

# A Proposed Model of Cognition

Janos Tapolczai

April 3, 2014

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Preliminary considerations &amp; justification</b>	<b>2</b>
<b>3</b>	<b>Diagram notation</b>	<b>5</b>
<b>4</b>	<b>Global architecture</b>	<b>5</b>
<b>5</b>	<b>Mathematical model</b>	<b>8</b>
5.1	Messages . . . . .	8
5.1.1	Structural notation . . . . .	9
5.1.2	Operational notation . . . . .	9
5.1.3	Plastic and non-plastic neural systems . . . . .	10
5.2	Invariants . . . . .	10
<b>6</b>	<b>Selected subsystems</b>	<b>11</b>
6.1	Sensory perception . . . . .	11
6.2	Counterfactual perception . . . . .	12
6.3	Affect . . . . .	15
<b>7</b>	<b>Errors in the model</b>	<b>17</b>



Figure 1: Relationship between the components of an organism without a nervous system.

## 1 Introduction

In this document, I will sketch a possible architecture of the human brain and a select few of its subsystems. The descriptions presented are supported by some empirical evidence, but I do not claim that they are straightforward transcriptions of neurological realities. The model is grounded substantially in evolutionary considerations, which provide the backdrop and the plausibility check for the claims presented herein.

Section 2 outlines the basic considerations that lead to the model. Section 4 sketches the proposed model of the mind. Section 6.3 then sketches the affective subsystem and lastly, errors and falsifiability are discussed in Section 7.

It should also be understood that everything in this document is, at best, a *rough* outline; it may be likened to a hexagon which approximates a circle: though (conjectured to be) basically correct, and useful, it is marred by significant incongruities with the object of its approximation.

## 2 Preliminary considerations & justification

In order to understand how our brain works or could work, we must possess conceptual clarity — we must conceive of it, not as a product of engineering, but as a historical artefact, and as one which was not produced “in one step”, but gradually, where each stage of its evolution had to be viable on its own. What, one might ask, is the consequence of such a view? Most importantly, it allows the distinction between GOOD systems and CLEAN systems. Since, at each stage of its evolution, the organism that carried the brain had to be viable, the end product is by definition guaranteed to be “good”. Because of that same fact, however, it is all but guaranteed not to be “clean”: for one, it was not possible to snap whole new components into the system; it would have also been impossible to combine old components in the elaborate and precise ways in which a human engineer might use parts. Worse, old components were almost certainly not discarded when new and better components came into being. A good exposition of this process in humans can be found in Paul MacLean’s seminal work *The Triune Brain in Evolution* [5]

**Origin of nervous systems** Let us imagine a microscopic organism without any sort of nervous system: all of its behaviour is hard-coded and mechanical. It can take in nutrients through its cell walls or through an opening; parts of it can contract or expand in response to stimuli like light or pressure; homeostatic conditions can influence its chemistry. Figure 1 shows this schema: if we enumerate the constituent parts or *components* of an organism is  $\{C_1, \dots, C_n\}$ , the organism’s behavior is caused by signals being sent between  $C_i$  and  $C_j$  (the case  $i = j$  is of course possible). Such an organism suffers from two disadvantages: (a) the behaviour is necessarily simple and (b) it is not very adaptable.



Figure 2: Relationship between the components of an organism possessing a nervous system.  $F$  can be understood as a simple signal transformer or a central coordinating mechanism.

Let us now imagine that such an organism develops a bundle of cells which transmit the signals from various parts of its body, modulate them in some way, and then send them to various parts, inducing changes. Schematically, this is shown in figure 2, where a function  $F$  is interposed between two components. The first such nervous systems were likely little more than signal transformers or magnifiers that expedited communication between parts: with a few neurons, an organism would have had the ability to coordinate movements or rely on sensing parts induce, say, movement.

The neuron bundles would have been quite malleable in the face of selection pressure: when the environment required it, they could, after several generations, start to compute different or more elaborate functions. For instance, an organism which had developed in an environment where food was abundant in bright places and which had now found itself in darkness would have benefited from a variety of plausible changes, such as

- an inversion of its light-seeking behaviour,
- switching off its metabolism in light places to conserve energy,
- accelerating its metabolism in dark places to make better use of the food there.

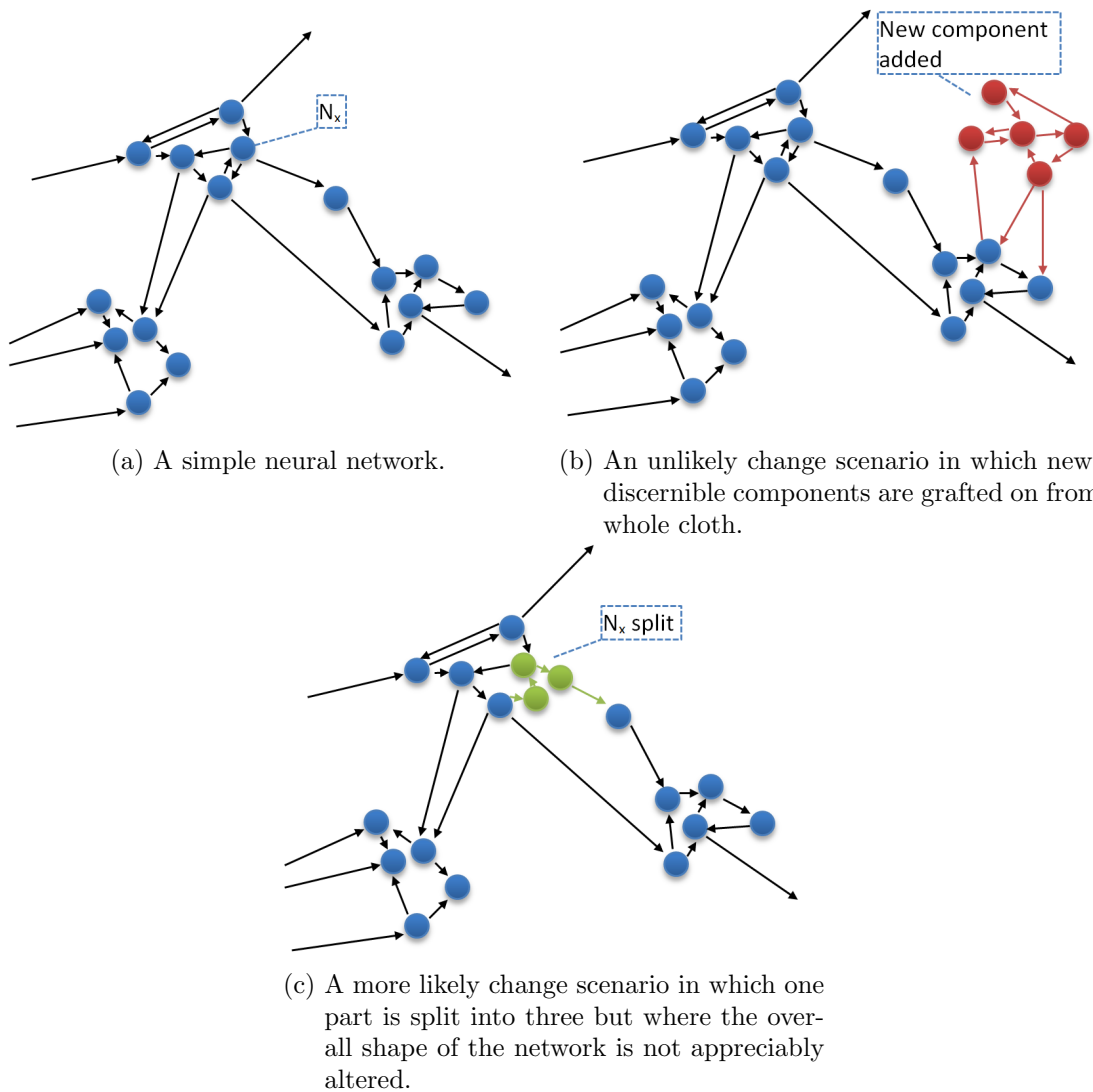
Of course, other changes would have also been possible, such as the metabolization of different food sources<sup>1</sup>, but we can see how the aforementioned three could have been effected through changes in a simple nervous system alone. Let us recall the beginning of this section and contrast such a malleable computational mesh with most products of human engineering: one cannot simply take out a piston in a car or replace a cogwheel in a mechanical clock with a differently sized one. Machines are designed to fit together perfectly and their complexity tends to be irreducible. Even programs, which are more open to mutation and which are often evolved in evolutionary algorithms, are easily broken by small changes.

