A Characterization of Processing Loops in AI and Biological Systems and its Implications for Understanding Consciousness

Malcolm J. Lett[1]

[1]No affiliation

*** Correspondence:**
Malcolm Lett
malcolm.lett@gmail.com

Keywords: Access Consciousness, Phenomenal Consciousness, Biosemiotic Process, Visceral Loop, Monitoring and Control, Recurrent Process, Non-physical Action, Conscious Content

Abstract

The claim is made that recurrent cycles of non-physical "mental" actions are required in agents that operate within complex environments, and that in order to maintain stability such actions require regulation through the use of a model. The *Visceral Loop* is proposed as a novel framework for characterizing three distinct kinds of processing within such a system in terms of how the system uses that model. It is shown how this can be used to characterize human thought, including about thought itself, and to explain the semiotic process when an individual concludes that they are conscious. A proof is given relating an upper bound on data available within access consciousness to the Visceral Loop characterizations of thought.

1 Introduction

Any computational system is limited in the complexity that it can handle within a single computational step. For embodied agents, this appears as a limit on the environmental complexity that they can sufficiently model and respond to within a single observe-infer-act cycle. For more complex problems, multiple iterations of processing are required in order to determine the next physical action. Such recurrency in processing may entail, for example, further analysis of the environment in order to better model its state; or consideration of alternative action plans.

In biology, this provides scope for evolutionary pressures to trade off between a more energy hungry complex brain and a simpler less energy intensive one that takes longer to make some decisions. Van Bergen and Kriegeskorte (2020) make the case that recurrency is indeed employed in biology for that very reason. A growing body of research in artificial intelligence is also now employing recurrency and is showing that complex results can be achieved with shallower networks when using recurrency, for example that of Kubilius et al. (2019) and Wen et al. (2018).

This paper makes the claim that an agent that employs *multi-step processing* (ie: multiple cycles of processing without producing physical action) also must employ a model of its own processing capabilities in order to regulate its *non-physical actions*. Different agentive architectures support different abilities for the agent to introspect the internal structure of that model. In this paper the *Visceral Loop*, is offered as novel framework for characterizing an agent's ability to introspect those models and then use them for drawing inferences about itself.

The descriptive power of the Visceral Loop will be illustrated through three human-centric examples: i) how an individual may reach the conclusion that they are conscious, ii) providing an upper bound on the data content of consciousness, and iii) providing a new interpretation of neurobiological studies suggesting that reported awareness occurs after the act of decision making.

In the rest of this paper, the Regulation and Model sections elaborate further the need for regulatory models within any biological or artificial agent. The Schemas section suggests how this is manifested within humans in the form of a hypothesized *mind schema*. The Visceral Loop section presents the core thesis of this paper in the form of a mathematical formalism over the inferences that may be drawn from different kinds of model. The Consciousness section examines how the Visceral Loop can be applied to understanding aspects of consciousness through the presentation of the aforementioned humans. Final thoughts are presented in the Summary section.

2 Regulation

An autonomous embodied agent, depending on its purpose, may need to control either its environment, itself, or both, towards some static or dynamically determined target. That agent can be described as a *regulator* of its target system.

For example, an agent that regulates its environment operates within a system containing environment state s_{env} which changes with some ambient dynamics $d_{env}(t)$. The agent must perform an action, a_{env}, against the environment in order to regulate itself towards some target. After an action has been executed the environment state outcome o_{env} is influenced by both $d_{env}(t)$ and a_{env}. This can be summarized as the following equation:

$$s_{env} + d_{env}(t) + a_{env} = o_{env}$$

According to the *good regulator theorem*, if the agent is to regulate the environment state it must be a "model of the system" (Conant & Ashby, 1970). Furthermore, we can say that the efficiency of the agent to regulate its environment depends on its accuracy in modeling the system. Errors in the accuracy of the model result in errors in the regulation of the system. In learning agents, those errors are used for subsequent training of the model.

An embodied agent with complex actions may require an additional level of regulation. For example, an animal must not only regulate its external environment, but also regulate its own physical state. This includes both maintaining homeostasis and controlling the efficiency and effectiveness of its actions against the environment. The agent performs action a_{body} against its body with the intent to regulate the body towards efficiently achieving environment action a_{env} while satisfying its requirement for body homeostasis. Such an agent thus operates in a system that additionally has body state s_{body} with ambient dynamics $d_{body}(t)$. The agent performs action a_{body} against its body, producing outcome o_{body}, summarized as follows:

$$s_{body} + d_{body}(t) + a_{body} = o_{body}$$

Agents that incorporate multi-step processing have a third kind of action: one that changes its internal data state without affecting its physical state. This system requires regulation for the same reasons as for environment and body, but such *non-physical* actions do not elicit any change to s_{body} or s_{env}. Thus the agent must regulate its non-physical state s_{mind}, having ambient dynamics $d_{mind}(t)$. The target state in this case is dynamically inferred based on its requirement for environment action a_{env}, body action

a_{body}, and possibly for some form of non-physical homeostasis of s_{mind}. In order to regulate towards that target it performs action a_{mind} producing outcome o_{mind}, summarized as follows:

$$s_{mind}+d_{mind}(t)+a_{mind}=o_{mind}$$

By way of example of the importance of such mind regulation, consider the case of fluent aphasia caused by damage to the Wernicke's area[1] of the brain. Individuals with fluent aphasia can easily produce speech, but it is typically full of many meaningless words and often unnecessarily long winded. Wernicke's area is associated with language comprehension. In a neurotypical individual, the comprehension of their own vocalizations provides a corrective mechanism. This illustrates the importance of feedback in the regulation of one's own actions, and by way of analogy extends to the regulation of non-physical actions.

3 Models

All of the systems described above are of the form $s+d(t)+a=o$. The production of the optimal action *a* for a given situation can be computed by a function, *f*, such that $a=f(s,o,t)=o-s-d(t)$. In this way, function *f* becomes a *model* of the system in exactly the way meant by Conant and Ashbey. There are many different ways of constructing such a function, with implications on how much its inherent model can be introspected for purposes other than merely computing the next action.

