Consciousness is a Semiotic Meta-management Feedback Loop

Sep 3, 2023 • Malcolm Lett

> **NOTE**
> A précis of the entire theory is included in chapter I.2.

Abstract

- Starting from first principles I develop a case for *why* consciousness evolved: to solve the meta-management problem of computational state trajectories in multi-iteration processing. I show that this is best solved through a meta-management feedback loop that presents a summary of computational state as an additional sense, enabling the first-order control process to meta-manage itself. I then explain *how* those mechanisms create the various phenomena associated with subjective experience: through the construction of Higher Order Observational States, a specific form of HOT. The “little person in the head” homunculus effect, the “stream of consciousness”, and the causal natures of consciousness all find a clear explanation. The theory shows how consciousness, intelligence, and meta-cognition are related, and paves a way towards human-level general intelligence.

Contents

Part I - Introduction

I.1 A Question of Consciousness

It has been said that the field of neuroscience is data rich, and theory poor. While vast quantities of data are available from the nature of individual neurons to the activity of the brain as a whole, we lack functional theories of cognitive function in order to make sense of that data.

That problem is no less poignant for the question of consciousness. What is the basis of consciousness? It's underlying mechanisms? What makes one thing conscious, while another is not? What is this ethereal "something other" that seems to be associated with conscious experience - the "what it feels like" to be conscious (Nagel, 1974)? We have ample information about so called *correlates of consciousness* - neurological data about high level brain activity at the time of conscious experience. But we lack an effective functional theory to compare that activity against. While many *theories of consciousness* (TOC) exist, they lack the details needed to sufficiently explain the neurological data.

I am interested in finding the mechanisms underlying conscious first-person subjective experience (*subjective experience*, hereafter), so that we can explain why those mechanisms and consequent subjective experience evolved, why subjective experience feels the way it does, and to explain the specific properties of subjective experience. This treatise seeks to do just that. Following a *design stance* that looks at the computational problems and solutions faced by increasingly complex artificial and biological organisms, I present a theory of both the computational and phenomenal aspects of subjective experience.

The theory presented is compatible with many existing theories of consciousness, including Higher-order Thought theories, Global Workspace Theory, and Integrated Information Theory. In fact, many of the ideas presented here are not new. But I believe this to be the most comprehensive attempt to ground those theories in a fundamental theory of why those mechanisms evolved in the first place. I suspect that in many respects, a key skill of anyone doing research in this area is to sieve through the many partial stories and see the path that links them together.

The immediately following chapter provides a concise summary of the theory from start to finish. This is followed by two background chapters, where the various terms used throughout are formally defined.

The rest of the treatise is divided into six additional parts. Parts II to IV examine the computational problems faced by increasingly complex embodied agents, along with a discussion of some of the solutions to those problems. Part V combines those discussions into a single coherent *Solution* and explains how that solution produces the properties of subjective experience. Part VI discusses how our intuitions produce the so called *explanatory gap*. Part VII completes the treatise with some discussion of further consequences of the theory and a final summary.

I.2 Précis of Thesis

I believe that three things go hand-in-hand: general intelligence, meta-management, and consciousness. Further, understanding these systems must proceed in that order. Thus, in order to understand why we have subjective experience, we need to begin to develop a theory of general intelligence and of the processes that restrain it from cascading into chaos. This treatise presents an end-to-end argument following that thread.

To save the impatient reader from the tedium of waiting till the end, the entire argument is first presented here in brief. Background explanations and some definitions are omitted for the sake of brevity. The rest of the treatise is devoted to detailed explanations of each of the logical steps in the argument thread here.

Any computational system is limited in the complexity that it can handle within a single execution of its computational process. For embodied agents, this appears as a limit on the environmental complexity that they can sufficiently model and respond to within a single executional iteration. For more complex problems, multiple iterations of processing are required in order to determine the next physical action. Such recurrency in processing may for example entail 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.

During the execution of a *multi-iteration* controller within an embodied agent, its control process (CP) passes through an internal state trajectory that is only occasionally associated with interaction with the physical environment. That internal state trajectory can become increasingly disassociated with the physical environment the more complex the problem space and the longer the time required for deliberation. If the multi-iteration processor must also learn through reinforcement then it is likely to exhibit chaotic and unproductive behaviour, particularly so during the earliest stages of learning. Reinforcement from the environment may be too sparse for efficient learning to take place, and simple rules that penalise longer deliberation time may be insufficiently flexible to cater for the complexity of problem domains that the agent may be faced with. Explicit meta-management (second-order) processes are required to observe the first order control process, to model its behaviours, to track its success rate, to act upon it to prevent chaotic behaviours that could harm the agent, and to participate in providing rewards and penalties with more advanced problem-domain aware knowledge than just a simple penalty for deliberation time.

In a complex biological brain that can operate its body effectively within the real world, including all the complexity of modern human life, the systems and processes required to model and understand the environment and to interact with them are immense. It turns out that the systems required to effectively carry out meta-management are similarly complex. More importantly, the kinds of systems required for meta-management are similar to those required for first-order control: observing, modelling, inference, planning, sequencing, controlling. Additionally, in order to effectively meta-manage the first-order control process, the domain knowledge required by the meta-management process is often strongly associated with the domain knowledge employed by the first-order control process at the time. The level of overlap between first-order control and meta-management implies a radical solution: that the first-order control process meta-manages itself.

In order to meta-manage itself, the control process needs to observe itself. This can be achieved through a *meta-management feedback loop* that observes the state of the control process, observes the recent trajectory of the control process, distils that into a high-level abstract representation with lower dimensionality, and makes it available to the control process as a sensory input signal. Thus, the fact that we have awareness of some aspects of our own mental state is a direct result of the need for meta-management. In other words, the meta-management feedback loop is an additional perceptual sense.

The existence of the meta-management feedback loop does not alone explain subjective experience. Two more ingredients are required: interpretation and meaning. In a complex organism such as a human being, the brain maintains *schema* that represent and track the characteristics of different aspects of the individual. This includes the body schema, which models and tracks the location and orientation of the limbs, their abilities, and whether any injuries have been acquired. This includes schema about regularly encountered external things, such as is required to mentally track the location and orientation of the wheels while driving a car over potholes. With the introduction of the meta-management feedback loop, this also includes schema for tracking the state and capabilities of the cognitive processes. And, importantly, it includes a strongly developed sense of self vs other. All external and internal sensory inputs (including the meta-management feedback loop), all schema states, are labelled as to their source. Different source labels imply significantly different meaning in terms of, for example, the level to which the agent can affect that state.

External sensory inputs, meaning association, and feedback loop together are interpreted by the brain and a decision made about the next action. Information about that action and/or the mental state that produced it is available via the feedback loop in the next iteration of processing. This creates a continuous cyclic stream of ever changing inputs, states and actions. As the control process decides to perform some body action, knowledge of that chosen action is immediately available as a sensory input before the action is even started. As the control process chooses to deliberate further on a problem at hand, knowledge of that deliberation is immediately available. At every moment in time, the control process can choose to attend to external matters, to the continuation of the current deliberation problem at hand, or to the very feedback sense that it receives continuously while it is usually distracted doing other things. If the control process stops to consider its own feedback sense, compares that to memory of recent deliberations, compares that to its schema of self vs other, it necessarily concludes that it has its own “stream of thought”. That stream of thought is subjective experience.

The metaphor of a *philosophical zombie* was introduced to hypothesise a human-like individual that has all the behavioural characteristics of a human, including voicing that it is conscious, without actually experiencing subjective experience. A philosophical zombie is *behaviourally indistinguishable* from a human with consciousness. I argue that, if the zombie employs the mechanisms given in the explanation above (and outlined in more detail in the chapters that follow), then it is also *computationally indistinguishable* from a human with consciousness. In other words, if we were to assume an ability to tap into all of the inner state and workings of a brain, and if we were to compare the philosophical zombie from the conscious human, we would find no difference. At that point I argue that the zombie is not a zombie after all, but instead is a fully conscious human being having subjective experience.

Finally, I argue that our disgruntlement with such an explanation is not an *explanatory gap* (Levine, 1983; Van Gulick, 2022), but an *intuitional gap*. Nagel famously pointed out that we have no conception of what it is like to be anything other than ourselves (Nagel, 1974), and Block argued further that we have “no conception of a ground of rational belief” that could enable us to develop such a conception, or even to know whether or not something unlike a human is conscious (Block, 2002, p. 408). But that does not stop us making assumptions about what things should and should not experience subjective experience. It took science a long time to accept that animals could have any form of consciousness and subjective experience, and likely many still deny that outcome. If we have no way to conceive of the experience held by an animal, why should we be so adamant about its properties? The answer simply is our deluded intuition. Our brains excel at finding patterns and extrapolating from them. This works well when the physical environment around us is there to provide an error signal. But when no error signal is available, we are prone to delusion. Our minds create such a strong sense of self vs other by keying that information into every sensory signal that we ever receive, so that our senses seem to take on an extra quality of realness, of subjectiveness, of *qualia* (Tye, 2021). That seeming extra quality is further processed by the same system that produced it, reinforcing the delusion that the qualia is something extra, beyond mere sensory information. And thus we are deluded into the intuition that subjective experience is somehow more than can be produced by mere computational processing.

I.3 Background - Consciousness

Many theories exist about the nature of consciousness. A very brief summary of such theories will be given here. This serves two purposes. Firstly, to provide a background to readers who are not familiar with the topic. Secondly, to define clearly the particular meanings that I will use for various key terms throughout the rest of the treatise.

The word "consciousness" is an overloaded term, ie: that it has many meanings, with the particular meaning depending on context. Furthermore, consciousness of the sort that I wish to talk about here is an ill-defined concept. For example, consider the following working definition:

> By consciousness I simply mean those subjective states of awareness or sentience that begin when one wakes in the morning and continue throughout the period that one is awake until one falls into a dreamless sleep, into a coma, or dies or is otherwise, as they say, unconscious. (Searle, 1990)

Firstly, this is not a definition, but an example. One that implores the reader to refer to their own intuition in order to guess the author's meaning. But this is typical of any research in this field, and I shall do no better. Secondly, the example leaves ambiguous the case of dreams, which many might argue also carry some extent of conscious-like awareness.

Here is my own attempt at defining consciousness of the form that shall be discussed within this treatise:

- Consciousness is the collection of first-person subjective experiences that we have, for example, while awake. Not only do we have such experiences, but we can be aware of the fact that we are having or have had those experiences. We also have first-person subjective experiences while dreaming, and thus consciousness refers to those experiences too. In contrast, consciousness is absent when in deep sleep, and presumably when in vegetative comas. Likewise we are not conscious of everything that happens in our brain, even at the best of times. For example, when you look across the room and observe a chair, you have no awareness of any of the processes that just occurred within the brain that took the visual sensory information and identified that a chair was present. So, consciousness includes the things that you aware of, and excludes those things that you are not aware of.

Unfortunately there is still much ambiguity and room for disagreement. But it will suffice as a start so that I can refine some ideas as we go along.

I.3.1 Questions of Consciousness

For the study of consciousness, we have three broad questions that we wish to answer, that can be paraphrased as "what, why and how?":

- **What:** what are the defining characteristics and properties of consciousness? ie: what do we look for in order to identify whether some thing or some creature experiences consciousness, or to distinguish whether a particular mental state is conscious? Also, what kinds of properties are we trying to explain through an explanation of consciousness?
- **Why:** why does it exist? ie: why did it evolve? What functional benefit does it offer over and above a creature that lacks consciousness?
- **How:** how does it occur? ie: what are the underlying mechanisms that produce consciousness?

The definitions given above are examples of attempts to answer the "what" question, but there are many more details yet to be elaborated on.

The "what, why, and how" questions can be focused on two different scopes:

- **Creature Consciousness:** what is it about certain things (humans, animals, maybe other kinds of things) that results in them having consciousness, while other things do not?
- **State Consciousness:** what is it about certain brain states that mean that the individual has conscious experience of those states, while not having conscious experience of other states? And which aspects of those states are associated with consciousness?

For example, this leads to different variants of the "what" question, two of which are a) what are the characteristics associated with creature consciousness? b) what kinds of things are we conscious of?

A good answer to the former requires us to answer the questions of why and how. But the latter is more open to investigation on its own. A few kinds of experience have been identified:

- **Perception:** specifically via the traditional five senses of sight, sound, smell, taste, touch that give us information about the world external to us (Gleitman, 2004, p. 204-237)
- **Bodily awareness:** this includes strong feelings such as pains and pleasures, and more subtle body-state awareness such as balance and proprioception (de Vignemont, 2020; Bermúdez, 2005; Mandrigin, 2021)

- **Memory:** recalling past experiences, where that past experience may be any of the sorts described here (Gleitman, 2004, p. 242-273)
- **Imagination:** conjuring images (and perhaps via other modalities) within our mind of things that do not exist (Liao and Gendler, 2020)
- **Emotion:** the experience of feeling one's emotional state (Scarantino and de Sousa, 2021)
- **Desire:** the awareness of ones goals and desires (Smithies and Weiss, 2019; Schroeder, 2020)
- **Action:** experience and awareness of one's actions, or of the intent to act (Tsakiris and Haggard, 2005; Bayne and Pacherie, 2007)
- **Thought:** an awareness of the sequence of thoughts, such as during problem solving, including but not limited to internal imagined worded vocalisations (Gleitman, 2004, p. 278-315)

The term "experience" itself is sometimes attributed to only a subset of the above, but I shall use it here to refer to any of the above where one could be said as being consciously aware of that experience. Or in other words, I take it that we can consciously experience all of the above "experiences".

We can also ask what it is about experiences that make them somehow *special*. Again an appeal to intuition is necessary. If a modern computer can produce interesting behaviour but we take it as granted that it has no conscious experience, and in fact nothing we would even call "experience" of any sort, then what are the properties of experience that are lacking in the computer? Another example uses a so called *philosophical zombie* (or p-zombie for short), which is identical to a conscious human in every respect except that they have no conscious experience (Chalmers, 1996). In particular, they behave in every way like a human including having all the same conversations as a human, but all of their behaviour is produced in some way more comparable to the modern computer - there is "no light on upstairs". We can say that the states of both the human mind and p-zombie include everything they need to include in order to produce behaviour, but that the state of the human mind also includes properties of experience that are lacking in the p-zombie.

These properties are the *phenomological properties* of experience, also known as *qualia* (Tye, 2021). Frustratingly, it has been painfully difficult to pin down these properties. They are often summarised simply as the "what it's like" to have experience (Nagel, 1974) or the "raw feels" of consciousness. Another problem with discussions of qualia is that such discussions often get caught up with the observation that perceptions don't always correlate exactly with reality; a case that is seen clearly with the way that our perception of colour of a particular object is only tenuously associated with the physics of light and its interactions with that object. How our perception correlates or fails to correlate with reality is an important question, and not entirely unrelated to the former question, but it is somewhat of a distraction when we are trying to define what qualia are in the first place.

One such property is that experiences seem to "look through" to the content of the underlying perception etc. that is being experienced (Siewert, 2004). While many debates exist, we seemingly always have experiences *of* something (Siegel, 2021; Siewert 2022). The something that experience is *of* is referred to as the *intent* of the experience (Crane, 2009). The word "intent" here does not take its usual English meaning; rather, it is best thought of as a target or a focus. A noticeable feature of experiential intent is that it is relatively easy to conceive that it could be constructed through entirely mechanistic processes. That stands in stark contrast to the experience itself. Other properties that could perhaps be associated with experience itself include the individual's awareness of themself, and a sense of agency or purpose (Smith, 2018).

A problem that arises is whether the experience of intent and the intent itself can be separated. This has led some to consider a conceptual separation, and perhaps even an actual separation, between the following (Block, 1995):

- **Access Consciousness:** awareness of the perception, body sensation, memory, imagination, emotion, desire, action, or thought, to a sufficient extent that the individual can in some way react to the experience
- **Phenomenal Consciousness:** the first-person subjective experience of those things.

For the most part these two forms of consciousness are tied together. Most would view access consciousness as excluding non-conscious processes by definition. In other words, access consciousness is restricted to awareness of experiences that have a phenomenal nature. Likewise, phenomenal consciousness has a *content*, and that content is usually what we refer to as access consciousness. But many debates exist, resting on more detailed analysis and on attempts to define these terms more accurately. For example, there are claims that some phenomenal conscious experiences carry no representational qualities of the sort that characterise access consciousness, and thus that it may be possible to have phenomenal consciousness without access consciousness (Block, 1995). Likewise, some definitions of access consciousness make it possible to have access consciousness without phenomenal consciousness.

I subscribe to the view that access consciousness and phenomenal consciousness are just different ways of looking at the same thing. Specifically, that access consciousness is a reference to the processes that construct the specific *content* of consciousness, including not just the intent but also its "raw feels", while phenomenal consciousness is a reference to the *question* of why that specific content became conscious in the first place. Mind you, this is not a standard view; most would associate the raw feels with phenomenal consciousness as something that is somehow distinguishable from access consciousness. It is for this reason that I predominantly use the term *subjective experience* throughout this treatise, instead of *consciousness*. I wish to be clear that my theory is about both access and phenomenal consciousness (as per Block's definitions), together as a single thing.

I.3.2 The Purpose of Consciousness

Why does consciousness exist? Presumably conscious experience confers some sort of benefit to the individual in order that it evolved. In order words, we assume that it serves some *function* or *functions*, that those functions are useful, and that the individual would be at a disadvantage if it were lacking those functions.

When attempting to decipher the *function of consciousness*, there are a number of different angles that should be considered (Niikawa et al, 2020; Rosenthal, 2008). Firstly, we need to distinguish whether we are asking about creature consciousness or state consciousness. On the one hand, we're asking whether the creature as a whole benefits from having conscious experience. On the other hand, we are asking whether a given process benefits from having conscious experience associated with part or the whole of that process.

A second important consideration is to distinguish whether we are talking about the *functional basis* of consciousness, or the *functional contribution of consciousness* (Niikawa et al, 2020). As we now understand, many brain processes occur unconsciously, including many processes that are associated with perceptions and thoughts that we consciously experience (Earl, 2014). The functional basis of consciousness is an otherwise unconscious process that *produces* conscious experience. In contrast, the functional contribution of consciousness is the effect that conscious experience itself produces - it is whatever behaviour or other function that conscious experience bestows upon the individual and that they would not otherwise have had.

Unfortunately, when we understand so little about the processes underlying consciousness, it can be very hard to tease those two conceptions apart. Thus most work that attempts to pinpoint a purpose for consciousness could be said more accurately to identify a correlation and to then build up a theory around that correlation.

A number of such theories have been proposed that either attempt to directly suggest a functional purpose for consciousness, or that merely suggest a possible purpose as part of larger theories into the *what* or *how* of consciousness. Some broad ideas follow.

Integration and Global broadcast: a predominant feature of consciousness is that it appears to pull together multiple streams of information into a single coherent representation and than make that available for further processing by other systems. This is held as the key purpose of consciousness within a number of theories, including Global (Neuronal) Workspace Theory (Baars and Franklin, 2007; Dehaene et al, 2003; Dehaene, 2014), IIT (Tononi and Sporns, 2003), and Supramodular Interaction Theory (Morsella, 2005).

Flexible behaviour: another predominant feature of consciousness is its apparent association with flexible behaviour. This can be characterised as an ability to adapt to novel situations more rapidly than would be expected from experimental "learning" alone. A closely related and somewhat ill-defined conception is that of "rational behaviour". Many have proposed that consciousness serves to enable adaptive and/or rational behaviour (Kotchoubey, 2018; Morsella, 2005; Earl, 2014; Humphrey, 2002; Shimamura, 2000; Tye, 1996).

Counter-factual reasoning: Some have proposed that consciousness enables flexible and adaptive behaviour through specific mechanisms. One such example is that of enabling counter-factual reasoning through the ability to imagine alternatives (Kanai et al, 2019).

Association learning: Another specific case of flexible behaviour is through the ability to learn a seemingly unlimited range of associations, both from direct experience and from second hand experience such as observing others or being told about the association. It has been suggested that consciousness directly correlates with that ability and thus must be an integral part of the ability (Birch et al, 2020).

Meta-cognition: A view that some have taken, this author included, is that consciousness is strongly associated with meta-cognition (Fernandez Cruz et al, 2016; Paul et al, 2015; Flemming et al, 2014; Flemming et al, 2012; Cleeremans, 2007; Shimamura, 2000). For example, perhaps the point of the integration and broadcasting associated with consciousness is to obtain enough information to enable the individual to determine how best to gain more information (Kriegel, 2004). Or perhaps it is for the detection and correction of errors encountered when performing long chains of reasoning (Rolls, 2004 and 2005).

Social interactions: A fascinating possibility is that consciousness evolved as an integral part of our nature as a social species. Consciousness enables *theory of mind* about one's own mind, and by extension, the minds of others (Frith, 2008; Bahrami et al, 2012; Flemming et al, 2012). Theory of mind plays a key part in enabling us to cooperate with other individuals.

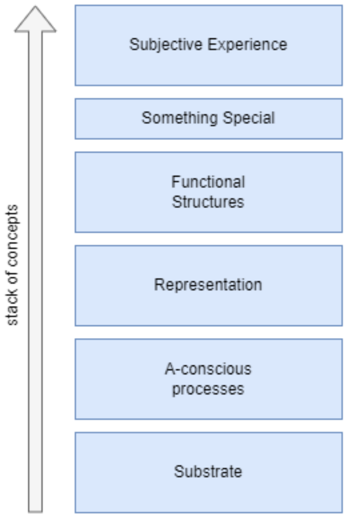
Volition: Consciousness may simply be the magic ingredient that enables mobile bundles of matter to have volition in a *free-will* sense (Pierson and Trout, 2017). For example, *Panpsychism* holds that consciousness is a fundamental property of the universe (Chalmers, 1996), and of all the matter within it. A commonly presumed feature of such panpsychist consciousness is that it has free-will, and thus it would confer free-will onto the behaviour of individuals.

Functionless: Another possibility is that consciousness confers no benefit to the individual. For example, this is implied by the claims of *Epiphenomenalism* (Robinson, 2019). A slightly softer stance is that consciousness occurs as a side effect of other functions, where those functions themselves provide significant benefit to the individual, but consciousness itself does not add any extra benefit (Rosenthal, 2008). Under this view, consciousness is a side-effect of behaviour rather than a driver of behaviour.

Other supposed purposes of consciousness identify strong correlations but fail to explain why consciousness is needed for that to occur. For example, consciousness appears to be essential for long-term episodic memory; we are unable to remember things that we were not conscious of at the time (Tulving, 1987; Edelman et al, 2011).

I.3.3 General Theories of Consciousness

I now describe some generic and specific theories of consciousness. To provide some grounding to the summary and to aid in comparing theories, a framework of a stack of conceptual layers is used, illustrated in the diagram that follows. This is not meant to imply a priori anything about the actual structure of consciousness and its underlying mechanisms. For example, none of the layers are assumed necessarily to exist a priori, nor is the particular illustrated order appropriate in every case.



- ***Conceptual layers used here to summarise theories of consciousness.*** *Substrate: the physical biology of the brain is viewed as a substrate that "hosts" other processes, such as computational processes. In some theories the substrate might potentially be replaced or emulated via some other substrate (eg: silicon neurons). A-conscious processes: all processes built upon the particular substrate (eg: neural signalling) that produce behaviour and/or affect the state of the brain. Some of these processes or their results are somehow associated with conscious experience, while others are not. Representation: some theories hold that particular kinds of representations are associated with conscious experience. Functional Structures: some theories hold that certain kinds of structure are key to conscious experience. Something Special: some claim that conscious experience is an effect that cannot be explained via purely materialistic means, or that we need some new kind of understanding to explain it. Subjective Experience: the ultimate effect of consciousness that we wish to understand.*

The *substrate* is the (generally physical) thing in which the processes of the mind are carried out. For example, biology and neuroscience has taught us that biological neurons are the substrate of the human mind. There is a conceptual difference between the thing that "hosts" the processes, and the processes themselves. Historically there has been three broad views with regards to the substrate. Perhaps the oldest view is the *dualist* view of Descartes. In their view there was the physical substrate, entailing the body and brain, and there was some other non-physical substance that hosted the mind, independent of the physical body (Descartes, 1911; Van Gulick, 2022). While few hold to that view today, it made more sense at the time when Descartes first proposed it. Reportedly Aristotle believed that the purpose of the brain was to cool the blood (Smart, 2022). Many today hold to a *monist* view that there is only one substrate, and that this substrate is physical, for example the biological neurons of the brain. More generally, such *physicalistic* or *materialistic* views hold that the properties of physical substances and their interactions are entirely sufficient from which to base an explanation of everything about consciousness, even if we don't know enough about those properties just yet (Stoljar, 2023). An alternative monist view, that of *idealism*, is that nothing is physical and that everything we perceive exists only within our minds (Guyer and Horstmann, 2023).

Returning to the materialistic view, for humans and for any other animal that happens to have consciousness, we can say that the conscious mind is *realized* upon the physical substrate of biological neurons. Is that the only physical substrate upon which consciousness can be realized? If consciousness is *multi-realizable* then perhaps it could also be realized upon silicon neurons. Even more radically, perhaps the operation of neurons could be *simulated* within a super computer, and a simulated brain could also by conscious. By extension, if multi-realizability is true, then any *functional isomorphism* of a biological human brain should be not only conscious, but conscious in a human-like way. In contrast, *psycho-physical identity theory* posits that there may be something special about biological neurons that give rise to subjective experience, and that non-biological functional isomorphisms would not have such experiences (Van Gulick, 2022).

