [2017, 2018], the built-in and plug-in operators in NARS provides the initial knowledge about what the system can do by including a special term *SELF* in the argument list of the operations. All these operations contributes to the meaning of *SELF*, and can be accessed from the concept. As the system starts to run, its experience gradually reveals the preconditions and consequences of the operations, so enriches the meaning of *SELF*, as well as adds non-operational aspects of the concept via inference.

The system's self-awareness also comes from its introspection, i.e. the examination of its own thoughts and feelings. When the system is running, its overall experience not only contains the system's external experience coming from the I/O channels, but also its internal experience coming from the inference engine, as described previously.

The internal experience of NARS is a sequence of events happening within the system's running process expressed as Narsese sentences. It includes the concepts processed and the conclusions derived, as well as the relationship between the premises and the conclusions. In this way, the system can answer questions like where a certain conclusion came from, and what approaches have been explored when solving a problem, etc.

Another important content of internal experience is the system's *feelings* about its status, including those listed in Wang *et al.* [2016]:

Satisfaction. The *achievement* level of a goal measures the closeness between the system's desire and its belief on the matter. The satisfaction level of the system is a weighted average of the achievement levels of the goals that have been considered recently. It indicates how successful the system is in adaptation at the moment.

Alertness. The *novelty* level of a new judgment measures its difference with the system's previous belief or anticipation on its content. The alertness level of the system is a weighted average of the novelty levels of the new judgments that have been considered recently. It indicates how insufficient the system's knowledge is in the current context.

Busyness. The *urgency* level of a task is indicated by its priority. The busyness level of the system is a weighted average of the urgency levels of the tasks that have been considered recently. It indicates how insufficient the system's time resource is at the moment.

Just like its knowledge about the outside world, NARS' knowledge about itself is also under AIKR, that is, being subjective, incomplete, and inaccurate. However, unlike external perception, internal perception starts at term level, though there is still a conceptual hierarchy formed by compound terms. The system gradually forms opinions about how its own mind works, using more and more general concepts to describe its thinking process, and to understand the other minds analogically.

The system's control to its thinking process is also incomplete and fallible. The system's working processes are mostly automatic and routine, determined by its built-in resource allocation mechanism. Self-control only happens in a limited scope, where the system's actions are mostly based on the conclusions and decisions coming from its experience, rather than its built-in mechanism.

In psychology, there is a well-known theory since the work of James [1890] that human thinking consists of both an automatic (unconscious) process and a controlled (conscious) process, where the former happens everywhere and without special effort, while the latter is focused and demands attention. NARS is designed with a similar distinction, though the details are not identical to what happened in the human mind.

As explained previously, I see consciousness as the system's awareness and control of itself, with a functional aspect and a phenomenal aspect. What is described above its the functional aspect in NARS. Though still in a preliminary stage, NARS does have knowledge about itself that is acquired from its experience, and this knowledge has impacts on the system's behavior. In this sense, the system knows what it is doing, though this knowledge is limited, and can even be wrong.

To be more specific, it is reasonable to say that NARS is conscious about part of the content of the overall experience buffer, including the recently perceived events in the external and internal environment, as well as the recent conclusions and decisions made by itself. Since items in the buffer have different levels of priority and obtain different amounts of attention, "conscious" should be taken as a matter of degree. However, it does not means that every process in the system is conscious. Instead, most processes are completely unconscious, meaning that they happen without corresponding representation in the system's overall experience, even though their cumulative effects may get into it when they become significant enough to be noticeable.

In this way, the processes within NARS can also be described from two perspectives: of NARS itself (first person) and of an observer (third person). In the terminology of NARS, the former includes (executable) *operations* and the latter are (non-executable) *events*. The two perspectives of a process are correlated when the system is considered as a whole, though are expressed at different levels of abstraction using difference vocabularies. They cannot be fully converted into each other, since there is no one-to-one mapping between their concepts.