Consider the following function. This function is, for example, effective at predicting the action required to regulate towards a target state of 3 by doubling the input signal and comparing to that target state. However, an agent that merely uses this function to calculate actions cannot inspect anything about the function other than the actions it calculates for different inputs.

$$f(x)=3-2x$$

Alternatively consider Figure 1, which shows an abstract syntax tree[2] (AST) of the function above, of the sort used by computer science to parse an expression within a software compiler. Instead of using the above function, a regulating agent could use this AST to calculate its next action and achieve the same outcome. However the AST is a more explicit model of the dynamics being regulated. The components of the original function are represented individually and thus they can be individually queried. So here the AST can be introspected and much more can be derived from it that may apply either to the system being modeled or to how the AST models that system.

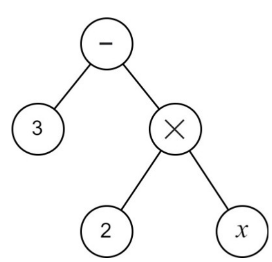

Figure 1: Abstract syntax tree of 3 - 2x

[1]https://en.wikipedia.org/wiki/Wernicke%27s_area

[2]https://en.wikipedia.org/wiki/Abstract_syntax_tree

To examine the introspective opportunities further, consider the task of constructing a set, *F*, that contains all beliefs that may be drawn from the model. In the case of the function, pairs of input and output action values are all that can be drawn from the model, ie: <0,3>, <2,-1>, <-3,9>, <-1,5>, etc. The AST supports the ability to draw those same pairs of input and action values. However the AST also supports that many other beliefs may be drawn from the model and added to *F*. For example that i) the target is 3, ii) input signal *x* is significant to the calculation, while *y* and *z* are not, and iii) the execution of the function depends on the operations of *subtraction* and *multiplication*.

So, it is clear that different architectures enable different levels of *introspection* of the underlying models. What about the case for neural networks? In the modern use of artificial neural networks (ANNs), it is commonplace to refer to ANNs as a *function approximator* (Goodfellow et al., 2016), and indeed many networks fall into the category of a function. For example, in *model-free* deep reinforcement learning (RL) an ANN is used to calculate either the next action or the expected value of all possible actions given the current state (Lazaridis, Fachantidis & Vlahavas, 2020). The architecture of the RL algorithm treats the ANN as a function without any introspective capabilities. See Figure 2(A) for an example. There is also *model-based* RL. One variant of model-based RL, illustrated in Figure 2(C), uses ANNs to predict the expected outcome of executing an action. The introspective ability here is the same as for model-free deep RL - the ANN is treated as a function. For the RL models mentioned so far, the set *F* of beliefs is of similar content: *F* is the set of <state,action> or <state,action,outcome> tuples. There do exist forms of model-based RL that use something more akin to the AST, usually where there is a known physics model that is represented mathematically, and which could potentially be used to introspect for more than just <state,action,outcome> tuples, such as is illustrated in Figure 2(B). However, a significant point to note here is that ANNs, and probably neural networks in general, do not lend themselves to introspection on their own.

This is a provisional file, not the final typeset article

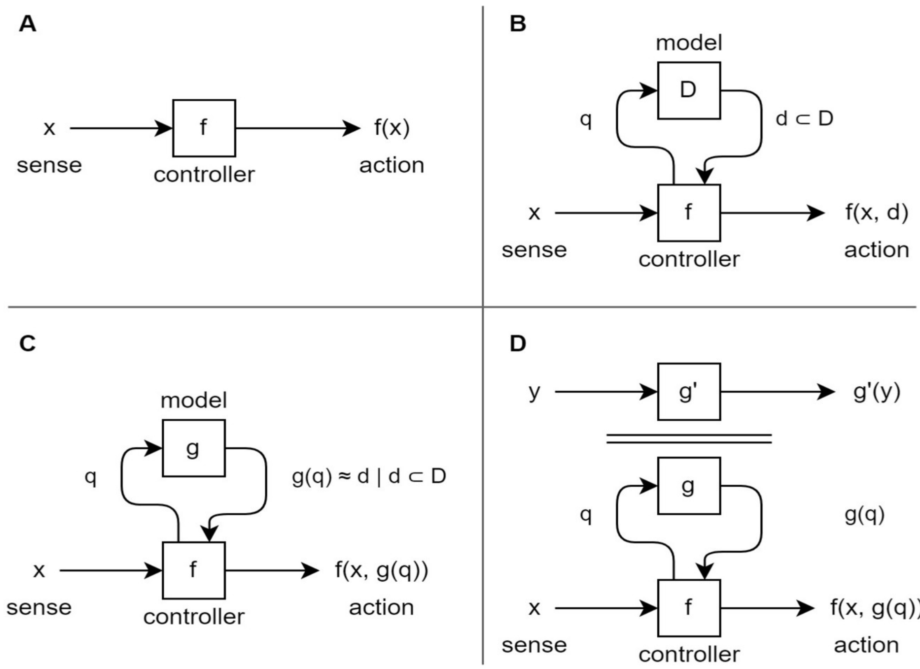

Figure 2: Different architectures for modeling regulatory actions against the environment. In **(A)**, the controller determines the next action by executing function *f* against sense input *x*. The function may, for example, be an ANN that is trained through many iterations of an appropriate learning algorithm. Function *f* merely models the best next action without modeling any other aspects of that environment and thus cannot be used to introspect anything other than the next regulatory action. In (**B**), set *D* holds an explicit model of the environment which can be arbitrarily queried (*q*) to gain insight about any aspect of the environment that is encapsulated within *D*. Controller function *f* uses that to determine the next action. In (**C**), set *D* is replaced by function *g(q)* which approximates queries against *D*. This architecture is commonly used in AI where the dynamics of the environment are too complex to determine a priori, and *g(q)* is built as a second ANN that is trained through exploration. In (**D**), the secondary system $y = g'(y)$ models some aspect of the environment other than the next regulatory action. For example, it may observe and predict long term trends in the environment state. Potentially further additional modeling systems are required for each additional aspect of the environment that needs to be modeled.