**Evolution of nervous systems** When discussing how an organism's nervous system can evolve and, in particular, *evolve to perform new tasks* and not just variations on old ones, explanations are again constrained by two criteria: (a) the change has to be small, or at least have a small cause<sup>2</sup> and (b) each change must be beneficial in the short term<sup>3</sup>.

<sup>1</sup>A current-day example is given by nylon-eating bacteria, which have developed in the last century and which now have an abundant food source and no competition.

<sup>2</sup>The effect does not have to be small — changes in singl

<sup>3</sup>Caveats apply: if the selection pressure on a group of organisms isn't too strong, changes which may be sub-optimal but perhaps beneficial at some later point may spread, and non-selective processes like genetic drift can also play a role.



To illustrate this, we can look at a simple neural network in figure 3a, with a marked node  $N_x$ . Figure 3b shows an unlikely change scenario in which some new component/function is cleanly grafted onto the system. Figure 3c then shows a much more likely scenario: a mutation causes  $N_x$  to be split and the new nodes take over some of its connections. In time, new functions can thus grow into the system, but never in the manner in which, say, an engineer would implement a new feature.

**The brain as a collection of functions** components don't just appear... is this in contradiction to the message space?






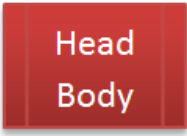

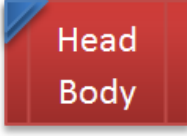
decentralized message space models the subsystems as white boxes, not as black-box functions (black-box functions emulate the white-box subsystems)

**Function abstraction** neural system encodes a function -; we look at the function, not at the underlying substrate computing it

TODO

### 3 Diagram notation

In the rest of this document, a number of diagrams appear. These will use the following notation:

Symbol	Description
 	..... Processing component
 	..... Choice
	..... Data container (Queue, List, etc.)
	..... Data
	..... Stream generator
	..... Counterfactual (imaginary) data

### 4 Global architecture

We can imagine the components of the mind as white boxes which inform other components by their very functioning — however, this does not lend itself to easy implementation. Instead, we can emulate this behaviour via a MESSAGE SPACE, from which individual components take their input and into which they put their output. A COMPONENT is then a local processing unit which continuously scans the message space, running messages through its FILTER. If the filter detects a relevant message, it is then passed to the INTERPRETER, which parses the message into the needed format and hands it over to the PROCESSOR. The processor, after having finished, puts its output back into the message space for other other components to read. Figure 4

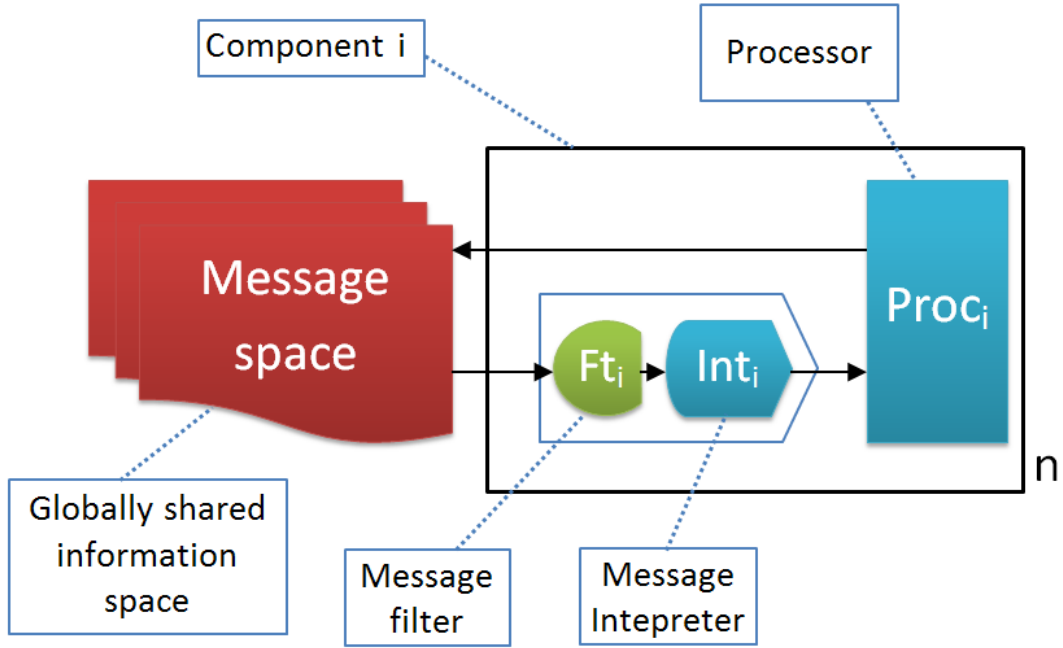


Figure 4: Global neural architecture.

illustrates this scheme. Note the lack of explicit hierarchical structure and central organising units.