4. Comparisons and Implications

In the following, the above theoretical positions and technical designs are compared with some other theories and approaches on consciousness, and their implications are discussed.

4.1. *Philosophical issues*

A highly influential theory on consciousness was proposed in Chalmers [1995] and his other writings. For this discussion, I summarize his opinions roughly into four major conclusions:

(1) There is an explanatory gap between the *phenomenal* (or *experiential*) aspect and the *functional* (or *access*) aspect of consciousness.
(2) It is possible (at least in theory) for something (like a zombie) to have functional consciousness but no phenomenal consciousness.
(3) The explanation of functional consciousness is relatively easy, but it is hard to explain how phenomenal consciousness arises from a physical basis.
(4) The solution of the problem is the "double-aspect principle," which postulates that "there is a direct isomorphism between certain physically embodied information spaces and certain phenomenal (or experiential) information spaces."

As explained previously, I agree with the distinction between the phenomenal and functional aspects of consciousness, and see it as similar to the first-person and third-person perspectives of thinking process, respectively. Because I consider them different aspects of the same underlying process, it is impossible to have one without the other. Consequently, I do not consider zombie a valid possibility, as the lack of phenomenal consciousness means that certain information fails to be obtained and processed by the agent, which will eventually lead to different behavior. The phenomenal aspect of consciousness is not independent of the functional aspect, as an option that the agent may have *accompanying* the functions. Instead, the phenomena are intrinsic features of consciousness, together with the functions. Though there is no exact isomorphism between the events of the phenomenal and functional aspects, there is still a rough correlation between them overall.

To me, the "hard problem" is not an appropriate problem, because "experience arises from a physical basis" is an incorrect presumption, so there is no "why and how" questions about it. As I explained in Sec. 2.1, the so-called "physical basis" does not consist of objective facts, but inter-subjective descriptions, so is in parallel with subjective experience, without causing the latter.

There are still people who think to understand consciousness means to explain it in neuroscience or even physics, though Chalmers himself explicitly rejected reductionism. I agree with Chalmers that experience should be taken as fundamental. However, I do not agree with his "naturalistic dualism," according to it only a physical theory and a psychophysical theory are needed to describe the world, and other theories, like that from biology, are judged as containing no fundamental principle. As explained in Sec. 2, in my opinion the differences among these theories are quantitative, rather than qualitative.

There are other strong criticisms to reductionism coming from neuroscience (such as [Freeman, 1999]) and physics (such as [Anderson, 1972]) that contain arguments I agree with, though my argument is most from the considerations of cognitive science and AI.

Another major philosophical issue related to this discussion is the long-lasting contrast between determinism and free-will. Most people believe that "Our world is either deterministic or nondeterministic, but not both" [Müller *et al.*, 2019], though there is also suggestion that "neuronal stochasticity may be a main prerequisite to keeping the brain in a flexible state, also for decision making" so that "It seem again to be the combination of chance and necessity determining the functions of life" [Braun, 2019].

As I have rejected the postulate that there is an objective description of anything, whether a process is "deterministic" depends on the observer or describer. The same process can be deterministic to one agent, but nondeterministic to another one, depending on their knowledge, levels of description, requirements on the accuracy and reliability of the conclusion, etc.

Under AIKR, I certainly do not accept the Newton-Laplace version of determinism, but still agree that as we know more and more about how the brain/mind complex works, we can take certain aspects or events as deterministic to the extent that the relevant knowledge can be used to build a computational model that makes more and more reliable predictions, even if the stochasticity of neural activities are taken into account. In principle, we may even be able to know how a person will act before the person is conscious about the decision herself/himself.

On the other hand, from the viewpoint of the person who is making the decisions, such predictions are impossible given AIKR. As far as the person sees the situation, she/he is the one that is about to make a decision freely. Therefore, free-will is typically from a first-person perspective, while determinism is typically third-person. To me, there is no contradiction to accept both determinism and free-will as different perspectives, or to see the world as both deterministic and nondeterministic, with proper definitions of the concepts involved.