For that reason, a third form of modeling system exists, whereby a secondary model predicts the behaviors of the former, such as is illustrated in Figure 2(D). The secondary model may, for example, be a second ANN that captures aspects of the same underlying system but at a more macro level, and it may be more suitable for integration with other data. This macro representation is at the core of the theory of Higher Order Thought Theory (HOTT) (Rosenthal, 1997 & 2006), and of recent theories based on it such as Hierarchical Active Inference (Giovanni et al, 2018) and Integrated World Modeling Theory (IWMT) (Safron, 2020).

4 Schemas

The lack of introspective ability of a simple function contrasts with the introspective ability of a human. Psychology has long identified in humans the existence of a model of the individual's body – known as the *body schema*. A good definition is given by Morasso et al (2015):

> *"In summary, we view the body schema as a set of fronto-parietal networks that integrate information originating from regions of the body and external space in a way, which is functionally relevant to specific actions performed by different body parts. As such, the body schema is a representation of the body's spatial properties, including the length of limbs and limb segments, their arrangement, the configuration of the segments in space, and the shape of the body surface".*

So the body schema is a model used in production of action control by integrating information from our main physical senses and the proprioceptive senses (Proske & Gandevia, 2012). That model can also be introspected – for example, we can know where our hands and feet are without seeing them – and those introspections can become the topic of subsequent thought. But there are aspects of the model that cannot be introspected – for example, we have no observability of the arrangement of the sense nerves used to infer the hand and feet positions, or of the effector nerves used to actuate their muscles.

This paper hypothesizes the existence of a second schema, the *mind schema*, that performs an analogous role for the regulation of the mind and non-physical actions. Anecdotally, this seems highly plausible within humans given our introspective ability towards our own mind's capabilities. For example, we can know that we are good at focusing, but struggle with math, that we are more creative when background music is present, and that we need the support of tools to help remember people's names (eg: a notebook). The underlying notion here is that the mind schema helps us to control, monitor, predict, and rationalize about our mental structure and actions in the same way that our body schema does that for our physical structure and actions. It is the regulatory model for our non-physical actions. Additionally, just as for the body schema, there is a distinct delineation between what can be introspected and subsequently thought about, and what cannot.

The suggestion of a mind schema has also been made in the form of *Attention Schema Theory* (Graziano & Kastner, 2011; Webb & Graziano, 2015; Graziano, 2017), although the meaning there is perhaps narrower than what is proposed in this paper.

5 Visceral Loop

This paper introduces the *Visceral Loop* as a novel framework for the characterization of inferences drawn by a processing loop within a biological or AI agent. The Visceral Loop is so named because it refers to an agent concluding that it experiences consciousness in a visceral way. It identifies that an agent with sufficient representational capabilities can, at the most optimum, conclude itself as conscious within three iterations of the processing loop. Each of those iterations have specific characteristics, and the Visceral Loop can be used to characterize any thought as falling into one of those three iterations.

Let:

- E be the agent's set of beliefs about the external world

- B be the agent's set of beliefs about its own physical body (drawn from the body schema) and of bodies in general, and if it has a concept of its identity then this set includes a belief that relates other body beliefs to its identity

- M be the agent's set of beliefs about its own mind (drawn from the mind schema) and of minds in general, and if it has a concept of its identity then this set includes a belief that relates other mind beliefs to its identity

- *f(..)* be the function executed by the agent on the specified inputs in order to draw inferences

M can be thought of as an agent's "theory of mind", because it relates not only to itself but also to its ability to predict the hidden mental state of others.

5.1 Iteration 1

Iteration 1 represents the most common kind of data processing, such as spending multiple processing cycles to refine the identification of something within the visual field. While an agent's mind schema may be used to regulate the thought process, the result of Iteration 1 never makes any reference to it.

Let x be an inference produced as the result of a processing step, such that it does not draw any reference to M (ie: $x \notin M$, and if x is a relation then $x = relation(\alpha,\beta)$ such that $\alpha \notin M$ and $\beta \notin M$ and $\alpha \not\subset M$ and $\beta \not\subset M$). Given some sense input or past state s, a processing step is characterized as Visceral Loop Iteration 1 if it is of the following form:

$$f(s, E \cup B \cup M) \rightarrow x$$

5.2 Iteration 2

Iteration 2 processing steps draw conclusions that relate past non-physical actions and conclusions to the agent's theory of mind and to the agent's concept of its identity. For example, concluding that a past data state or non-physical action is classified as "thought", concluding whether the primary source of a past data state was external or internal, or relating the fact of an internal source to the agent's concept of its identity.

Iteration 2 requires an agent to have sufficient representational capabilities to produce inferences that represent relations involving M. Given some prior inference y, a processing step is characterized as Visceral Loop Iteration 2 if it is of the following form, and the relation with respect to M is non-empty, and it can not be characterized as Iteration 3:

$$f(y, E \cup B \cup M) \rightarrow relation(y, M)$$

5.3 Iteration 3

Iteration 3 is a special case of what would otherwise be Iteration 2, but it implies stricter requirements on the agent's introspective and representational capabilities. Iteration 3 covers the ability for the agent to develop a summary of its own mental capabilities (ie: some subset $m \subset M$), and to consider that in relation to its conception of mental capabilities in general or to its identity (ie: M). Iteration 3 is involved in an agent concluding itself as conscious, as will be seen in the section below.

Given some prior inference *relation(z, M)*, and some subset of beliefs $m \subset M$, a processing step is characterized as Visceral Loop Iteration 3 if it is of the following form and the relation with respect to M is non-empty:

$$f\left(relation\left(z, M\right), E \cup B \cup M\right) \rightarrow relation\left(m, M\right)$$

6 Consciousness

The Visceral Loop has implications for understanding consciousness; three examples of which will be detailed shortly. For example, based on the Visceral Loop one can make an argument that associates the behaviors of the underlying modeling mechanism to limits on the ability for self-reference and to the ability for an individual to recognize their own consciousness. However, it is necessary first to define what is meant here by the term *consciousness*.