However, as we'll show in the next section, this model is generic enough to accommodate such special-purpose structures. Figure 4 sketches a kind of graph in which the nodes are the components and the edges are accepted messages; the edges are the components and through their filters, they control the shape of the graph. By imposing invariants on these filters, we can have the graph take any shape we desire. In particular, we can model the kinds of structures that occur in other cognitive models and in empirical research: central organisers, sequences ("pipelines") of components, localized messages affecting only a small part of the mind, et cetera.

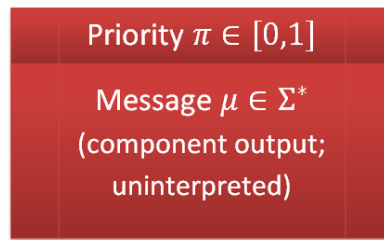


Figure 5: Structure of a neural message.

**Messages** We may now ask how such messages between components are structured. Here, I make two empirical claims:

1. messages have a priority and

2. they are effectively unstructured.

The my knowledge, the veracity of either has thus far not been determined by neuroscience. For the first, Marvin Minsky’s “The Emotion Machine” provides some circumstantial evidence [6, p. 222]:

Of course, when one activates two or more Critics or Selectors, this is likely to cause some conflicts, because two different resources might try to turn on a third resource both *on* and *off*. To deal with this, we could design the system to use various policies like these:

1. Choose the resource with the highest priority.
2. Choose the one that is most strongly aroused.
3. Choose the one that gives the most specific advice.
4. Have them all compete in some “marketplace”.

The selection strategies Minsky lists imply that there is some mechanism in the brain to determine the urgency of a signal. While it is possible that higher brain functions like reasoning or affect make an additional, rational evaluation, sensations like intense pain, bright lights, or great sadness can likely be communicated most easily by the appropriate components causing a flood of activity which, by its very intensity, informs other components of the urgency of their messages.

The second claim — that messages are essentially unstructured — means that there is no common, agreed-upon format in which they are stored. In addition to the evolutionary implausibility of such a format being created, an unstructured message format is in line with the white-box nature of components: since components merely “listen in” on others, and since each components will have its own pattern of activity, a listener would simply have to try and make sense of this activity as best it could. The proposed structure of messages is thus shown in figure 5: every message comprises a priority header, together with an unstructured body which, for our purposes, is simply a string of bits.

**Filters** Before a component can respond to a message by another, such a message must be assessed for the presence of relevant information. Conceptually, this happens via a FILTER in each component, which pattern-matches incoming messages and, if a certain threshold is reached, signals relevance and hands the message over the INTERPRETER for parsing. Figure 6 shows such a filter: it is composed of a directed graph of nodes, and a node is activated if it detects some specific content in the message. Nodes, in turn, are connected via edges of strength  $\in [0, 1]$ . When a node is activated, it sends a charge proportional to the strength of its link to its neighbours, contributing to their activation as well. Some nodes are marked as *output nodes*; if enough such output nodes become activated, the message is deemed to be sufficiently relevant. This model of filters is inspired by the *spiking neural P Systems* of Georghe Paun et al. ([7, p. 337] and [3]), in which charges sent along directed graphs of neurons are used to compute functions.

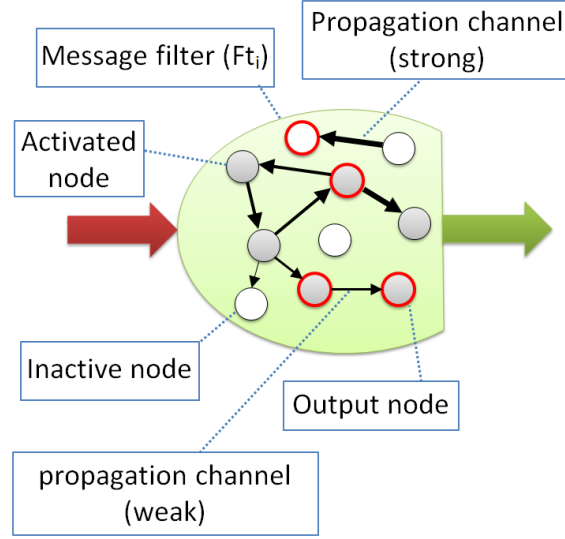


Figure 6: A pattern-matching filter for a component  $C_i$ .

## 5 Mathematical model

We now create a mathematical model for the description of the architecture of section 4. This model will be split into two parts: the structural and the operational semantics. The structural semantics encode the static properties of neural systems, whereas the operational semantics describe the behaviour of such a system at runtime.

**Definition 1.** A neural system is a tuple  $\langle \mathbf{Co} : \text{Set}[C], M : \text{Set}[T], I \rangle$ , where

$$C = \langle \text{name} : I, \\ \text{ft} : T \rightarrow \mathbb{B}, \\ \text{int} : T \rightarrow \text{Maybe } T, \\ \text{proc} : T \rightarrow T \rangle.$$

$\mathbf{Co}$  is the set of components,  $M$  is the set of messages (with type  $T$ ) and  $I$  is an index set by which elements of  $\mathbf{Co}$  are indexed.  $C$  is the type of a component, consisting of a NAME, a FILTER, an INTERPRETER and a PROCESSOR.

The filter, interpreter, and processor of a component  $C$  are denoted by

$$\begin{aligned} &Ft_C, \\ &Int_C, \\ &Proc_C. \end{aligned}$$

### 5.1 Messages

We now give a notation for the sending and receiving of messages in a system. Here, we distinguish two aspects: first, the structural, which describes how messages *can* travel in a system and the operational, which describes how the *do* travel in some given scenario.



