

Available at SSRN: <https://ssrn.com/abstract=6911039> or
<http://dx.doi.org/10.2139/ssrn.6911039>

Consciousness is all you need

John E. Stewart

Evolution, Complexity and Cognition Group, Center Leo Apostel, Vrije Universiteit Brussel,
Brussels, Belgium

future.evolution@gmail.com

Abstract

An acceptable information-processing theory of consciousness should be able to identify the adaptive advantages that drove the emergence of consciousness during the evolution of life. It should also predict the specific dynamical architecture of information processing that would need to be instantiated in AI to produce consciousness and the superior adaptation it enables. Whether such an instantiation produces AI that is actually conscious and also more adaptable would provide the ultimate test of the theory. A prime candidate for such a theory is the Subject-Object Emergence Theory of consciousness. It argues that consciousness first evolved because it enabled organisms to achieve adaptive body-environment coordination without extensive trial-and-error learning. It postulates that the subject in an appropriate Subject-Object subsystem would be able to use depictive (iconic) visual representations of the relative positions of its body and the environment to guide motor actions that will produce adaptive body-environment coordination. The depictive representations will 'light up' for such a subject, producing subjective experience that is used to deliver adaptive benefits. Hand-eye coordination is a familiar example in humans—novel and intricate coordination tasks can be undertaken without additional reinforcement learning, provided focused conscious attention is employed to provide us (the subject) with relevant depictive images. The paper identifies how such a conscious Subject-Object subsystem could be instantiated in AI systems, enabling hand-eye and other body-environment coordination without the extensive reinforcement learning or complex computational programming needed at present. Drawing further on the Subject-Object theory of consciousness, the paper also identifies how these simple conscious subsystems evolved further in organisms to establish the conscious modelling that enables conscious planning, imagining, abduction and other higher cognitive functions. It demonstrates that current approaches to incorporating world modelling in AI will fail to achieve key elements of the general intelligence found in humans that require consciousness.

Keywords

consciousness; blindsight; conscious vision; hard problem of consciousness; conscious robots; adaptive advantages of consciousness; conscious AI.

1. Introduction

Currently, hundreds of theories of consciousness have been developed, but none have yet been accepted widely (e.g. see Kuhn, 2024). Very few of the theories make novel, bold, surprising predictions that are readily testable and that would justify elevating them above other hypotheses if the tests were met.

This paper focuses on a relatively recent theory which does make such predictions: the Subject-Object (S-O) Emergence theory of consciousness (the S-O theory) (Stewart, 2022). The theory has major implications for how the capabilities of both AI and humans can be enhanced.

These implications include two sets of novel and surprising predictions which can be tested reasonably easily. Passing these tests would provide a sound basis for elevating this theory of consciousness above all other current approaches.

The first set of tests are derived from the theory's claim to identify a 'minimally-complex' information-processing architecture that can produce conscious functioning. A decisive test of the theory is whether the instantiation of such a dynamical architecture in an AI system is capable of producing conscious functioning and the consequent enhancements in adaptability that the theory predicts. Examples of these tests are whether such an instantiation will enable an AI to learn to drive an autonomous vehicle consciously with as little specific pre-training as an 18-year-old, or to consciously navigate complex, changeable environments with as little specific pre-training as a human toddler.

Current AI systems provide an ideal opportunity to test information-processing theories of consciousness such as the S-O theory. This is because their current functioning proceeds unconsciously 'in the dark'. Furthermore, their information processing systems can be readily modified.

The second set of tests of the theory rely on direct observational evidence of how consciousness functions in humans. Of course, due to the subjective nature of consciousness, these observations can only be undertaken by an individual who observes their own functioning. Furthermore, the observations can only be made effectively by individuals who have the capacity to be mindful and meta-aware while they are undertaking tasks and activities consciously. In such a state, their conscious attention will not be fully absorbed in the details of their actions. Although they continue to give conscious attention to the details of the tasks, they are simultaneously aware of themselves, their actions and internal conscious functioning from a broader meta-perspective.

An example is an individual who observes in a state of uncontracted, mindful awareness what is actually occurring when they are undertaking a novel, unpractised, complex, hand-eye coordination task, such as threading a needle in windy conditions. Such an individual will be able to observe that they accomplish the task by using conscious images of the relative

position of their fingers, the needle, and the thread. As they proceed with the task, they will observe that the conscious images identify how the relative positions of fingers, needle and thread change in response to their motor actions. They will see that they use real-time feedback from these ‘lit up’ conscious images to guide their motor actions and to accomplish the desired hand-eye coordination. They do not rely on the extensive trial-and-error reinforcement learning that many robotic systems would need to accomplish such a novel, complex task (e.g. see Tang et al. 2025), or on mathematical modelling that depends on precise positional data, such as SLAM and visual servoing systems (e.g. see Yue et al. 2024), or on the extensive training needed to produce comprehensive world models (e.g. see Ding et al. 2025)

The remainder of the paper considers these issues in detail:

Section 2 outlines the key elements of the S-O theory of consciousness and the implications of the theory for enhancing the adaptive functioning of AI systems.