Firstly, as so little is known of the nature of consciousness within non-human animals or in artificial agents, the discussion will be limited to that of humans. Within that constraint, this paper focuses on an individual's ability to have subjective experience of its observations (eg: of the external environment), to be able to rationalize (a.k.a. "think") about things, and to have subjective experience of those rationalizations. That domain of consciousness is typically split into two closely incident but distinct forms:

- *Access consciousness* (A-Cs) refers to state representations that are "(1) inferentially promiscuous, that is, poised for use as a premise in reasoning, (2) poised for rational control of action, and (3) poised for rational control of speech." (Block, 1995). In other words, A-Cs is the data content that has a direct influence on subsequent thought. For example, generally the data content of access consciousness can be directly consciously thought about and reported on via speech or other actions.

- *Phenomenal consciousness* (P-Cs), in contrast, refers to subjective experience and to the question of why a physical bundle of matter should have such a thing. It is often described as "what it is like to be" (Nagel, 1974) a conscious organism. In approximate terms, P-Cs is the phenomenal aspect of having a subjective experience of the data content of A-Cs.

The discussions in this paper refer to the data content of both A-Cs and P-Cs, but in general does not draw an explicit distinction between A-Cs versus P-Cs, under the assumption that there is a single shared physical mechanism underlying them both. Here, *data content* is used merely to draw a distinction between what is and what is not present within A-Cs or P-Cs. It is not intended to imply any particular representational structure. For example, whether or not someone can see and report on something within their visual field determines part of the data content of A-Cs, and whether or not someone phenomenally experiences pain determines part of the data content of P-Cs.

6.1 Concluding oneself as conscious

In this first example, the Visceral Loop is applied to understand the thought processes whereby an individual concludes themselves as conscious.

Consider the following sequence of internal mental observations:

1. "What's that red blob in the tree? Oh, it's an apple".

2. "Those thoughts just came from my mind, and not from the outside world".
3. "That's what consciousness is. I am conscious".

The first observation is a straightforward example of Visceral Loop Iteration 1 that does not make any reference to the agent's theory of mind (of their own mind or of others). With reference to the formal definition of the Visceral Loop in the prior chapter, the concepts of "red", "blob", "tree" and "apple" are all contained within the set E, and thus the inference in relation to the visual field sense input s is of the form $x_1 = relation(s, E)$.

The second observation contains two examples of Iteration 2 inferences. In the first, the individual's processing capabilities have selected attentional focus upon the prior Iteration 1 inference, and have drawn a subsequent inference about it as being data that can be classified as a "thought". As beliefs about "thought" are contained within M, this is an inference of the form $x_2 = relation(x_1, M)$. In the second, the individual draws a subsequent inference about the source of the Iteration 1 inference as being their own mind. The individual's ability to classify inferences in relation to themselves also depends upon M, and the inference is of the form $x_3 = relation(x_1, M)$.

The third observation draws upon the individual having an a priori conception about consciousness in general, denoted here by $m_c \subset M$. The individual compares its prior Iteration 2 inferences x_2 and x_3 to m_c, and produces an inference that x_2 and x_3 together satisfy the requirements for consciousness. This is another iteration 2 inference of the form $x_4 = relation(x_2 \wedge x_3, m_c)$. Finally, the individual relates m_c, the belief of consciousness in general, to itself, which again depends on M. That final inference is thus an Iteration 3 inference in the form $x_5 = relation(m_c, M)$.

6.2 Content of conscious thought

As a second example of the descriptive power of the Visceral Loop, a theorem is now presented that makes the claim that the Visceral Loop explains the data content of conscious experience. An exact distinction of how it applies to A-Cs versus P-Cs is omitted here, with some discussion of its issues included in a later section.

First an axiomatic baseline must be established. It seems reasonable to expect that there is no way in which an individual may consciously experience something and yet be unable to subsequently think about that experience and to know that they are thinking about that experience. Indeed, this is consistent with the *transitivity principle* of Rosenthal (1997). Thus, it would seem that being able to knowingly think about our conscious experiences is a fundamental component of consciousness. The following claims are derived from this statement without further proof:

Claim 1:

- All conscious experience is subsequently available for further thought.

Claim 2:

- For all thought about conscious experience, the individual can identify that thought as being their own.

Note that these claims do not assume that all conscious experience is subsequently thought about; only that it is in principle available for such thought.

Theorem 1:

- the data content of conscious experience is upper bounded by the data about which Visceral Loop iteration 2 inferences can be produced.

Proof:

- Thought is a computational process, and thus is a series of inferences.
- As per claim 1, all of conscious experience must be available for producing subsequent inferences about those conscious experiences.
- As per claim 2, the individual must be able to identify that they produced those inferences.
- In order for an individual to identify an inference as being their own, they must have some beliefs about their inference capabilities and how they relate to themselves as an individual entity. This is included in the set *M*, which iteration 2 produces inferences in relation to, and which is not directly accessible for inferences within iteration 1.
- Imagine some supposed experience, and an inference *i* produced about that experience. Additionally imagine that an iteration 2 inference cannot be produced about *i*, for example, due to some incompatibility of structure, lack of data path to iteration 2 processing capabilities, or inherent limitation in iteration 2 processing capabilities. The inference *i* cannot be identified in relation to the individual. As such, the supposed experience fails on Claim 2 and *i* must be in actual fact an inference about some sort of non-conscious experience.
- Thus, an experience is not a conscious experience if it can only lead to inferences which cannot be included in an iteration 2 inference.

6.3 Delayed awareness of decisions

The Visceral Loop can be used to understand other aspects of thought. For example, fRMI and EEG studies have suggested that we become aware of a decision after it is made (Soon et al, 2008; Libet et al, 1983). At first glance this might seem to suggest that our conscious thought is non-causative, instead being just some sort of after-the-fact passive summarization, such as is claimed by the theory of Epiphenomenalism[3].

The framework of the Visceral Loop provides a different interpretation. It explains that the act of making a decision and the conscious consideration of having made that decision occur in different processing cycles. First one or more processing cycles are used to reach the decision and to trigger the resultant action. Subsequently the individual may use one or more additional processing cycles to examine their most recent inference. So it is entirely predictable that the individual would activate brain regions for reporting the decision after a measurable delay from the decision itself being made and a resultant action initiated. Importantly, the same underlying system can produce both sets of

[3] https://plato.stanford.edu/entries/epiphenomenalism

processing cycles, and thus there is no reason to conclude from fRMI or EEG delays that consciousness is non-causal.