### 5.1.1 Structural notation

When a component  $C$  can, in principle, output every message in  $\{m_1, \dots, m_n\}$ , we write

$$C \succ \{m_1, \dots, m_n\}. \quad (1)$$

Analogously, when a component  $C$  can, in principle, receive all messages in  $\{m_1, \dots, m_n\}$ , we write

$$\{m_1, \dots, m_n\} \succ C \quad (2)$$

The opposite statement — that a component  $C$  cannot receive any message in  $\{m_1, \dots, m_n\}$  — is denoted by

$$\{m_1, \dots, m_n\} \not\succ C \quad (3)$$

The set of components which can receive a message  $m$  is denoted by

$$\text{Rec}(m) \equiv \{C \in \mathbf{Co} \mid \{m\} \succ C\}. \quad (4)$$

$\text{Rec}$  can also be overloaded to refer to the set of components which can receive & interpret at least some message of a component  $C$ :

$$\text{Rec}(C) \equiv \{C_i \in \mathbf{Co} \mid \exists m : C \succ \{m\} \wedge \{m\} \succ C_i\} \quad (5)$$

### 5.1.2 Operational notation

When a component  $C_i$  outputs a message  $m_{out}$  that another component  $C_j$  receives and interprets as message  $m_{in}$ , we write

$$C_i \rightarrow [m_{out}, m_{in}] \rightarrow C_j. \quad (6)$$

If it's clear that the message  $m$  doesn't change, we just write

$$C_i \rightarrow [m] \rightarrow C_j. \quad (7)$$

Components can be mutated by messages they receive or output<sup>4</sup>. When a component  $C$  is changed into  $f(C)$  by a message  $m$  it receives, or changed into  $f'(C)$  by a message  $m'$  it sends, we write, respectively:

$$\dots \rightarrow [m] \rightarrow \langle f \rangle C \quad (8)$$

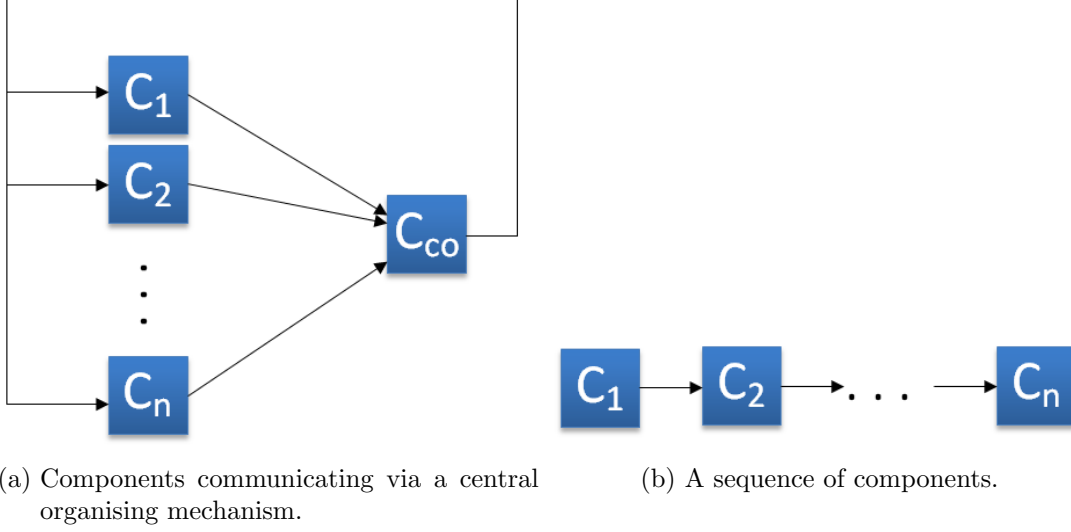
$$C \langle f' \rangle \rightarrow [m'] \rightarrow \dots \quad (9)$$

Lastly, as a shorthand, function application can be denoted by  $\$$ , which has lower precedence than any other operator:

$$f \$ x = f(x). \quad (10)$$

---

<sup>4</sup>This concept corresponds, roughly, to the idea of neuroplasticity.



### 5.1.3 Plastic and non-plastic neural systems

**Definition 2.** When, for all messages  $m_{out}, m_{in}$  and all components  $C_i$  and  $C_j$ , the following holds

$$C_i \langle \rangle \rightarrow [m_{out}, m_{in}] \rightarrow \langle \rangle C_j.$$

we call the system non-plastic. Otherwise, the system is called plastic.

## 5.2 Invariants

Such a model does not necessitate the existence of special structures, such as central organizers or sequences of components, one activated after another<sup>5</sup>, but it does not preclude them either. In fact, we can enforce certain features via first-order invariants. For example, a central organizing units for the components  $C_1, \dots, C_n$  can be emulated by a component  $C_{co}$  which accepts messages and transforms them into an appropriate format for the some other components.

**Invariant 1** (Central organiser).

$$[\forall i \in \{1 \dots, n\}][\forall m] : (C_i \mapsto \{m\} \Rightarrow Rec(m) = \{C_{co}\}) \wedge \left( (Proc_{C_{co}} \circ Int_{C_{co}} \$ m) \in \bigcup_{1 \leq j \leq n} Rec(C_j) \right)$$

Figure 7a depicts such an organizer. Similarly, sequences can be created by components  $C_1, \dots, C_n$ , where each components reads the message of the last one.

**Invariant 2** (Sequence).

$$[\forall i \in \{2 \dots, n\}] Rec(C_{i-1}) = \{C_i\}$$

<sup>5</sup>An example of such a sequence is found in [8], where the authors model the emotion process as a four-step pipeline of relevance, implication, coping and normative significance.

## 6 Selected subsystems

The global architecture now specified, we will introduce three related subsystems and fit them into this global framework: sensory perception — the processing of raw sensory input into a format intelligible to other brain components —, counterfactual perception — the imagination, which mimics the output of the senses —, and affect — broadly speaking, the emotional component of cognition.

### 6.1 Sensory perception

The model presented herein is inspired by Marvin Minsky’s “The Emotion Machine”. Therein, Minsky proposes a layered mental structure where each successive layer operated on more and more abstract representations of the world, starting with primitive sensations and proceeding all the way to self-conscious reflection and rational planning. Figure 8 shows such a layered structure.

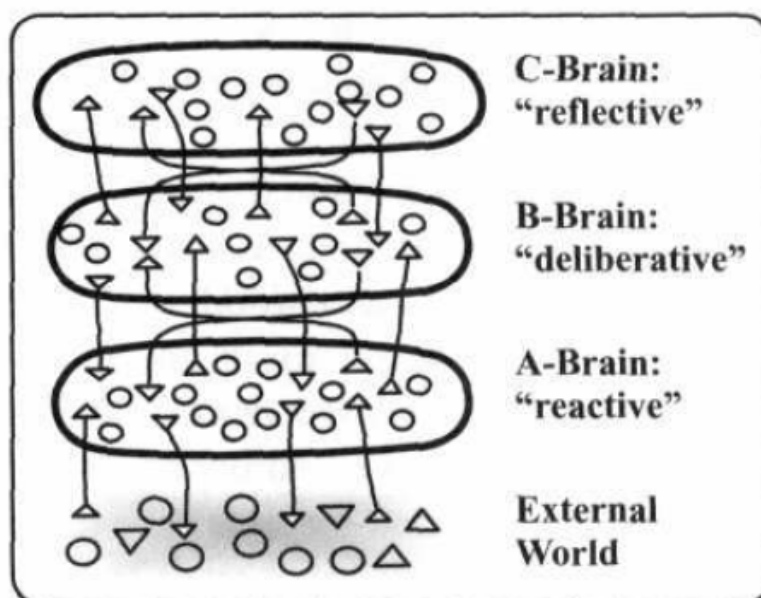


Figure 8: Layered perception of the world, from [6, p. 100].