Section 3 identifies how conscious functioning could be instantiated in robotic and other AI systems.

Section 4 considers how the existence of consciousness in an AI system could be assessed, despite the inability of the AI to self-report about any conscious experience that it may have.

Section 5 outlines how the S-O theory can also be tested through the mindful self-observation of novel, complex hand-eye coordination tasks and other actions which require close conscious attention if they are to be performed effectively.

2. The Subject-Object (S-O) theory and its implications for instantiating consciousness in AI

2.1 The adaptive advantages that consciousness provided

The primary focus of the S-O theory of consciousness is to identify the nature of the information-processing subsystems that first gave rise to conscious experience during the evolution of life on Earth. More specifically, S-O theory sets out to identify the simplest (‘minimally-complex’) subsystem that can produce consciousness, how its functioning generates conscious experience, and the adaptive benefits that such a subsystem can provide (Stewart, 2022).

Prior to the emergence of consciousness, organisms were able to discover effective new behaviours during their lives through various forms of reinforcement learning, such as operant conditioning. Novel behaviours that enabled organisms to achieve their goals in particular circumstances were reinforced and learnt. A potentially unlimited variety of new adaptive behaviours could be discovered and adopted in this way.

For example, given sufficient reinforcement learning, a pre-conscious organism could learn by trial-and-error how to navigate through a complex environment to obtain, manipulate

Consciousness is all you need

and ingest a food item. However, if an organism is equipped only with reinforcement learning, it must continually repeat this trial-and-error learning process as circumstances change. It needs to discover and learn the specific sequences of actions that will be effective in the changed circumstances e.g. the organism is likely to need to learn different sets of actions for different food items that are in different positions in environments that contain different obstacles that are positioned differently, and so on.

Organisms that have a capacity to substantially reduce the extent of trial-and-error learning needed to achieve their goals in complex, changing circumstances would possess a major evolutionary advantage.

S-O theory sets out to demonstrate that conscious functioning was able to provide this advantage, enabling the emergence and spread of consciousness. The theory shows that conscious functioning can greatly reduce the extent of trial-and-error needed for the organism to discover the actions and behaviours needed to achieve its goals. Conscious S-O systems can achieve this by using real-time feedback from visual representations of the organism's interactions with its environment.

2.2 The functioning of a conscious 'minimally-complex' Subject-Object subsystem

S-O theory argues that the simplest form in which conscious functioning can operate effectively in this way is what it refers to as a minimally-complex Subject-Object (S-O) subsystem. I will now describe the dynamical architecture of such a S-O subsystem and demonstrate how it can use real-time visual feedback to significantly reduce the need for costly trial-and-error.

The architecture of such a S-O subsystem which processes visual inputs comprises a depictive representation (the Object) which is interpreted and used by the Subject. The depictive representation which is assembled from visual processing depicts the body of the organism as it interacts with elements of its environment (a 'depictive' representation of a scene is one in which the spatial relationships between elements in the scene are preserved in the representation itself. Representations of this kind have also been referred to as 'iconic' or 'analogue' representations). As the representation is updated through time, it provides the Subject with continuous, real-time feedback about the consequences of motor actions that the Subject initiates in order to navigate the organism through its environment. Importantly, the relative positions of elements of the environment and of the body of the organism are represented depictively as they change in response to actions initiated by the Subject.

Equipped with such S-O subsystem, the Subject does not have to learn how to navigate to a food item in a novel environment through extensive trial-and-error. Instead, it can navigate using real-time feedback from depictive representations of the consequences of its actions.

Consciousness is all you need

Hand-eye coordination provides a specific example of the functioning of a simple S-O subsystem that is familiar to humans. It can serve as an effective 'intuition pump' for understanding the operation of a S-O subsystem. When, for example, we set out to thread a needle in challenging circumstances, we find that we must give close conscious attention to the thread and needle. Given the novel circumstances, we will not have already learnt the specific actions that we would need to take if we were to complete the task automatically, without conscious attention (if we had, we could thread the needle without conscious attention, blindsightedly). The depictive representation of our hands, the needle and the thread is the Object in our S-O subsystem, assembled from disparate visual processing. As we (the Subject) initiate motor actions intended to progressively achieve the movements that will thread the needle, we use continuous feedback from the representation to guide our actions, not extensive trial-and-error.

Prior to the emergence of S-O subsystems and consciousness, organisms were like robotic systems that are trained only by reinforcement learning. Fully equipped with visual sensors, these organisms could learn to identify salient objects in their environment and to learn how to navigate their way to them and manipulate them. But this processing occurred 'in the dark'. Nothing 'lit up' for a subject. Consciousness did not arise. Like reinforcement-trained robots or self-driving vehicles, the organisms were capable only of blindsight. But like robots, if given sufficient training, they could learn much the same behaviours as those exhibited by conscious organisms. From the outside, their behaviour could be virtually indistinguishable from the behaviour of conscious organisms. But they would not be conscious of anything they were doing.