In short, we can only think about one thing at a time, so the decision itself and thought about the decision require separate processing steps.

7 Analysis

The examples above show that it is possible to mathematically reason about thought processes, their sequencing, and what can and cannot be thought about depending on the capabilities of the underlying thought processing system. This section examines the Visceral Loop against some existing theories and points out some existing weaknesses.

7.1 Visceral Loop as a Biosemiotic Process

The Visceral Loop views thought as a *biosemiotic process*[4]. Semiotics is the study of signs and their interpretations, and biosemiotics looks at how semiotics plays out within biology, including neural cognition. A semiotic process has three components: a *referent*, the object for which a sign or representation will be made; a *representamen*, which is a representation of the referent, a.k.a. a "sign"; and a *interpretant,* the interpretation made from the representamen. The intepretant may or may not accurately reflect the original referent, depending on the quality of representamen and on the ability of the system (ie: the intepreter) to infer information about the referent from the representamen. For example, in Figure 1 the referent of someone liking something is converted into a thumbs-up representamen, but it has multiple possible interpretants depending on the context and background of the interpreter.

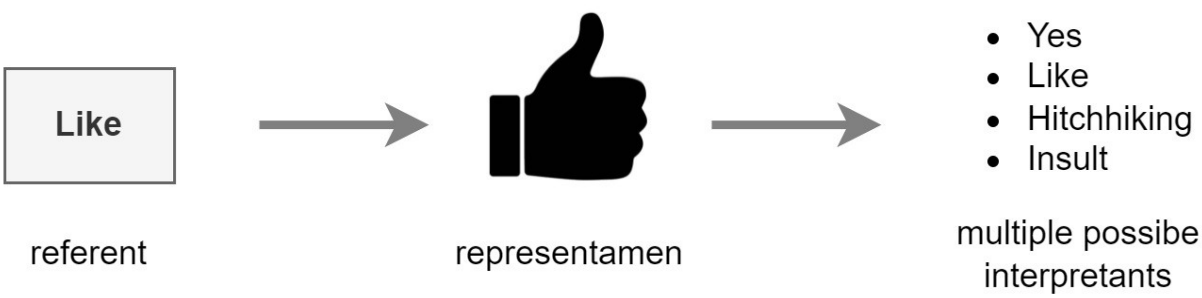

Figure 3: An example semiotic process. Someone wishes to indicate that they like something - their liking of that something is the *referent* that they wish to convey through a sign. The sign they use, the *representamen,* is the icon of a hand with thumb pointing up. The representamen is then interpreted by someone else, forming the *intepretant* within the mind of that second individual. Depending on the context and the cultural background of the second individual, they may form one of multiple possible interpretants (eg: a thumbs up gesture is considered an insult in some cultures).

From an external point of view, the brain acts as a semiotic process against the environment. As illustrated in Figure 4, the environment is the referent, with a concrete and actual state that the brain does not have direct awareness of. Instead, the brain uses the physical senses as a representamen of that environment, and from that representation produces an interpretant.

[4] https://en.wikipedia.org/wiki/Biosemiotics

The Visceral Loop is interested in a more internal semiotic process within the brain. As illustrated in Figure 5, the interpretant from before must itself be encoded in a representation. And that representation is subject to the constraints imposed by the specific characteristics of the underlying biological substrate. In order for the brain to subsequently use that representation, it must decode its meaning. Thus an inner semiotic process is activated, where the prior inference becomes the referent, which is encoded as a representamen, and subsequently decoded as a new interpretant, possibly after combining with additional information from other sources.

