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

Malcolm J. Lett1

1No affiliation

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

Keywords: keyword1, keyword2, keyword3, keyword4, keyword5. (Min.5-Max. 8)

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.

# Introduction

*tbd: should this be rephrased in terms of recurrency?*

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 & 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.

# 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 *Senv* which changes with some ambient dynamics *Denv(t)*. The agent must perform an action, *Aenv*, against the environment in order to regulate itself towards some target. After an action has been executed the environment state outcome *Oenv* is influenced by both *Denv(t)* and Aenv. This can be summarized as the following equation:

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 *Abody* against its body with the intent to regulate the body towards efficiently achieving environment action *Aenv* while satisfying its requirement for body homeostasis. Such an agent thus operates in a system that additionally has body state *Sbody* with ambient dynamics *Dbody(t)*. The agent performs action *Abody* against its body, producing outcome *Obody*, summarized as follows:

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 *Sbody* or *Senv*. Thus the agent must regulate its non-physical state *Smind*, having ambient dynamics *Dmind(t)*. The target state in this case is dynamically inferred based on its requirement for environment action *Aenv*, body action *Abody*, and possibly for some form of non-physical homeostasis of *Smind*. In order to regulate towards that target it performs action *Amind* producing outcome *Omind*, summarized as follows:

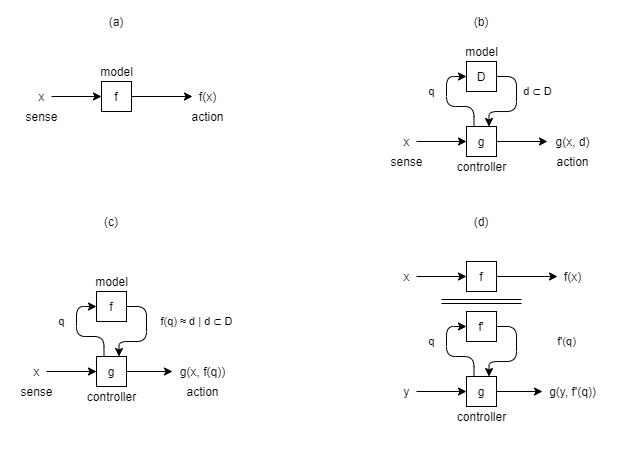
By way of example of the importance of such mind regulation, consider the case of fluent aphasia caused by damage to the Wernicke's area[[1]](#footnote-2) of the brain. Individuals with fluent aphasia can easily produce speech, but it is typically full of many meaningless words and often unnecessarily 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.

# 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.

Alternatively, consider the following diagram, which shows an abstract syntax tree[[2]](#footnote-3) (AST) of the function above, of the sort used in 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: (tbd: needs caption)

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. There is also *model-based* RL. One variant of model-based RL 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. However, a significant point to note here is that ANNs, and neural networks in general, may not lend themselves naturally to introspection on their own.

For that reason, a third form of model exists, whereby a secondary model predicts the behaviors of the former. 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 (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).

# 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:

"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" (2015).

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 kind of 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.

# 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.

## 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* ∉ ℳ, and if *x* is a relation then *x = relation*(α,β) such that α ∉ ℳ and β ∉ ℳ and α ⊄ ℳ and β ⊄ 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:

## 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:

### 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* ⊂ 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* ⊂ 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:

# Author Contributions

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

# Funding

No funding was received.

# References

1. https://en.wikipedia.org/wiki/Wernicke%27s\_area [↑](#footnote-ref-2)
2. https://en.wikipedia.org/wiki/Abstract\_syntax\_tree [↑](#footnote-ref-3)