4.2. *Psychological issues*

Among the related psychological issues, the most prominent one is the distinction between conscious and unconscious processes in thinking.

Since Freud [1965], the function of unconscious processes has been studied in various ways, though there are still many questions to be answered [Cleeremans, 2014]. Instead of surveying the literature, here I only explains how these questions are answered in NARS.

First, the distinction between conscious and unconscious processes exists in NARS, though there are boundary cases. As consciousness is considered as self-awareness and self-control, not all processes in the system are conscious, but only those that become the object of the reasoning process. For example, if an input event E_1 triggered an output event E_2, the system is conscious about these two events only when their relation becomes part of the system's experience, such as in the form of implication judgment "E_1 implies E_2." In this aspect, our approach is similar to the opinion that conscious processes are higher-order [Carruthers, 2016] meta-cognition [Cox, 2005].

Under AIKR, NARS cannot know every event within itself, nor can it control all of them. Consequently, self-awareness and self-control happen selectively. For example, if event E_3 merely enters the system's input buffer, it is not necessarily conscious; if its priority is high enough, it may be picked up to generate event "E_3 is happening," and becomes conscious.

The perception of internal events and execution of internal activities are all carried out by mental operations [Wang, 2013], so the scope of consciousness in NARS is restricted by the set of mental operators. This is one sense of the conclusion that consciousness is a matter of degree. If a version of NARS has a larger set of mental operators, its consciousness will be richer and more complicated than another version of NARS that has a smaller set of mental operators. Since we have not seen any reason for all intelligent systems to be equipped with the same operator set, they may have different scopes (or levels) of consciousness.

As far as actions are concerned, the "conscious versus unconscious" distinction largely overlaps with the "controlled versus automatic" distinction among the processes. As described in Sec. 3.3, there is a default mechanism managing the routine reasoning/learning activities of NARS, so the system can carry out these activities without explicitly thinking about them, as someone doing free association or day dreaming. On the other hand, the mental operators allow the system to deliberately control its own thinking process to a certain extent. Conscious thought is closely associated to controlled thinking, which adds flexibility to the unconscious/automatic processes, though the latter are more efficient and reliable, just like the relation of the two in the human mind. Of course, I am not claiming that the consciousness/unconsciousness distinction in AI systems will be identical to that in the human mind, but that they are very similar from a functional point of view.

4.3. *AI issues*

At the present time, it is still quite common to find the extreme opinions with respect to the relationship between AI and consciousness: there are people who completely dismiss the possibility for AI system to be conscious, while on the contrary, there are people who consider everything conscious to different degrees. A major reason for this diversity of opinions is that the concept of consciousness has been used with very differently meanings.

Nevertheless, there are also researchers taking a position somewhere between these two extremes. To them, the ordinary computer systems are not conscious, though it is possible to build conscious AI and robots using some new theory and technique. This type of opinions can be found in Chella *et al.* [2008]; Baars and Franklin [2009]; Cleeremans [2014]; Perlis and Brody [2019]. Unlike philosophers and psychologist, the focus of AI researchers is the cognitive functions of consciousness, though some of them also addressed its phenomenal aspects.

Baars and Franklin [2009] describes how to realize a well-known theory of consciousness, the Global Workspace Theory (GWT), in an AGI system LIDA. According to this theory, "consciousness is associated with a global workspace — a fleeting memory capacity whose focal contents are widely distributed ('broadcast') to many unconscious specialized networks" [Baars and Franklin, 2009]. As this approach makes the conscious versus unconscious distinction by the *location* (type of memory) where the processing happens, it is different from how this distinction is made in NARS. However, there is still a strong correlation between these two approaches.