One view of brain activity is that it is computational in nature (Rescorla, 2020). Initial versions of this idea likened such computation to that of the *Turing machine* (McCulloch and Pitts, 1943), with its serial computation, finite set of states, and random access memory. Later variations accepted that the brain wasn't exactly like a Turing machine, but was still Turing-like. In particular, Turing machines operate upon a finite set of symbols. Fodor's *representational theory of mind* (Fodor, 1975, 1981, 1987, 1990, 1994 and 2008) extended that idea to more generic *representations*. Importantly, these representations could be hierarchically composed, enabling their computational model to create infinitely many variations of states, more closely mirroring our ability to form larger ideas by composition of smaller ones. Nowadays, most apply a *connectionist* lens to computation (Marcus, 2001; Smolensky, 1988; Kriegesgorte, 2015), following advances in both our neuroscientific understanding of the brain and in artificial intelligence (Krizhevsky, Sutskever, and Hinton, 2012; LeCun, Bengio, and Hinton, 2015). Another lens recognises that the brain must deal with a significant amount of uncertainty and thus might actively include that within its representations and computations (Ma, 2019; Rescorla, 2019).

One particular line of criticism levelled at computational models draws issue with the idea of representation (Rescorla, 2020). Initial views of the brain as a Turing machine, or at least Turing-like, suggested that the brain would operate against symbols in the same way that a computer does. This drew obvious criticism. But the modern connectionist view of computation is not immune. One particularly strong issue lies with the distinction between *syntax* and *semantics*. Here, syntax refers just to the *form* of the representation, while semantics refers to its *meaning*. A clear example of the distinction can be seen in the case of computers. In a modern computer, all information is represented via the *form* of zeros and ones, regardless of its meaning. For example, a particular string of zeros and ones might *mean* a particular base-10 number, or it might *mean* the colour of a pixel. Furthermore, all computation is performed against the syntactic form alone. This can be seen in the way that two large base-10 numbers can be multiplied by a mechanistic process that lacks any knowledge that it is multiplying base-10 numbers. It sees only a sequence of binary bits and performs only a series of simple per-bit operations (Booth, 1951) that do not involve any multiplication. There exist arguments that all computation is of this form (Rescorla, 2020), including that of the brain (Fodor, 1981). It is intuitively problematic that the inherent meaning of the representation is discarded, and thus there is a reluctance to accept representation as being a complete story of mind.

My own stance is that meaning is not lost. For simple habitually repeated operations it is encoded within the process that performs the computation, and for complex operations the meaning is itself represented. But more on that later.

It is now known that many brain processes operate that neither directly nor indirectly lead to any kind of subjective experience [citation]. And for those processes that do have an impact on subjective experience, the vast majority of the details of those processes are still hidden from that subjective experience, including the operation of the processes and the content that they operate on (Nisbett and Wilson, 1977a). While earlier work assumed that most brain processes are conscious, there is increasingly strong evidence in fact that almost all mental processes are non-conscious, even for mental processes that are associated with attentive subjective experience [citation].

An obvious question thus arises about why certain processes and/or representations would be associated with subjective experience and others would not. Is it sufficient that a particular kind of representation exists for it to be associated with subjective experience, or does that representation need to occur in conjunction with particular functional structures? For example, imagine that all aspects of the brain were understood to the point that we could identify exactly which representations are associated with subjective experience. Now imagine that an exact replica of a particular representation was encoded within the gates of a silicon memory chip within a computer. Most would argue that the silicon memory chip does not subsequently have subjective experience of that representation, because a representational state alone is insufficient for subjective experience. Some kind of functional structure presumably must be required to observe that representation. But if functional structures are indeed required, what are those functional structures? And what distinguishes those functional structures that are associated with subjective experience and those that are not?

A deeper philosophical question exists about whether representation and functional structures alone are sufficient to produce subjective experience, or whether something else is required. A representational state is just a state. It doesn't do anything. So it can't be the source of subjective experience. But a functional structure is just static (biological) machinery. More clearly, while the functional structure may change as the result of learning, at the moment that it produces any given behaviour its structure is static. So that couldn't produce subjective experience either. Thus perhaps qualia is a case of an extra *something special* that we do not yet understand about physics, or that exists in a dualistic relationship to physics. These are the questions posed by the distinction between access consciousness and phenomenal consciousness. Consider this example provided by the Britannica online article on philosophy of mind:

> Suppose that, in order to avoid the risks to his patient of anaesthesia, a resourceful surgeon finds a way of temporarily depriving the patient of whatever nonfunctional condition the critic of functionalism insists on, while keeping the functional organization of the patient's brain intact. As the surgeon proceeds with, say, a massive abdominal operation, the patient's functional organization might lead him to think that he is in acute pain and to very much prefer that he not be, even though the surgeon assures him that he could not be in pain because he has been deprived of precisely "what it takes." It is hard to believe that even the most ardent qualiaphile would be satisfied by such assurances.

If all the representational and functional structures are in place for an individual to have both the external behaviours and internal mental behaviours of an individual in intense pain, can we even conceive it to be possible that they would not have the typical associated subjective experience of the pain?

I.3.4 Specific Theories of Consciousness

Higher-order Thought (HOT) and Higher-order Perception (HOP) are a group of theories that focus on the form of the representation. It is claimed that there are broadly two types of cognitive state (Carruthers and and Gennaro, 2020; Rosenthal, 2004). *First-order states* are the primary states of the brain involved with control of behaviour in the absence of subjective experience, such as visual and auditory information about the outside world. In contrast, *higher-order states* represent things about those first-order states.

Where a higher-order state represents that a first-order state was *experienced*, then we have conscious experience. Several variations exist in this group of theories. HOP theories focus on perception, and in particular an idea that we have an explicit perceptual inner-sense that observes our cognitive state, for example in the same way that our visual sense observes the world outside (Armstrong, 1994; Armstrong & Malcolm, 1984; Lycan, 1996 and 2004). HOT theories are computational theories that propose that higher-order states are *constructed* from or about the first-order states. Some HOT theories propose that first-order states become conscious through the presence of associated higher-order states, and thus that we only experience such first-order states at the moment that a HOT in constructed about them (Rosenthal, 1986, 1993, and 2005). A subtle variation holds that certain first-order states are inherently *disposed* to have associated higher-order states and that that is sufficient for the first-order states to be experienced as conscious (Carruthers, 1996, 2000, and 2005). Some view that first-order states and higher-order states are related but independent, and thus that one kind can occur without the other, while others take a *self-representational* view that higher-order states are somehow always constructed inline with their intentional first-order states (Gennaro, 1996 and 2012; Kriegel, 2003, 2006 and 2009).

Global Workspace Theory (GWT) sees the brain as a system of individual computational processes that compete, and sometimes collaborate, to gain a winner(s)-take-all right to *broadcast* their results to a *global workspace* (Baars, 2021; Baars and Franklin, 2007). The state of the global workspace then forms the context upon which subsequent processing occurs by all those same computational processes. This occurs in a continuous and dynamic way typical of the parallel processing that we've come to understand about the brain. GWT is primarily a theory of cognitive processing that focuses on its a-conscious processes, representations, and functional structures. In doing so it offers a partial explanation of the underlying mechanisms of rational thought, or so called *general intelligence*. The link to subjective experience is by way of claiming that the global workspace is the content of subjective experience, based on correlations with the apparent integrative nature of conscious perceptual attention and decision making, forming an apparent single stream of consciousness (Baars, 2021). GWT is an abstract computational theory that composes high-level concepts such as a workspace, processes, frames, and competition and collaboration between processes. While it makes some claims about regions of brain involved, it doesn't explain how groups of neurons produce such behaviour. In other words, it omits important details about the substrate from its explanation.

GNT was formed by Baars (1988), but has been taken up and extended by many others, such as to provide a more detailed computational model (Franklin and Graesser, 1999), the addition of internal simulation (Shanahan, 2005), or to model via spiking neurons (Shanahan, 2008). Of particular note is *Global Neuronal Workspace* theory (GNW). GNW takes inspiration from a growing view of neuronal interactions as *dynamical systems* (Miller, 2016; Favela, 2020; Shapiro, 2013). It proposes how non-linear interactions between populations of neurons can create a sudden "ignition" effect, where multiple independent stimuli suddenly become mutually *salient*, particularly due to long-range axons from sensorial stimuli (Dehaene, Sergent, and Changeux, 2003). In other words, it proposes a specific mechanism of how groups of neurons can cooperate to form a single representation that is then broadcast to others.

The massive number of neurons within the substrate of the brain makes it hard to study how specific computational theories may or may not apply, both from the point of view of empirical neurological investigations and in analytical attempts to formulate how neurons interact to form those higher-level computational abstractions. One avenue is to embrace the combinatorial complexity and to examine the activity across *populations* of neurons - ie: groups of neurons that are spatially close and/or are activate together in some sense. This is seen in theories that view the brain as a *dynamical system*, mentioned earlier, use *information theoretic* techniques, or that look at waves and oscillations in the pattern of activity.

The work on the *Theory of Neuronal Group Selection (TNGS)* (Edelman 1987; Edelman 2003), also known as *Neural Darwinism*, provides one such example. As part of developing TNGS, the authors developed the *Dynamic Core* hypothesis (Edelman and Tononi 2000; Tononi and Edelman 1998), which provides a quantitive measure of *neural complexity* (Tononi et al, 1994). The measure is based on the information theoretic idea of *mutual information*. Consider two random variables *x*, and *y*. If knowing the value of *x* informs me of anything about *y*, then we say that they have mutual information. The more *x* and *y* correlate, the more mutual information there is between them. In contrast, if they are statistically independent, then they carry zero mutual information. Mutual information is important if you care about finding correlations or relationships between things (more mutual information is good), and if you care about being able to store or represent large amounts of detail (less mutual information is good). Within the brain there will be regions of activity that are *differentiated*: to a large extent uncorrelated perhaps because they are busy with different functions, or perhaps because the brain is in a state of disorganised chaos. Likewise there will be regions of activity that are *integrated*: strongly correlated because for example that the brain is organised and focused on one task, or perhaps because it has a lot of unnecessary duplication. Neural complexity measures the extent to which activity within the network is both integrated and differentiated by computing the average mutual information between bipartitions of the network.

Integrated Information Theory (IIT) provides another closely related measure that claims to calculate an objective quantity of consciousness, indicated as Φ (phi). This looks not just for mutual information, which is inherently bidirectional (ie: just a correlation without any idea of cause-effect relationships), but attempts to measure the extent to which certain information is *causually effective*. It then computes the amount of causally effective information that can be integrated across the weakest link of the the system (Tononi and Sporns, 2003; Tononi, 2004). This has since received two major revisions, first in 2008 to measure based on active dynamics rather than static configuration (Tononi, 2008), and then again in 2014 with the introduction of *maximally irreducible conceptual structures* (MICS) (Oizumi, Albantakis, and Tononi, 2014).

Another approach is to look at the dynamics of waves and oscillations of activity within the brain. For example, in older studies it was found that consciousness coincided with so called *gamma-waves*, in the frequency range 30 to 80 Hz, in electroencephalogram (EEG) readings. That is now understood as being the outwardly measurable effect of the micro-interactions at the neuronal level (Llinás, Ribary, Contreras & Pedroarena, 1998; Friston et al 2014; Hunt and Schooler, 2019). In particular, a productive line of thought has been to treat such brain activity as harmonics, with different global and sub-global populations of neurons dynamically forming groups of closely synchronised activity (Atasoy, Donnelly, and Pearson, 2016; Atasoy et al 2018). The ever changing groupings of neurons that create those harmonics is proposed to be self-organising, with one theory describing *self-organizing harmonic modes* (SOHMs) (Safron, 2020). The larger *Integrated World Modeling Theory* of consciousness (IWMT) (Safron, 2020), where SOHMs were introduced, combines IIT and GNW with the *Free Energy Principle* (Friston, 2019) and *Active Inference Framework* of Friston (Friston, Kilner, and Harrison, 2006; Friston et al 2017; Sajid et al 2021).

It is important to note that while such theories find useful correlations to consciousness, known as *neural correlates of consciousness* (NCC), they may not explain the mechanism by which those correlated activities or measures produce subjective experience.

Other noteworthy theories include the *Orch OR* theory of consciousness (Hameroff & Penrose, 1996a, 1996b, 2014; Hameroff, 2021), and pan-psychism (Chalmers, 2013; Goff, Seager & Allen-Hermanson, 2022).

I.3.5 More Reading

For those who wish to learn more, the *Stanford Encyclopedia of Philosophy* has excellent articles covering many topics related to this area. In particular I recommend articles on: Consciousness, The Unity of Consciousness, The Contents of Perception, Perceptual Experience and Perceptual Justification, Representational Theories of Consciousness, The Computational Theory of Mind, Neuroscience of Consciousness, Higher-Order

Theories of Consciousness, Introspection, Mental Causation, Epiphenomenalism, and Animal Consciousness. The Wikipedia article on Experience also provides an excellent summary of various concerns.

For more background on the various theories of consciousness, I recommend Seth (2022) and the Scholarpedia article on Models of Consciousness.

Many of the broad topics on consciousness require a background in philosophy to really understand them. For some readings, I recommend the Stanford Philosophy articles on Metaphysical Explanation, Phenomenology: 1. What is Phenomonology, Functionalism, and the Wikipedia article on Intentionality.

I.4 Background - Biology and Neuroscience

I shall not attempt to explain any of structure of the brain itself. However there is one important feature of the neurological aspect of brain function that needs to be described more fully, that of its dynamical systems nature. Many of the discussions later in this treatise combine simplistic functional descriptions with simple box-and-line diagrams using clearly demarked components. I want to dispel any misconceptions about the interpretation of those simplistic descriptions before they arise.

“Boxology” functional descriptions that, for example, include single individual boxes for processing and memory are simply wrong (Dennett, 1991, p. 271). The brain does not operate that way. Processing is a distributed product of many different systems (Vision: Spillmann et al, 2015; Livingstone and Hubel, 1987; Zeki and Shipp, 1988. Attention: Luo and Maunsell, 2019). Even memory may be a distributed effect, with a single memory actually being produced by different systems focused on the various different modalities that make up the memory (Postle, 2016; Carter et al, 2019, p. 156-159). However, simple functional descriptions give us something understandable to work with while we’re trying to figure out the broad ideas. Remapping that onto a dynamical systems substrate can come later.

Nevertheless, it will helpful to the reader to have some intuition of the brain’s true underlying dynamical systems nature. To that end, what follows is a brief description of one attempt at explanation, known as *Predictive Coding*.

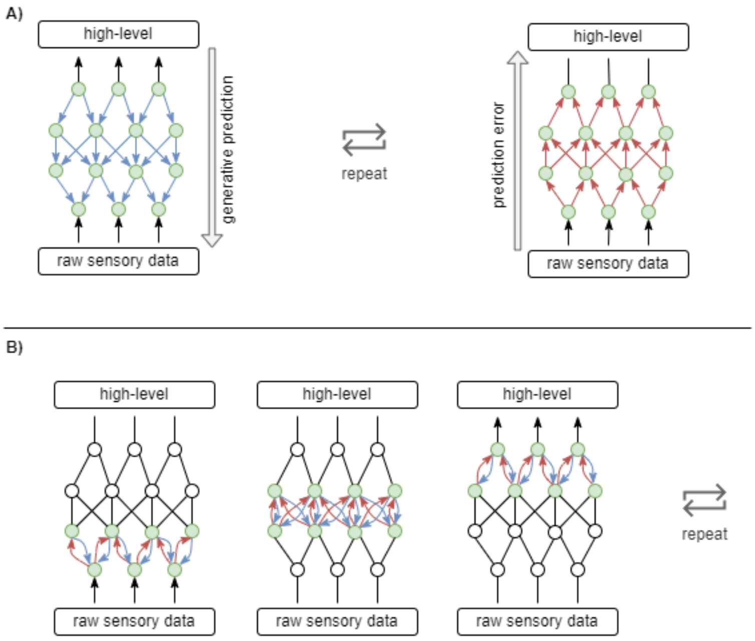
I.4.1 The Predictive Coding theory of brain function

A traditional conception of brain processing of senses can be characterised as “perception by representation”: that the brain attempts to use the senses to accurately represent what is observed. A typical assumption associated with that characterisation is that sensory perception is a largely “feed-forward” process: raw low-level sensory signals are hierarchically interpreted into higher and higher-level representations, eventually identifying specific objects, their boundaries, and other properties such as location, pose, and motion (Buckley et al, 2017; Walsh et al 2020).

An alternative conception is characterised as “perception by inference”: that the brain attempts to infer the (hidden) state of the environment, known as the *latent state*, from sensory signals. In this conception, rather than predicting a representational model that *correlates* to the sensory signals, the brain attempts to model the underlying structure that *caused* the sensory signals (Friston, 2005). Furthermore, rather than producing this inference within a single forward pass, it is derived through an iterative process employing both feed-forward and feed-back signals (Rao and Ballard, 1999).

Predictive Coding is one such theory. It holds that much of brain function is the result of such inference (Friston, 2010; Clark, 2013 and 2019; Kilner, Friston, Frith, 2007), including not just perception but also action generation (Friston, 2010). The explanation stems from the observation that Bayesian inference is analytically intractable for most problems, but can be solved through the *empirical bayes method* by *factorising* the problem space (Buckley et al, 2017). A so called *generative model*, which models the the causal structure from environment state to sensory signal, can be approximated by factorising it in three ways (Buckley et al, 2017; Millidge, Seth, Buckley, 2021). Firstly, the state of the environment at a given moment in time can be factorised as a multi-variate combination of independent gaussians. Secondly, the time-dependent dynamics of state can be factorised into the current value, its first-order derivative, its second-order derivative, and so on. Thirdly, the unknown relationships between latent causes can be modelled and learned as a hierarchy of layers (Friston, 2008), with each successive layer acting as a generative model of the layer before.

This factorisation can be distributed across the neural structure of the cortex (Mumford, 1991; Rao and Ballard, 1999), with activity of each individual neuron representing the mean of the gaussian distribution of the particular variable that it models (Buckley et al, 2017), and possibly also representing the variance (Feldman, Friston, 2010). Some aspects of this 3-dimensional factorisation have already been identified within the structure and activity within the brain. For example, the hierarchical nature of brain processing can be seen in the way that the primary visual cortex processes visual information at a lower level of representation than the visual association area [citation], and similarly for the primary and secondary somatosensory areas (explained in *Neuroscience Online, “Chapter 3: Motor Cortex”*). Additionally, there is some evidence that neurons within the cortex are formed into *columns*, where the neurons in each column together model a single multi-dimensional variable (Mountcastle, 1997), likely modelling the full covariance matrix between those dimensions [citation].



- ***Predictive coding in action.*** *A) Top-down computations predict sensory data based on high-level priors. Bottom-up computations indicate prediction errors, leading to updated priors. The process is repeated until prediction errors are sufficiently reduced. B) The same top-down and bottom-up interactions occur at the micro level between adjacent pairs of the hierarchical layers, and at the macro level across the entire system.*

Counter-intuitively, under the theory of predictive coding, the forward computational direction from sensory signal to higher-level representation, also known as the *bottom-up* calculation, conveys only prediction error. The main computation is performed by the generative model during *top-down* computation, ie: in backward direction from high level representation towards low-level sensory input. Each layer within the hierarchy represents a *prior* on the layer below, *conditioned* on the layer above. At time of inference, bottom-up prediction error is used to identify priors that don't fit reality, which triggers further prediction errors up to higher levels. That is eventually returned with new top-down conditioning adjusting the priors, ultimately resulting in each layer representing its best guess of the latent state at its level of representation. Over longer timescales, bottom-up prediction errors are also used to learn better generative models (Friston, 2008).

This leads to a lot of activity. A novel sensory signal is likely to immediately trigger prediction errors against priors in low-level layers that were *framed* by previous contextual information. Thus there is immediate short-range waves of generative prediction and prediction error activity (see panel B in the diagram above). In order to completely resolve the prediction errors, higher-level priors may need to be revised, resulting in long-range waves of activity (panel A in the diagram above). Activity eventually settles once prediction errors are sufficiently minimized across all layers.

Predictive coding offers an explanation of various otherwise puzzling features of perception (Millidge, Seth, Buckley, 2021), including so called "end-stopping" in visual perception, bistable perception effects under right/left-eye competition, repetition suppression, attentional modulation of neural activity, and of hebbian learning. The suitability of predictive coding as a larger theory of brain function is debated (Walsh et al 2020, and see commentaries on Clark, 2019), but the basic idea behind it may yet prove to be a good explanation of the waves of activity that we see in EEG and fRMI recordings.

Part II - Problems in Simple Synthetic Control Processes

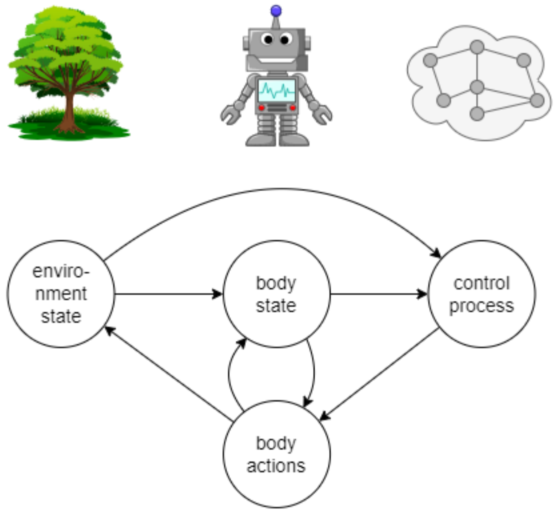
This part and the two that follow it describe a series of problems that are faced by *control processes*, computational processes that control the behaviour of the larger system of which they are a component. This part is focused on simple artificial *agents*, while Part III is focused on more complex artificial agents. Part IV takes that up a notch to look at control processes in humans. Interspersed between chapters on control process problems are the occasional *interlude* chapter that provides additional background to the discussions immediately following.

II.1 Interlude: Environment, Body, and Control Processes

I shall start by defining some terminology and concepts.

While the latter sections and chapters of this treatise focus on human brain function, the earlier sections refer to agents in a more generic sense. Here I draw inspiration from artificial intelligence (AI) research, particularly from Reinforcement Learning (RL) settings (Schmidhuber, 2015; Lazaridis, 2020). In RL research, an agent may be nothing more than a computational device running on a computer that incorporates an artificial neural network (ANN) plus a hand-coded learning algorithm. The agent may be operated within a virtual environment, simulated within the same computer that executes the agents computations. Often the hand-coded learning algorithm has direct access to ground truth information (the true state of that virtual environment and the agent's coordinates and state within that environment), whereas the agent has only specific simulated senses through which it must infer those states. In such a simulated environment, the agent may be embodied - having a (simulated/virtual) physical form; common examples include animal shape (eg: Eysenbach et al, 2019), cars (eg: Gosh et al, 2019), and robotic arms with a fixed base (eg: Nair et al, 2018; Colas et al, 2019; OpenAI, 2021). AI agents may also exist in the real world. Often in RL settings these are agents that are initially trained within a virtual environment that closely mimics the real world, and then the partially trained agent is transferred into the real world equivalent body. Humans and other animals can be considered as very complex natural biological agents. Many of the complex processes within the brains of humans have their origins in simpler organisms. Studying simple artificial agents is a first step towards understanding those processes.

For an embodied agent that exists within an environment, there are three independent but interacting components at play: the environment, the body of the agent, and the agent's collection of control processes (CPs). Illustrated in the following diagram.



- ***Relationships between environment, body, and control in an embodied agent.*** *Body actions influence the state of the environment and of the body state, while body state influences how those body actions carry out. Environment state affects body state and motivates some of the body actions. Control processes observe body and environment state and govern the state of the body via body actions.*

The environment is everything external to the agent, potentially including other agents. In simple synthetic AI settings, the environment is usually innocuous. In the real world it is highly dynamic and includes many dangers. The embodied agent that exists within the environment must monitor and predict the state of that environment so that it can a) benefit from the environment, b) modify the environment to meet its own needs, and c) avoid dangers inherent within the environment.