The emergence of conscious functioning enabled what has been referred to as 'one-shot learning'. But consciousness accomplishes this not by providing superior learning capacities, but by enabling the achievement of complex body-environment coordination without significant learning.

2.3 How a S-O subsystem gives rise to subjectivity and conscious experience

But why does the information-processing architecture embodied in a minimally-complex S-O subsystem produce consciousness? Why do the Subject's interpretation and use of the representations which constitute the Object 'light up' for the Subject?

The central function of the Subject is to use its interpretation of the depictive representations (the Object) to initiate and guide its motor actions to achieve the desired coordination of its body with the relevant parts of the environment. The Subject needs to monitor, assess and act upon the representations continually until the desired outcome is achieved. If the Subject's interpretation of the representations is to be effective for this purpose, it must be able to distinguish the body from the relevant parts of its environment within the representations, and to track their relative positions in space as these are changed by motor actions.

Consciousness is all you need

Selection and learning can be expected to favour useful changes to the processes which construct the depictive representations as well as useful changes to the way in which the Subject inspects the representations i.e. changes which enhance the ability of the S-O subsystem to produce adaptive body-environment coordination.

For example, changes to the Subject and/or the Object that enable the Subject to distinguish more effectively the relevant parts of the body and of the environment. can be expected to be favoured/reinforced. This will be the case even where the changes produce representations and interpretations of representations that are not veridical i.e. that do not reflect actual differences in the body or environment. For example, changes that aid the Subject to distinguish between the body and the environment would be expected to be favoured, even if they do not reflect differences in the original sense data.

As a consequence, selection and learning will tend to ensure that both the nature of the representations that constitute the Object and also the nature of the Subject's interpretation of the representations will tend to coevolve. The arbiter of what emerges will be determined by the ability of both, acting together, to enhance the adaptive functioning of the S-O subsystem.

Furthermore, the Subject's continual interpretation of the Object means that the Subject continually experiences the relevant depictive representations which constitute the Object. This continual inspection produces the subjective experience of the Subject.

However, with our current state of knowledge, it is not possible to explain in detail the nature of the experiences that arise for the subject in particular circumstances. We do not know, for example, how specific patterns of neuronal firings produce particular experiences in the Subject of aspects of its environment. We cannot explain why the colour we label 'blue' looks the way it does to us.

But what we can assert with confidence is that the Subject and Object will coevolve so as to ensure that the depictive representations that 'light up' for the Subject will be such as to enable the subsystem to undertake effective body-environment coordination. Evolution and learning processes will shape subjective experience (the experience of the Subject) so as to enhance its ability to contribute to these adaptive capabilities.

2.4 The further evolution of Subject-Object subsystems – conscious modelling

S-O theory also considers how minimally-complex S-O subsystems and consciousness evolved additional capacities as evolution proceeded. A central challenge facing S-O theory is to explain the evolution and functioning of the more complex forms of consciousness found in humans. Conscious functioning in humans goes considerably beyond the ability to achieve body-environment coordination more efficiently than through reinforcement learning.

Consciousness is all you need

Perhaps the most significant enhancement of functioning in human consciousness is the ability to consciously build and manipulate mental models and simulations of interactions with the environment. Humans can use these models to predict the outcomes of alternative behaviours. This enables them to identify and implement the actions that will achieve their goals most efficiently and effectively. Ultimately, the development of this capacity has enabled science and technological development.

How did this functionality evolve and develop from minimally-complex S-O subsystems? What is it about these subsystems that facilitated the development of modelling capacities?

The Subject in a simple S-O subsystem has the ability to interpret depictive representations as they are updated through time. The Subject also has the capacity to use its interpretations to assess whether its motor actions are achieving its goals—it can read-off from the depictive representations the consequences of motor actions that it initiates. As such, the Subject in a simple S-O subsystem fulfills the role of an assessor/interpreter of depictive representations. It uses its assessments to evaluate the adaptive impacts of alternative actions. It is the ability of the Subject to be conscious of depictive representations that enables it to achieve the coordination that is useful to it.

But simple S-O subsystems use depictive representations of the organism's current interactions with its environment, not of possible future interactions. Furthermore, the representations are updated primarily by the senses in real time, not by predictions about the consequences of possible future actions. Additional functionality that can enable these predictions is essential if simple S-O subsystems are to evolve a capacity to consciously build and manipulate depictive models. How was this functionality incorporated? What were its antecedents?