The diagram is explained in this paragraph [6, p. 100]:

Now suppose that your A-Brain gets some signals from the external world (via such organs as eyes, ears, nose, and skin) — and that it also can react to these by sending signals that make your muscles move. By itself, the A-Brain is a separate animal that only reacts to external events but has no sense of what they might mean. For example, when the fingertips of two lovers come into intimate physical contact, *the resulting sensations, by themselves, have no particular implications*. For there is no significance in those signals themselves: their meanings to those lovers *lie in how they prepresent and process them in the higher levels of their minds*.

If we apply this to the architecture of section 4, we can devise a system in which each sense  $S$  has an associated component  $C_S$  which does two things:

1. Consume the raw sensory information delivered by various organs and output processed input for higher brain functions;
2. as a side effect of this processing, cause instinctive, low-level reactions in the body, such as pulling away from pain or jumping at a sudden fright.

In figure 9, a slice of just such a system is shown for visual, auditory, olfactory/gustatory and tactile sensation. The perceivable data is of two kinds: one is more abstract facilitates deliberative action, and the other contains instructions for instinctive behaviour for the body.

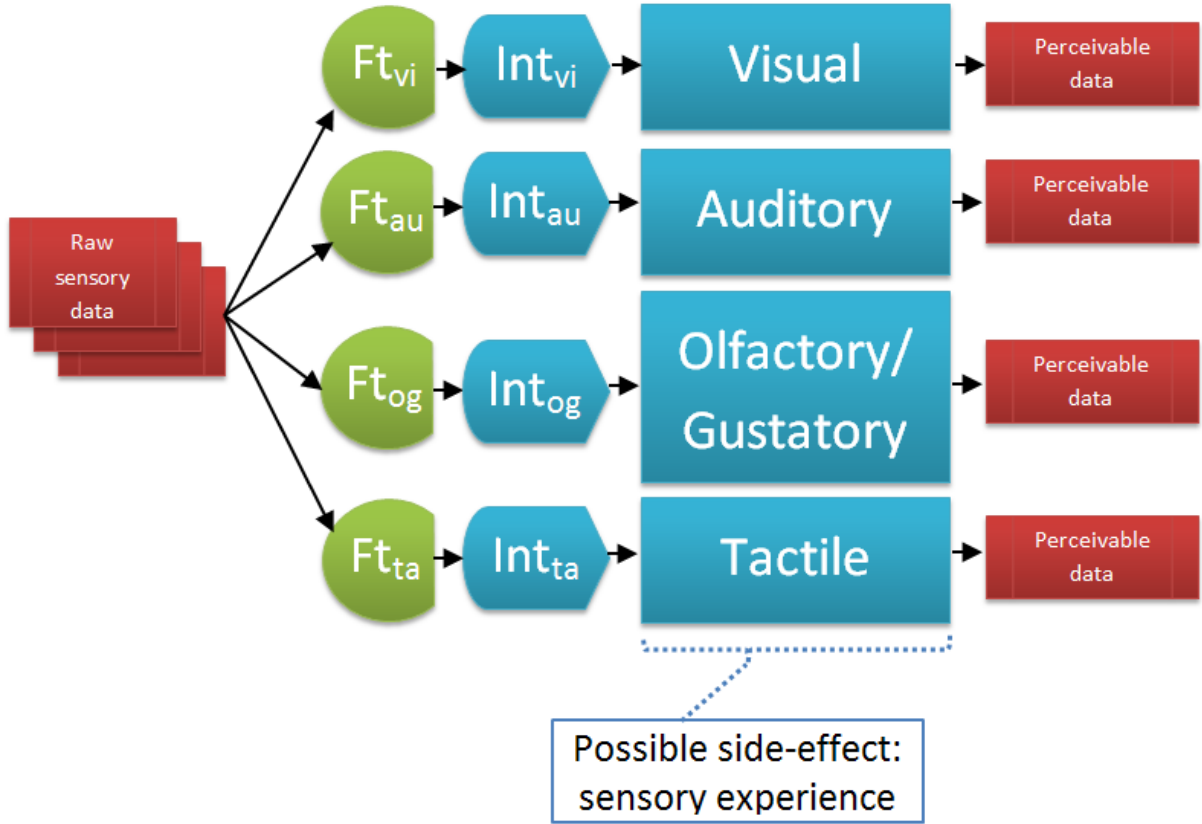


Figure 9: Partial structure of sensory perception - raw sensory data is processed and made available to higher functions such as the affective subsystem. The comment “Possible side-effect: sensory experience” signifies the fact that conscious and sub-conscious sensory experiences might occur as a side-effect of this processing. Whether this is indeed the case is unknown, however.

## 6.2 Counterfactual perception

Broadly speaking, counterfactual perception can be described as “imagination”, and is closely related to sensory perception and world simulation. In examining the system, we might broadly classify its processes into three categories:

1. Counterfactual perception — imagining sights, sounds, etc. Such experiences have much in common with those caused by our sensory organs, yet are marked not as real. In particular, imagined experiences evoke only parts of the conscious experience that accompanies real perceptions. Research by Berthoz and Lotze et al. suggests that (a) the brain indeed uses similar circuitry for real and imagined experiences and that (b) imagined experiences are prevented from being confused with real ones via inhibitory signals. Lotze et al. write [4]:

The results of cortical activity support the hypothesis that motor imagery and motor performance possess similar neural substrates. The differential activation in the cerebellum during EM and IM is in accordance with the assumption that the posterior cerebellum is involved in the inhibition of movement execution during imagination.

From the abstract of Berthoz’s paper [1]:

(...) experimental evidence suggesting that the brain can use the same mechanisms for the imagination and the execution of movement. In particular the fact that adaptation of the vestibulo-ocular reflex can be obtained by pure mental effort and not solely by conflicting visual and vestibular cues has been suggestive of the fact that the brain could internally simulate conflicts and use the same adaptive mechanisms used when actual sensory cues were in conflict.