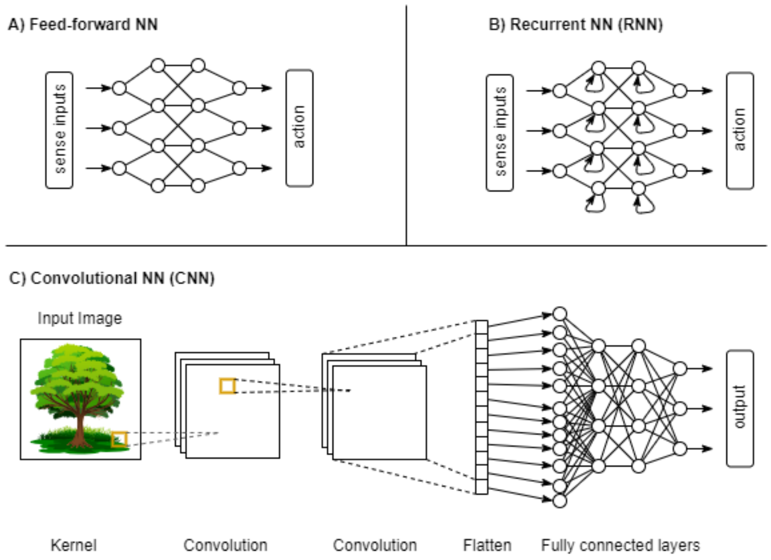
Within this treatise, *body* refers to the entire (real or virtual) physical form of the agent, including its outer surfaces and its internal structures and processes. The state of the body captures its externally visible components such as limbs, and its internal mechanical or biological processes such as those involved with homeostatic temperature regulation and energy production. Actions by the body include externally visible events such as moving a limb or producing audible communication, and internal processes such as those involved with temperature regulation and energy production. Body actions may be performed for the purpose of merely changing the body state, or they may be performed with the intent to change the environment (eg: put a plate on the table, or lift the object held by the robotic claw). There are three broad and somewhat overlapping reasons why the agent would perform body actions: i) to meet immediate homeostatic needs, ii) towards meeting a goal, and iii) exploration for the benefit of learning and gaining additional information that would be useful in the future.

For the sake of convenience of the discussions that follow we consider the control processes (CPs) of the agent as distinct from its body, particularly those that can be said as being computational in nature. In a typical AI agent, this refers to the entirety of its neural network(s) and/or other computational mechanisms of control and learning. In a biological organism such as humans, an approximation is to consider it as referring to the organism's central and peripheral nervous system. A significant component of the control processes are required to monitor, predict, and to tune the body's static state (eg: it's current location and energy levels) and dynamic state (eg: speed and acceleration of limb movement, adapting to resistance in movement due to external or internal factors).

II.2 Complexity and the need for Processing Loops

In most artificial neural network (ANN) based reinforcement learning (RL) agents today, each input is associated with an immediately produced output. This means that in an embodied agent the choice of the next physical action is made by a single pass through its ANN(s): input nodes are populated with current sensory signals, matrix operations are carried out that permute and transform those input node values through the multi-layer network of weights, and the values produced by the output nodes are immediately taken as the chosen next action. This is true even for

Recurrent Neural Networks (RNN). RNNs are *recurrent* in the sense that state from a previous pass is made available to influence the output on the next pass with the next input value. In this way, when a time-bound signal stream is fed into the RNN, it produces an output stream where each value in the output stream is influenced not just by the current input but by all inputs received up until that point. However, the RNN still produces exactly one output for every input, and each output is produced by a single pass through its network.



- ***Single-iteration Artificial Neural Networks (ANN).*** *Each of these networks produce one output for each input, via a single pass pass through the network. In the context of an embodied agent, this means that the agent has no option for further deliberation of the same input.*

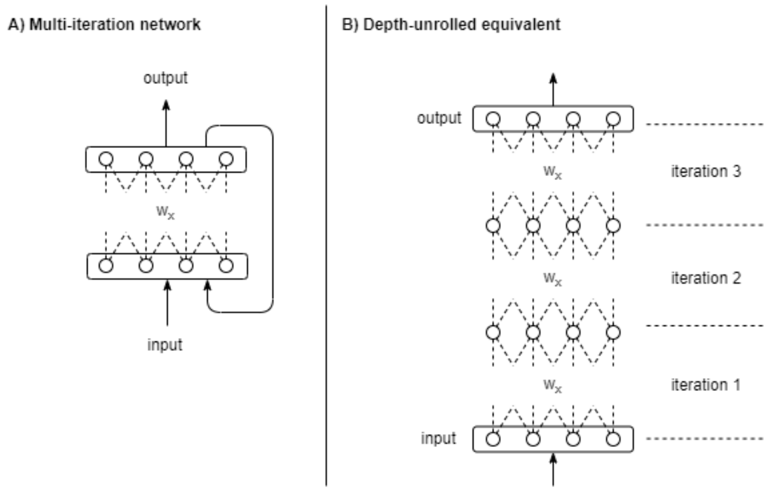
Another form of recurrency is to execute multiple passes through the same network before producing an output. This form is common in hand-rolled algorithms, where it is usually referred to as *processing loops*. When an algorithm employs a processing loop, a single output may be produced for each input, but only after a variable length delay. Some inputs may lead to updates of internal algorithm state only, without producing an output. Or a single input may produce multiple outputs. Examples abound, but one familiar to those in the AI research community is the Expectation Maximisation algorithm. It takes as input a set of data points, produces as output a set of parameters that describes the input data set, and employs multiple iterations of alternating calculation of log-likelihood expectations and parameter optimisation. The alternating expectation and parameter optimisation loop is stopped according to a *halting rule* that is either based on detecting diminishing returns in the improvement of log likelihoods or on completing a fixed number of iterations.

Some have begun to experiment with loops in ANNs. Complex results can be achieved with shallower networks when using a loop-style of recurrency (Kubilius et al, 2019; Wen et al, 2018). Loop architectures have been used to adaptively vary the amount of computation time allocated to problems, as Adaptive Computation Time (Graves, 2016), which has been suggested as an important component of next generation language decoder-encoders known as Universal Translators (Dehghani et al, 2018).

There is a practical limit to the complexity that a single-iteration processing architecture can achieve. The network can be made broader (more nodes in each layer) and deeper (more layers), but that increases the number of parameters that need to be optimised during learning. In earlier versions of ANNs, where smooth non-linearity functions such as sigmoid were used within hidden layers, the vanishing gradient problem (Hochreiter et al, 2001; Schmidhuber, 2015) meant that practical networks could not be more than a few layers in depth. Current state of the art ANNs obtain non-linearity through piecewise linear functions (Glorot et al, 2010) and enable many more layers before the vanishing gradient problem becomes an issue. While theoretical work has shown successes with as many as 10,000 layers (Xiao et al, 2018), most ANNs use around 100 layers or less. Even Chat GPT-3 only uses 96 layers (Brown et al, 2020).

Another problem with a single-iteration processing architecture is that its fixed depth implies a trade-off between the maximum complexity that the architecture can handle and the cost of training in order to cater for the average complexity of situations that the agent must cope with. Additionally, if we consider that such processing may entail multiple stages of processing, the order in which those stages is executed is fixed.

An architecture that employs multiple passes through its network can be conceptualised by unrolling its iterations into a much deeper single-iteration network, as illustrated in the diagram below. But the multi-iteration architecture has a number of advantages. Its depth varies dynamically as needed, for example that it is deeper for more complex problems. If processing is made up of multiple separable stages, the order in which those stages are executed can now be dynamically varied. It is additionally quite natural to imagine that for certain problems, some stages will be simply omitted entirely.



- ***Multi-iteration network.*** *Panel A: a multi-iteration network with the result from its output layer fed back as input. Panel B: an equivalent single pass network by unrolling the iterations into a deeper network assuming 3 iterations. Notice that in the depth-unrolled network, weights are shared between sections.*

So, it can be said that there is a limit on the complexity that can be handled by a single pass through any computational process. While that computational process can be extended with more parameters, there are practical limitations to how much it can be extended. For embodied agents, this appears as a limit on the complexity of the environmental and of their own body that they can sufficiently model and respond to within a single processing iteration. In biological terms, this practical limit is manifested in terms of both the energy costs of larger brains and in terms of the time required to reach maturity of brain function.

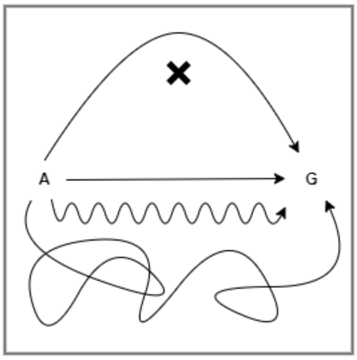
To adapt to more complex environments, an embodied agent must employ multiple iterations of processing. This enables, for example, further analysis of the environment in order to better model its state; or further deliberation of alternative action plans before proceeding. In biology, this provides scope for evolutionary pressures to trade off between a more energy hungry complex brain and a simpler less energy intensive one that might take longer to reach a decision for more complex problems. Van Bergen & Kriegeskorte (2020) make the case that recurrency is indeed employed in biology for that very reason.

The term *recurrency* can mean many things because recurrency can occur at any level. For example, in the case of Recurrent Neural Networks (RNNs) as used within AI, recurrency occurs at the level of a single neuron in order to hold state. Thus I shall continue to use the term *multi-iteration processing* in order to avoid confusion about the level at which the recurrency occurs in the context of discussion.

II.3 State Trajectories in a Multi-iteration Processor

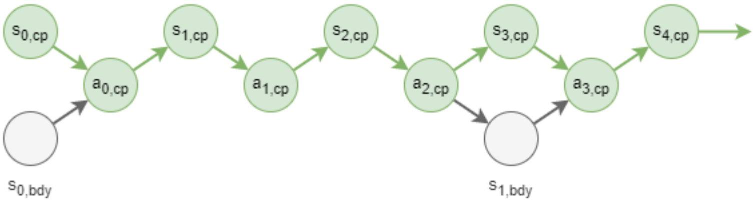
The course taken by an agent to get from a past state to its current state is its *state trajectory*. Analogous to the path taken by an agent while walking through a maze, the state trajectory describes the path of the agent through state space. Here the state space can refer to its possible locations in physical space, such as in the maze example, or to more abstract possible states, such as an encapsulation of all measurable aspects of the agent's body parts.

Not all state trajectories are good ones. The figure below illustrates a number of possible state trajectories from start state A to goal state G, while avoiding obstacle X. Each trajectory successfully reaches the goal, but they vary in other ways that may have significant impact to the agent. The length of the trajectory may indicate energy efficiency, which is important for an agent with limited energy reserves. The length may also indicate the time taken, which impacts whether or not the goal is reached "in time". The smoothness of the trajectory can be important. A jagged trajectory might indicate that the agent's physical body is moved in a chaotic way with abrupt stops and starts, causing damage to delicate moving parts from the stresses of that chaotic movement. A smoother trajectory may be easier for the agent to subsequently learn from and reason about in order to improve its later attempts; whereas a more chaotic path may add so much noise to the observations of the trajectory that the agent is unable to detect the most important patterns for such learning.



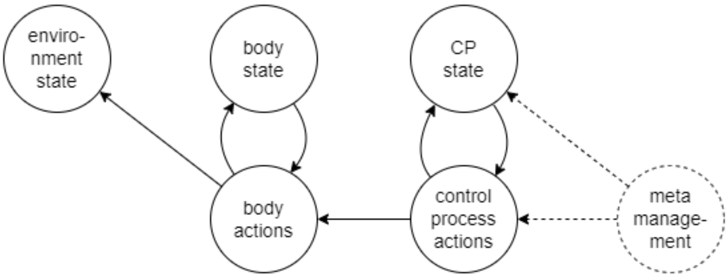
- ***Good and bad state trajectories.*** *Examples of some possible state trajectories from start state A, to goal state G, while avoiding obstacle X. The shortest and smoothest trajectory is assumed to be the best: the most energy-efficient, the quickest, the least stresses applied to the mechanics of the agent.*

In a multi-iteration control process, there are periods where the controller traverses a state space that is independent of the state of the body that it controls, as illustrated below. It needs to incorporate mechanisms to control its own computational state. Those mechanisms are referred to as *meta-management*, because they relate to management of the controller's own processes, rather than to management of the primary thing that the controller acts against (the agent's body in this case).



- ***Control process trajectories.*** *With multi-iteration processing, the control process (CP) has its own state trajectory ($s_{i,cp}$), influenced by its actions ($a_{i,cp}$). Control process actions only occasionally produce changes to body state ($s_{j,bdy}$).*

Within a learning setting, the control processes must learn to manage the state of the agent's body. Typically this is influenced by feedback received in association with the outcome of some sequence of actions. That feedback must be interpreted and used to infer the best way to optimise the parameters of the control process. In a synthetic RL setting, that feedback and parameter optimisation is performed via hand-coded learning algorithms, often incorporating back-propagation and gradient descent. In a biological organism, the corresponding learning processes may be somewhat more complex and are certainly much less understood, but their effect is the same: that parameters of the control process are optimised such that future attempts would be more successful or efficient. This is a first concrete example of meta-management.

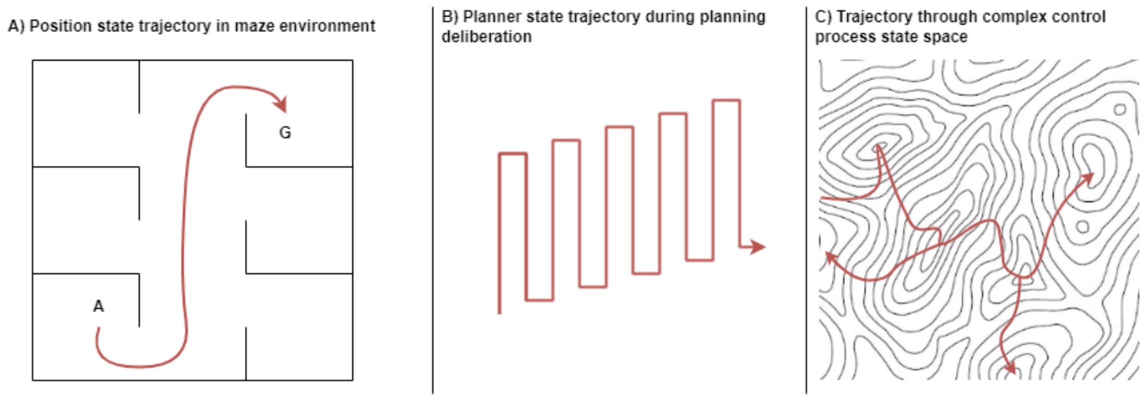


- ***Control Process with state.*** *A control process (CP) that has state needs to act to manage its own state as well as the actions and state of the body that it controls. In some cases, this may require an additional _meta-management* process. Some interactions omitted from the diagram for simplicity._

The learning processes involved with a multi-iteration processor may be very similar to those involved with multi-step bodily actions. Each body action plays out over time, with complex dynamics affecting the speed and trajectory taken. The body actions required to reach a particular target body state may involve the sequencing and coordination of multiple actuators or muscles. Feedback about the relative success or failure may be sparse - only received as certain points in time, with no specific details about the relative effectiveness of steps in between, and even then the meaning of the information and how it relates to the state trajectory may be ambiguous. Any learning algorithm resolves that by assuming some distribution of the effects of the feedback over the length of the state trajectory and by averaging over multiple attempts. Some of the feedback received by an agent can be more frequent and detailed, such as endogenous feedback produced by evolved low-level mechanisms that encourage smooth and efficient movements.

For simple control processes, those same simple mechanisms can be applied to parameter optimisation, ultimately improving the control process state trajectories. For example, the same endogenous low-level feedback processes can encourage efficient CP state trajectories by attempting to minimise the number of CP actions that occur without body actions, provided that it doesn't noticeably degrade the quality of the body state trajectories. Likewise, they may encourage "smooth" CP state trajectories in order to avoid disorganised chaotic processing.

But the simplicity of these suggested low-level CP state controls limits the capacity of the control process. Some computational problem spaces will require much more extended computational time with much more divergent state space trajectories. If a real physical world includes not just straight lines, but obstacles, walls, mazes, and other complex environmental constructions involving complex sequences of actions, then so too might a "computational world" that an advanced control process might have some need to operate within. This is illustrated in the following diagram, which contrasts the physical position trajectory through a real-world maze (panel A) against trajectories through two hypothetical CP state spaces (panels B and C).


A) Position state trajectory in maze environment
G
A
B) Planner state trajectory during planning deliberation
C) Trajectory through complex control process state space

- ***Complex state trajectories***. *A) An example of agent position state trajectory as the agent navigates a maze environment. B) A hypothetical 2D representation of the trajectory of the internal state of a planner as it considers different possible paths for navigation within the maze environment. Here, pairs of upward and downward sub-trajectories represent the forward and backward pass of each considered path, and the general motion from lower-left to upper-right signifies the planner progressing as it finds increasingly better solutions. C) A hypothetical 2D terrain map of a complex computational state space, and some trajectories that a search algorithm may take in attempting to find the globally best solution.*

An example of a complex CP state space is illustrated by a path planner. For navigation through a maze, an agent can use a learned model of the maze to run multiple simulations of possible paths before taking action. So, while the output of the path planner is to produce a trajectory through physical space, the planner itself progress through a trajectory of state space in the action of finding that physical path. Various algorithms exist for path planning but they often involve some combination of forward and a backward pass, planning or refining a plan from the start towards the goal and again from the goal backwards towards the start. Each simulation that is attempted is compared against the prior best path and eventually the planner picks a single best path, according to some measure of "best". Panel B in the diagram above illustrates a cartoonised state space trajectory of the planner in action, with its forward and backward passes while it gradually identifies better and better paths.

The planning algorithm is typically hand-written by AI researchers and thus progresses in a fairly linear fashion towards its goal. However, planning algorithms are complex, there are multiple ways to do it, and researches are often finding improvements. In a biological setting, path planning is something that we learn to do, and learning is fraught with mistakes and inefficiencies. In practice, the CP state space trajectory of a biological path planner is probably something more akin to the trajectory in Panel C of the diagram above.

In conclusion, a multi-iteration processor requires meta-management. For the simplest multi-iteration processors, meta-management may be in the form of simple parameter optimisation algorithms applied during an offline learning phase. For more advanced multi-iterations processors, a much more advanced and active form of meta-management may be required, one which might model the behaviours of the first-order processes and which might even have comparable complexity to the first-order control process itself.

The next part looks in more detail at some possible forms of meta-management.

Part III - Problems in Complex Synthetic Control Processes

We now turn our attention to deliberative artificial embodied agents. The reasoning here follows a "design stance" where we look at what problems might need solutions if we were to try to design these systems ourselves. Heavy inspiration is taken from current AI techniques.

III.1 Meta-management in Deliberative Systems

Why might we need to add meta-management processes to connectionist architectures? Deep AI techniques have had many successes of late (Lazaridis et al, 2020; Brown et al, 2020). However, these networks still lack some of the most basic adaptive capabilities that we see in many biological organisms. A key feature lacking in AI today is *deliberation*. Deliberation can be thought of as an extension of multi-iteration processing to more human-like thought that incorporates modelling of multiple problem domains, selection of goals, the ability to break problems into smaller sub-goals, and the ability to select between multiple strategies for problem solving.

A number of potential control problems have been identified in systems with such deliberative capabilities (Beaudoin, 1994):

- **Oscillation between decisions.** Wasteful re-assessments of decision points, leading to a meta-stable (oscillating) but stagnant state (ultimately achieving nothing useful).
- **Insistent goal disruption.** Repeatedly getting distracted by competing goals that have been previously disregarded.
- **High busyness.** Attempting to multi-task between too many goals, leading to poor outcomes.
- **Digressions.** Choosing to deliberate over some sub-goal, and then loosing track of the "big picture" by forgetting to return to the overarching goal.
- **Maundering.** Getting stuck deliberating over the details of a goal without making a decision.

The following subsections list some specific ways in which meta-management plays a part in computational systems, with a particular focus on connectionist implementations. Many of these examples highlight areas that have existing solutions for the most simple cases, but which require more research to cope with the more complex cases.

III.1.1 State Trajectory Control during Body Action

Actions by an embodied agent occur over time. During the time it takes for an agent to move its arm through space from the arm's initial position to target position the agent will make many observations about the environment and body states. The agent's goal and action plan must be relatively persistent during that time, otherwise the agent's behaviour will be chaotic, with rapid goal and action changes.

Thus, while the computational control process manages (controls) the trajectory of its body state, it must also meta-manage the trajectory of its CP state (eg: the given goal and action-plan at the time). In this case, the agent's CP state must to some extent resist change influenced by new observations.

III.1.2 State Trajectory Control during Multi-iteration Processing

As introduced in chapter II.3, during multi-iteration processing the control process navigates through computational state space, without performing body actions.

This state trajectory needs to be managed just the same as for the body state trajectory. In order to maintain stability the agent needs to i) observe the CP state trajectory, ii) apply some objective measure to decide upon the relative effectiveness of the trajectory, and iii) act to change the trajectory if a better one is available.

III.1.3 State Trajectory Control during Iterative Inference

A special case of multi-iteration processing is that of *iterative inference*, where the control process takes multiple iterations to interpret some input signal. Here a representation of the sensory input signal may need to be held persistent for the duration of the inference, even if the original input signal has ceased. For example in an animal context where a fleeting glimpse of a potential predator has been observed but that observation needs re-consideration before being certain.

In that case, some portion of the state must be held stable, while the rest is free to change significantly. This requires some form of meta-management. In simple cases that may develop as an implicit product of the learned connectionist control process. In more complex agents, such as those with attention, explicit meta-management is required.

III.1.4 Objective learning

How does a continuously learning embodied agent know which actions are better than others? This decision is tied to the agent's *objective*: it's ultimate goal that influences all other goals. For example, to eat and stay healthy in order to survive. Or to produce as many paperclips as possible (Bostrom, 2003, 2014; Gans, 2017). If the agent is not pre-configured with its objective, then it must learn that objective.

An agent in the human world requires the use of inedible metal tokens (coins), which are used in complex ways for the purpose of life preservation. The involvement of such an inedible metal token as part of some process (eg: doing a job and being paid) does not necessarily immediately result in a life sustaining outcome. Thus, without any other information, it is hard for the agent to learn the relationship between that inedible metal token, the processes that it must be involved in, and the life sustaining result. This is known in the AI community as *sparse feedback*, and it poses a particularly difficult problem for continuously learning agents (Armstrong et al, 2020).

Another problem for a continuously learning agent is known as the *exploration-exploitation dilemma* (Kaplan & Friston, 2018). The agent gains knowledge about its world and itself by exploring places and things, and by experimenting with novel behaviours. When the agent needs to achieve a goal, it may by able to achieve the goal by *exploiting* its existing knowledge, but it may be able to achieve that goal in some better way if it were to *explore* more first. However, further exploration may not yield better results, and may even put the agent in harms way. The dilemma concerns how the agent chooses between exploration and exploitation at any given moment.

Sparse feedback and the exploration-exploitation dilemma make objective learning difficult. One solution is for the agent to build high-level simplified models of its environment, itself, the behaviours it can perform, and how those behaviours influence different outcomes. High-level models have fewer degrees of freedom than found in the raw first-order signals. This means that the models can be built up from fewer examples, and they are easier to change as learning progresses. These models become the agent's "knowledge", and somewhere within that knowledge a continuously learning agent builds a structure that ultimately governs its behaviours and goals – that is, an objective that it infers over time.

For example, some success has been shown using model-based RL to learn the structure of goals and combining that to train a model-free RL agent (Krueger and Griffiths, 2018). The authors of that paper suggest how objective learning feeds into behavioural automatization: "our knowledge of the world doesn't just provide a source of simulated experience for training our instincts, but that it shapes the rewards that those instincts latch onto." (p. 1)

Objective learning can be seen as another form of meta-management for a number of reasons. Firstly, even with a learned objective function, first-order behaviour continues to use low-level representations, where the high-level objective function is used only in the generation of endogenous feedback as part of subsequent behaviour learning. Thus the learning and use of the objective function is a second-order process. Secondly, learning of objective functions is hard (Armstrong et al, 2020), and in complex environments it may too require deliberative involvement.

III.1.5 Mode control

A number of seemingly distinctly different behavioural outcomes share a single principle, referred to here as *mode control*. Mode control involves a decision being made between multiple alternatives and that decision influencing the way in which a subsequent process or decision is carried out.

Examples of mode control include:

- **Strategy selection.** Choosing between multiple previously learned strategies (ie: sequences of processing) that may be useful for solving the particular problem at hand. The selected strategy may affect goal selection and/or it may bias the outcomes of certain processes.
- **Goal selection.** Choosing the next target state, for example based on an interpretation of external signals, or from weighed up options in an ambiguous situation. The chosen target state thus becomes the reference point for inference of appropriate actions.
- **Context.** Context plays a huge part in the interpretation of sparse signals. A visual patch of yellow with dark spots, when seen in the Savannah, may indicate a leopard, but the same patch on the beach may simply indicate sea shells. Context is not always available from direct sense of the external environment. Most perceptual interpretation also receives context from short-term and/or long-term term memory. Thus meta-management plays a role in inferring that context from a mixture of current sensory signals plus memory.
- **Attention.** As suggested in Part II, the bandwidth of any computational system is limited, and the complexity of the environment may exceed the agent's computational bandwidth. One solution is to focus on only the most salient features of the environment, ignoring the rest. What the agent considers salient differs depending on things in the environment, the context in which the agent is operating, and on the agent's knowledge. Attention has a significant impact on the processes executed by the main control process – a change in attention changes the input to the control process, and thus to its output.
- **Exploration vs exploitation.** Already introduced in an earlier discussion on objective learning, the choice between exploration and exploitation affects sub-goal selection and the actions taken by the agent. For example, an agent can measure its actions according to the certainty associated with the predicted outcome; in an exploration mode, the agent will prefer actions that have less associated certainty (Kaplan & Friston, 2018).

III.1.6 Mode identification

For mode selection to be possible, the agent must identify the modes that can be selected from, whether they be discrete or a range of continuous values. This requires two important features of the meta-management system: i) that it has sufficient access to *observe* the things that it needs to control, the outcomes of the control, and the values used in control; and ii) that it can *model* those observations and later use that model to choose the control mode.