Predictive-processing theory provides plausible answers to these issues. The theory explores the emergence and functioning of processing that anticipates the sensory and other effects produced by motor actions (e.g. see Friston, 2010). This processing uses internal 'simulations' of motor actions to predict the sensory consequences that will result from taking the actions. If the simulated predictions are not met, the organism modifies its actions until the outcomes of its actions enable it to achieve its goals (Pezzulo, 2011).

Initially, these predictive processing simulations were part of the Subject in simple S-O subsystems. They operated 'in the dark.' However, once the simulations became an object of consciousness that 'lit up' for the Subject, they could be interpreted and manipulated consciously e.g. in 'the mind's eye' in the case of humans. The predictive processing which began as part of the Subject become Object to a new, higher-level Subject. The new Subject could use its control over motor actions to also control the predictive processing simulations in the expanded Object. The models that were produced would 'light up' for the Subject.

From these foundations, the Subject developed the ability to consciously build and manipulate the mental models which were part of the expanded Object. The eventual

Consciousness is all you need

outcome was depictive mental models that could be used consciously to predict and evaluate the outcomes of alternative motor actions, including in counterfactual circumstances (Pezzulo, 2011; Clark, 2015; Stewart, 2022).

More generally, Kegan (1982, 1992) suggests that as humans develop, an increasing variety of brain processes become object to consciousness, enhancing cognitive and social-emotional capacities. This occurs through a recursive developmental process in which brain processes that were initially part of the Subject (and operated outside conscious awareness), become Object to a new, higher-level Subject (the new Subject is conscious of and manages a wider range of processes--these constitute an expanded Object). This enables these processes which previously unfolded automatically to be adapted and manipulated consciously. It is likely that such a recursive process that produces a hierarchy of levels has also driven the evolution of consciousness in other organisms, not just in humans.

Stewart (2022) uses Kegan's framework to examine four levels in the evolution/development of consciousness that follow the initial emergence of a simple subject-object subsystem and that are particularly significant in human development. Three of these levels concern cognitive development, and the fourth involves social/emotional development as well. Each transition in cognition enabled the development of depictive simulations that were more complex and abstract. Broadly, the first two levels correspond to Piaget's Concrete Operations and Formal Operation's levels. The third which is referred to by Stewart (2022, 2024 and 2026) as Metasystemic Cognition enables the building of conscious mental models of dynamical complex phenomena. Very few individuals attain this level currently.

The instantiation and development of these key levels in AI is likely to prove indispensable in attempts to develop Artificial General Intelligence and Superintelligence.

3. How conscious functioning can be instantiated in AI systems, enhancing their capabilities and testing the S-O theory

3.1 Existing robotic AI are blind-sighted, not conscious

Current AI, including robotic systems and autonomous vehicles, do not contain conscious S-O subsystems. They do not achieve body-environment coordination consciously using depictive representations of their body as it interacts with the environment.

Nevertheless, equipped with appropriate capacities for reinforcement learning, they can be trained to undertake tasks requiring complex body-environment interactions, including hand-eye coordination. As noted above, these learnt behaviours can appear very similar to the body-environment coordination tasks achievable by conscious humans and other biological organisms. However, the extent of training required by AI far exceeds that required by conscious organisms. For example, for its robotic systems, Generalist AI currently uses a foundational pretraining dataset containing over half a million hours of real-world human physical interaction data (Generalist AI, 2026). This estimate of training time would

Consciousness is all you need

be massively increased if robots rather than humans were used to build the dataset (i.e. if the time taken to train robots to the human level of performance was included in the estimate).

Commonly, human observers without a deep understanding of AI functioning will tend to intuit that such a highly-trained robot 'sees' what a conscious human sees when it performs similar tasks. The 'Theory of Mind' of human observers will tend to provide them with the same intuition they have when they observe other humans undertaking similar tasks i.e. that they are consciously aware of what they look at (Gray et al. 2007).

Few human observers are likely to have any experience themselves of successfully undertaking demanding hand-eye coordination tasks without visual consciousness. Furthermore, they are likely to find it difficult to imagine how a human or a robot could successfully undertake complex hand-eye coordination tasks 'in the dark' without any conscious visual awareness of the elements of the task.

Yet it is clear that robotic systems equipped with reinforcement learning operate with what has been referred to in humans and other organisms as blindsight (Azzopardi and Cowey, 1998). Like a personal computer, their information processing occurs 'in the dark.' Neither does their information processing actually assemble the kinds of internal depictive representations of objects that are perceived consciously by humans.

The same is the case for AI that is equipped with SLAM or visual servoing systems which undertake body-environment coordination by using sensors to locate relevant elements of the environment and the body in space, and then using the resultant map and appropriate algorithms to calculate mathematically what movements need to be taken in order to achieve the appropriate coordination.