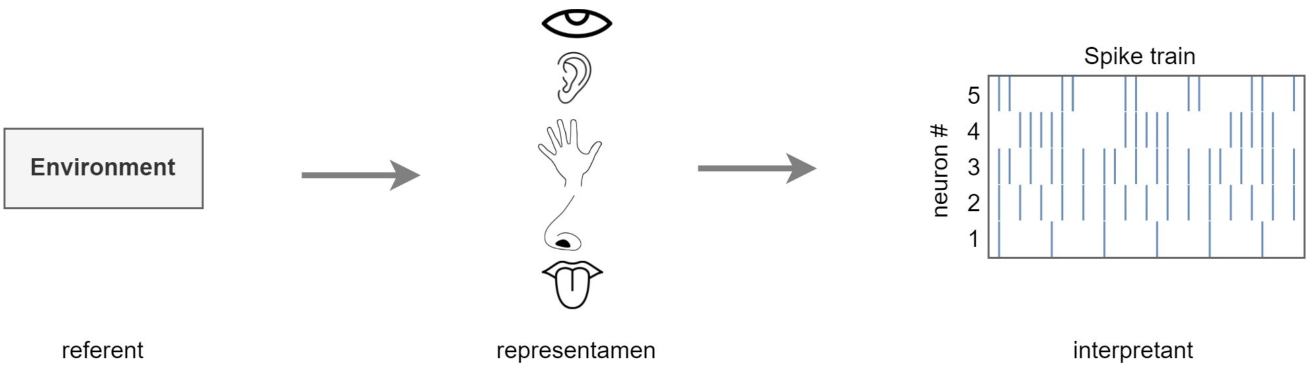

Figure 4: Semiotics of environment interpretation

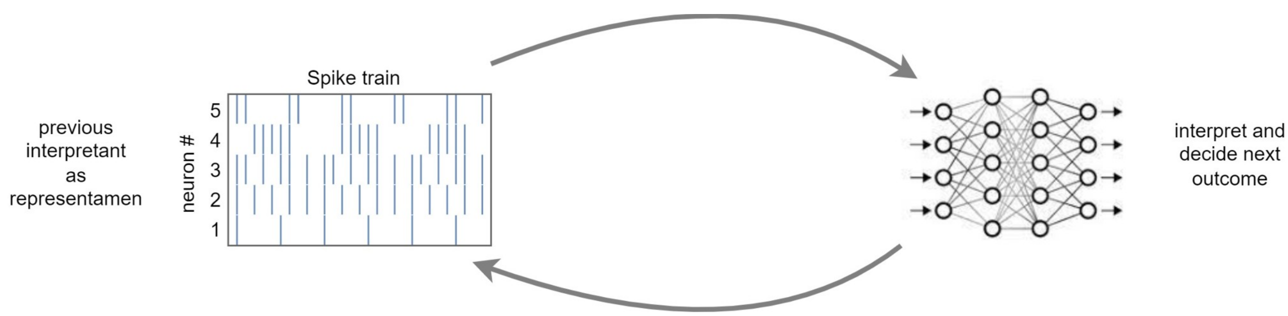

Figure 5: Semiotics of internal thought

The semiotic interpretation of cognition is important because it explains that a data state is meaningless without a process that interprets that data state; even if the process generated the data state in the first place, that data state is still meaningless to the process without re-interpreting it. Thus, while the brain must include considerable machinery for the interpretation of external stimuli, it must also include considerable machinery for the interpretation of its own outputs. The Visceral Loop builds on this by describing the kinds of interpretations that may be generated depending on the what information is available within the representamen and on the capabilities of the process.

7.2 Relationship to Higher Order Thought Theory and Global Workspace Theory

The Visceral Loop and the biosemiotic process that it operates upon examine thought at a high-level, but the Visceral Loop does not presuppose a particular physical or computational structure. In that sense it is applicable to many theories of cognition and brain function. However, it is easiest to see its consistency with other theories that also examine thought at a high-level or that provide theories

This is a provisional file, not the final typeset article

about the underlying mechanisms of such high-level thought. For example, those of Higher-Order Thought Theory (HOTT), and Global Workspace Theory (GWT).

HOTT describes mental processes as hierarchical (Rosenthal 1997; Rosenthal 2006). In HOTT, consciousness forms the highest layer of the hierarchy, and has access to the output from the immediate layer below, but not of further lower layers. Thus the information available for conscious thought is a unified high level abstract representation formed from the output of many sub-processes. The Visceral Loop does not require a HOTT architecture, but it explains how a HOTT architecture results in the specific nature of human experience in terms of their being a distinction between those internal processes which are directly observable, versus those which are not.

The Visceral Loop can also be seen as a monitoring and control device that governs the actions of the rest of the agent's cognitive function. This suggests why a higher-order layer is requisite. The internal processes of the entire brain are too complex to monitor at their native low-level neuronal states. The global monitoring and control function of the brain needs to know only the broad trends in behavior in order to predict whether the existing behavior is suitable for the larger objective. Thus it is more efficient that it operates over a higher-order unified representation that contains only the information that is pertinent for the purpose of self-governing.

Conveniently, this finally explains why we are conscious of certain processes within our own brains and not conscious of others – that our brains create a higher-order representation for the purpose of self-governing, and only that self-governing function has the full semiotic process capable of producing conscious experience, and thus we are only conscious of the data content that is presented to that self-governing process.

7.3 A step towards understanding General Intelligence

The problem of solving Artificial General Intelligence (AGI) has received renewed interest of late (Adams et al., 2012; van Gerven, 2017; Cervantes et al., 2021). The Visceral Loop may provide some insight.

Firstly, an AGI operates within a complex environment, has complex responses, and thus it needs to incorporate recurrent non-physical actions, as per the arguments presented within this paper. This implies the need for the agent to introspect and monitor its own mental behaviors in order to control them for stability and efficiency. That monitoring and control operates as a semiotic process against the agent's own internal mental representations. Thus an AGI should be expected to exhibit the characteristics that are measured by the Visceral Loop, and one aspect of the level of advancement of the AGI can be measured by whether it can execute Visceral Loop Iterations 2 and 3.

This suggests a particular AGI architecture whereby a discrete sub-component provides monitoring and control and governs the actions of multiple other sub-components, each with individual specializations, such as is illustrated in Figure 6. Were those sub-components to operate against each other through competitive processes alone, they may be at risk of chaotic behavior and runaway instabilities. Provided with an appropriate learning algorithm and an appropriate objective that measures the agents overall stability and efficiency, an aggregate system that incorporates a global monitoring and control function can efficiently learn to govern itself and to avoid chaotic runaway instabilities. For example, one such learning algorithm and objective is that of Active Inference (Friston, Daunizeau, and Kiebel, 2009; Tschantz, Baltieri, et al, 2020; Sajid, Ball, et al, 2021).

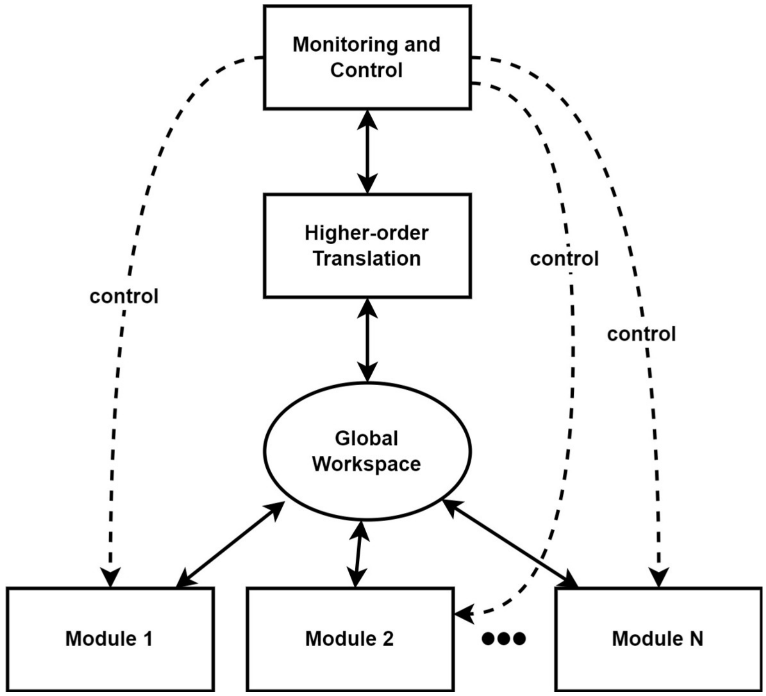

Figure 6: A potential AGI architecture. A Monitoring and Control unit governs the execution of individual specialized modules. Based on discussions within Section 7.2, the monitoring and control unit views the behaviors of the individual modules through a higher-order translation, and applies control either directly or through reverse higher-order translation. A global workspace model is naively assumed for sharing of information between modules.