In some cases this may involve modelling the relationships between different components of the first-order control process. For example, Timmermans et al (2012) identify that meta-cognitive processes appear to learn cause-effect relationships between the supplementary motor cortex and the primary motor cortex. They suggest that this model is used to infer the most appropriate signals to send from higher order areas.

III.1.7 Distributed cooperation

Some theories of brain function describe the brain as having multiple independent processes that are in constant competition. For example the biased-competition theory of attention (Turova and Rolls, 2019) assumes multiple processors, each interpreting their own local sub-scene out of a larger visual scene. It pits those different sub-scene interpretations against each other, until a single unified scene interpretation wins out. Global Workspace Theory adds the option for groups of otherwise competing processes to cooperate (Baars, 2021; Baars and Franklin, 2007), with the outcome being that a group of processes can collectively win the competition for attention when each process individually would loose.

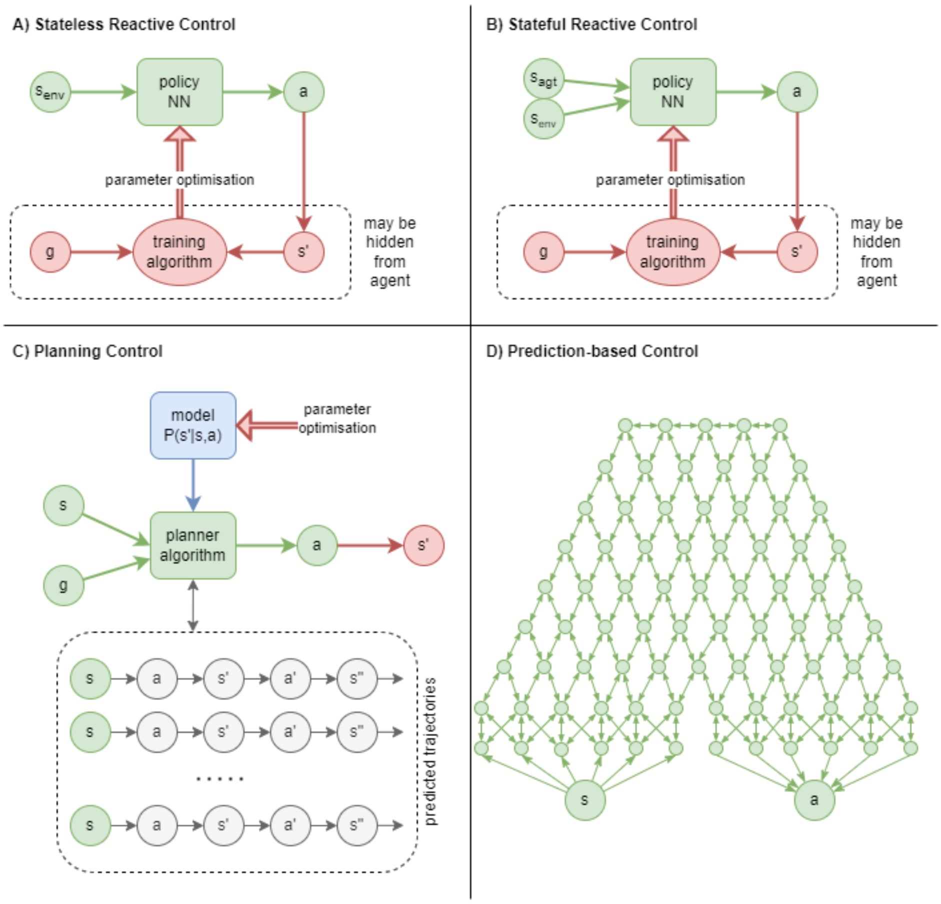
This is an obvious situation in which meta-management has a part to play – in managing the competition and cooperation between those processes. One possible mechanism is the same as discussed in the section above on Mode selection – by adjusting priors.

Curiously, as observed by Baars, humans don’t appear to have experiential awareness of this competition / cooperation process. Rather, we observe only a sort of stabilized outcome. So perhaps this is a first-order concern, at least in humans. But in principle it could also be a meta-management concern.

III.2 Interlude: Mechanisms of First-order Control Processes

In order to elaborate more fully on the possible implementation details of meta-management, I need first a more detailed description of the first-order control processes that it might operate against. This chapter examines a small selection of different architectures for first-order control.

In AI, a common scenario is to train a simulated robot to navigate within a virtual environment. It is common to incorporate an artificial neural network (NN) as part of the control system, known as a *policy* network (Sutton and Barto, 2018), that chooses each next action, and to use Reinforcement Learning (RL) to train that neural network. Several broad options exist for the architecture of the control algorithm. These can be framed as a progression of improvements that lead to increasingly better adaptability. The progress is illustrated in the diagram below, and outlined in what follows.



- ***Control algorithms.*** *A graduation of increasingly adaptive control algorithms in an embodied controller. Some training algorithm connectivities have been simplified or omitted for the sake of simplicity. A) A so called “model-free” _policy* network that produces action without awareness of the dynamics of the agent itself, with an additional training algorithm that optimises its parameters. B) Model-free policy network that is aware of the state of its body. C) A so called “model-based” planner that simulates entire trajectories in order to choose the best action, with or without awareness of its own state. D) Hierarchical prediction-based control process that may not need explicit meta-management._

III.2.1 Stateless reactive control process

In the most simple case, the policy network simply predicts the best action given a sensory input about the environment. The policy network is a learned approximation of a probability distribution $P(a|s)$, representing the best action for each given state. A typical example is a robot car with very simple choices of actions: stop/go, left/right/forwards. The robot observes its surroundings using, for example, vision, sonar, or laser.

The training is done by the RL algorithm, which is hand-written by researchers. The RL algorithm is really the thing with the smarts here. It has access to much information that the agent does not. For example, it knows what the training goal is (eg: to navigate a race course), it knows the true position and orientation of the robot at all times (the *ground truth*), and it knows how “costly” each robot action was relative to the goal. From that information it computes a *loss function* as the time-discounted sum of those action costs (Sutton and Barto, 2018, p. 54-62). The loss function effectively computes a unit-neutral measure that has the characteristic that lower loss values indicate behaviour is closer to ideal, without needing to know exactly what ideal is. That loss function is then used via gradient descent and back propagation to update the weights within the NN. Over many training iterations the policy network *converges* towards producing ideal behaviour.

In this simplest version, the dynamics of the agent itself are also ignored. For example, a car robot is assumed to produce the required action immediately, the policy has no way to take into account how quickly the steering angle can be changed. In the earlier days of neural network research, including deep learning, many of the problems addressed were of this very simple form.

III.2.2 Stateful reactive control process

In a first small improvement, we give the agent *some* information about its own physical state, enabling the agent to cope with its own dynamics. For example, a car robot may have information about its current steering angle and speed of its wheels. Through multiple iterations of RL, the policy network learns to take the agent's physical state into account when predicting the next action, effectively incorporating knowledge of its own dynamics.

Both this and the former kind of agent are known as *model-free* (Sutton and Barto, 2018; Lazaridis et al, 2020), because they lack an explicit causal model of the space in which they operate. For example, they have no ability to predict the expected outcomes of actions and to detect when actual outcomes diverge from expectation. In more realistic real-world scenarios, there is a particular scalability problem that arises. In the real world the best trajectory is a function not just of the initial state, but also of the goal at the time. Here the policy network effectively must learn a probability distribution $P(a|s,g)$. The potential range of goals could be large, and thus the dimensionality of the distribution is exponentially larger than $P(a|s)$ alone. Furthermore, the policy network only generalises to new goals that are similar to ones seen at training time.

III.2.3 Planning control process

A significant improvement to adaptability and reduction in training time is seen in AI research by incorporating a causal model with a *planner* into the control process, known as *model-based* RL (Sutton and Barto, 2018; Lazaridis et al, 2020). Here, a model is learnt that predicts the effect of an action on the state of the environment and on the agent's own physical state. Rather than predicting a single best action, the agent simulates a trajectory from its current state via a sequence of actions to see where it may end up. It does that multiple times with multiple trajectories. Finally it chooses the best trajectory, and the executes the first action from that trajectory. Then it repeats the whole process again for the next action step. The rest of the computed best trajectory is typically discarded at each step.

Compared to reactive control processes, such a solution has a significant advantage in the real world where the most appropriate action depends on the goal at the time. The planner learns a model $P(s'|s,a)$. The model is not parameterised by goal, as the goal needs only be considered at the time of planning. Thus the same model suffices to work with many goals, including goals never seen during training. Additionally, during training the same model parameters can be updated from experience regardless of the goal that was being followed at the time.

A key feature of this control algorithm is the use of *simulation*, to consider different possible trajectories. A simulated sequence of actions can often be run orders of magnitude faster than actually carrying out the same sequence of actions. Negative outcomes in a simulation have no impact on the agent except for the time spent running the simulation.

Unfortunately, this planning approach can also be computationally inefficient. In a naïve implementation an extensive amount of computations are performed that are completely discarded and repeated again. It also doesn't scale well into long trajectories of high-resolution. Various optimisations exist. One particularly relevant optimisation is to use a planner to produce a course-grained high-level trajectory, and to use a reactive control process for the fine-grained motion control at time of action execution. In such a setup, the next point in the high-level trajectory sets a dynamic goal that is fed into the reactive control process as an additional input (Zazaridis et al, 2020, p. 1438-1440).

Note that the structure and algorithm of the planner is far from given. It too has many parameters. In AI research the planning algorithm is typically chosen beforehand and hand-written. In a biological setting, the planning algorithm is likely learned from experimentation and instruction.

III.2.4 Prediction-based Control

A fourth control strategy exists that doesn't fit neatly into the supposed sequence of adaptive improvement, but it needs to be described because it may be the significantly more biologically plausible strategy, and this is as good a place as any to describe it.

A feature of the descriptions of the first three first-order control processes is that those descriptions incorporate a lot about how the control processes are meta-managed. For reactive control processes, the first-order process that actually drives behaviour is just the policy network. The RL learning algorithm is meta-management. It should also be pointed out that there are many different RL algorithms that have been developed, and many more improvements under active research (Lazaridis et al, 2020). The description of the model-based control process included a hand-written planning algorithm that simulates trajectories (how?) and then chooses (how?) one of those trajectories over the others and then finally chooses (how?) one action out of the chosen trajectory. Those interspersed "how?" comments identify points where many different algorithms exist. These are all steps that don't produce behaviour but rather are involved in determining how to produce the behaviour - meta-management again. It should be becoming clear that there are a great many variations and choices to be made in *how* to meta-manage a first-order process.

A question worth considering is why these systems require meta-management. One answer is that they require meta-management in order to *converge*. The RL algorithm is what causes the policy network's weights to be adjusted so that over time the behaviour of the agent more closely approximates our desired "ideal" behaviour. The processes of the RL algorithm are complex and operate independently from the policy network.

An entirely different strategy may, at least for simple cases, avoid the need of explicit meta-management by being *inherently convergent*. This can be achieved through *prediction-based control*. The reactive control process infers one direction of causation - from state to action - but ignores the other direction - from action to state. A prediction-based control process learns to infer in both directions by continually attempting predictions and using prediction errors to improve its predictions. Before performing an action, it predicts the expected outcome, and then observes the actual state that results. Likewise, it infers actions by identifying a target state and predicting the action that will get it there. When it observes the actual state that arises, it uses that to adjust the prediction of action that it *should have* taken. That is made possible by modelling the world, again learned through prediction and prediction errors. Effectively this approach combines the best aspects of the reactive and planning based controllers.

It also needs to simulate trajectories, but where the planner above does that through iterative analysis, the predictive process does that through hierarchy. Multiple layers of granularity are built up on top of each other, with the top-most layer representing the most high-level abstraction. The interactions between layers is again one of bidirectional prediction. Think of it like how you might navigate to a distant location using a map in a new city. The first step is to draw a straight line from your start to your target, effectively assuming a single action. That is subsequently refined into finer and finer granularity as finer and finer detailed aspects of the terrain are taken into consideration. That can be solved iteratively. It can also be solved in a hierarchical architecture.

Additional explicit meta-management processes are not required in the system described above because each individual component inherently converges towards its ideal state: there is nothing more ideal than being the best at predicting what it needs to predict. In a neural network, learning from prediction errors occurs locally, without requiring an overarching system to compute and backpropagate weight changes. The system as a whole converges due to the convergence of all of its parts, and due to mutual interactions between them.

An example of such an architecture is that of Predictive Coding, described in chapter I.4. This approach has been generalised further into the concept of *free-energy minimisation* (Kaplan & Friston, 2018; Friston, 2010; Friston et al, 2017; Friston et al, 2006). Prediction error is considered as one kind of "free energy". Another kind is uncertainty in the prediction, ie: lack of confidence in a prediction, prior to measuring the error. On a longer timescale, another source of free energy is expectation of negative outcomes in the future, or just uncertainty about the future at large. It has been theorised that significant behavioural complexity can be achieved through free-energy minimisation alone, and that it plays a large part of driving behaviour in biological organisms. For example, free-energy minimisation balances between exploitation and exploration. Uncertainty leads to exploration. Certainty of negative outcomes (eg: hunger) leads to avoidance of those negative outcomes (eg: by finding food). Uncertainty in negative outcomes leads to cautiousness.

It remains to be seen how complex a system can get via such a hierarchical mutual-predictive process alone, and whether explicit meta-management processes may also provide benefit.

III.3 Interlude: Planning in AI and Biology

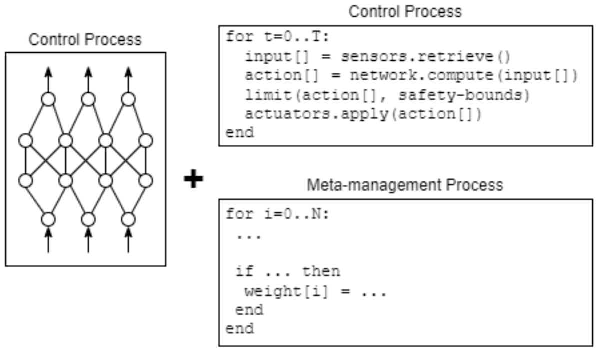
III.3.1 Hand-written Control Processes in AI vs Biologically Plausible Control

In the above I have taken substantial inspiration from contemporary AI research into artificial connectionist computational paradigms. I shall continue to use AI research for inspiration and to draw comparisons. For that comparison to be legitimate I need to address a major difference between contemporary AI's use of connectionist architectures and that of biology.

In contemporary AI, many of the processes are hand-written *imperative code* - written in a chosen computer language that executes steps one at a time. It could be argued that in most AI agents the most important control process is that very imperative code, rather than the neural network. The connectionist part of the architecture is only a small component of the overall solution. Somewhat confusingly, the AI community refer to that artificial NN as a "model" regardless of whether they are implementing model-free or model-based RL or some other form of AI. However this reference as a "model" is appropriate - it is not a real neural network, but a data structure on a computer that represents certain simplified aspects of how such a neural network might be structured. As it is just a data structure, something else must *execute* it. And the thing that executes it is the hand-written imperative code.

For our purposes, the significant example applicable to many implementations today is that the main cyclic processing loop - the thing that takes the current sensory input, processes it, and then actions the output - is governed entirely by that imperative code, rather than by the NN. The decision of whether or not to feed the output into actuators is governed by hand-written imperative code. The decision of whether to start the loop in the first place is governed by hand-written code. The decision of *when* to stop the loop is governed by hand-written code. That last, most important decision, is not based on anything related to the connectionist model, nor the outputs of the connectionist model, nor even of any sense of "lifespan" of the agent. Rather, it is typically related to either a training "epoch" (an entirely arbitrary time period, usually of only a few seconds) or to a "buffer size" (ie: of how long before some data buffer becomes full and where those hand-coded meta-management processes must kick in).

The following diagram illustrates some of the components used in AI RL agents.



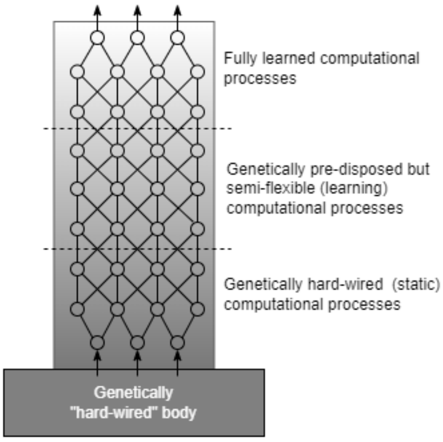
```
for t=0..T:
  input[] = sensors.retrieve()
  action[] = network.compute(input[])
  limit(action[], safety-bounds)
  actuators.apply(action[])
end
```

```
for i=0..N:
 ...

 if ... then
  weight[i] = ...
 end
end
```

- ***Control and meta-management processes in contemporary AI.*** *A mixture of imperative code hand-written by AI researchers plus connectionist models implement the control and meta-management processes of the agent. A hand-written control process executes a loop that repeatedly takes input data, uses the neural network "model" to compute output data, and applies that output against whatever actuators lead to action. Separately, another hand-written process performs meta-management by calculating and applying weight updates to the neural network model.*

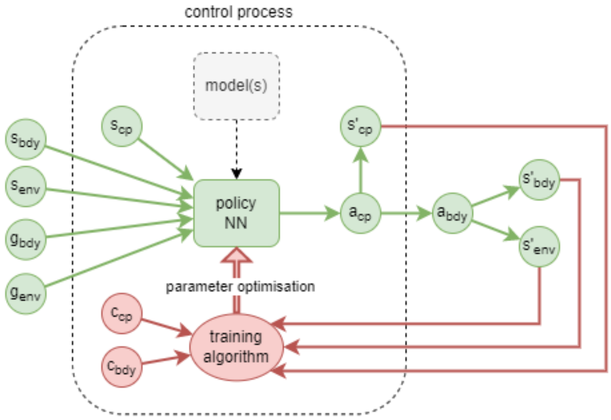
In a narrow way there are aspects of the above that are not entirely dissimilar to biology. Biology employs aspects that are "hard-coded" by an individual's genetics. But significantly more of the first-order control and of meta-management processes use connectionist solutions that *learn*. In that way, contemporary AI today avoids some of the difficulties faced by biology, but it also misses out on the opportunities provided by that flexibility. In practice, as illustrated in the following diagram, biological brains have a graduation of inflexible to flexible computational processes. At the most low-level there are entirely genetically hard-wired processes such as reflexes. At the most high-level there are processes that are entirely learned through experience. And in between there are computational processes with strong genetic influence, but which can be adjusted over a lifetime through experience.


Fully learned computational processes
Genetically pre-disposed but semi-flexible (learning) computational processes
Genetically hard-wired (static) computational processes
Genetically "hard-wired" body

- ***Control and meta-management processes in biology.*** *Layers of increasingly flexible and learning connectionist networks run on top of more static (genetically pre-determined) connectionist networks, all of which are "hosted" within a largely genetically "hard-wired" body.*

III.3.2 A Biologically Plausible Planning Control Process

In order to more accurately motivate further discussions of meta-management in a way that might be applicable to biology, we need a control process that is itself more biologically plausible. A biological version of a trajectory planner provides a good such example. It is reasonable to consider that an ability to do planning is very important for biological agents just as much as it is for artificial agents. Biological agents are not imbued with a fully-formed pre-built planning engine. And even if such a thing was partially formed it would still be constructed using the same kind of neural network structures found throughout the rest of the brain.



- ***A biologically plausible planner.*** *A NN-based policy executes as a multi-iteration control process (CP). CP state (s_{cp}) represents everything that a planner may need to hold onto, including the partially complete trajectory being considered at the time and information about other trajectories already attempted. The policy predicts control process actions (a_{cp}) that change the control process state (s_{cp}). CP actions sometimes also cause body actions (a_{bdy}) that lead to new body state (s'_{bdy}) and environment state (s'_{env}). One or more separately trained models could feed into the policy, or the policy itself could effectively represent those models. The training algorithm optimises policy parameters in order to achieve the right body state trajectories while meeting CP state constraints (c_{cp}) and body state constraints (c_{bdy}). Parameter optimisation of models not shown.*

The above diagram presents a rough structure of a possible biologically plausible planning control process. The hand-written planner is replaced by a policy network that controls the computational behaviour of the planner in exactly the same way that a policy network would normally control the outwardly visible behaviour of an embodied agent. The same policy network also controls those outwardly visible behaviours. Actions produced by the control process (a_{cp}) are for the most part hidden - they modify the state of the control process (s_{cp}) without producing any outwardly visible behaviour. Depending on certain properties, some CP actions additionally produce body actions (a_{bdy}), this being the key goal of the planner. Those body actions have the effect of changing the state of the agent's body (s_{bdy}) and of the environment around it (s_{env}).

The policy network is trained through RL to produce suitable CP actions and body actions. On average the right body actions should be generated that produce suitable body state trajectories from current state to goal. In order to avoid unreasonably lengthy deliberation, the number of CP actions that don't produce body actions should be minimised, without significantly reducing the appropriateness of body action. In a biological setting, suitable low-level constraints would be optimised through evolution in order to achieve those outcomes, such as in the form of "impatience".

A traditional AI planner incorporates one or more models in order to predict the effect of actions. We assume that a biologically plausible connectionist planner incorporates such models too. For example this could perhaps be along the lines of recent work that incorporates aspects of both model-based and model-free learning, enabling both the policy and the model to be learned as neural networks in so called "Value Iterative Networks" (Tamar et al, 2016) and "Value Prediction Networks" (Oh et al, 2017).

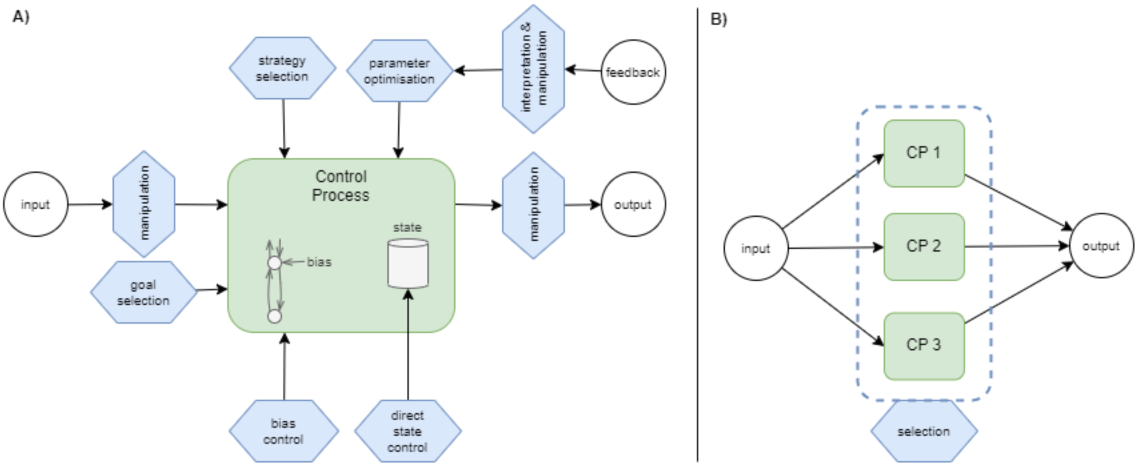
For the sake of simplicity we shall assume that the processes of the RL training algorithm are somehow hard-coded through evolution. There are many other details that could be considered but they don't affect what shall be discussed subsequently and so we shall stop at this high level description.

One feature that may not be initially obvious is that the architecture described here is not limited to trajectory planning. We assume that it will learn to perform trajectory-based planning in order to improve its accuracy of body action, but it is not restricted to that. It could also produce other kinds of behaviour, including iteratively refined interpretation of observed state, other AI algorithms such as Expectation Maximisation (EM), or more human-like behaviours like problem solving. The same policy network could even exhibit multiple behaviours, depending on the need at the time.

This planning controller exhibits many of the potential problems with multi-iteration processing that have been discussed so far. The next three chapters look at ways of adding additional meta-management to reign in those problems.

III.4 Meta-control Options in Meta-management

What options are available to meta-manage a first-order control process? This includes for the biologically plausible control process described above, and for other potentially biologically plausible architectures such as hierarchical predictive coding. We have already mentioned parameter optimisation. Here we shall look at some other options. The goal is not to provide an exhaustive list, but to build up a case for the need to observe the control process and to draw out what kinds of observation might be needed.