Humans and primates who suffer brain injuries later in their life that cause blind sightedness apparently have no conscious visual experience, yet they can pick up and manipulate objects in their environment to some degree. Nothing 'lights up' visually for them, yet they can detect objects visually. Apparently, before the onset of blindsight and while they still possessed visual consciousness, they learnt to undertake some body-environment coordination tasks which could be undertaken automatically, without conscious awareness (Azzopardi and Cowey, 1998). Furthermore, after the onset of blindsight, they can learn new tasks if trained with appropriate reinforcement learning, just as robots can (Humphrey, 2023).

Some researchers hoped that this contrast between conscious vision and blindsight would provide significant opportunities to investigate several key issues about consciousness (Humphrey, 2023). These include:

- (a) What adaptive functionality (if any) would consciousness add to a blind sighted individual?

Consciousness is all you need

- (b) What additional information processing would be needed to produce consciousness and any additional functionality in a blind-sighted individual?
- (c) How would this additional information processing produce visual conscious experience?

However, this potential has not yet been fulfilled. In significant part, this is because researchers are unable to investigate these issues by, for example, instantiating alternative information processing systems in human or animal subjects.

But the recent emergence of critical new opportunities has the potential to overcome these obstacles. First and foremost, this is because robotic AI now exists whose information processing can be modified in ways that enable these issues to be tested and explored. For example, robotic information processing can be changed to test competing hypotheses about the processes and systems that could enable visual consciousness.

3.2 Instantiation in AI of a simple S-O subsystem

In order to be capable of conscious visual functioning, including body-environment coordination in complex circumstances, robotic systems and autonomous vehicles will need to be equipped with appropriate S-O subsystems. Significantly, the Object in these subsystems will need to comprise depictive representations. Representations that are mathematical structures such as tensors (or simulations of them), will not enable the same capacities as the S-O subsystems that produce consciousness and that organise body-environment coordination in organisms. Representations that are not depictive will fail to suffice. If the Subject is to be able to read-off directly from the representations the extent to which its motor actions are achieving the desired coordination, the representations must be depictive, showing the relative positions of body and environment as they change through time in response to the actions.

The advantage of representations which are depictive is not just a conclusion derived from S-O theory. Extensive evidence is accumulating that human consciousness relies in large part on the assembly and processing of depictive representations: the conscious visual perceptions experienced by humans are depictive representations, as can readily be confirmed by mindful self-observation; as we navigate our environment, areas of the cortex in which our current location is represented are themselves organised as spatial grids (Doeller et al. 2010). The spatial grids can also represent information about imagined movement and spatial orienting, as well as thoughts and mental simulations (Jacobs & Lee, 2016). Haun and Tononi (2019) note that most of our conscious experiences feel spatially extended and argue that this is due to the organisation of our posterior-central cortex as a hierarchy of grids; the mental models that humans build and manipulate consciously are often perceived in 'the mind's eye' as being organised spatially, as can be confirmed by mindful self-observation (Kunda, 2018); human metaphors for phenomena that are not themselves organised spatially often make use of spatial terms e.g. narratives and other events in time (Jaynes, 1976).

Consciousness is all you need

In humans as well as in AI systems, the initial processing of visual sensory information is undertaken in a distributed fashion using neural networks that, for example, are trained to detect particular features and combinations of features. In humans, the results of this disparate, distributed processing of visual inputs are then reassembled into internal depictive representations. This final step does not occur in current AI systems. Typically, they do not re-assemble the distributed outputs of their visual processing to produce internal depictive representations, although they are often capable of reassembling processed visual inputs to produce depictive images on external screens.

In order to instantiate in AI a minimally-complex S-O subsystem which is comparable to those in conscious biological organisms, the distributed outcomes of visual processing in AI must first be re-assembled to produce internal depictive representations. These would then need to be scanned and interpreted by the Subject of the S-O subsystem. This functioning would need to be tuned so that relevant aspects of the depictive images ‘light up’ and are experienced by the Subject, and so that the subject is thereby provided with real-time feedback about the extent to which its motor actions are achieving appropriate body-environment coordination.

There are no obvious technical reasons why such a simple S-O system could not be instantiated and trained in AI. First, current AI equipped with deep neural networks (DNNs) could be trained to assemble the outcomes of distributed visual processing into internal depictive images of the relevant interactions between body and environment. These depictive images could be displayed in internal spatial arrays and grids, instantiating the Object. Second, suitable DNNs could be trained to use interpretations of these internal depictive images to guide motor actions to achieve appropriate body-environment coordination. These processes would instantiate the Subject.

However, several requirements need to be emphasised: First, as underlined already, the actual internal representations which constitute the Object must be depictive and instantiated, not distributed or simulated images, or mathematical structures.

Second, the other key elements of the S-O subsystem also need to be instantiated, not simulated. If the S-O subsystem is to actually function consciously, it must contain not only instantiated depictive representations, but also a functioning Subject that interprets the representations as depictive representations. If this condition is not met, there will be no Subject that actually experiences the Object—subjectivity and consciousness will not arise.