In the architecture of NARS (as shown in Sec. 3.2), the overall experience buffer, the internal experience buffer, plus the storage space in the inference engine play the role of a "global workspace," while the sensorimotor buffers and the concept memory mostly correspond to the "unconscious specialized networks," because the information stored in the latter is partitioned into specific concepts or sensor/actuator, while the information in the former is shared across the whole system. The overall experience buffer holds the tasks to be processed, the internal experience buffer contains new tasks just generated, and inference engine is where the processing happens. In this way, NARS shares the feature stressed by Baars and Franklin [2009]: "A striking feature of GWT is that it accounts for both the massive parallel processing of the human brain, most of which is not conscious (i.e. not reportable), and the surprisingly narrow moment-to-moment capacity of the conscious stream of thought."

The difference between the two systems is that in NARS being in the above workspace is only an approximate *necessary* condition for a data item to be conscious, but not a *sufficient* condition for it, as data in those buffers will not be accessed if their priority values are not high enough. Even the necessary condition is only an approximation, as the other types of memory may also have tasks under conscious processing.

There are several approaches that treat consciousness as higher-order or metacognition [Chella *et al.*, 2008; Chatila *et al.*, 2018; Reggia *et al.*, 2018; Kwiatkowski and Lipson, 2019; Perlis and Brody, 2019], which are similar to the position we take in NARS. Since each of these systems has different architecture and mechanism, what is considered as "higher-order" is quite different, so it is not easy to compare them with each other. What I want to highlight about NARS is its *unified* representation (in Narsese) and processing (using NAL) of first-order and higher-order knowledge, as well as of first-person and third-person perspective. This unification makes the design more consistent, efficient, and elegant.

NARS also unifies conscious and unconscious processes, in the sense that the same representation language and inference rules are used for both. Their differences are more in attention allocation and whether the process is automatic. Since NARS is a reasoning system, many people may consider it a traditional "symbolic AI." However, with experience-grounded semantics, reasoning-based learning, dynamic resource allocation, etc. NARS actually shares many properties with connectionist models, including the features listed in Cleeremans [2014] — "active representation, emergent representation, graded processing, and mandatory plasticity," which are argued to be necessary for unconscious processing. Consequently, the challenge raised in Cleeremans [2014], "how the symbolic representations characteristic of conscious information processing can emerge out of the subsymbolic representations characteristic of unconscious information processing" does not apply to NARS, as the "symbolic versus subsymbolic" distinction no longer exists there, so it cannot be aligned with the "conscious versus unconscious" distinction, which does exist.

In NARS, the inward and outward sensorimotor mechanisms follow the same principles, though the sensors and actuators are different. Consequently, there is also an explanation gap. In principle, NARS can describe its internal processes from both a first-person and a third-person perspectives, though the two descriptions are not isomorphic, though correlated, as discussed previously. It is quite likely that the system will also be puzzled by its own mind-body problem, or even believe that only itself is conscious, because it cannot feel what it is like to be anyone else. In this case, to merely share input data among AI systems will not be enough, because the first-person perspective also depends on the mental structure that has been formed within the agent.

Finally, there is the issue of moral and ethical consequences of machine consciousness. Like all major technical breakthroughs, progress on this topic may produce both positive and negative effects in the human society, and we should do our best to avoid undesired results. However, I do not agree with the popular belief that as soon as an AI becomes conscious, it will necessarily attempt to dominate the world or become evil in some other way. In general, I consider consciousness a necessary feature for an advanced intelligent system, and believe the human society will benefit from progress in the study of machine consciousness, both by getting a better understanding about our own thinking process and by obtaining more powerful techniques to solve the problems we are facing.

5. Conclusion

In my opinion, *consciousness* is not a necessary feature of *intelligence*, but that of an advance type of intelligent system, where the environment the system can perceive and act on includes the system's own cognitive processes.

Consciousness refers to the self-awareness and self-control of a cognitive system, or agent, and its phenomenal and functional aspects correspond to its first-person and third-person descriptions, respectively. Since these two types of descriptions consist of different concepts, there is no perfect translation between them, but an "explanation gap." However, it does not mean that consciousness cannot be studied scientifically, nor that it cannot appear in AI systems.