- ***Meta-control options.*** *Identification of some ways in which a control process may be meta-managed. A) A control process that takes sensory input, incorporates goal and processing state and bayesian-style inferences with biases in order to produce outputs. The result of actions produce feedback that is used in parameter optimisation. All of those components pose as avenues for meta-control. B) Where multiple control processes "compete", some mechanism needs to choose the winning outcome.*

Let's examine some options:

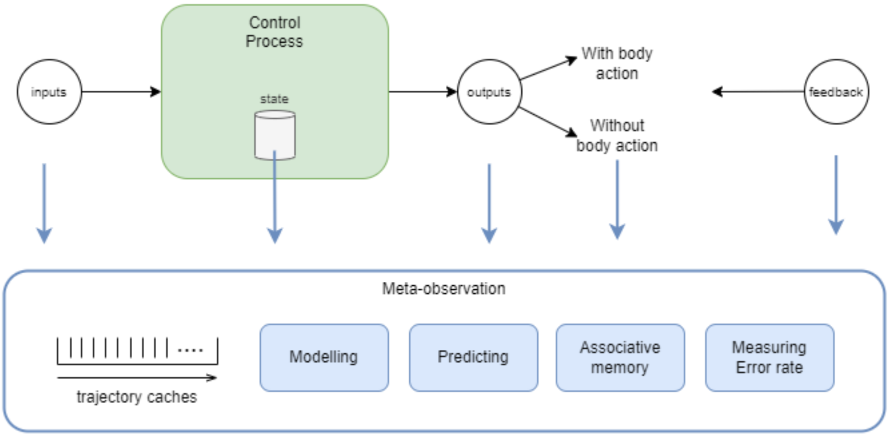
- **Parameter optimisation.** Our first example of meta-management was "after the fact" (aka *offline*) parameter optimisation. This occurs as a training process guided by feedback following execution of the control process against some problem. Through processes such as gradient descent and back-propagation the parameters of the control process can be optimised so that future attempts are improved.
- **Strategy selection.** The control process may develop multiple strategies for solving different kinds of problems. For example there are multiple ways to do path planning (Wang et al, 2019). Selection of the most appropriate strategy for a given problem at hand is an example of meta-management.
- **Goal selection.** Real world agents don't have hard-coded goals. They change goals according to situation.
- **Bias control.** The description of the biological brain as a hierarchical predictive coding architecture (Friston, 2010; Clark, 2013 and 2019; Kilner, Friston, Frith, 2007) (discussed in chapter I.4) suggests that individual inferences at one level are affected by biases that are inferred at an adjacent level. Those biases can be manipulated by some explicit meta-management process. A possible example is seen in mammals with interactions between the sympathetic and parasympathetic nervous systems influencing thought processes in different ways when calm versus when anxious [citation].
- **Direct state control.** Perhaps it is possible to directly influence the state of the control process.
- **Input manipulation.** It is possible to change the input in order to change the behaviour of the control process. This could be, for example, through attention. A more elaborate example is to infer what input manipulations are necessary in order to produce a desired CP behaviour. This would require the meta-management process to model CP behaviours and how various inputs affect those behaviours.
- **Output manipulation.** Another possibility is to directly manipulate the output from the control process before it takes effect on other systems. One example is to use this to entirely veto the control process in some situations. Another, perhaps more realistic example, is to attenuate the strength of signals from the control process while the control process is in its earliest stages of training. When the control process is untrained, it is likely to produce chaotic behaviours that might be detrimental to the survival of the agent. Some measurement of its level of stability could be used to gradually increase the strength of its output signals over time.
- **Feedback manipulation.** The outcome of the control process causes feedback, such as is used by RL to calculate parameter updates. Meta-management can be involved in the interpretation and even manipulation of that feedback. A simple example is to infer what parameter optimisations are required based on the feedback. This may include mechanisms for handling sparse feedback by somehow averaging and distributing the feedback over the sequence of actions that were carried out. Another example is to produce the feedback itself. In AI this is known as learning the reward model (Armstrong et al, 2020). A more elaborate extension is for the meta-management process to develop its own model of innate motivation and to produce feedback from that.
- **Controller selection.** Similar to the case of strategy selection. If multiple different control processes are available, a meta-management task is to choose which control process should take effect in a given scenario, or perhaps to choose a relative weighting of effect across the controllers.

A few general notes can be said about the above. Firstly, parameter optimisation is the only example listed of an "after the fact" meta-management process. The rest all take effect during online execution of the control process against a current problem. This is significant because it suggests that a) meta-management processes need to be actively involved during execution of control processes, and b) meta-management processes need immediate live observation of the behaviour of the control process as it executes.

Many of the above meta-management processes could be implicit or explicit. Implicit meta-management occurs as a side-effect of the reactive mechanisms of the control process. Explicit meta-management is driven by a separate process that influences the first-order control process in some way. For example, in AI, parameter optimisation is typically carried out as an "offline" process by a learning algorithm that is entirely separate from the processes used when executing the control process. In contrast, the prediction-based control process described in an earlier chapter illustrates implicit meta-management. Likewise it is believed that *hebbian learning* occurs as the primary learning mechanism in brains, which occurs naturally from local interactions between neurons without any explicit global orchestration.

III.5 Meta-observation Options in Meta-Management

In order to carry out any of the meta-control mechanisms described in the section above, those meta-management processes need to observe the behaviour of the first-order control process. We look now at a brief review of some of those *meta-observation* options.



- ***Meta-observation options.*** *Illustration of the sorts of things that may need to be observed in order to perform meta-management, and the processes that might be involved to draw inferences from that information. Behaviour of the control process can be observed, modelled, and predicted over time. Behaviour can be described in terms of the trajectories that it takes through a state space that incorporates inputs, CP state, and outputs. For rapid adaptive learning from single experiences, memory could associate salient situational features to urgent recall of behaviours that need to be avoided or repeated in the future. The effectiveness of first-order behaviour control needs to be observed, in terms of prediction error, reward signals, etc.*

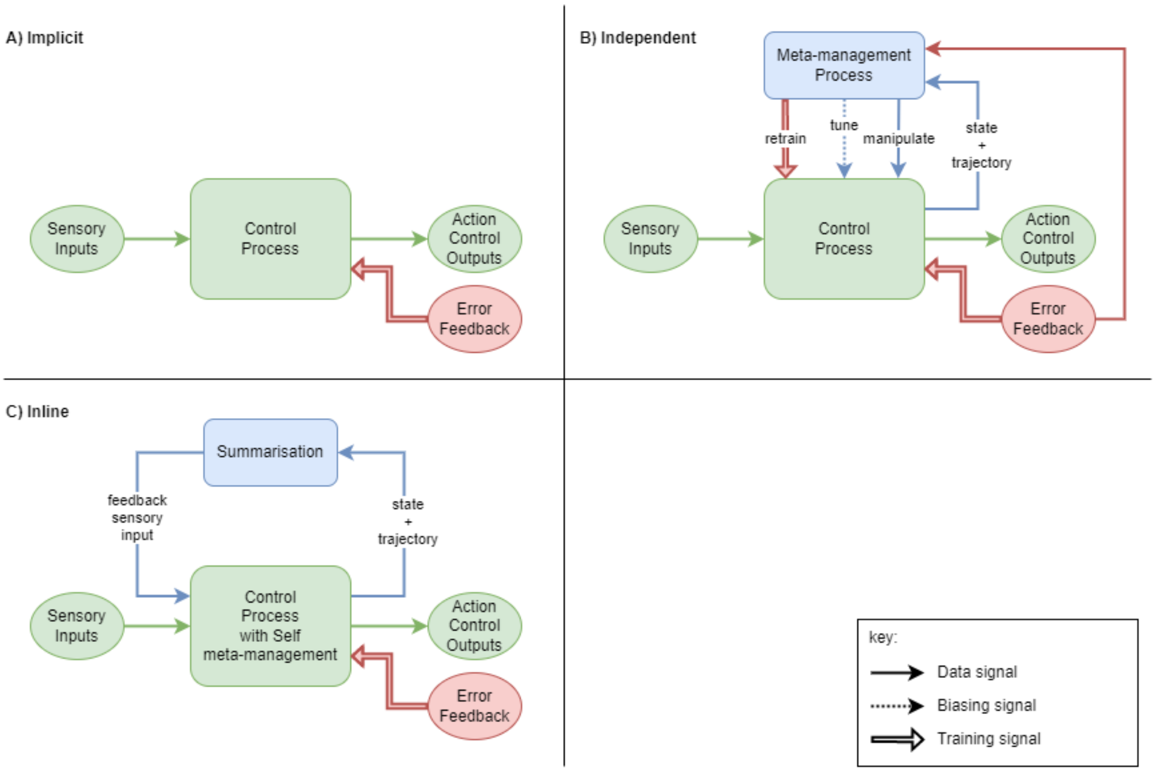
Mechanisms for meta-management processes to observe the control process include:

- **Inputs.** Inputs to the first-order control process need to be observed, including any goal selection supplied from systems outside of the scope of discussion. This is needed in order to associate CP behaviour with certain kinds of input.
- **Outputs.** CP behaviour is defined by its output for given inputs, so its outputs need to be observed and recorded or modelled w.r.t. to the inputs at the time.
- **State.** For multi-iteration control processes, their internal hidden state may be the only thing that changes from step to step. The trajectory of the CP state is what we first called out as needing meta-management back in part II.
- **Whether outputs lead to body action or not.** Useful in order to measure the “efficiency” of the multi-iteration process for producing useful body actions.
- **Feedback.** Generation of low-level feedback signals such as to indicate efficiency and “smoothness” of state trajectories. Generation of feedback signals based on higher order understanding of the problem domain (eg: that the path planner considered paths in the wrong order).
- **Trajectory caches.** Unlike the control process itself, meta-management may need to track the trajectory of CP behaviours over time. Likely across multiple timeframes. This will involve some mechanism to represent those trajectories. For example, the trajectory in the context of the current problem at hand in order to monitor whether it is leading towards a solution. And for example tracking of the control process’s overall abilities and over time, and whether it tends to produce useful results or tends to be “wrong” (for some definition of “wrong”).
- **Modelling.** Modelling of CP behaviour and how meta-control signals affect them, in order to infer the most appropriate meta-control signals.
- **Predicting.** For example predicting whether the current CP state trajectory is likely to lead towards a beneficial outcome or not.
- **Associative memory.** Used to recognise frequent and infrequent CP behaviours that need specific meta-management. For example, situations that lead to maundering.
- **Measuring error rate.** Recording and tracking how beneficial the CP behaviours are, such as would be needed to attenuate CP output strength in early stages of CP training.

Many of the mechanisms described above would already be required in an advanced agent for the observation, inference, and manipulation of interactions between the agent’s limbs and between the agent and the environment.

III.6 Architectural Options for Meta-management

Now that we have considered various options for meta-control and for meta-observation, we can tie them together by looking at how a meta-management process may be integrated with a first-order process. The diagram here illustrates three such options, with details discussed in the sections that follow.



- ***Meta-management architecture options.*** *Three broad architectures for meta-management. A) Implicit - the first-order control process converges towards stable behaviour without any explicit meta-management processes acting upon it. B) Independent - explicit meta-management processes acting upon the first-order control process. C) Inline - control process acting upon itself to self meta-manage.*

III.6.1 Implicit Meta-management

Some control processes are structured in such a way that they are inherently convergent and thus do not need explicit meta-management, as mentioned already in the discussion on first-order control processes. This is the null-hypothesis of meta-management.

One example is a simple mechanical thermostat that uses temperature feedback to control a heater. No meta-management is required as the system's control strategy is static - it does not learn. Another example is that of prediction-based control processes, described in chapter III.2. That does incorporate learning, and with the extension to free-energy minimisation may even produce relatively complex behaviour. For example, perhaps it is a good description of much of insect behaviour.

Through the various arguments presented in this treatise, I claim that such control processes are limited in their adaptability. For example, Part IV discusses model-based rational thought in humans, which goes well beyond what is possible with a single hierarchical predictive mechanism.

III.6.2 Independent Meta-management

Independent meta-management uses explicit processes that operate separately from the first-order processes.

A training process that records positive and negative feedback, derives a loss function, computes the gradients, and optimises the parameters of the control process is a simple example of an independent meta-management process. It is the most common example within AI research today. In AI research the training process is typically hand-rolled, but research has begun to look at how some parts of that process can be replaced with artificial NNs. This includes using NNs to estimate gradients (Bengio, Léonard, Aaron Courville, 2013) and using NNs to calculate parameter updates (Andrychowicz et al, 2016).

One can imagine the possibility of a neural network that calculates and executes all of the meta-management needs against a target control process. This could include all of the meta-control options discussed in a section above. A challenge with this architecture is how to train the meta-management NN.

III.6.3 Inline Meta-management

A somewhat radical suggestion is that perhaps the first-order control process can meta-manage itself, given the right conditions. This suggestion stems from the observation that the processes involved to observe, infer, and act as part of meta-management are very similar to those processes that make up the first-order control process in terms of interactions with its own body and the environment. In a complex environment we would expect those processes to be very complex. In a biological brain, due to the complexities and difference between the different sub-processes (modelling vs memory, for example) we expect some degree of brain region specialisation. Correspondingly, if two different problem domains need the same kind of processing capability, perhaps the same brain region might handle that processing capability for both problem domains. That would be the most neurally efficient solution because neurons cost a lot (Herculano-Houzel, 2012; Fonseca-Azevedo and Herculano-Houzel, 2012).

What conditions might make this possible? Firstly, the control process needs to observe its own behaviours. This could be achieved via a feedback loop that captures that behaviour and feeds it back as input. That feedback may capture the control process' current state, its recent trajectory, and its inputs and outputs.

In some architectures, the overall state of the control process may be held as state within each individual network node (eg: in the case of RNNs). Any attempt to capture the whole of that state leads to an infinite regress on the size of the control process. Thus, necessarily, the feedback loop would provide a high-level, dimensionality reduced, summary of that state and behaviour.

Importantly, the output of the feedback loop needs to frame observations about the control process not from the perspective of the control process's current state on its immediate current task, "from the inside" as it were. Rather, it must carry the perspective of the control process as part of a larger system, and as it interacts over a larger period of time. In other words, framed with a perspective as "from the outside". This difference will be discussed in more detail in chapter IV.6.

Lastly, training of the whole system needs to be bootstrapped. Initially the self meta-management capabilities of the control process will be as chaotic as its first-order behaviours. This could be achieved through a separation of meta-management concerns into a) an evolutionarily hard-wired process that applies simple domain agnostic constraints, and b) the complex, learning, an adaptive self meta-managing control process.

In some respect this is a description of one particular solution for implicit meta-management, but I find it useful to treat it separately for two reasons. Firstly, inline meta-management incorporates a feedback-loop with the very explicit purpose of aiding in meta-management. Secondly, the ways in which the control process can meta-control itself include all of those explicit meta-control mechanisms described chapter III.4. Thus inline meta-management is better described as a form of explicit meta-management than it is as a form of implicit meta-management.

III.6.4 Chapter Summary

All of the meta-management architectures described above may be valid. In fact, biological brains probably incorporate a mixture of all of them, and perhaps with other mechanisms not listed above. However, we are interested in the most significant form of meta-management that might be at play in the context of subjective experience. Subjective experience is strongly correlated with higher-order executive control, and so we are interested in the most significant form of meta-management as it pertains to that higher-order executive control.

When looking at generic control problems, none of the above meta-management architectures is clearly more effective or more realistic than the others. Perhaps the answer depends too much on how the control process itself functions. The devil is in the details, as they say. The functioning of the control process depends also on what kinds of control are required, with the more interesting kinds of control occurring in more complex agents.

Another sobering issue is that despite my best efforts the distinction between these three architectures is still somewhat poorly defined. The distinction between implicit and inline meta-management is that there are components of "explicit" meta-management in the latter. Could one look at a brain connectome and understand which parts are "explicit" meta-management? In the diagram above, what really distinguishes the "Summarisation" component in panel C from the independent meta-management process component in panel B?

For those reasons, the next part drills deeper into more complex control process requirements and mechanisms. After that we will be in a better position to circle back to re-address the question of meta-management architecture.

In the interim, however, there is some anecdotal evidence to suggest that the inline meta-management architecture is at play within humans. It is said that humans have five senses of touch, taste, smell, sight, and hearing. That list has since grown with the recognition of proprioception (location of limbs), equilibrioception (balance), pain (nociception), kinaesthesia (movement), environmental temperature (thermoception), time (chronoception), and others. While it is not often classified as such, our awareness of our own mental state behaves like any other sense. The extent to which we can attend to any sense varies by modality (eg: while we have proprioceptive sense, it doesn't carry the same ability to flood our consciousness than vision), but they are clearly all available for observation, integration, differentiation, correlation, modelling, and reaction-to within that higher-order executive control process that is associated with subjective experience. This applies equally to our sense of cognitive state, and suggests the feedback loop described within the inline meta-management architecture.

Part IV - Problems in Biological Control Processes

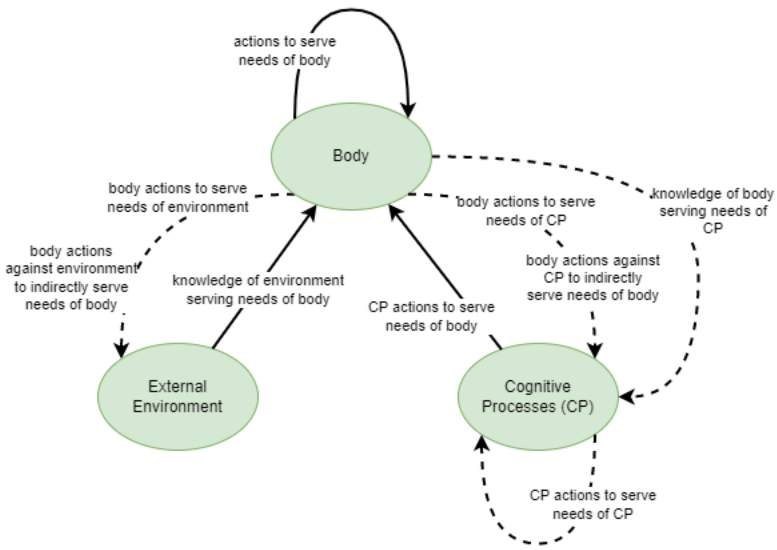
The focus of this treatise now shifts from generic "agents" of any form to biological organisms with complex brains. I shall start by reviewing some human concerns that relate to meta-management and for which suggestions have been made about their connection to human consciousness. This will be followed by an examination of potential meta-management processes.

Like many others, I focus on humans in order to avoid the ambiguity and uncertainty inherent with topics such as intelligence and subjective experience within the larger animal kingdom. This should not be taken to imply any statement about that larger animal kingdom.

Now that this treatise is focused on biology, and humans in particular, the reader will also notice that some of the terminology shifts to be more human-centric. One such shift, with a happy coincidence, is that the abbreviation CP will now be used interchangeably to stand for both *Control Process* and *Cognitive Process* - because a cognitive process is just a computational-style of control process, implemented upon a substrate of wet biological neurons.

IV.1 Embodiment

To fully capture all the ways in which cognitive processes are involved to sustain life would require a considerable review well beyond the scope of this treatise. In particular, such a review would need to discuss in detail the many different interactions between the external environment, the body, and the cognitive processes. Just a small selection of such interactions are illustrated in the diagram that follows.


actions to serve needs of body
Body
body actions to serve needs of environment
body actions to serve needs of CP
knowledge of body serving needs of CP
body actions against environment to indirectly serve needs of body
knowledge of environment serving needs of body
body actions against CP to indirectly serve needs of body
CP actions to serve needs of body
External Environment
Cognitive Processes (CP)
CP actions to serve needs of CP

- ***Some pathways of how the environment, body, and cognitive processes are leveraged to sustain the individual's life.*** *Solid arrows indicate the most significant relationships, with dotted arrows indicating supporting processes.*

What follows is a brief discussion of a few salient features that bear significance to the discussions of meta-management and subjective experience.

IV.1.1 Identification of Source

As the brain learns about the world around it and about its own body, it builds up models of the various objects within those spaces of environment and body. An important aspect of those models is to capture what ability the individual has to affect the state of the different objects within its models. A chair that obstructs motion can be easily lifted and moved. A tree cannot. The modality of action varies by the form and location of the object. The chair can be moved by performing physical actions involving the entire body. An itch may be scratched via motion of only a single arm and hand. A racing heart can be calmed by modulating one's breathing. A pain in the eyes from a blinding light can be resolved by closing the eyelids or averting gaze.

The same applies to actions against our cognitive processes. Different modalities of action are required for different situations. To remain focused on a particular task in the presence of distractions requires us to continually re-asses whether we are still focused on that task. In contrast, if we are "stuck in a rut" unable to solve a problem, the action required is to intentionally remove the focus in order to allow broader ideas to surface. On the topic of distractions, different actions are more suitable depending on the source of the distraction. If someone is repeatedly interrupting our thought by talking to us, a suitable action is to remove ourselves from the situation or ask them to desist. In contrast, if our own wondering thoughts are the source of distraction, we can only perform physical or mental actions against ourselves in order to solve the problem. The distraction examples are particularly interesting because in both cases some sort of thought is projected into our conscious processes, but the ultimate source differs. The ability to identify source, at least for external events, is acknowledged in metacognitive source monitoring studies, for example where subjects are asked to remember when or where some event occurred or to remember who presented the information (Shimamura, 2000). Source monitoring of memories in particular has been heavily studied (Johnson et al, 1993; Mitchell and Johnson, 2000, 2009).

More generally, I suggest that the brain identifies the source for the majority (if not all) of perceptions and thoughts that pass through our cognitive processes. This includes identifying whether the perception (of external senses or imagined) and thoughts were sourced from external factors or from the result of the self's cognitive processes. The identification may be in the form of an extra meta-data label that is *attached* as it were to the original perception or thought, or perhaps just the nature of the underlying representation is sufficient for the brain to identify the source wherever it needs to (Johnson et al, 1993). This *source labelling*, as I shall refer to it, would be useful for many components of cognitive processing, including i) helping to distinguish between independent objects when developing models of the environment, body, and of cognitive behaviours, ii) identifying the range of suitable actions, and iii) would form the basis of the concept of "I".

Source labelling is consistent with the suggestion that the feeling of *conscious volition* is the result of a an after-the-fact inference process that identifies whether action was caused by the self. Experiments have found that such inferences can be easily misled, leading to the individual falsely acknowledging or rejecting their causal involvement in the action (Wegner, 2003). Likewise, our identification of the source of memories is sometimes inaccurate or completely confabulated (Johnson et al, 1993; Mitchell and Johnson, 2000, 2009).

IV.1.2 Schema

Psychology has long identified in humans the existence of a model of the individual's body – known as the *body schema*. It is used in production of action control, and integrates information from our main physical senses and the proprioceptive senses (Proske & Gandevia, 2012). A clear 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. (p. 1)

Why should the brain need a body schema? There are a few obvious reasons, such as for tracking of limb position (proprioception) and for tracking of the capability and health status of various body parts. With reference to the prior section, another reason for a body schema is to build up a statistical model of the range of suitable actions a) that those body parts can themselves perform, and b) that can be performed against those body parts by other body parts. One way in which such a model may develop, at least in part, is to perform a clustering against sensory and effector signals into separate objects, through a measure of (relative) statistical independence. If the brain treats all sensory and effector signals as part of one single system, then the scale of joint probabilities that must be learnt is enormous. In contrast, the total scale of learnt joint probabilities is significantly reduced by modelling the fine-grained sensory and effector aspects of objects independent from each other. This follows from a more general result that the total number of free parameters within a generative model are reduced by introducing additional layers of latent variables, provided that the structure of the resultant generative model reflects the true causal structure (Rigoli et al, 2017).

A consequent effect is that those objects acquire their own *identities*. Cognitive control processes can represent knowledge and planning at a higher-level of abstraction that just refers to the identity of those objects. Those modelled objects of (relative) statistical independence will in general coincide with what we refer to as the different *body parts*.

I propose the existence of a second schema, the *cognitive schema*, that performs an analogous role for the modelling and regulation of cognitive processes. Anecdotally, this seems highly plausible within humans given our introspective ability towards our own cognitive 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 cognitive schema helps us to monitor, predict, control, and rationalize about our cognitive structure and behaviours in the same high-level abstract way that our body schema enables for our physical structure and behaviours. It can be indirectly observed in individuals where this regulation breaks down, such as in the case of *anosognosia* where individuals are unaware of their own cognitive deficiencies.

The idea of some sort of cognitive schema has also been made in the form of Attention Schema Theory (AST), whereby a model of attention is developed and used for control of attention (Graziano & Kastner, 2011; Webb & Graziano, 2015; Graziano, 2017). The idea of a cognitive schema is similar but has a broader focus.

In the same way that the body schema likely creates explicit *identities* for individual body parts, the cognitive schema likely creates an explicit *identity* for the "mind". When we say that we had a thought, we identify the source of that thought to our "conscious self", the inner part of us that experiences subjective experience. For neuro-typical individuals, we are under no delusion that the thought arose from perception of some external source, nor that it arose from our limbs. Furthermore, in the same way that we identify multiple independent body-parts, it's possible that our cognitive processes create not just one "mind" identity, but multiple. This may not be as far fetched as it sounds. An old idea of consciousness was of a little person, known as a *homunculus*, watching from the inside (now more commonly referred to as the Cartesian Theatre). While the idea of a little person inside our heads has been thoroughly dismissed in modern times, it remains an accurate description of what it "feels like" to be conscious. It seems quite reasonable to me to conclude that the brain thus forms at least two separate identities for its own cognitive processes: i) the homunculus, which clusters together cognitive senses and actions that are captured under higher-order monitoring, and ii) the non-consciousness, for all the rest. Another possibility is that cognitive processes are divided along more functional lines, such as having a cognitive identity for "memory", explaining why we consciously perceive no obvious difference between recall of episodic and semantic memories although their underlying structures are very different.

The exact specifics are not important here, as the most important point has already been made, summarised as follows: a cognitive schema creates a causal association between the outcome of mental processes and the source of those outcomes as coming from the individual, and it strongly demarks the boundary between the individual as source versus other possible sources. This is a major component of our first person experiences of consciousness.