Third, the processes that produce both the internal depictive representations and the interpretations of those representations need to be co-trained, not trained separately. This is because in a functioning S-O subsystem, the representations and the processes that interpret and utilize them must work together to achieve the goals of the subsystem. For example, the feature-detection processes that operate in initial visual processing cannot be optimised in isolation. Instead, they need to be co-trained with the Subject so that their

Consciousness is all you need

combined ability to provide relevant real-time feedback about progress in coordination is optimised.

Of course, it is possible that the DNN forming the Subject might learn to perform a particular coordination task without developing a functioning S-O subsystem. Instead, for example, it might learn to perform the task through basic reinforcement learning.

However, the DNN and training process could be set up in ways that bias the training in favour of producing a functioning S-O subsystem. For example, the AI could be trained on tasks that each involve variable, complex circumstances which cannot be mastered easily and simply through trial-and-error learning.

However, by themselves, these kinds of methods will not provide certainty about whether any particular attempt to instantiate a simple, conscious S-O subsystem in AI has succeeded. A key challenge remains to be met: what tests can be used to assess whether any particular attempt has succeeded in producing conscious functioning? This challenge is addressed in Section 4 below.

3.3 Further development of S-O subsystems in AI to enable conscious modelling

Once simple S-O subsystems are instantiated successfully in AI, the next challenge will be to develop their capacity for conscious modelling. It is beyond the scope of this article to identify in detail how this might be achieved. But it is worth making several key points here.

Work on this challenge has been seriously limited by a number of factors. To date, mainstream human cognitive science has been unable to develop adequate explanations of how humans build and use conscious modelling. Furthermore, it has not identified practices and methods that can escalate human cognition through the levels referred to briefly in Section 2 above.

Broadly, this is because these capacities are too dynamically complex to be understood by the analytical/rational (formal operations) thinking which spread during the European Enlightenment and that drives current mainstream science and technological development (Stewart 2024, 2026). Currently, very few individuals have developed the metasystemic level of cognition which is essential for understanding dynamical complex systems as they unfold and interact through time.

However, individuals who develop the capacity to observe mindfully in real time their own thought processes and model-building will begin to overcome these limitations. The acquisition of this capacity will significantly boost their ability to understand the conscious modelling and associated cognitive processes that need to be instantiated in AI. This will particularly be the case as individuals begin to develop the ability to consciously and intentionally escalate their own cognition by evaluating their current modelling abilities and improving them recursively.

Consciousness is all you need

Stewart (2024) deals in detail with how capacities for meta-aware self-observation and for metasystemic cognition can be developed in humans. The book identifies how meta-aware self-observation can be used to develop the ability to 'see' one's thinking and model building consciously as Object, to develop models of one's modelling capacities, and to use these to evaluate and improve one's cognitive processes recursively. These abilities will enable individuals to continually remake themselves cognitively, moving through the hierarchy of levels of cognition and model building mentioned in Section 2 and considered in more detail in Stewart (2022, 2024).

Conscious AI should be able to use similar practices and methods to escalate their own cognition, including their capacities for conscious modelling. AI that develops these abilities will have obvious advantages over AI that is capable of attempting to improve its intelligence only by modifying its source code. In particular, source code can only be re-engineered intelligently to escalate cognition if it is guided by an understanding of the effects that possible changes will have at each level in the complex hierarchy of emergent levels of cognitive development.

Mindful observation of model building will also confirm that attempts to install higher cognition in AI using current approaches to world modelling will fail to produce AI with human-like general intelligence. This is despite the fact that existing approaches to world modelling can be extremely effective at identifying patterns and regularities in data, including sense data, given enough training. They can also predict the consequences of these patterns, if the patterns continue (e.g. see Ding et al. 2025; Zhang et al. 2025; Feng et al. 2025).

However, world modelling is not effective at doing what advanced human science can achieve. It is poor at imagining and modelling the underlying generative processes that can explain the patterns and regularities that manifest in existing data. This is because the patterns embodied in the generative processes themselves are not present in this data (e.g. see Bhaskar, 1975; Pearl, 2019).

In contrast, conscious modelling in humans can use counterfactual and abductive modelling to build novel explanations of great power and generality that go well beyond existing data (Pearl, 2019).

4. Assessing whether an attempt to instantiate a S-O subsystem in a robotic system or other AI has produced conscious functioning

The current lack of knowledge about consciousness and its causes is reflected in our lack of ability to assess whether a particular organism or other entity is conscious. In the case of other humans, reliance on self-reporting is often the best we can do. But self-reporting is not effective for other biological organisms or for AI.

Consciousness is all you need

This is obviously a central issue for attempts to test the S-O theory by instantiating simple S-O subsystems in AI systems. How can we assess whether the instantiation has in fact produced conscious functioning?