Consciousness can be specified abstractly, and at this level the mechanism in the human brain and in advanced AGI systems can follow the same principles, though the concrete contents of consciousness, as well as the separation between conscious and unconscious processes, will be different from system to system, due to their different natures and nurtures. Therefore, I do not agree with the conclusion that "AGI is attempting to reproduce all human behaviors linked with intelligence" [Gamez, 2008], because I think it is neither necessary nor possible to reproduce *human behaviors*, which depend on human-specific features. On the contrary, the *principles* behind human behaviors, like "adaptation under AIKR" and "uniformly treat external and internal environments" can be followed in computer systems.

The constructive model of consciousness in NARS is exactly built in this way. We abstract the functions observed in the human mind to a level of description where they become desired and feasible for computer systems that are designed to work in certain types of environments. In this way, I give consciousness a functional explanation without involving neuron or quantum level concepts.

Based on the understanding that functional and phenomenal consciousness correspond to the same underlying process, the above cognitive functions also credit NARS with a phenomenal consciousness, though it is still very simple and basic. I agree the subjective and private nature of this aspect of consciousness, which means that we will never get the feeling of what it is like to be a NARS. However, it is not a reason to deny its consciousness, since NARS has the feeling of what it is like to be a NARS — actually it is all what it can feel.

Given the complexity of all the "self" related phenomena [Hofstadter, 1979], the work presented in this paper is just the first step towards a satisfactory explanation of consciousness. Even so, its constructive nature makes the ideas directly testable, so as to make it possible for them to be gradually refined and enriched into a competitive theory of intelligence, cognition, and consciousness.

Acknowledgments

Thanks to Fanji Gu and Karl Schlagenhauf for profound discussions on the related topics, to Karl Schlagenhauf and Peter Isaev for commenting and proofreading the draft, and to the OpenNARS team for implementing the NARS model and refining the details of the design.

References

Anderson, P. W. [1972] More is different, *Science* **177**(4047), 393−396.

Aristotle [1989] *Prior Analytics* (Hackett Publishing Company, Indianapolis, Indiana), translated by R. Smith.

Baars, B. J. and Franklin, S. [2009] Consciousness is computational: The LIDA model of global workspace theory, *Int. J. Mach. Conscious.* **1**, 23−32.

Blackmore, S. [2004] *Consciousness: An Introduction* (Oxford University Press).

Braun, H. A. [2019] “Stochasticity versus determinacy in neurodynamics — and the questions of the ‘Free Will’,” in *Proc. 7th International Conference on Cognitive Neurodynamics (ICCN2019).*

Carruthers, P. [2016] Higher-order theories of consciousness, in E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, Fall 2016 ed. (Metaphysics Research Lab, Stanford University).

Chalmers, D. J. [1995] Facing up to the problem of consciousness, *J. Conscious. Stud.* **2**(3), 200−219.

Chatila, R., Renaudo, E., Andries, M., García, R. O. C., Luce-Vayrac, P., Gottstein, R., Alami, R., Clodic, A., Devin, S., Girard, B. and Khamassi, M. [2018] Toward self-aware robots, *Front. Robot. AI* **5**, doi: 10.3389/frobt.2018.00088.

Chella, A., Frixione, M. and Gaglio, S. [2008] A cognitive architecture for robot self-consciousness, *Artif. Intell. Med.* **44**, 147−154.

Chella, A. and Manzotti, R. [2012] AGI and machine consciousness, in P. Wang & B. Goertzel (eds.), *Theoretical Foundations of Artificial General Intelligence* (Atlantis Press, Paris), pp. 263−282.

Cleeremans, A. [2014] Connecting conscious and unconscious processing, *Cogn. Sci.* **38**, 1286−1315.

Cox, M. T. [2005] Metacognition in computation: A selected research review, *Artif. Intell.* **169**, 104−141.

Freeman, W. J. [1999] Consciousness, intentionality and causality, *J. Conscious. Stud.* **6**(11−12), 143−172.