IV.1.3 Meaning Attachment

It has been observed that thoughts and experiences are experienced in conjunction with the meanings of those thoughts and experiences (Van Gulick, 1992 & 2022). For example, we don't just see a lion, we see a lion plus with an instantaneous association of the level of danger that we're in, and perhaps the beauty of such a wild animal. When we see a toothbrush on the bathroom sink, we don't just see it a patch of colourful pixels and edges mapped into a three-dimensional shape, we immediately experience an identification of whether it is our toothbrush or someone else, and perhaps displeasure that it is no longer safely in its holder.

I refer to this process as *meaning attachment*. The source labelling discussed above is one form of meaning that may be attached to the experience or thought. Information regarding the thought or experience's association to our body or cognitive schemata is another form of meaning attachment. Emotions may be a third form. More generally, meaning attachment includes any contextual information that is associated to the thought or experience.

For the topic of subjective experience, the effect of meaning attachment is that we never truly experience a raw perception. Every perception is contextualised with source labelling, knowledge, and emotional affect. If the meta-management feedback loop led to construction of a higher-order thought representing that the individual had just observed their own thought, it would be experienced as far more than just information about a thought. It would also be contextualised by the cognitive schema, with concomitant emotional significance.

IV.2 Interlude: Meta-cognition

One recurring question in research on consciousness in general, and on subjective experience particularly, is what utility it provides over and above other brain processes that are not associated with subjective experience. Many theories of consciousness make reference to the need for flexibility or adaptation, but it is hard to pin-point what that entails. How do you quantify flexibility? What kinds of processes produce flexible behaviour? Behavioural scientists have been asking these very questions in the study of *meta-cognition*. This area looks at the ability of an individual to monitor and control their own mental processes, and how that relates to flexibility, learning, and other capabilities. The study of meta-cognition tries to incorporate behavioural studies with our growing understanding of brain function from neuroscience.

Meta-cognition does not have a single agreed definition, but in general covers the ability for an individual to *examine* their own knowledge, memories, and physical and mental abilities; to *respond* to that information in some way, such as to act to improve their knowledge or abilities; and to *report* on that information (Winkielman & Schooler, 2012; Fleming at al, 2012; Koriat, 2007). Meta-cognition is sometimes paraphrased as "knowledge about knowledge" (Karmiloff-Smith, 2012), "behaviour about behaviour" (Koriat, 2007), or "cognition about cognition" (Fleming et al, 2012a; Timmermans et al, 2012). The main purpose of meta-cognition is that it "adds flexibility to cognitive processes, making them less dependent on external cues" (Fernandez-Duque et al, 2000, p. 289).

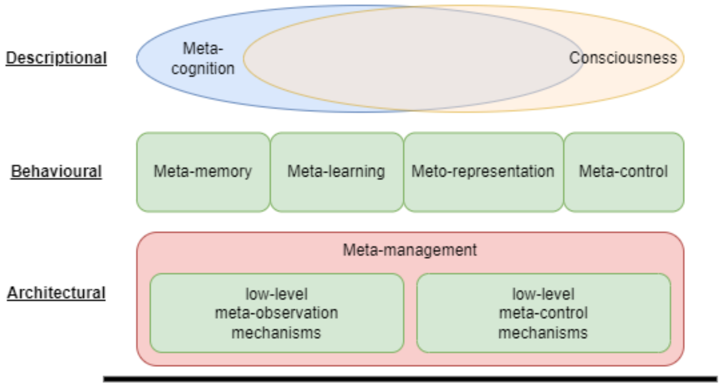
Meta-cognition has been variously studied in terms of:

- "feelings of knowing" or "tip of the tongue" situations, where one has a feeling that they know the answer before being able to recall the specifics of the answer itself (Rosenthal, 2012; Shimamura, 2000; Metcalfe & Shimamura, 1994),
- its contribution to successful learning as a whole (Paris & Winograd, 1990),
- memory of the source of knowledge or the source of other memories (Dunlosky & Bjork, 2008; Shimamura, 2000; Fernandez-Duque, 2000; Benjamin et al, 1998; Metcalfe & Shimamura, 1994),
- judgements of certainty and error detection (Carruthers & Williams, 2022; Cleeremens et al, 2020; Whitmarsh et al, 2017; Fernandez Cruz et al, 2016; Paul et al, 2015; Fleming et al, 2012a; Fleming et al, 2012b; Shimamura, 2000; Fernandez-Duque, 2000),
- classification of first-order outcomes into knowledge, hope, fear, regret, etc. (Cleeremans et al, 2007),
- identification of links between separately obtained knowledge (Clark & Karmiloff-Smith, 1993; Karmiloff-Smith, 1992),
- representing the absence of knowledge (Fleming et al, 2012a),
- selection of strategies for memory, learning, lifespan approaches (Marković et al, 2021; Shimamura, 2000),
- learning higher-level objectives (Timmermans et al, 2012),
- trading off between exploration and exploiting existing knowledge (Marković et al, 2021),
- balancing effort vs benefits of possible behaviours (Carruthers & Williams, 2022; Marković et al, 2021; Peters, 2010; Fernandez-Duque, 2000),
- planning (Marković et al, 2021; Cleeremens, 2020; Fernandez-Duque, 2000),
- monitoring and predicting first-order dynamics (Cleeremens, 2020; Fleming et al, 2012a; Timmermans et al, 2012; Cleeremans et al, 2007; Peters, 2010),
- control of attention (Whitmarsh et al, 2017; Shimamura, 2000),
- control over working-memory (Whitmarsh et al, 2017; Shimamura, 2000),
- internal conflict resolution (Shimamura, 2000; Fernandez-Duque, 2000),
- maintenance of cognitive homeostatic needs (Peters, 2010; Shimamura, 2000),
- emotion regulation (Shimamura, 2000),
- theory of mind and its ability to support social cohesion (Carruthers & Williams, 2022; Cleeremens, 2020),

- and in support of social cooperation by enabling a group to identify the individual who is most certain about some decision point (Cleeremens, 2020; Fleming et al, 2012a; Fleming et al, 2012b; Cleeremans et al, 2007).

IV.2.1 Analysis of Meta-cognition

In discussing meta-cognition together with meta-management, I view the term *meta-management* as referring to the lower-level architectural and network-level mechanisms that underly the higher-level behavioural forms. And I view *meta-cognition* as those higher-level behavioural forms. An analysis of the various studies of meta-cognition reveals a few common factors that provide some insight into the lower-level architectural aspects of meta-management. These are summarised in the following diagram and discussed below.



- ***Layers of Meta-cognition.*** *At a level of scientific description, meta-cognition and consciousness are presumed to closely coincide, although there may be some exceptions. At a behavioural level, meta-cognition consists of meta-memory, meta-learning, meta-representations, and meta-control. In contrast, the term "meta-management" is used to refer to the low-level architectural mechanisms underlying those behaviours, which is made up of low-level mechanisms of meta-observation and meta-control.*

There are four broad *behavioural forms* of meta-cognition:

- **Meta-memory.** Meta-cognitive awareness of memory. Memories appear to be encoded or are accessible in at least two forms: one for their specific content, and a second form that identifies meta-data such as source, certainty, and mere familiarity. The second form appears to be used extensively in judgements about memory, including the "tip of the tongue" effect.
- **Meta-learning.** Individuals can assess their own learning progress. They can use that to make judgements of certainty and to take corrective actions to improve their learning progress.
- **Meta-representation.** Various aspects of meta-cognition appear to depend on developing new representations about first-order processes and knowledge. These new representations are assumed to be both a) at a higher-level abstraction than used in first-order processes, and b) *about* those first-order processes and knowledge, known as *second-order* or *higher-order* representations.
- **Meta-control.** While some aspects of meta-cognition merely enable reporting on first-order processes, others elicit direct control over first-order processes as a result of that meta-cognitive awareness.

Through extensive use of fRMI and comparative studies, meta-cognitive studies are beginning to tentatively identify specific brain regions involved with those different behavioural forms. The reader is encouraged to review the papers referenced in the prior section for details.

From an architectural perspective, we can identify some low-level meta-observation mechanisms that play a significant part in meta-cognitive abilities. Including for example:

- identification of familiarity,
- identification of prediction uncertainty and conflicts,
- knowledge of current attentional focus,
- emotional state,
- and judgements of informational source and form.

Additionally, we can identify some low-level meta-control mechanisms, including for example:

- control of attention,
- conflict resolution,
- and apparent direct manipulation of first-order processing.

IV.2.2 First-order vs Conscious

Meta-cognitive studies attempt to divide our cognitive processes into *first-order* and *second-order* (Fleming et al, 2012a; Winkielman & Schooler, 2012; Koriat, 2007). First-order *behaviour* is considered to be produced by "mindless" sense-action processes, for example merely repeating learned actions for a given sensory perception. These are the *habitual* or *automatized* behaviours. Likewise, first-order *knowledge* is gained directly from our experiences, particularly from our perceptions. In contrast, second-order knowledge is considered to be gained by introspecting our own mental processes - it is knowledge *of* our mental processes. In addition, it is common in meta-cognitive studies to tie second-order knowledge to both *rational* cognition and to *consciousness* (Fleming et al, 2012a; Koriat, 2007; Nelson, 1996; Snodgrass et al, 2009).

In practice, however, it is difficult to distinguish whether a given behaviour is truly meta-cognitive (Fleming et al, 2012a; Winkielman & Schooler, 2011; Koriat, 2007). Many of the claimed second-order behaviours might be explained by unconscious first-order processes. Lab results are hard to interpret. For example, many meta-cognitive studies depend on verbal report and it is difficult to separate whether an identified brain region is activated because of its involvement in the original meta-cognition task or from the separate production of verbal report subsequent to the meta-cognitive task (Norman, 2020; Turvey & Crowder, 2017). Consequently, some have questioned the link to higher-order processes and consciousness in general, suggesting rather that meta-cognition is itself just part of first-order processes (Rahimian, 2021, Overgaard & Kirkeby-Hinrup, 2021; Cleemans et al, 2021).

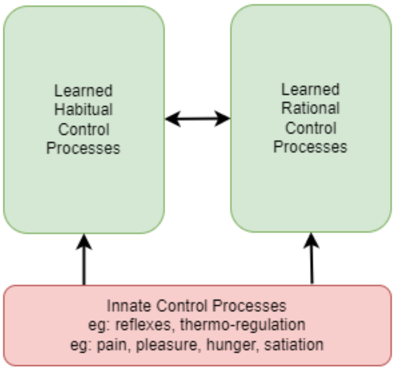
A deeper understanding of the low-level mechanisms underlying meta-cognition would help significantly to untangle the confusion. For example, many of the debates revolve around trying to correlate identified behaviours with whether they were produced habitually or through rational control. I suspect many of the debates are missing a more subtle nuance: that "first-order versus second-order" may be orthogonal to "habitual versus rational". This is a point to which we will return, but first we need to look a little deeper at habitual, rational, and other modes of control.

IV.3 Systems of Behavioural Control

By the end of Part III it was clear that complex multi-iteration control processes require meta-management, and we were able to identify some possible architectures that capture the interaction between control processes and meta-management, but we are left with an outstanding question: how do we identify which meta-management architecture is more likely in humans? The question is important to our end goal because it impacts our ability to explain subjective experience. To my knowledge, there is no research investigating the relative merits of the proposed possible meta-management architectures within the context of connectionist computational architectures. Thus we are left to postulate based on a theoretical analysis.

The controversies surrounding meta-cognition research stems from one problem: that it's not just one system that controls meta-cognitive processes. The meta-cognitive research discussed above generally acknowledge two systems: first-order non-conscious processes and high-order conscious processes. Debates about whether a behaviour is meta-cognitive or not hangs on an assumption that only conscious processes can be classified as meta-cognitive. We can avoid getting caught up in such debates when we acknowledge that all behaviour is an outcome of interactions between all systems. What we need to investigate is what those systems are, their underlying mechanisms, and how they interact to produce that behaviour.

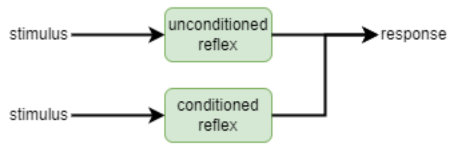
To help alleviate some of these confusions, we can examine behavioural control from a developmental perspective.


Learned Habitual Control Processes
Learned Rational Control Processes
Innate Control Processes
eg: reflexes, thermo-regulation
eg: pain, pleasure, hunger, satiation

- ***Three Systems of Control.*** *Innate control and innate feedback interpretation bootstrap learning of habitual and rational systems. Habitual and rational systems interact in order to produce learned behaviour.*

The literature in behavioural science identifies three broad categories of behaviour, delineated by presumed differences in underlying control mechanisms: *innate*, *habitual*, and *rational* Innate behaviour is predetermined through our genetics and is characterised by a (semi-)fixed response to stimulus irrespective of goal. Habitual behaviour is characterised by *automatized* stimulus-response associations. Rational behaviour is characterised by considered stimulus-goal-action associations. As will be seen, these three systems of behaviour follow a path of evolutionary development that combines their respective strengths and weaknesses: speed of adaptation, accuracy, the resources required for computing control, and whether they are modulated by current goals.

IV.3.1 Innate behaviour



- ***Innate System.*** *Evolutionarily pre-determined (unconditioned) responses are produced for certain stimulus. Some adaptability is possible with the same reflex response being associated with (conditioned against) other stimuli.*

Often referred to as *Pavlovian* behaviour in the literature (Dayan, 2008), innate behaviour has two forms: unconditioned responses and conditioned responses. Unconditioned responses encompass inborn, inflexible and genetically pre-determined *reflexes* that often cannot be consciously overridden (Dayan, 2008; Dolan & Dayan, 2013; Gęsiarz & Crockett, 2015). For example the knee-jerk reflex, salivating at the recognition of food, thermo-regulation, heart-rate, and the fight-or-flight response to danger. Innate behaviour is considered to arise because it was significant in our evolutionary history (Dayan, 2008; Gęsiarz & Crockett, 2015). In that sense it can be seen as providing the basic life-sustaining backbone of behaviour upon which all the rest of our behaviour is built.

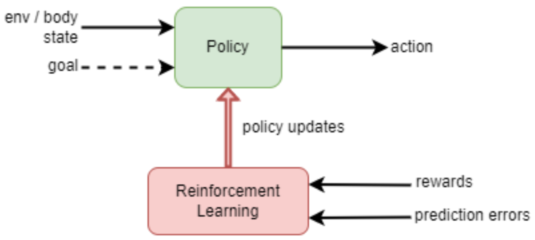
Despite the genetic pre-definition, some minimal experience-driven adaptation does occur. A *conditioned response* occurs where an otherwise neutral stimulus becomes associated with a reflex response (Dayan, 2008; Dolan & Dayan, 2013; Gęsiarz & Crockett, 2015). The famous example of Pavlov's dog describes an experiment where a bell was rung every time that food was presented. The dog had a pre-existing unconditioned response of salivating at the sight of food. Some time later, the bell would be rung without food and the dog would salivate at the sound of the bell alone. The dog was *conditioned* to elicit an innate reflex from a stimulus that was not directly genetically associated with that response.

A significant feature of both unconditioned and conditioned responses is that they are not modulated by the individual's goals at the time. There is a direct stimulus-response association, irrespective of other context.

IV.3.2 Habitual behaviour

Unlike innate behaviour, habitual behaviour is a) learned through experience, b) can be very complex, and c) can be goal-modulated. Habitual behaviour is learned to the point of being *automatic*: we no longer have to think about how to carry out the behaviour (Snow, 2006). However habitual behaviour is slow to be learned, and slow to be unlearned (van Es, 2019). There is some disagreement in the literature as to whether it is affected by contextual priming such as goals (Snow, 2006) or not (Gęsiarz & Crockett, 2015; Bernacer & Murillo, 2014; Dayan, 2008). It is likely that this reflects the specific focus of the research undertaken at the time and as an attempt to identify distinguishing factors between habitual and rational behaviour, rather than a claim that habitual behaviour is truly inflexible to the point of ignoring the individual's goals.

As illustrated in the diagram below, control of habitual behaviour is believed to be structured as an *implicit model* (Sutton, 1998) that directly associates (or *computes/infers*) a response to a given stimulus (de Wit, 2009). In Reinforcement Learning literature this same architecture is referred to as *model-free* because it lacks a structural model of the environment that can be queried in arbitrary ways - discussed further in the section below on rational behaviour. The reader will notice the similarity to mention of *reactive control* in earlier sections. I use the word habitual now that the context is biologically focused, but the two terms should be treated synonymously. Indeed, the behavioural science literature often uses the *reactive* term.



- ***Habitual System.*** *A policy network learns a mapping from stimulus to action, sometimes referred to as an implicit-model. The stimulus-action mapping may or may not be modulated by the current goal. Adaptation occurs through reinforcement learning that takes rewards and punishments and/or prediction errors and uses them to update neural weights.*

The learning of habitual behaviour is thought to develop concurrently with rational behaviour (de Wit, 2009; Dayan, 2008). While rational systems control behaviour, the habitual system learns those same behaviours. As the behaviour becomes more practiced, the habitual system takes over (Dolan and Dayan, 2013; Dayan, 2008; Sutton, 1998; Gęsiarz & Crockett, 2015). The mechanism by which the habitual system is "judged" as being trustworthy enough to take over is unclear. A commonly accepted working theory is that both the habitual and rational systems track their predictive uncertainty, and this uncertainty is used to select the system for control (Daw et al, 2005; Sutton, 1998).

Habitual and rational control of the same behaviours enable a trade-off between effort and time costs versus accuracy (Daw et al, 2005; Gęsiarz & Crockett, 2015; Douskos, 2018; Sloman, 1998). As will be discussed, rational behaviour control is a) computationally expensive, thus taking considerable time to infer the next action, and b) prone to mistakes or inaccuracies. For a well practiced behaviour, habitual control captures fine-tuned improvements that cannot be attained through rational control, and thus it is more accurate than rational control for such behaviours. Indeed, one particular form of habitual control, described as *Episodic Control*, repeats wholesale a complete sequence of actions (Dayan, 2008), such as the complex sequence of actions required in a golf swing. Many such sequences of action occur too quickly for deliberative planning to have any benefit.

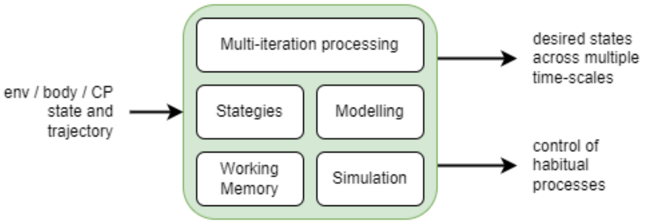
Historically, habitual and rational control were seen as a binary choice, but more recent work has identified ways in which they can be combined for control of a single behaviour, leveraging their relative strengths (Douskos, 2018). For example, the goal used by the rational system may be selected by a habitual process, or the rational controller may select a particular habitual sequence of actions to carry out (Cushman & Morris, 2015).

IV.3.3 Rational behaviour

Often referred to as *goal-directed*, *deliberative*, *executive control*, or *voluntary*, rational behaviour is characterised by multi-iteration processing with conscious awareness of at least some part of the decisioning process. Unlike for habitual behaviour, rational behaviour control adapts quickly (Dayan, 2008; van Es, 2019), and is the optimal behavioural control strategy for novel situations (Cushman & Morris, 2015). Behavioural sciences describe it as learned outcome-action associations (de Wit, 2009; Gęsiarz & Crockett, 2015).

There are two key presumed features of the systems underlying rational behaviour that distinguish it from habitual behaviour. Firstly, rational behaviour is thought to employ an *explicit model* (Sutton, 1998) that captures the causal structure of the world (Cushman & Morris, 2015). For example, it can capture likely outcomes for different possible actions. It might also capture relationships between things in the world, such as the shape of a maze that the individual needs to navigate. This model can be queried in almost arbitrary ways. Importantly, the model can be used to obtain information without immediately triggering action. In other words, the explicit model can be used to *simulate* a sequence of events and to identify the most likely result. Thus the outcomes for different actions can be considered, or the most likely action to achieve a certain goal can be selected, and from that information the rational system can choose how to act next. Within AI research, this is known as *model-based* reinforcement learning (Carruthers & Williams, 2022; Kanai et al, 2019), and there is considerable knowledge on how it can be implemented within an AI setting.

Secondly, there is evidence that only rational behaviour employs *working memory* (van Es, 2019). For example, it has been proposed that the explicit model is combined with working memory in order to perform *planning*, by searching through a decision tree (Dayan, 2008; Dolan and Dayan, 2013).


env / body / CP state and trajectory
Multi-iteration processing
Stategies
Modelling
Working Memory
Simulation
desired states across multiple time-scales
control of habitual processes

- ***Rational System.*** *Some of the features that might underlie rational thought: deliberative strategies, explicit modelling, working memory, and simulation. All driven through multi-iteration (a.k.a. deliberative) processing.*

While rational behaviour confers considerable adaptive advantage for novel situations, is believed to be complex, resource intensive, and inefficient (Cushman & Morris, 2015; Sutton, 1998; Dolan & Dayan, 2013). Where habitual and innate behaviour can be controlled immediately upon stimulus, rational behaviour requires deliberation - multiple iterations of processing. For example, where the model is extremely complex or where the search space is too large, approximations may be required in order to avoid spending too much time deliberating. It has been proposed that habitual processes may play a part in helping to carry out some of those deliberative approximations (Dayan, 2008; Cushman & Morris, 2015). Another limitation of rational behaviour is that it tends to be inaccurate. For example, where the choice of behaviour requires searching a deep tree of possible actions, it is thought that working memory is unable to accurately track of all of the intermediate states, and thus introduces noise (Dayan, 2008). Another possibility is that the explicit model can only capture relationships within the world to a course level of granularity, as it represents information in a more complex way than the stimulus-action mapping of habitual behaviour.

Planning through decision tree searches is just one particular *strategy* that's possible for rational determination of appropriate behaviour. There may be many others. In particular, connectionist learning methods are quite adept at mixing-and-matching different informational sources and control mechanisms. So in addition to the above, I would propose that rational behaviour is also characterised by many different deliberative strategies, and that those strategies are *learned* through experience and reinforcement.

Notice that, while rational behaviour is generally associated with conscious experience, it is not necessarily meta-cognitive in nature. For example, a goal-driven or explicit model-based decision process that focuses on external needs can be carried out without any recourse to self-appraisal. Where self-appraisal becomes necessary is for after-the-fact analysis of performance. Thus, second-order processes and rational processes should not be seen as the same thing.

IV.3.4 Emotion

There is another fascinating aspect of control that has for a long time been left on the periphery of studies, but which has more recently gained more traction. Damasio and others are beginning to make strong cases that emotion is a key part of “truly rational” behaviour (Damasio, 2006; Picard, 1997; Sloman, 2001). Studies are finding that individuals with brain lesions that prevent emotions influence over control struggle to make suitable long-term life choices. It appears that they lack fear of their own loss and suffering and thus cannot “rationally” weigh up options, often choosing short-term wins that have significant long-term deleterious effects.

Damasio’s 1994 book *Descartes’ error : emotion, reason, and the human brain* provides an excellent discussion of the importance of emotion in rational decision making, as well as discussions of how cognitive processes, emotion, and body are intertwined. Clearly, a discussion of human behavioural control would not be complete without a discussion of emotions. Unfortunately such a discussion is beyond the scope of the present focus.

IV.3.5 Co-development of behavioural control systems

From an evolutionary perspective, innate behaviour is seen as being most ancient. It developed initially as simple localised reflexes in single celled organisms, using simple chemical signalling (Godfrey-Smith, 2016, p27-41) and non-nervous electrical signalling (Erulkar & Lentz, 2023). Later, as multi-celled organisms evolved, their reflexes needed to be globally choreographed, leading to the evolution of nerves (Erulkar & Lentz, 2023). This eventually led to the diffuse nerve nets found in cnidarians (hydroids, jellyfish, sea anemones, corals) and ctenophores (comb jellies), and later to central nervous systems. Habitual action, specifically the ability to learn from experience, is necessary for the coordination of limbs (Paulin & Cahill-Lane, 2019; Godfrey-Smith, 2016, p27-41). There is clear evidence that this had evolved by the time of the “Cambrian explosion” (about 542 to 485 million years ago), where there was an explosion in the forms of animal bodies and where there was significant predation between species (Godfrey-Smith, 2016, p27-41). There is some evidence that complex nervous systems (though not necessarily central nervous systems) may even have evolved a little earlier, during the Ediacaran period (635 to 542 million years ago), where there was apparently little predation but there were scavenging creatures with complex bodies (Paulin & Cahill-Lane, 2019), such as the trilobite (Godfrey-Smith, 2016, p27-41).

The evolution of rationally controlled behaviour is hard to identify, but it is a significantly much more recent development than that of habitual control. Rational behaviour is believed to correlate with so called “symbolic thinking”, evidence of which is searched for in the forms of cave drawings and in the construction of tools that would require planning ahead. Those traits are identified as evolving somewhere between 70,000 and 164,000 years ago (Wayman, 2023; Henshilwood et al, 2001).

That is from the perspective of development of a species. What about for the development of an individual? How do these three systems interact in order to help the individual learn to control its own behaviour?