Strong support would be provided for conscious functioning if instantiation of a S-O subsystem enabled an AI to perform the additional functionality predicted by the theory, with minimal additional trial-and-error learning. Provided alternative explanations for the additional functionality could be ruled out, this would suggest that the new capabilities resulted from the functioning of the instantiated S-O subsystem.

For example, consider the case of an instantiated S-O subsystem which successfully uses visual inputs to achieve novel body-environment coordination with little additional training. This would support the conclusion that the coordination was achieved by the Subject as it inspected depictive representations and used the results of this inspection to guide its attempts at coordination in real time. A positive outcome in such a test would suggest that the Subject in the AI's S-O subsystem had experienced the depictive representations in a similar way to humans who undertake novel hand-eye coordination tasks.

It is worth emphasising that the tests would need to be designed and set up so that the only way the AI could successfully complete novel tasks is through the operation of its S-O subsystem—the experimental design must rule out the possibility that the AI is able to use any other learnt, programmed or hard-wired mechanisms to complete the novel tasks with minimal trial-and-error.

Of course, a positive outcome for such a test would not provide direct, conclusive evidence that the AI is conscious. It does not prove that it 'feels like something' to be the subject as it inspects the depictive representations continually through time. It does not prove that depictive representations 'light up' for the Subject. Direct evidence of the subjective experience of other entities is impossible. But a positive outcome would be persuasive.

5. Testing the S-O theory using mindful self-observation of conscious body-environment coordination

Individuals with a capacity to be mindful and meta-aware while undertaking conscious body-environment coordination have the potential to test key predictions of the S-O theory of consciousness. With a still mind which is not embedded and contracted in thought and feeling, they will be able to observe themselves as they undertake relevant tasks consciously. This will enable them to collect data about conscious functioning that is not available to individuals without the capacity. Individuals can develop this ability through appropriate meditation-like practices (e.g. see Stewart 2007, 2024).

An example of a task that can be used for such a self-observational test is the undertaking of a novel, untrained, difficult, hand-eye coordination challenge, such as threading an unfamiliar needle with a limp thread. Significantly, this is an example of a task that current

Consciousness is all you need

robotic systems would be unable to perform effectively without extensive pre-programming and/or specific training.

The individual will note that successful completion of the task requires focused and sustained conscious attention. Unlike body-environment coordination tasks that the individual has undertaken extensively in the past, such as driving a vehicle in familiar circumstances and grasping and manipulating tools whose use has been mastered, it cannot be completed successfully with scant conscious attention. It cannot be undertaken in an unconscious, blindsight-like state.

If the individual is in a state of spacious meta-awareness while they observe themselves attempting to thread the needle, they will be able to see directly how they complete this untrained, novel task. They will see that as their attention moves between the needle, the thread, and their hands, they are using conscious, depictive images of these to assess in real time the effectiveness of movements that they are initiating. The depictive images 'light up' for them. When they make a movement, they read-off from the depictive images whether or not the movement has advanced the successful completion of the task. The outcome of this assessment is used to correct for the previous movement or to build upon it, and so on until the needle is threaded successfully. They will notice that in undertaking the task, they rely primarily on direct feedback from their conscious perceptions.

It will be obvious to them that their undertaking of this novel task is not a result of trial-and-error reinforcement learning. Nor is it produced automatically by a SLAM-like system which uses mathematical representations of the precise positions of the components of the task to compute the specific movements that are needed to undertake the task successfully. Instead, they will notice that as they undertake the task, they are relying in the main on real-time feedback from what 'lights up' in their consciousness. They will also notice that the actions they perform are experienced as being under their voluntary control, not automatic.

Finally, the individual will find it unimaginable that such a novel coordination task which demands focused, conscious attention could be performed effectively without continuous, real-time feedback from appropriate depictive representations that 'light up' for them.

Another example that individuals who have a capacity for mindful self-observation can explore is how they navigate an unfamiliar environment strewn with obstacles.

6. Conclusion

Current robotic systems and other AI are not conscious—when they 'look' at an object, an image does not 'light up' in their awareness. They are capable only of blindsight. This limits their ability to exploit the adaptive advantages of conscious functioning which drove the emergence, spread and development of consciousness in biological organisms during the evolution of life on Earth.

Consciousness is all you need

However, the absence of consciousness in current robotic systems and other AI whose information processing is readily modifiable provides major new opportunities for consciousness research. An understanding of minimally-complex information-processing architectures that are capable of producing conscious functioning opens the door for experimentation involving the instantiation of these dynamical architectures in robotic systems.

This experimentation can avoid ethical issues by focusing on the instantiation of simple conscious architectures that only enable conscious experience of depictive representations assembled from visual inputs—as such, it would not generate subjective experience of feelings.

Because conscious functioning can enable adaptive capacities that are far beyond the capabilities of existing AI, the development of conscious AI has the potential to provide superior functionality. This ranges from body-environment coordination in robotic systems that can be achieved with far less training, to the kind of conscious modelling capacities exhibited by humans that are the foundation of Human-like General Intelligence and that enable, for example, abductive science.