Freud, S. [1965] *The Interpretation of Dreams* (Avon Books, New York), translated by James Strachey from the 1900 edition.

Gamez, D. [2008] Progress in machine consciousness, *Conscious. Cogn.* **17**, 887−910.

Hofstadter, D. R. [1979] *Gödel, Escher, Bach: an Eternal Golden Braid* (Basic Books, New York).

James, W. [1890] *The Principles of Psychology* (Henry Holt and Company).

Kowalski, R. [1979] *Logic for Problem Solving* (North Holland, New York).

Kwiatkowski, R. and Lipson, H. [2019] Task-agnostic self-modeling machines, *Sci. Robot.* **4**(26), eaau9354, doi: 10.1126/scirobotics.aau9354.

Müller, T., Rumberg, A. and Wagner, V. [2019] An introduction to real possibilities, indeterminism, and free will: Three contingencies of the debate, *Synthese* **196**, 1–10.

Nagel, T. [1974] What is it like to be a bat? *Philos. Rev.* **83**, 435–50.

Newell, A. and Simon, H. A. [1976] Computer science as empirical inquiry: Symbols and search, *Commun. ACM* **19**(3), 113–126.

Perlis, D. and Brody, J. [2019] "Operationalizing consciousness," in *Papers of the AAAI Symposium* "*Towards Conscious AI Systems*", Stanford, CA.

Reggia, J. A., Katz, G. E. and Davis, G. P. [2018] Humanoid cognitive robots that learn by imitating: Implications for consciousness studies, *Front. Robot. AI* **5**, doi: 10.3389/frobt.2018.00001.

Wang, P. [2005] Experience-grounded semantics: A theory for intelligent systems, *Cogn. Syst. Res.* **6**(4), 282–302.

Wang, P. [2006] *Rigid Flexibility: The Logic of Intelligence* (Springer, Dordrecht).

Wang, P. [2012] Theories of artificial intelligence — Meta-theoretical considerations, in P. Wang & B. Goertzel (eds.), *Theoretical Foundations of Artificial General Intelligence* (Atlantis Press, Paris), pp. 305–323.

Wang, P. [2013] *Non-Axiomatic Logic: A Model of Intelligent Reasoning* (World Scientific, Singapore).

Wang, P. [2019a] On defining artificial intelligence, *J. Artif. Gen. Intell.* **10**(2), 1–37, doi: 10.2478/jagi-2019-0002.

Wang, P. [2019b] Toward a logic of everyday reasoning, in J. Vallverdú & V. C. Müller (eds.), *Blended Cognition: The Robotic Challenge* (Springer International Publishing, Cham), pp. 275–302, doi: 10.1007/978-3-030-03104-6_11.

Wang, P. and Goertzel, B. [2007] Introduction: Aspects of artificial general intelligence, in B. Goertzel & P. Wang (eds.), *Advance of Artificial General Intelligence* (IOS Press, Amsterdam), pp. 1–16.

Wang, P. and Hammer, P. [2018] "Perception from an AGI perspective," in M. Iklé, A. Franz, R. Rzepka & B. Goertzel (eds.), *Proc. Eleventh Conference on Artificial General Intelligence*, pp. 259–269.

Wang, P., Talanov, M. and Hammer, P. [2016] "The emotional mechanisms in NARS," in B. Steunebrink, P. Wang & B. Goertzel (eds.), *Proc. Ninth Conference on Artificial General Intelligence*, pp. 150–159.

Wang, P., Li, X. and Hammer, P. [2017] "Self-awareness and self-control in nars," in T. Everitt, B. Goertzel & A. Potapov (eds.), *Proc. Tenth Conference on Artificial General Intelligence*, pp. 33–43.

Wang, P., Li, X. and Hammer, P. [2018] Self in NARS, an AGI system, *Front. Robot. AI* **5**, doi: 10.3389/frobt.2018.00020.