7.4 A step towards understanding Consciousness

Within section 6 the discussion was restricted to human consciousness, but the Visceral Loop provides a mathematical basis for understanding the potential for consciousness in other agents, including artificial ones. Where an agent has a processing architecture sufficient for producing Visceral Loop Iteration 2 inferences, then it has the potential for conscious thought.

The recurrent nature of the semiotic process underlying the Visceral Loop explains the need for a computational architecture that incorporates i) recurrency at the high-level scale, ii) persistence of state between recurrent cycles, and iii) representational integration of diverse computational outcomes.

Another similar theory that attempts to mathematically reason about levels of consciousness is Integrated Information Theory (IIT) (Tononi, 2004; Oizumi et al., 2014). IIT provides a mechanism for quantifying the amount of recurrent integration in a system, symbolized in its amount of Φ. Importantly, IIT is applicable to both biological and artificial agents. The Visceral Loop is thus compatible with and enhances IIT, in that it gives a higher-level reason for the need for recurrency.

This is a provisional file, not the final typeset article

Phenomenal experience is generally considered to be the hallmark of consciousness, making the distinction between an agent that is truly conscious versus a philosophical zombie of the same. The theory presented here offers some insight into the mechanism behind the ability for an individual to examine their own thought. This is assumed to be closely related to the content of the individual's phenomenal experience. Thus, the Visceral Loop is an important aspect of consciousness, and understanding of it may help lead towards artificial consciousness.

7.5 Limitations

The Visceral Loop offers significant insight into consciousness, but with the introduction of any new theory there are limitations and areas which may need improvement. It is thus worth briefly discussing some of the limitations with the Visceral Loop as stated, and with the examples of its use given above.

Theory 1, presented above, is weakened somewhat by the fact that it avoids clarifying whether it applies to A-Cs, P-Cs, or both. Unfortunately such an analysis is hampered due to presently irreconcilable differences in opinion about the relative natures and definitions of A-Cs and P-Cs, and about whether A-Cs or P-Cs can exist without the other[5]. For example, some authors make the claim that P-Cs can exist without A-Cs (Block, 1995; Armstrong, 1995). This would imply some kind of representational content within P-Cs that is independent of A-Cs and thus cannot be subsequently thought about (ie: violating Claim 1). There is also research suggesting that some A-Cs information can influence our subsequent thought without us having P-Cs of it, such as in the example of blind-sight (Block, 1995).

In both cases there remains heated debate with no obvious end in sight. The best we can say is that A-Cs and P-Cs largely correlate, although there may be exceptions. The author believes that stronger claims can yet be made about the natures of both A-Cs and P-Cs in terms of the Visceral Loop, but that shall remain the topic of a separate investigation.

The theories presented in this paper make no claim over the so called "hard problem of consciousness" – the question of why we have phenomenal experience at all (Chalmers, 1995). That remains an open question. The definition of the Visceral Loop itself is compatible with many of the existing arguments that attempt to address the hard problem, although Theory 1 is challenged by some such arguments.

One area in which the Visceral Loop does not suffer from a common limitation is that it avoids any risk of a recursive definition of conscious thought. Based on the Visceral Loop, a tentative definition of consciousness might be that:

> We are conscious of a perception or thought if that perception or thought is in principle capable of being subsequently thought about via Visceral Loop Iterations 2 or 3, and we have subjective experience of it at the moment that an Iteration 2 or 3 result is produced.

An important point here is that the perception or thought does not have to be immediately processed by Iteration 2 or 3; it can pass via memory and be processed by Iteration 2 or 3 at a later moment. In fact, anything that is accessible from memory and which is capable of being processed through

[5]For a good example of the issues with defining A-Cs and P-Cs, see the responses to Block (1995), accessible at https://www.nyu.edu/gsas/dept/philo/faculty/block/papers/1995_Function.pdf

Iteration 2 or 3 (subject to the availability of the necessary data pathways) can be subjectively experienced.

8 Summary

This paper makes the claim that agents operating within complex environments and with complex bodies require a trade-off between the inferential computing power of a single processing step versus the use of multiple iterations of a recurrent processing loop. Processing loop iterations that do not lead to physical actions thus have non-physical actions, and in order to be a good regulator of those non-physical actions, the agent must model its non-physical behaviors.

It has been shown that some architectures for non-physical behavioral modeling enable introspective abilities, such as that enjoyed by humans. The Visceral Loop provides a mechanism to mathematically reason about those introspective abilities, and to classify the capabilities of different architectures. Where Integrated Information Theory (IIT) provides a mechanism for quantifying the amount of recurrent integration in a system, the theory behind the Visceral Loop explains why that recurrency is important in the first place. The Visceral Loop provides a mechanism for qualifying the introspective capabilities in a way that is simultaneously at a coarser grain and simpler to calculate than what IIT targets.

The Visceral Loop provides insight into consciousness in general. It shows that the content of consciousness is defined by a biosemiotic process that operates against a higher-order summary of the internal processes of the rest of the brain, and that it is necessary for stable functioning of a complex cognitive system. It also suggests a pathway towards building an Artificial General Intelligence. It suggests the introduction of a self-monitoring and control process that governs the execution of multiple specialized modules which.

9 Author Contributions

The author confirms being the sole contributor of this work and has approved it for publication.

10 Funding

No funding was received.

11 References

Adams, S., Arel, I., Bach, J., Coop, R., Furlan, R., Goertzel, B., et al. (2012). Mapping the Landscape of Human-Level Artificial General Intelligence. AI Magazine, 33:1, 25-42, doi:10.1609/aimag.v33i1.2322.

Armstrong, D. M. (1995). Perception-consciousness and action-consciousness? Behav. Brain Sci., 18:2, 247-248, doi:10.1017/S0140525X0003819X.