But consciousness is not needed only to enable the further development of AI. Higher levels of conscious functioning are also needed by researchers and engineers if they are to contribute effectively to the development of conscious AI. They will need to develop capacities such as meta-awareness, mindful self-observation, and metasystemic cognition. This will enable them to observe and understand their own conscious functioning and to enhance it recursively. These capacities will also assist their understanding of how similar information processing capabilities might be instantiated in AI.

In particular, meta-aware self-observation will give individuals access to relevant data that is not accessible to those without this capacity. Individuals who do not develop such a capacity will be as handicapped as individuals at the time of Galileo who wanted to confirm and understand his astronomical observations, but who did not have access to a telescope.

7. Acknowledgements

I have had the benefit of stimulating discussions about these issues with Clement Vidal, Wilson Kennel, William Davis, and Jake Perry.

8. References

- Azzopardi, P., & Cowey, A., (1998). Blindsight and visual awareness, *Consciousness and Cognition*, 7: 292-311.
- Bhaskar, R., (1975). *A Realist Theory of Science*. Verso, London.
- Clark, A., (2015). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press, U.K.

Consciousness is all you need

- Ding, J., Zhang, Y., Shang, Y., Zhang, Y., Zong, Z., Feng, J., Yuan, Y., Su, H., Li, N., Sukiennik, N., Xu, F., & Li, Y., (2025). Understanding World or Predicting Future? A Comprehensive Survey of World Models, *ACM Computing Surveys*, 58: 1-38.
- Doeller, C. F., Barry, C., & Burgess, N., (2010). Evidence for grid cells in a human memory network, *Nature*, 463: 657-661.
- Feng, T., Wang, W., & Yang, Y., (2025). A Survey of World Models for Autonomous Driving, *arXiv*, arXiv:2501.11260.
- Friston, K., (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.*, 11: 127-138.
- Generalist AI., (2026). GEN-1: Scaling embodied foundation models to mastery, Generalist AI Blog, April 2. <https://generalistai.com/blog/apr-02-2026-GEN-1>
- Gray, K., Gray, H. M., & Wegner, D. M., (2007). Dimensions of mind perception. *Science*, 315: 619.
- Haun, A., & Tononi, G., (2019). Why does space feel the way it does? Towards a principled account of spatial experience, *Entropy*, 21: 1160.
- Humphrey, N., (2023). *Sentience: The Invention of Consciousness*. MIT Press, Cambridge, MA.
- Jacobs, J., & Lee, S., (2016). Spatial Cognition: Grid Cells Support Imagined Navigation, *Current Biology*, 26: R277-R279.
- Jaynes, J. (1976). *The origin of consciousness in the breakdown of the bicameral mind*. Houghton Mifflin, Boston, MA.
- Kegan, R., 1982. *The evolving self: problems and process in human development*. Harvard University Press, Cambridge, MA.
- Kegan, R., 1994. *In over our heads: The mental demands of modern life*. Harvard University Press, Cambridge, MA.
- Kunda, M., (2018). Visual mental imagery: A view from artificial intelligence, *Cortex*, 105: 155-172.
- Kuhn, R. L., (2024). A landscape of consciousness: Toward a taxonomy of explanations and implications, *Progress in Biophysics and Molecular Biology*, 190: 28-169.
- Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62: 54-60.
- Pezzulo, G., (2011). Grounding Procedural and Declarative Knowledge in Sensorimotor Anticipation, *Mind & Language*, 26: 78-114.

Consciousness is all you need

Stewart, J. E., (2007). The future evolution of consciousness. *Journal of Consciousness Studies*, 14: 58-92.

Stewart, J. E., (2022). The Evolution and Development of Consciousness: The Subject-Object Emergence Hypothesis, *Biosystems*, 217: 104687.

Stewart, J. E., (2024). *Human Superintelligence: How you can develop it using recursive self-improvement*. The Glenroy Press, Melbourne.

Stewart, J. E., (2026). The Meaning of Life in a universe whose ultimate origins are unknown, *Biosystems*, 262: 105733

Tang, C., Abbatematteo, B., Hu, J., Chandra, R., Martín-Martín, R., & Stone, P., (2025). Deep Reinforcement Learning for Robotics: A Survey of Real-World Successes. *Annual Review Control, Robotics, and Autonomous Systems*, 8: 153-188.

Yue, X., Zhang, Y., Chen, J., Zhou, X., & He, M., (2024). LiDAR-based SLAM for robotic mapping: state of the art and new frontiers, *Industrial Robot*, 51: 196–205.

Zhang, P. F., Cheng, Y., Sun, X., Wang, S., Li, F., Zhu, L., & Shen, H. T., (2025). A Step Toward World Models: A Survey on Robotic Manipulation, *arXiv*, arXiv:2511.02097.