2. World simulation — the imagination of future states. Simulating worlds goes beyond the imagination of sensory experiences; it involves constructing models of worlds and simulating their behaviour. The details of this process are unknown, but we can assert that it is capable of a number of things:
  - a) construction of non-physical worlds, such as mathematical models,
  - b) extrapolation into the future and the past
  - c) simulation of the minds itself and other agents.
3. Executive planning — humans can plan both both in immediate & concrete terms (such as body movement) and in the abstract. It is likely that different circuitry is used for movement planning and for planning involving abstract reasoning, both in both cases, it is necessary that the brain simulate the world in some way. The simulation of the consequences of body movement is likely older than humanity and distinct from the kind of world simulation described above, but both share their function: the agent proposes as series of actions to take, inserts them into some mental world and judges the utility of those actions based on the predicted consequences.

Needless to say, that this process in all its subtleties is immensely complex and thus, I simply endeavour to sketch its possible structure only in extremely rough outlines. This sketch is shown in figures 10, ??, and 12: the world simulation is an ordinary component with a filter and interpreter which outputs, for simplicity’s sake, messages marked as counterfactual. We can imagine such messages to be very much like ordinary sensory ones, with the exceptions that they have no accompanying sensation and, more importantly, that we are aware of their non-reality. The planning component receives instructions about desirable states and outputs hypothetical actions which the world simulator incorporates. The world simulator’s output is in turn read by the planner, which then abandons the plan or decides to pursue it further.

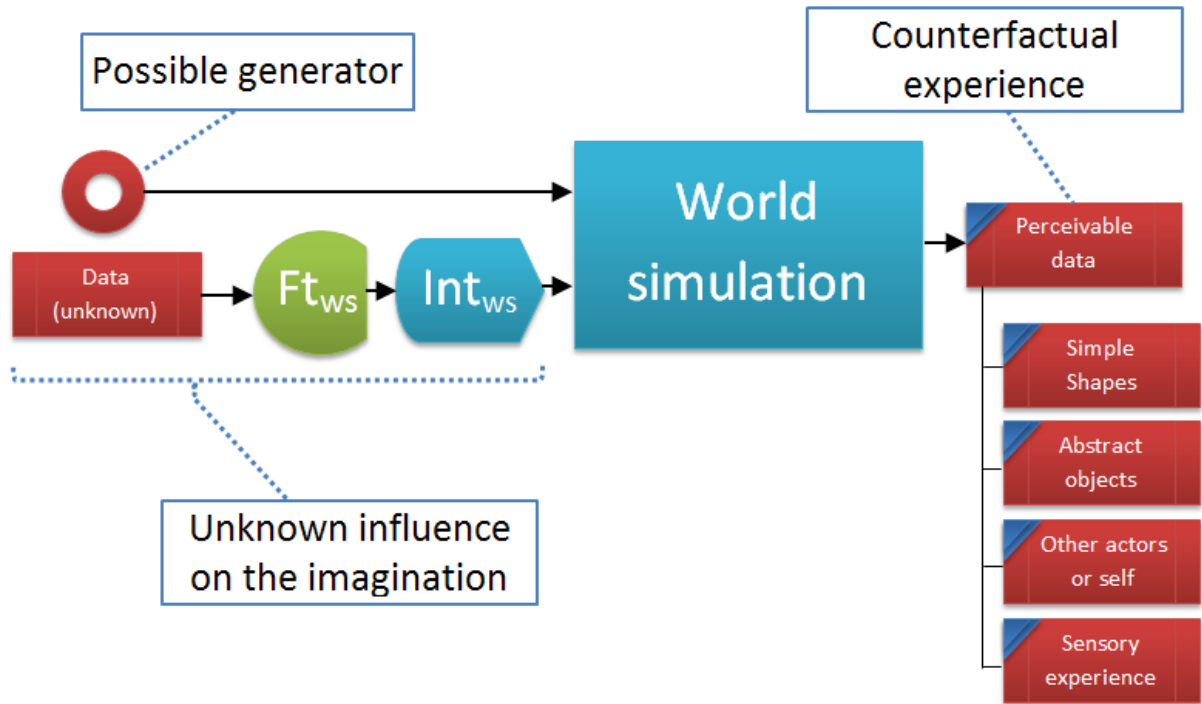


Figure 10: Structure of of counterfactual perception & world simulation: messages emulating the output of sensory perception are generated, but are marked as counterfactual by unknown means.

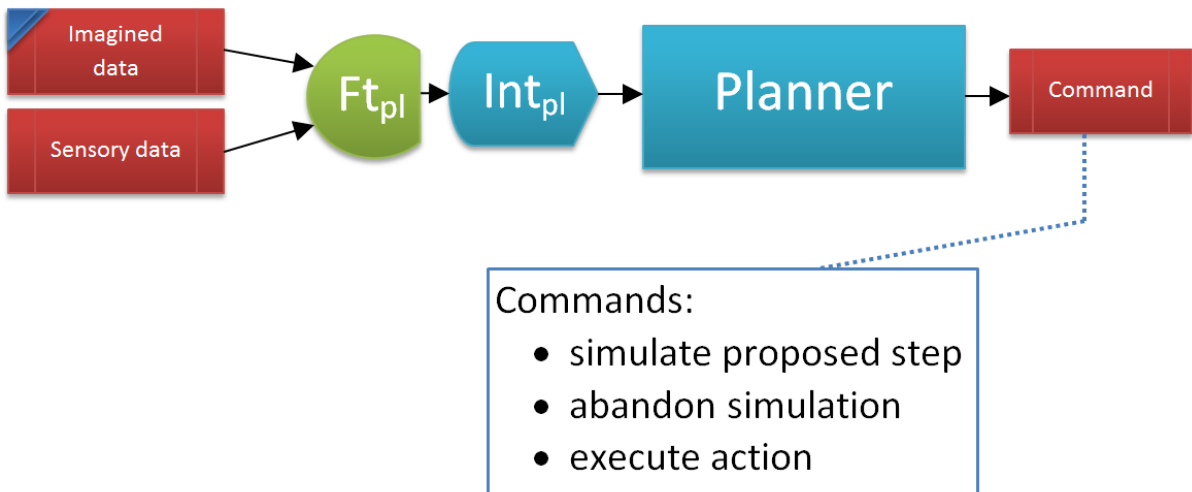


Figure 11: Planner with two kinds of inputs: (1) on the basis of real and counterfactual sensory data, possible steps are developed. (2) Projected consequences of actions lead to the readjustment of planned steps.