Block, N. (1995). On a confusion about a function of consciousness. Behav. Brain Sci., 18:2, 227–247. doi:10.1017/S0140525X00038188.

Cervantes, J. P., Martin, L., Dounce, I. A., Avila-Contreras, C., and Corchado, F. F. (2021). Methodological aspects for cognitive architectures construction: a study and proposal. Art. Intel. Rev., 54, 2133-2192. doi:10.1007/S10462-020-09901-X.

Chalmers, D. J. (1995). Facing up to the problem of consciousness. J. Conscious. Stud., 2:3, 200-19. doi:10.1093/acprof:oso/9780195311105.003.0001.

Colagrosso, M. D., and Mozer, M. C. (2004). Theories Of Access Consciousness. NeurNIPS 2004.

Conant, R. C., and Ashby, W. R. (1970). Every good regulator of a system must be a model of that system. Int. J. Systems Sci., 1:2, 89-97. doi:10.1080/00207727008920220.

Giovanni, P., Rigoli, F., and Friston, K. J. (2018). Hierarchical Active Inference: A Theory of Motivated Control. Trends in Cognitive Sciences, 2:4, pp 294-306. doi:10.1016/j.tics.2018.01.009.

Friston, K. J., Daunizeau, J., Kiebel, S. J. (2009). Reinforcement Learning or Active Inference? PLOS ONE, 4:7, doi:10.1371/journal.pone.0006421.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). Deep learning. MIT Press.

Graziano, M. S. A., and Kastner, S. (2011). Human consciousness and its relationship to social neuroscience: a novel hypothesis. Cogn. Neurosci. 2, 98–113. doi:10.1080/17588928.2011.565121

Graziano, M. S. A. (2017). The Attention Schema Theory: A Foundation for Engineering Artificial Consciousnes. Front. Robot. AI. doi:10.3389/frobt.2017.00060.

Kubilius, J., Schrimpf, M., Kar, K., Rajalingham, R., Hong, H., Majaj, N., Issa, E., et al. (2019). Brain-like object recognition with high-performing shallow recurrent anns. NeurIPS 2019, 12805-12816.

Lazaridis, A., Fachantidis, A., and Vlahavas, I. (2020). Deep Reinforcement Learning: A State-of-the-Art Walkthrough. J. Artif. Intell. Res., 69, pp 1421-1471. doi:10.1613/jair.1.12412.

Libet, B., Gleason, C. A., Wright, E. W., and Pearl, D. K. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). The unconscious initiation of a freely voluntary act. Brain: a journal of neurology, 106:3, 623–642. doi:10.1093/brain/106.3.623.

Morasso, P., Casadio, M., Mohan, V., Rea, F., and Zenzeri, J. (2015). Revisiting the body-schema concept in the context of whole-body postural-focal dynamics. Front. Hum. Neurosci. 9:83. doi:10.3389/fnhum.2015.00083.

Nagel, T. (1974). What Is It Like to Be a Bat?. The Philosophical Review, 83:4, 435–450. doi:10.2307/2183914.

Oizumi, M., Albantakis, L., and Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: integrated information theory 3.0. PLoS Comput. Biol. 10:e1003588. doi:10.1371/journal.pcbi.1003588.

Proske, U., and Gandevia, S. C. (2012). The Proprioceptive Senses: Their Roles in Signaling Body Shape, Body Position and Movement, and Muscle Force. Physiol. Rev., 92:4, 1651-1697. doi:10.1152/physrev.00048.2011.

Rosenthal, D. M. (1997). "A Theory of Consciousness," in The Nature of Consciousness: Philosophical Debates, ed. N. Block, O. Flanagan, and G. Güzeldere (Cambridge, MA: MIT Press/Bradford Books), 729-753.

Rosenthal, D. M. (2006). "Consciousness and Higher-Order Thought," in Encyclopedia of Cognitive Science, ed. L. Nadel, doi:10.1002/0470018860.s00149.

Safron A. (2020). An Integrated World Modeling Theory (IWMT) of Consciousness: Combining Integrated Information and Global Neuronal Workspace Theories With the Free Energy Principle and Active Inference Framework; Toward Solving the Hard Problem and Characterizing Agentic Causation. Front. Artif. Intell., 3:30, doi:10.3389/frai.2020.00030.

Sajid, N., Ball, P. J., Parr, T., Friston, K. J. (2021). Active Inference: Demystified and Compared. Neural. Comput., 33:3, 674–712, doi:10.1162/neco_a_01357.

Seth, A. K., Suzuki, K., and Critchley, H. D. (2012). An interoceptive predictive coding model of conscious presence. Front. Psychol., 2:395. doi:10.3389/fpsyg.2011.00395.

Soon, C. S., Brass, M., Heinze, H. J., & Haynes, J. D. (2008). Unconscious determinants of free decisions in the human brain. Nat. Neurosci., 11:5, 543–545, doi:10.1038/nn.2112.

Tschantz, A., Baltieri, M., Seth, A., and Buckley, C. (2020). Scaling Active Inference. IJCNN, 1-8, doi: 10.1109/IJCNN48605.2020.9207382.

Tononi, G. (2004). An Information Integration Theory of Consciousness. BMC Neurosci., 5:42, doi:10.1186/1471-2202-5-42.

van Bergen, R. S., and Kriegeskorte, N. (2020). Going in circles is the way forward: the role of recurrence in visual inference, Curr. Opin. Neurobiol., 65, 176-193, doi:10.1016/j.conb.2020.11.009.

van Gerven, M. A. (2017). Computational Foundations of Natural Intelligence. Front. Comp. Neurosci., 11. doi:10.3389/fncom.2017.00112.

Webb, T. W., and Graziano, M. S. A. (2015). The attention schema theory: a mechanistic account of subjective awareness. Front. Psychol. 6:500. doi:10.3389/fpsyg.2015.00500.

Wen, H., Han, K., Shi, J., Zhang, Y., Culurciello, E. and Liu, Z. (2018). Deep Predictive Coding Network for Object Recognition. Proceedings of the 35th International Conference on Machine Learning, in Proc. Mach. Learn. Res., 80, 5266-5275

This is a provisional file, not the final typeset article