One area that has received only minimal research to date is the extent to which innate behaviour may *bootstrap* learning of habitual and rational behaviours. An old idea proposed by John Locke was that infants start with a “blank slate” and learn everything. Only more recently have we started to find clear evidence that evolution predisposes individuals with innate capabilities that bootstrap learning (McCarthy, 2008; Chappell and Sloman, 2007). For example, there is research suggesting that the same mechanisms underlying innate behaviour are also involved in the prediction of outcomes (Dolan & Dayan, 2013) such as future rewards and punishments (Dayan, 2008). I suggest that innate behavioural control systems influence habitual and rational control in more fundamental ways.

Firstly, it prevents untrained habitual and rational systems from causing the individual harm through incorrect action or through inaction. For example, the reflex to flinch and withdraw from pain is so powerful that to counteract that reflex requires an individual to carefully coordinate calming effects against the parasympathetic nervous system and tensing of counteracting muscles. Likewise, young children are often unable to prevent themselves from taking food when they have been told to wait. Habitual and rational systems can counteract innate control, but only once they have reached sufficient maturity. So, while habitual and rational systems are under-developed and prone to error, the innate system maintains behaviour within the bounds of a “safe zone”.

Secondly, the innate control system participates within the ongoing training of those more advanced systems through its interpretation of primitive feedback signals such as pain, pleasure, hunger, and satiation. Effectively, it bootstraps learning of rewards. The more advanced systems learn through reinforcement learning by seeking to maximise the integral of positive valence interpreted by the innate control system.

Thirdly, it provides additional primitive feedback in the form of indications of energy and time cost that should be minimised by reinforcement learning. Any action should be carried out in the most energy conserving way - don’t use more muscles than necessary; don’t tense counteracting muscles so that motion is difficult and more effort required than necessary. Mental processes should produce effective outcomes quickly.

Much of that plays a direct influence on habitual learning in particular. Thus innate control provides a bottom-up *restriction* on habitual behaviour as well as a learning pressure that leads to improved habitual control over time. At the same time, rationally controlled actions provide a top-down reinforcement learning pressure against habitual behaviour. In that way, habitual behaviour *converges* towards the best possible behaviours through a combination of bottom-up and top-down restrictive and learning pressures.

To complete the circle, as it were, I wish to briefly mention how the same might occur for the rational system. However I have not studied this at length and so this is just a conjecture. The rational system must surely also need to *learn* to perform its duties. I have suggested above one such learning in the form of strategies, but there may be more. Thus the rational system is likely prone to gross errors in its early development during early childhood. Furthermore, as a far more complex system than for habitual control, it is likely that the rational system takes longer to develop. I suspect that the rational system may be “held to account” by three factors: i) the innate system, which prevents the rational system from sending the individual outside of a safe operating range, ii) the habitual system, that may develop earlier than the rational system, and iii) developmental processes, that only let the rational system take control once it is ready to do so. I elaborate further on this speculation in chapter VII.1.

IV.4 Habitual and Rational Meta-management

In comparing innate, habitual, and rational behavioural control we see that the innate system primarily provides a bootstrapping role for the more advanced systems of habitual and rational control, and we see that habitual and rational control are combined in ways that strike a balance between their relative strengths. Habitual systems are slow to adapt, but efficient to execute. Rational systems are fast to adapt, but complex and slow to execute. Independent habitual systems are distributed throughout the human brain, as multiple habitually controlled behaviours can be executed simultaneously, eg: talking while driving. In contrast, there appears to be only one rational system, which can only operate against one task at a time. Clearly there is an advantage to combining them wherever possible.

Further to that, we have seen that there is evidence that habitual processes may play a role in rational control. This fits well with the prevalence of debates about whether apparently rational behaviour is actually executed by habitual first-order processes. Perhaps rational behaviour *always* involves significant habitual control.

What about for meta-management?

I have made the claim earlier that rational behaviour is not the same thing as second-order or meta-cognitive behaviour. I now make the claim that meta-management, the low-level architectural aspect of meta-cognition, is itself a combination of habitual and rational control mechanisms and that these systems develop in conjunction with each other in the same way as they do for first-order control.

Rational mechanisms of meta-management would leverage all of the modelling and deliberative capabilities discussed above, but operate at a higher-order meta-cognitive level. Some examples might include:

- **Higher-level modelling of reinforcement feedback.** For example, reinforcement learning struggles where reinforcement is sparse. By observing and developing explicit causal models of the world, including eventual reinforcements, the individual can infer whether a given action just taken is likely to be a step towards or away from a desired future reinforcement. Thus the individual can generate their own dense reinforcement feedback for individual actions, filling in the gaps as it where between the sparse rewards provided by the environment.
- **Behaviour modelling.** Developing higher-order models of first-order habitual and rational behaviours, including tracking of abilities (in other words, effectiveness) against different problem domains.
- **First-order strategy selection.** Through deliberative processes involving the above models, the meta-management process may help to select the strategy used within first-order rational deliberation.

As discussed earlier, rational processes are computationally expensive, but it gets worse. As rational processes are executed in a serial fashion, if an individual is carrying out some rational meta-cognitive thought, then they are preventing themselves from carrying out rational first-order control. An example of this is seen when we are doing something that needs constant attention but we become distracted. Our default mode network (DMN) takes over our rational processes, leaving our primary task under solely habitual control.

Thus, ideally, even meta-management would be done through habitual processes wherever possible. This is entirely plausible. Habitual control of meta-management would occur for meta-management tasks that have been sufficiently repeated that they can become automatized. Some examples might include:

- **Generation of dense feedback.** Learned from regular use of higher-order models.
- **First-order rational strategy selection.** Learned from repeated similar decisions.
- **Error detection.** Detecting when habitual first-order control behaviour has resulted in unexpected outcomes that require rational systems to take over in order to correct the situation
- **Learned meta-learning behaviours.** For example, as observed by others as an individual’s position on the sliding scale of *learning mindset* to *fixed mindset*.

Before I wrap up the discussion of habitual and rational control, I need to be up front about an issue that is emerging with my description of meta-management. The distinction between first-order control and meta-management is getting blurry. At first I claimed that meta-management is something entirely different from first-order control. Then I claimed that rather than being independent, meta-management processing may be performed by the same underlying system that performs first-order processing. I have now elaborated on that by claiming that meta-management makes the same distinction between habitual and rational control. So is meta-management really distinct from first-order control? In order to answer that, we need to look at the details of the representations used and of how they are processed. And in order to do that, I need to first introduce the concept of *semiotics*.

IV.5 Semiotics

The claim so far is that the meta-management feedback loop generates a (high-level, and filtered) representation of the state of the goings-on within the brain. That representation is then available for processing, which may produce some computational actions as part of meta-management. Assuming an inline meta-management architecture, a logical extension is that the outcome of that meta-management processing result is a new state that is subsequently captured by the meta-management feedback loop and fed back as a new representation.

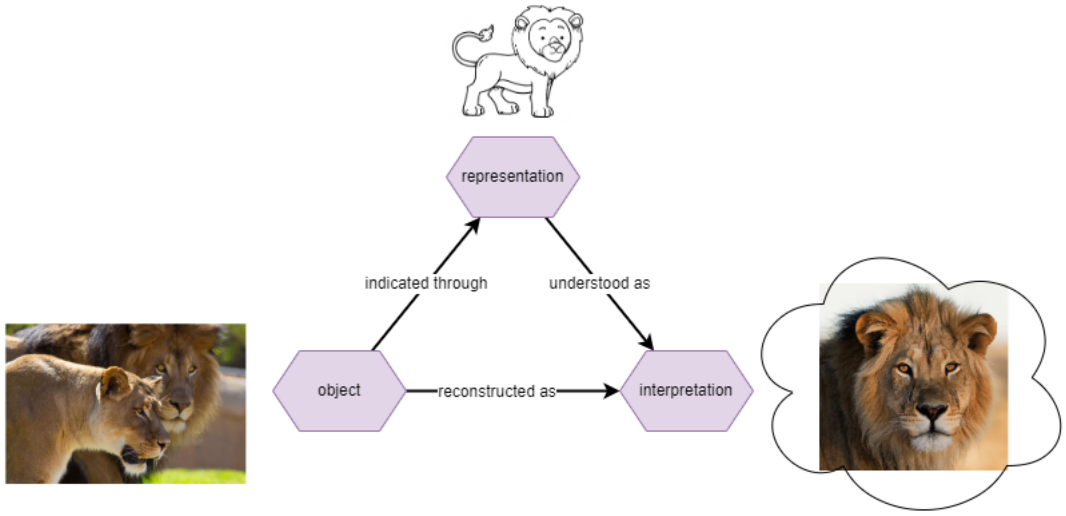
This description begins to mimic what we'd expect of the basis for subjective experience: it provides a self-referential observation, along with a choice about simply observing or reacting to that observation. But the meta-management process only has a few basic things going on: capturing of data, generation of a representation of that data, processing of that representation, and generation of actions in response to that representation. Which of those could possibly produce subjective experience? The answer is that it's all of those basic steps, taken together, and repeated over and over. But to explain that clearly we need the right language to talk about it.

Another issue that has arisen from the discussion so far is that the proposed mechanisms of meta-management overlap so much with those of first-order control that it appears hard to disentangle them. This mirrors the difficulty in behavioural studies to distinguish between first-order and meta-cognitive behaviour. As it will turn out, this confusion can be alleviated by looking more closely at the representations in use by the different processes. But again, to explain that we need a language that will unable more precision in our statements.

In both cases the language that will help us is that of *semiotics*: the study of signs and their interpretations.

IV.5.1 Objects, Representations, and Interpretations

Semiotics studies signs, meaning making, and the processes involving them. There are different formulations but the one that we shall use is through the triad of *Objects*, *Representations*, and *Interpretations*.



- ***Components of Semiotics.*** *Semiotics identifies three components to communication and meaning-making: the original object, its representation, and the interpretation of that representation. For good communication, the interpretation should closely resemble a re-construction of the original object, or at least a mental image of the original object. In that sense, the interpretation is itself just a representation.*

Representation. Also known as a *signifier*, *representamen*, *sign-vehicle*, and just *sign*. A representation is anything that does or could communicate meaning through some aspect of its presentation. It can be of any form, eg: visual, auditory, tactile, electrical, or as some kind of virtualised information. Three forms of representation are typically identified. *Icons* share some important quality with the object that they represent, such as by bearing a direct resemblance to that object. For example a cartoon picture of a lion represents an actual lion through nature of resembling the appearance of a lion. *Indices* share some linkage of fact with the object, such as through a causal relation. Smoke represents fire because fire causes smoke. *Symbols* represent their objects through conventional or cultural knowledge, and unlike the first two have no natural relation to the object that they represent. For example, in phonetic languages (and semi-phonetic languages such as English) written characters represent the sounds they make, not because there is any resemblance in the way that the characters look to the way they sound or because the shape of the characters are somehow caused by their sounds, but because we've all agreed to the convention that these characters make those sounds.

Object. Also known as the *referent*, the object is the underlying thing or concept being represented. For example, the lion, the fire, the sounds that written characters make. It may be some simple physical thing, or be a deep and broad concept, or a thought.

Interpretation. The interpretation or *interpretant* is the meaning obtained through the act of interpreting or translating that representation. In some cases the representation exists with the intent to convey the original object as the interpretation. For example, a stop sign on the road is a representation erected with the intent to be interpreted to convey the concept of the original object: all of the traffic rules around stopping at stop signs in order to prevent accidents. In other cases the object and interpretation are different. This is particularly the case for physical objects. For example, the interpretation of smoke is that a fire is present, it does not clone the original fire within the mind of the person seeing the smoke. Likewise, the interpretation of the word "lion" elicits the recall of the hearer's conceptualisation of a lion, rather than an actual lion. That conceptualisation is itself just a representation, and depending on other contextual information available, may or may not be very accurate in relation to the particular lion that the speaker was referring to.

A few points are worth drawing out.

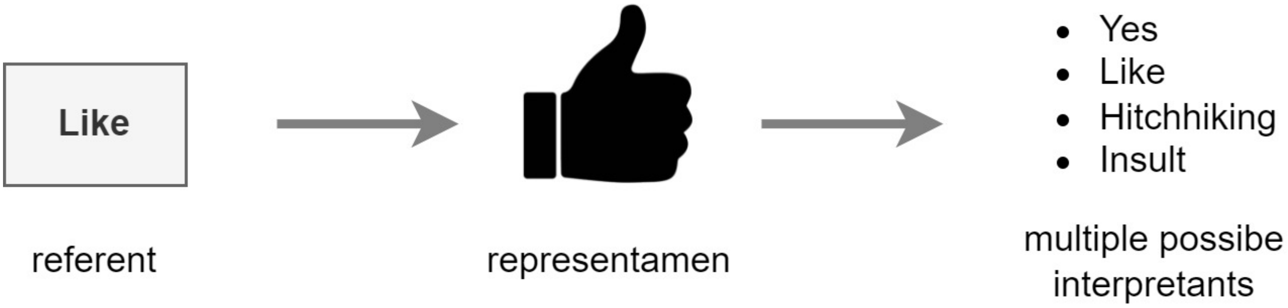
The representation is an *informational bottleneck*. The object typically has far greater detail and nuance than is ever captured by a representation of it. A cartoon image of a lion captures only a superficial resemblance to the outward appearance of a real lion, omitting all of its internal biological structure, its behaviour, and the fear induced in someone who hears it roar. The varieties of intentions behind the single spoken word "happiness" far exceed the informational qualities of the sound waves that represent it to the listener. Indeed, in many cases the purpose of a representation is that it indicates the object without consuming the same physical and informational resources as the object it represents. How short our life would be if we needed to explain to a friend that a lion is approaching by explaining in minute detail every molecular property of all the cells in the near (and rapidly approaching) vicinity.

If the object is a detailed and nuanced entity with high physical and informational bandwidth, and the representation is a bandwidth bottleneck, the interpretation can be viewed as an up-scaling of bandwidth through *re-construction*. This requires that information be added, where should that information come from? An example will help.

While driving, it is common to see signs that indicate that certain traffic features are ahead by some distance, such as an intersection. Often the features themselves are around a corner out of sight, and the only information one has about the approaching feature is within the sign that says "Intersection ahead, 500m" - an informational bottleneck. At the moment of observing the sign, the rest of the information comes from the observer's own knowledge - of the road rules involving intersections, of past experience with intersections, and perhaps with memory of that specific intersection.

So the full construction of the interpretation is a little bit from the sign, and a lot from the individual doing the interpretation. In other words, interpretation requires an *interpreter*.

Moreover, the particular interpretation obtained from a representation depends on the knowledge and disposition (a.k.a. state) of the interpreter at the time. As example, consider the thumbs up hand signal illustrated in the following diagram. A thumbs up hand signal can mean several different things. When used in conversation or while interacting with social media platforms, it can mean that you appreciate or support the statement made by someone else. When used while standing on the side of the road it can mean the request for a ride. In some cultures it makes a sexual reference intended as an insult.



All of these issues with informational bottlenecks and misinterpretations arise because the object is unknown at the time, or even ultimately unknowable in a practical sense. When smoke is observed in the distance, you do not have any other information about the object itself at the time; though it is possible to move closer to investigate and learn more about the source of the smoke. Conversely, when someone speaks, they are translating (representing) the complex sequence of ideas and thoughts within their mind. While it is possible to ask questions and gain more information, it is practically impossible to know every nuance of those original ideas and thoughts within the mind of the speaker.

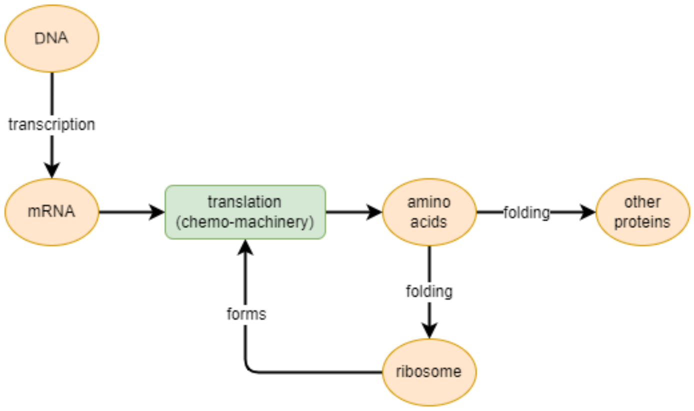
Semiotics provides a tool for understanding meaning-interpretation processes in different domains, regardless of whether those domains include obvious signs. Today semiotics is used in many fields, either explicitly or implicitly. It can be said that medical diagnosis is the act of interpreting a sign (the symptoms) in order to identify (diagnose) the original object (the underlying condition) (Burnum, 1993; Nessa 1996). It is used extensively in marketing to understand what meanings people are likely to add to different messaging (Borţun & Purcarea, 2013). A listing available from the Wikipedia page on Semiotics indicates specific sub-fields for biosemiotics, cognitive semiotics, computational semiotics, literary semiotics, cultural semiotics, social semiotics, and many more.

The study of Semiotics has a long history, but it's modern structure is most commonly traced to the works of Ferdinand de Saussure and Charles Sanders Peirce. Peirce was the first to introduce it via its triadic formulation. It is informative to note that in Peirce's earlier works he was focused on semiotics as the process underlying cognition (Peirce, 1982, vol 2, pg 56, 213). To that end, Peirce supported an idea of *infinite-semiosis* within the mind: representations being interpreted, producing new representations that need to be further interpreted, and so forth. This is convenient to us, as it is an idea to which we shall now turn.

IV.5.2 Infinite semiosis

The examples mentioned above describe single iterations of a semiotic process: an object exists in some way, a representation of the object is presented, an interpreter constructs an interpretation based on the representation. Infinite-semiosis is the continued execution of the semiotic process, with the interpretation itself becoming an object that must be represented and further interpreted. Infinite-semiosis is not necessarily infinite in an absolute sense, and in practice never can be, but the process is in principle capable of continuing "for a long time" before it breaks down in some way.

A fascinating and informative example comes in the form of the life sustaining processes surrounding DNA replication. A biosemiotic description of DNA and related processes was described eloquently by Pattee (2007), following inspiration from a much earlier attempt by Von Neumann (1966). I myself take heavy inspiration from that in what follows.



- ***Steps in the DNA Transcription and Translation Process.*** *DNA is transcribed into messenger RNA (mRNA) through a direct copy process and chemical attractive processes that do not require any explicit machinery. mRNA is translated into amino acids via expicit chemo-machinery. Amino acids fold into various proteins, including the ribosomes that form the translation machinery. Thus the translation machinery can create more of itself, according to the description encoded within the DNA.*

If you want to produce an exact clone of some system there are roughly two ways to go about it. The first, *inspection*, involves examining each and every part of the original and producing a copy of each part as it is examined. The second strategy is to produce a *description* of the original, and then to *interpret* that description in order to produce the copy. Inspection is a good choice for simple systems, but there are problems with trying to inspect a large, complex, and dynamic system. If the system being inspected continually changes, at what particular point in time should the copy be made? Also, if the system is large, then the machinery to inspect it will also be large. For the second strategy, the description might for example be something akin to a recipe, that contains a sequence of steps necessary to construct the original; or it might be more of a declaration that certain things have certain characteristics. In either case, the description is a *code*, and it requires considerable machinery to do the interpreting of that code. Thus the second approach is somewhat more complex than the former, but it benefits from the fact that the description is separated from the original. More concretely, the specific physical form of the description does not have to be the same as the physical form of the original system. Thus, for a large and complex animal made up of countless kinds of proteins folded from combinations of 23 amino acids, you can have a tiny DNA strand made up of combinations of only 3 base pairs.

Protein synthesis from DNA is a complex process that involves both methods of cloning. In the first step, the base pairs of a DNA strand split and the sequence is *transcribed* to messenger RNA, with each messenger RNA usually representing a copy of only a short sub-sequence of the entire DNA strand. Next, ribosome machinery scans along the base triplets of each messenger RNA as it encounters them, *translating* those base triplets into corresponding amino acids which it then joins together into a sequence of amino acids. The last step is that the sequence of amino-acids spontaneously *folds* into a 3-dimensional protein structure according to relative energy levels across the various weak bonds within the molecular structure. The *transcription* phase is a cloning operation by direct inspection of the original. The *translation* phase is an example of cloning by description plus interpretation. Messenger RNA is the description, and the product of interpretation is a sequence of amino acids. The forms of RNA and amino acid chains are quite different. Even more different are the RNA and the folded proteins that are produced as the final product.

Now, in some cases that folded protein goes onto become part of basic structure and form of the body, such as in skeletal, muscular, or nervous tissue. In other cases the protein is of a special sort known as an enyzme, which has the capacity to manipulate other proteins. One particular form of enzyme is the ribosome machinery discussed above that translates messenger RNA into proteins. So RNA has the ability to describe the machinery that can interpret itself. Furthermore, because RNA uses the same bases as DNA and is itself just a copy of one half of a DNA strand, we can say that DNA has the impressive ability not only to describe many different proteins, including being able to describe the construction of the cells that subsequently will contain copies of that DNA, but it also describes the machinery that copies and interprets itself.

The processes described above have a clear semiotic explanation within an individual. DNA and RNA are codes that require interpretation, and so from a semiotic point of view they are *representations*. The *objects* that they represent are the whole or parts, respectively, of the animal. The *interpretation* of those representations results in new cells that replace old cells and in the continual production of all of the micro and macro machinery needed to keep the whole organism alive. But there is more. With each newly formed cell, there is a clone of the original DNA. And that new cell begins to carry out all the same processes that were carried by its parent cell, but this time all driven off a copy of the DNA. At the level of a cell, the object, representation, and interpretation together produce another cell, and that new cell becomes the next object in a repeated cycle. This forms a pseudo-infinite semiotic chain.

At the evolutionary level, another aspect of DNA becomes important. Not only is the form of DNA somewhat independent of the form of thing that is produced from its instructions, but the particular form of DNA itself is arbitrary. In the words of Pattee, "Many other copolymer strings or even bit strings in a computer could be interpreted or translated by a suitable coding mechanism to synthesize the same proteins as a DNA sequence." (2007, pg 120). That arbitrariness is an important feature for evolution to work. Firstly, because it enabled unguided primordial chemical interactions to experiment with many different chemical forms, and to eventually stumble upon one such experiment that happened to become self-sustaining. Secondly, it enables evolution to experiment in an open-ended way, trying out many different DNA sequences, exploring the search space, and finding better sequences (more "evolutionarily fit"). In that way, the changes in DNA from one individual to another as a species evolves creates another pseudo-infinite semiotic chain.

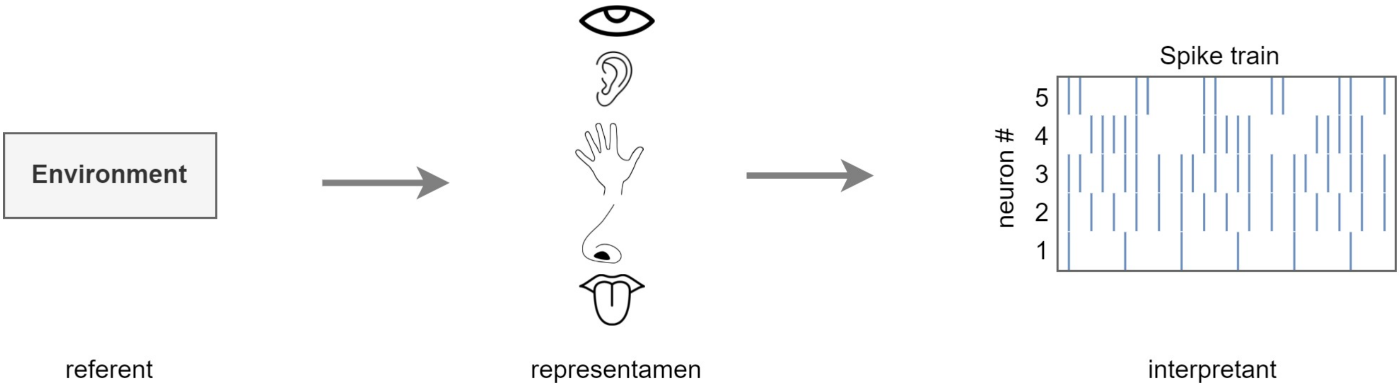
There is an important a lesson here about representations: they don't have any meaning on their own. A representation only has a specific meaning in the context of a specific interpreter. The DNA that becomes messenger RNA that becomes sequences of amino acids would only become that particular sequence of amino acids thanks to the ribosomes that do the translation. A different interpreter would produce a different result. And a different form of representation would not be translated by the ribosome. The form of representation and its interpreter must evolve together.

It is also worthwhile to note that no cognitive entity is required to do any of the *interpretation* in any of the steps discussed above. This perhaps stretches the original meanings of the semiotic jargon. Indeed, in a later treatise Pattee (2021) discusses the fact that the science of semiotics was originally designed for interpretation of thought and human artifacts, and thus the terminology used generally assumes some cognitive entity that does the interpretation. However, they go on to make the case that the underlying concepts are just as relevant to mechanical processes such as RNA transcription.

IV.5.3 Thought

Things get more complex and nuanced when you look at computational semiotic processes in connectionist paradigms. When the entire brain is a massive multi-layered network involving many different kinds of processes, some of which talk to each other in some way, but all part of a single larger system, where are the boundaries of object, representation, and interpretation? One answer is that the brain involves many objects, representations and interpretations, layered in deep hierarchies. More generally, semiotics is best viewed as a tool that can be applied at whatever layer or part of the process that we are interested in.