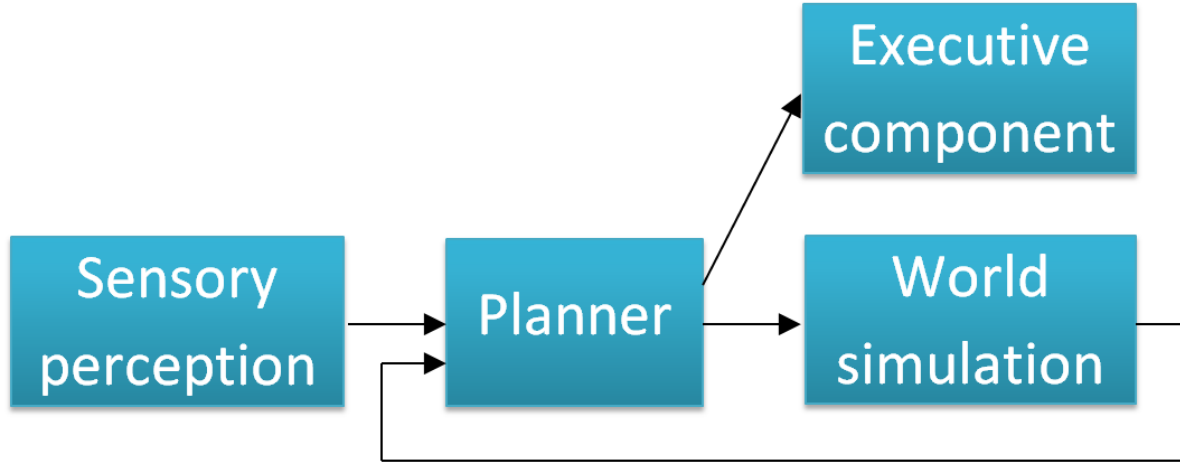


Figure 12: Interaction between world simulator and planner: the planner devises possible steps and feeds them into the world simulator, which, in turn, tries to calculate their effects. The results are fed back to the planner, which uses them as a basis for other steps.

### 6.3 Affect

When discussing human affect, one can mean various things: the causation of emotion, its internal mechanisms, the expression of emotion, social communication of emotions, etc. In this document, we restrict our attention just to the internal mechanisms — that is, to the means by which emotions are evoked in an agent and how they shape its thinking.

Furthermore, the issue will only be the causative mechanism itself; taxonomy and hierarchy of emotions are deferred to future versions of this document.

The model presented herein is adapted from Gadanho and Hallam [2], who employed it in the context of robot learning. They constructed a system of FEELINGS and SENSATIONS  $\mathcal{F}$ , EMOTIONS  $\mathcal{E}$ , and a hormone storage  $H$ .

Figure 13 shows this model: SENSATIONS enter the system and are connected to the FEELINGS. They, in turn, determine the agent’s EMOTIONS. The emotions then feed into a HORMONE STORAGE, the contents of which influence, together with the SENSATIONS, the agent’s FEELINGS. In the context of their paper, this model had a very restricted application. Its purpose was to merely help guide a robot through a world, and accordingly,  $\mathcal{F}$  and  $\mathcal{E}$  were only defined as [2, p. 47]:

$$\begin{aligned}\mathcal{F} &= \{\text{Hunger, Pain, Restlessness, Temperature, Eating, Smell, Eating, Proximity}\} \\ \mathcal{E} &= \{\text{Happiness, Sadness, Fear, Anger}\}\end{aligned}$$

The main advantage of Gadanho’s and Hallam’s model is that (a) it is sufficiently generic to accommodate various schemas and (b) posits an internal state (the hormone storage), giving agents a certain inertia. For example one can imagine integrating a many-dimensional model like Brazeal’s [TODO: CITE BRAZEAL 2003] detailed taxonomy of emotion like Ortony’s OCC model [TODO: CITE OCC (ortony 2003)]. The existence of an internal state is necessitated by the simply observation that our internal world is not solely dependent on momentary stimuli, but merely influenced by them. The idea of a hormone storage might be a simplistic approximation

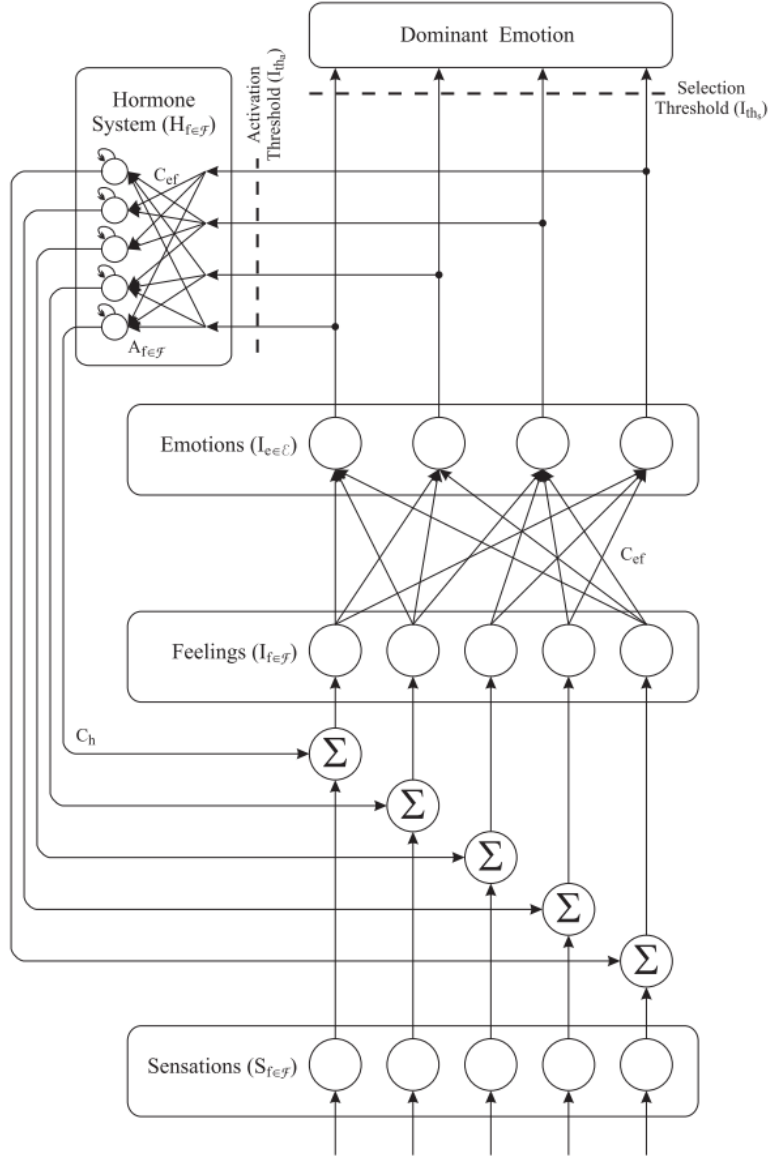


Figure 13: Emotional model of Gadanho and Hallam [2, p. 46].

but it, too, can be refined as needed<sup>6</sup>. Figure 9 shows the adapted model. The general structure was retained, but the set of sensations was replaced by the sensory processor described in section 9 and, instead of a single dominant emotion, competing emotions simply emit messages for executive components. TODO: redo graph to account for this.

<sup>6</sup>It might be tempting to simply replace the hormone storage, but doing so would ignore the role that neurotransmitters like dopamine and serotonin play in cognition, irrespective of the purely computational activity of brain components.



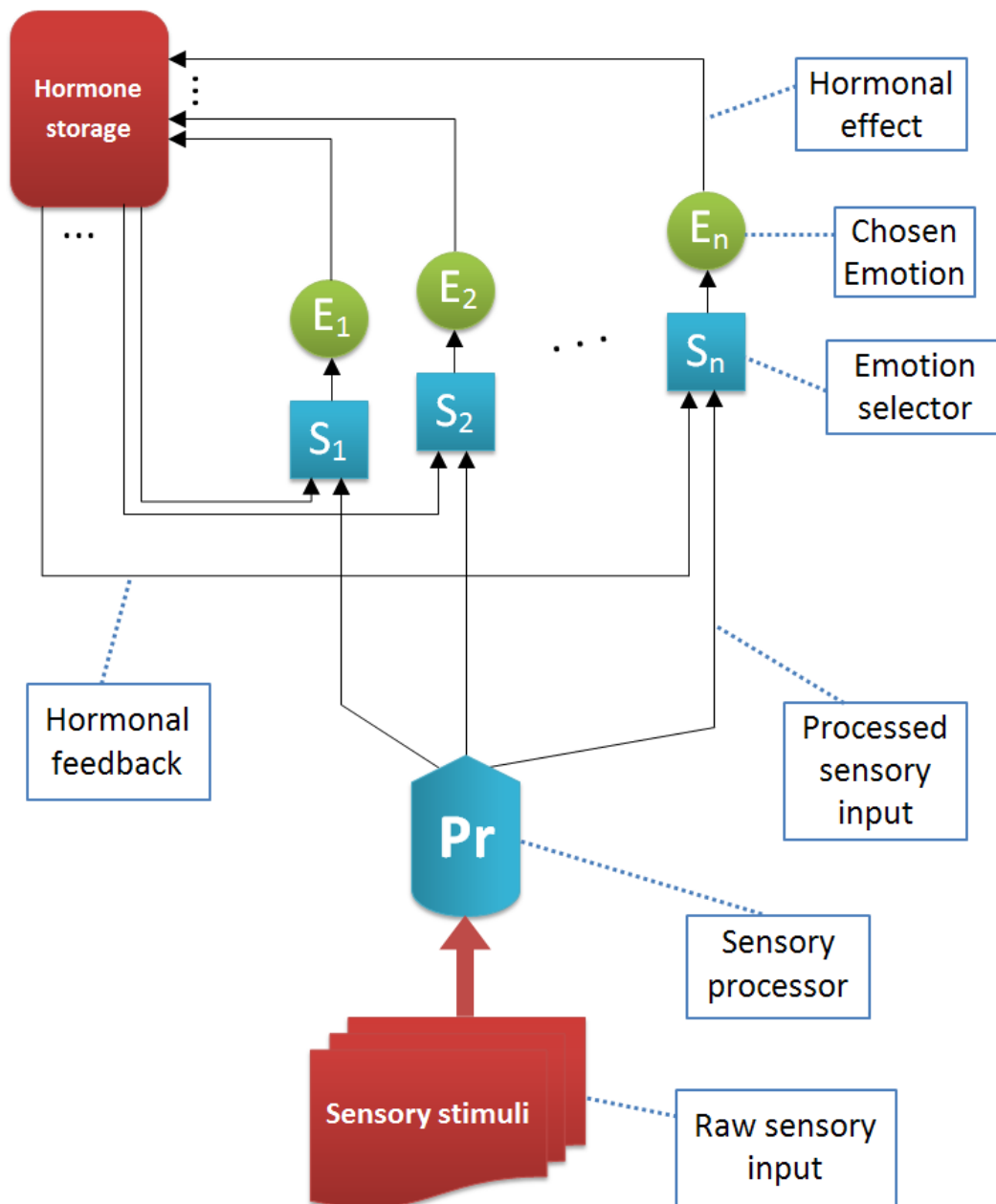


Figure 14: Affective subsystem; specialisation of the global neural architecture. In plastic neural systems, selections may change over time.

## 7 Errors in the model

### References

- [1] Alain Berthoz. The role of inhibition in the hierarchical gating of executed and imagined movements. *Cognitive Brain Research*, 3(2):101–13, 1996.

- [2] Sandra Clara Gadanho and John Hallam. Robot learning driven by emotions. *Adaptive Behaviour*, 9(1):42–64, 2001.
- [3] Mihai Ionescu, Gheorghe Păun, and Takashi Yokomori. Spiking neural p systems. *Fundam. Inf.*, 71(2,3):279–308, February 2006.
- [4] Martin Lotze, Perdo Montoya, Michael Erb, Ernst Hülsmann, Herta Flor, Uwe Klose, Niels Birbaumer, and Wolfgang Grodd. *Journal of Cognitive Neuroscience*, 11(5):491–501, 1999.
- [5] Paul MacLean. *The Triune Brain in Evolution: Role in Paleocerebral Functions*. Plenum Press, New York, 1990.
- [6] Marvin Minsky. *The Emotion Machine*. Simon & Schuster, New York, 2006.
- [7] Gheorghe Paun, Grzegorz Rozenberg, and Arto Salomaa. *The Oxford Handbook of Membrane Computing*. Oxford University Press, Inc., New York, NY, USA, 2010.
- [8] David Sander, Didier Grandjean, and Klaus R. Scherer. A systems approach to appraisal mechanisms in emotion. *Neural Networks*, 18(4):317–352, 2005.