As a first example of semiotics in the brain, consider how the body gathers and interprets information about the environment, illustrated in the diagram below. The environment is the *object*, and its true state is what the individual hopes to attain knowledge of. But the individual will never gain such knowledge as they do not directly perceive the environmental state. Signals obtained from the classic five senses of sight, sound, smell, touch, taste *represent* the environment. That collective representation captures only a narrow and shallow aspect of the true state of the environment. For example, our visual sense only reports on a small field of view at any given time, and that too only about the external surface of whatever thing is being looked at. That representation is interpreted by the network of brain neurons, producing an inference about the state of the environment. What form does this inference take? Millions of *spike trains* in some particular pattern of activity (Krüger & Aiple, 1988; Paiva, 2010; Deco, 2008).

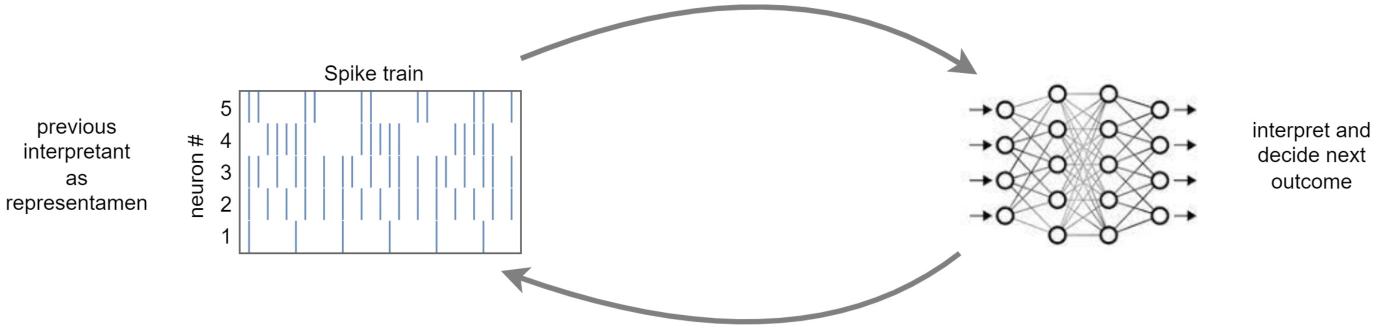


Those millions of spike trains are not an end to themselves. Spike trains that went only to dead-end neurons would result in nothing happening. Spike trains are themselves just a representation, and thus must be interpreted.

In order to respond to the environment, the individual interprets those spike trains by applying knowledge of the environment and of the individual's own body. The resultant interpretation: an action plan that is sent as further spike trains via effector nerves to muscles. Or perhaps the individual simply chooses to think more about the environment. Thus the semiotic process forms a repeating cycle of representation and interpretation, illustrated in the figure below. This plays out in the communication between brain regions. It also plays out at the whole system level during multi-iteration processing.

Importantly, spike train representations have the same informational bottleneck effect as discussed earlier. For example, if we consider some localised region of brain, whatever processing that occurs there is through the activity of many layers involving hundreds or thousands of neurons.

The activity of those neurons, and the synapses between them, is potentially impacted by dynamic static held within the neurons and synapses. In contrast, only a small subset of synapses convey messages to other regions of the brain. Whatever region receives those messages re-constructs meaning through the addition of its own knowledge - knowledge that is inherent through its structure and the learning processes that have tuned it throughout the individuals lifetime.


previous interpretant as representamen
Spike train
neuron #
5
4
3
2
1
interpret and decide next outcome

One question arises. Where is the object in such cyclic semiotic processes? From a computational point of view, the question is not an interesting one as the processes will continue to function regardless of how we understand them - after all, semiotics is just a tool for analysis, rather than a statement about how the world must be. But perhaps from a philosophical perspective we may gain some extra insight, or at the least another reference point for further discussion. From the outside world we have the example of a real physical lion, a cartoon image representing it, and the mental concept created in the mind of the interpreter upon seeing the cartoon image. The object, the lion, is a thing that exists and has a reality extending through time. In a multi-iteration cognitive semiotic process, the object of the next iteration is the interpretation of the current interpretation. But even the interpretation seems nebulous.

The discussion one paragraph before gives the example of a brain region producing a spike train representation that it supplies through its synaptic connection to other brain regions. That is a representation of its interpretation. Where is the interpretation? The interpretation never exists. At least, it never exists as a single whole entity with all its component parts being present at the same time. Rather, as the individual neurons fire, performing their individual proto-computations, they construct individually some small part of the whole interpretation and then immediately forget it. The immediately next neurons pick up from there and construct, individually again, other small parts of the whole interpretation. At the same time, each synaptic connection between individual neurons supplies only a spike-train representation, as part of a semiotic process at a finer level of granularity. By the time that the wave of neural activity reaches the edge of our arbitrarily chosen brain region, only the finer-grained representation of the final outgoing synaptic connections exist.

So, in computational infinite-semiosis, objects are past interpretations, but interpretations are nebulous and exist only as the sequence of activities spanning many neurons over a course of time. And representations are no better - they exist only as a sequence of electrical spikes, spanning many synapses, over a period of time. More brutally, in computational infinite-semiosis, at any given instantaneous moment in time, nothing ever exists in its complete form except the interpreters with their background (learned) knowledge ready to be applied as representations pass transiently through them.

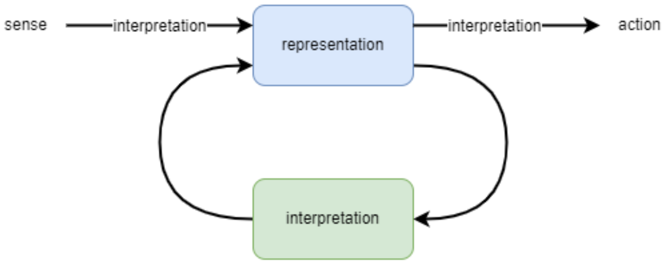
IV.5.4 Trains of Thought

There is another key takeaway from the semiotic analysis above that needs a more thorough discussion. This regards the kinds of thoughts that are possible.

A common turn of phrase is to refer to a “train of thought”. For example, when someone suddenly speaks to us after a period of silence we complain that they ”broke our train of thought”. There are two notions behind that statement. One is that thought progresses. For example when trying to solve a problem, thought progresses from problem to solution through a series of steps. Another notion refers to there being one particular basic grounding idea, or *topic*, being considered which is held onto for the duration of a period of thought. For example, the topic may be the underlying problem that one is attempting to solve. So that while thought shifts from step to step, and between different possible solutions, the topic remains the same for the duration of that sequence of thoughts. When our train of thought is “broken”, our attentional focus shifts from that topic to some other, and we loose some of the details held as temporary state in the progression from problem to solution.

For the discussion that follows, I shall define a *train of thought* as the finite sequence of processing that occurs while a given topic is attended to, with that train of thought ending and another starting when the attended-to topic changes. The *topic* is be to defined intuitively from the example given above. The definition here is somewhat vague but that is not important for the discussion to follow. More generally, we can say that a thought belongs to the same train of thought as before to the extent that the topic is similar to before.

Now, given a computational system, and given some particular potential topic and associated train of thought, what is required for that train of thought to inhabit the computational system? We cannot expect that a given computational system can hold just any arbitrary thought, there must be representational and computational limits.



- ***Semiotic cycle.*** *All brain processes are the result of a constant cycling between representation and interpretation. A train of thought is a semiotic loop where the interpretation produces a new representation that is further interpreted by the same system.*

Illustrated in the above figure, we have seen that all thought is the act of interpreting representations, producing more representations requiring interpretation, in an endless cycle. The questions around trains of thought translates to questions around the capacities of those representations and interpretations. By this I mean that a representation only captures a certain amount of information, about a certain thing. Something that is never represented is never available for interpretation. A representation with a coarse granularity makes less information available to the interpreter than a fine-grained representation. Likewise, an interpreter can only produce interpretations (or rather, interpretational representations) according to its degree of computational complexity (eg: the number of neurons and synapses that make it up) and the learned knowledge available to it.

Furthermore the question of capacities of those representations and interpretations comes down to their purposes. Despite many historical claims of evolution producing features without utility [citations], most have been eventually debunked [citations]. Thus we have a strong reason to believe that evolution would only produce representational and interpretational capacities that meet the needs of the individual. More specifically, evolution would only produce representational and interpretational capacities that increase the individual's likelihood of producing progeny that also go onto to produce more progeny.

So, what kinds of trains of thought can be entertained by a given computational processing system? Answer: those that provide utility according to the given system's function in maximising evolutionary survival.

There are fascinating philosophical discussions that can stem from this idea, like asking what kinds of thought we have never had and will never have simply because our evolved structures are incapable of entertaining them (McGinn, 1989, 1991). But I am more interested in something a little more concrete.

The first key message I wish to convey is simply the fact that representational and interpretational capacity is not arbitrarily large - it is finite, and tuned according to specific needs. Thus the trains of thought that are possible according to a given system are *constrained* by that system.

Secondly, successive iterations of interpretation, representation, and re-interpretation will tend to *attenuate* the train of thought down to the capacity of some weakest, least complex stage. As we have stated before, representation is almost always an informational bottleneck, so whatever receives that representation has only the informational capacity of the representation plus its own (learned) knowledge. The brain is an extremely complex system, with many sub-processes interacting in complex ways. Any information that is obtained but not retained through all those interactions, is lost. We can see this effect in the apparent limitation on the number of things that can we can retain within working memory.

Another concrete example of attenuation at play can be directly experienced. It is experienced anytime that we are unable to focus. Imagine this scenario: it's 2pm, you finished your lunch an hour or two ago, and you forgot to have your coffee. You're attempting to complete the work you started earlier in the morning. Something mentally difficult, like doing the accounts for your tax return. You're finding, like many around this time of day, that you're struggling to remain focused on the task at hand. You start to think about the problem immediately in front of you, you start to form the idea in your mind, and then you realise you can't remember what the next step is. You shift focus to recall the next step, but then can't remember the details of the result you'd just obtained from the prior step. Rather than there being anything wrong with our fundamental potential for such thought, the underlying issue here is one of attention, metabolic processes (Serin and Tek, 2019), and the way our circadian rhythm impacts our cognition (Valdez, 2019). However, it illustrates what our experience might be if we had a certain amount of the required computational and representational capacity for such a task, but that during multi-iteration processing it quickly became too attenuated to be operated against.

The last key message I want to draw out here is that a given thought only exists for the duration that it is repeatedly cycled between representation and interpretation. As we have seen, representations in the brain are fleeting transient sequences of energy spikes travelling along synapses, and interpretations are merely the action of target neurons in receiving those signals and producing new signals. Once those spike trains stop representing a given thought, it ceases to be a thought.

"But what about memory?", I hear you say. Indeed, through mechanisms that we are only beginning to understand, individual neurons do hold temporary state that modifies their behaviour [citations]. So interpreters could "hold onto" the thought representations that they receive or produce. Likewise, we easily recall past thoughts, indicating that they are somehow recorded within our episodic memory. However, the point I wish to make is that these representations are not "thought" while they are stationary. It is only through re-interpretation that they become thought.

The argument here may seem somewhat tautological to some (thought only exists while thought exists), but the message is more fundamental: that thought only exists as a fleeting interaction between representation and interpretation and that a given thought collapses (ceases to exist as a thought) when that cyclic interaction ends or changes topic. This argument will make even more sense when used in the discussion of subjective experience that follows in the next part.

IV.6 Representations

From the discussion on semiotics we see that representations must always be interpreted in order for them to have any effect. Each interpreter process has some specific focus, and the representations that it uses must have a similar focus for it to have any benefit to the individual. Evolution and individual developmental processes produce representations that are targeted to the needs of the interpretive processes that use them. The representational spaces used by different processes are thus different. These spaces can differ in terms of i) what thing is represented, ii) represented properties of that thing, and iii) the form of representation used.

In this chapter we shall look at some specific examples of the ways in which different representations can be produced about the same domain. Finally from here the differences between first-order and meta-management processes will become clear.

IV.6.1 Representation Redescription Theory

The theory of *Representation Redescription* (RR) proposes that the brain develops multiple representations for the same domain (Karmiloff-Smith, 1992 & 1994; Clark & Karmiloff-Smith, 1993). It looks at the development process of an individual from infant, and sees the first representation as habitual in nature. The literature uses the term *implicit*, which is meant to signify a particular observation about the nature of first-order connectionist architectures. A simple connectionist architecture (predictive or not) encodes a mapping from sense plus state to action. The mapping is learned from multiple examples, with each repeated exposure subtly modifying connection weights. The behaviour that the mapping produces is *emergent*, resulting from the interactions of all those many connection weights. Importantly, the details of that mapping are hidden to every part of the system. Even the particular sub-system that encodes a particular mapping cannot identify the rules that it encodes. Thus a first-order network encodes "knowledge *in* the system", but it does not make it available as "knowledge *to* the system" (Clark & Karmiloff-Smith, 1993, p495).

That is contrasted with an *explicit* representation that captures the individual parts of the causal chain from input to action. Where implicit representations *compute* an action for a given input, explicit representations *model* causal relations and as such can be queried and manipulated: "No system in which rules are always merely implicit and emergent can, in our view, exhibit the kinds of higher order flexibility and creativity found in humans. Only explicit rules have the genuine, systematically manipulable components that make radical flexibility possible." (Clark & Karmiloff-Smith, 1993, p504). RR proposes that implicit knowledge is *re-described* into an explicit representation through an offline process. Subsequently, habitual behaviour continues to be driven by the original implicit representation, but rational cognitive processes can make use of the explicit representation. Explicit representations may be formed that directly cover some particular habitual domain, or may integrate across multiple domains. In that way, rational cognition is able to identify and use associations across different domains and time-scales. Importantly, it is able to do all that independently from actioning the behaviour that the original implicit representations encoded.

RR does not attempt to explain the underlying details of how redescription takes place. The *Radical Plasticity Theory* of consciousness (RPT) of consciousness proposes a predictive mechanism, whereby one system in the brain learns to predict the behaviour of another and in so doing builds up a representation of that other system's behaviour (Cleeramans, 2007; Cleeramans, et al 2007; Pasquali et al, 2010; Timmermans et al 2012; Cleeremans et al, 2020). Clearly, in order to observe and predict the behaviour of a first-order system, a second-order process must occur. RPT even takes this one step further to claim that consciousness itself is a second-order process that we learn to do.

There are two important points to make here about the predictive mechanism. Firstly, every representation encodes information about a given domain in a certain way that is applicable for the way in which that representation will be used, or in other words, according to its *purpose*. The process that attempts to predict internal behaviours would almost certainly have a different underlying purpose, and thus would produce a representation that captures a different aspect of the same domain. Secondly, we should not assume that the *form* of representation would be the same either. As we have already said, a first-order connectionist representation is *implicit*. Many predictive networks may also be inherently *implicit* in their representation. In contrast, the assumption here is that redescribed representations are *explicit*, and thus capture the *causal structure* of the domain.

IV.6.2 Semiotic Analysis of Control Processes

I shall now return to the question of control processes and meta-management, by examining them in terms of semiotics. I shall present three examples of processing, and in so doing we will see that their objects, representations, and interpretations have some important differences:

- first-order body action control
- task-focused CP state trajectory meta-control
- meta-learning

In any computational processing, the purpose of the process is to produce its result. In evolved systems, the information made available to the process is partly a result of the needs of the process to produce those particular results. We shall mirror that fact by working backwards from the interpretation to the object. For this exercise, we will define the interpretation produced by a computational process as the direct *effect* of the process under discussion: any changes to state, any actions that are triggered by the process, and any new representation that it makes available for subsequent processing. We will consider the representation to be the synaptic signals received by the process in question. The object is the (mostly unknowable) state of the thing that the process is “focused on” - ie: the thing that the processing attempts to understand in some way in order to produce appropriate results.

Body action control

For first-order control of body actions via a multi-iteration control process, we can look at habitual and rational processing separately. The generated interpretation by habitual processes are simply the outcomes that we’d expect of a body control process: body actions that affect the state of the body and the environment around it, plus CP state updates as it executes its multi-iteration processing.

Interpretation from rational processes is a little more interesting. It includes modelling of the problem space for later use (ie: modelling of environment and body), deliberative outcomes in terms of environment and body, strategy selection with regards to the overall type of response required against the current environment/body problem (eg: the individual is getting wet in the rain, so built a hut or find a cave), and body actions.

To make that work, the habitual and rational control processes require information about the environment and body. This includes knowledge about their dynamics. Furthermore, the more the individual knows about the future of the environment and body, the better it can make decisions. These are the attributes of the object.

Notice that the object is dominated by state that the individual needs but will never have, or will only have after the fact. This is a feature of the object in any semiotic system where the interpreter must learn about the object from scratch (more or less). We gain wonderfully complex and nuanced understandings of the world around us and of our own body through experience and through education. But the truly amazing thing about that learned knowledge is just how little we have and yet manage to survive.

The representation presented to the first-order control process includes the individual’s raw senses that sample some practically insignificant portion of information about the environment and body. It includes those learned approximate models of the dynamics of the environment and body produced through prior interpretations. And it includes working memory, for example so that a given aspect of state can still be responded to even when no longer in sight.

As a concrete example, consider the biologically plausible path planner introduced earlier. This leverages the modelling and simulation capabilities of rational control systems. While for entirely novel situations it may need considerable meta-management oversight, let us consider just the first-order elements of this process. The state that this planner would need to hold and use in its computations would be focused on the specific problem at hand at the time, for example:

- real-world paths already simulated
- identification of choice points along those paths, where alternative paths could be simulated
- factors helping to decide whether to continue to search for better real-world trajectories or to settle on the current best guess.

Notice that largely these representations are focused on direct objects in the particular problem domain, in this case paths through a real-world environment.

Meta-management of CP state trajectory

For meta-management of the state trajectory of a multi-iteration control process, the generated interpretation most predominantly includes some action that affects the first-order control process (CP) state trajectory. This could be through any of the mechanisms described in section III.4, including but not limited to direct state manipulation, adjusting bayesian biases, or through manipulation of input signals to the first-order control process.

Arguably another form of meta-management against CP state trajectory includes directly initiating body action. This can be seen in the example when someone consciously chooses to avert their eyes to aid in distracting themselves from a thought that was associated with the previous visual scene (as opposed to an instinctual revulsion aversion response).

A key difference to the previous example is that the object in this case is the first-order process itself, as it performs processing against a particular task. That includes its state, the trajectory that it has taken, and its progress in relation to the task. The object also includes the CP state trajectory that it will take going forwards on the current task. Knowing that is important for selecting appropriate meta-management actions that need to be taken now. Of course the future is unknown, so models stand-in as a means to estimate the most likely future.

The representation provided to this meta-management process includes that provided by the meta-management feedback loop: a high-level summary of the current CP state, and its trajectory so far on the current task. It includes learned dynamic models of CP behaviour. The identification of the current task would presumably be represented within the state of the first-order process itself. For inline meta-management, that state could be easily accessed directly or somehow directly influence the meta-management processing. Or perhaps the meta-management feedback loop includes a representation of the current task with a higher-level abstraction that is more suitable for the needs of meta-management processing.

For a concrete example, let us return to the planner. In contrast to first-order control, the state information required for meta-management of that same planner has the planner itself as its target of representation. This may include for example:

- the computational state trajectory of the first-order process itself (ie: the three bullet points above, together, form a single point in the computational state trajectory), including how long it has been running for on the current problem

- the *class* of problem that the first-order process is currently operating against
- modelled information about how effective the first-order process is against that class of problem
- indication of whether the first-order process is currently trending towards finding an effective solution
- indication of whether the first-order process appears to be repeating any past negative behaviours that would require intervention

Meta-learning

The idea of *meta-learning* captures the ability for an individual to examine their abilities and to make decisions that directly influence their subsequent learning. Through teacher instruction and/or by observation of others, an individual can develop a model of how good they *should be* at a particular task. For example, after observing that their friends have become proficient in use of a skipping-rope after spending a day practising, an individual may set a goal to do the same. Meta-learning is thus about long term skill observation and development. Here the meta-management process not only observes the state of the first-order process, but generates additional state information in the act of comparing the current behaviour of the first-order process against past actual behaviours and future possible behaviours.

For meta-learning applied against the first-order process, the interpretation produces state changes and actions. To do this, its processing includes a number of things. Firstly, it requires the construction of a broad understanding of the behaviours and abilities of the first-order process across many problem domains. This includes strengths and weaknesses. It also includes building models of the comparative behaviours, strengths and weaknesses of other individuals. Finally, interpretation includes meta-cognitive decisions that drive further learning, such as through setting goals, acknowledging and remembering certain behaviours that need to be avoided, or likewise for behaviours that should be repeated.

The object is once again the first-order process itself, but the focus is on a much longer term than in the example in the prior subsection. It now encapsulates a measurement of the relative effectiveness of the first-order process against the range of all potential problem domains. Like for the other examples, the object has much that the individual is entirely unaware of, including what the optimum set of behaviours really is, and what the future holds. And once again, learned models stand in as approximations of dynamics in order to make predictions about those optimum behaviours.

Contrast this to the object in the case of CP state trajectory meta-management. There the object was the immediate state trajectory. Here, the object is about the collection of learned abilities and how they can be used in different scenarios.

The meta-management process infers what it can about the object through the representations made available via the meta-management feedback loop, and via models. Here the models are built up over time to capture the CP behaviours and abilities on different kinds of problems. Separate models capture the individual's understanding of how other individuals fit against those measures.

IV.7 Review of Meta-management

To wrap up this part I shall briefly lay out the salient points that have been made about the architecture and processes of meta-management. I shall also present some final points that I have not found a better way to introduce earlier.

IV.7.1 Similarities between first-order processes and meta-management

Meta-management processes likely leverage many of the same systems used for first-order control. I have made the case that the balancing of strengths and weaknesses between habitual and rational control equally benefit meta-management. This enables well practiced meta-management actions to be carried out automatically. Anecdotally I believe this makes every bit of sense and correlates with with our general lack of needing to continuously consciously monitor our thought processes.

Another aspect that I believe is shared between first-order control and meta-management is *domain knowledge*. The world that we interact with is tremendously complex, and so are our bodies. We must devote considerable neural mass to learned implicit and explicit models of our environment and of the control of our own bodies. That knowledge obviously plays an extremely important role in first-order control. But it is also must surely play an important role in meta-management - how else can meta-management model and operate against those first-order processes without understanding the domain in which they operate? We have discussed that different purposes require different representations, and Representation Re-description theory even formalises that as a key part of developmental processes, but multiple representations take up more neural mass. There must surely be ways in which the brain optimises the form of representations in order to minimise the number of re-representations required, and it seems likely that the shared domain knowledge required by both first-order and meta-management are a contender for such optimisations.

IV.7.2 Differences between first-order processes and meta-management

One key difference between first-order processes and meta-management has been identified as the representations used by these respective processes. First-order processes use representations that focus on the immediate task at hand. Meta-management uses representations about the first-order processes themselves.

Other likely areas in which they differ include the *pathways* taken for such processing, and associated *learning* mechanisms. Meta-management requires two key differences in pathway from first-order processes: i) it needs to observe and model behaviours over a longer period of time than for first-order control, and ii) it requires a higher-level abstraction in order to avoid an infinite regress on the neural count. This is encapsulated in the idea of a meta-management feedback-loop. Likewise, within a reinforcement learning paradigm, meta-management needs different reinforcement signals; though a detailed discussion is beyond a scope of this treatise.

IV.7.3 Meta-management Architecture

We are now finally in a position to place our bets on the most likely architecture of meta-management, that of *inline meta-management*, incorporating both habitual and rational systems.

This stands from i) the evolutionary need to avoid duplicating the complexity of rational control systems, ii) the benefit in sharing domain knowledge, and iii) anecdotal evidence that we can freely switch between consciously considering a task at hand and consciously thinking about our own awareness.

But that is not the all of it. We started with the suggestion that complex systems require explicit meta-management and then took ourselves on a journey to investigate the component parts of that meta-management system. By the end of our journey we have identified that key first-order systems equally play a part in meta-management. While we have identified some differences in terms of representations, pathways, and learning pressures, these differences are not necessarily more fundamental than the differences required to cope with different sensory modalities and different problem domains as part of first-order control. Innate, habitual, and rational control systems already incorporate many different representations, explicit models, pathways and learning pressures.

We are thus at a point where a re-description of meta-management is warranted. Meta-management is an emergent phenomena, evolutionarily enabled by one key modification: the addition of a meta-management feedback-loop. Everything else being equal, with that one adjustment the same habitual and rational systems that provide automatized and deliberative first-order control can perform automatized and deliberative meta-cognitive management of their own processes. While there are differences in representation and modelling, those can easily be incorporated. There are also differences in reinforcement learning pressures, but these are likely bootstrapped by the same innate feedbacks: pain, pleasure, hunger, satiation, time cost, energy consumption, etc. In practice, rational control and meta-management likely evolved together, developing all of these representations and innate feedbacks in conjunction.

Part V - Solution

The sections before laid out a sequence of logical inferences. Working from a basic need to cope with complexity, we have seen that this concludes with meta-management processes superimposed over the same systems that perform primary control of the body, enabled through a feedback loop that captures a high-level summary of the state of that control process. And we have seen that this results in the control process state having both a “from the inside” representation and a separate “from the outside” representation.

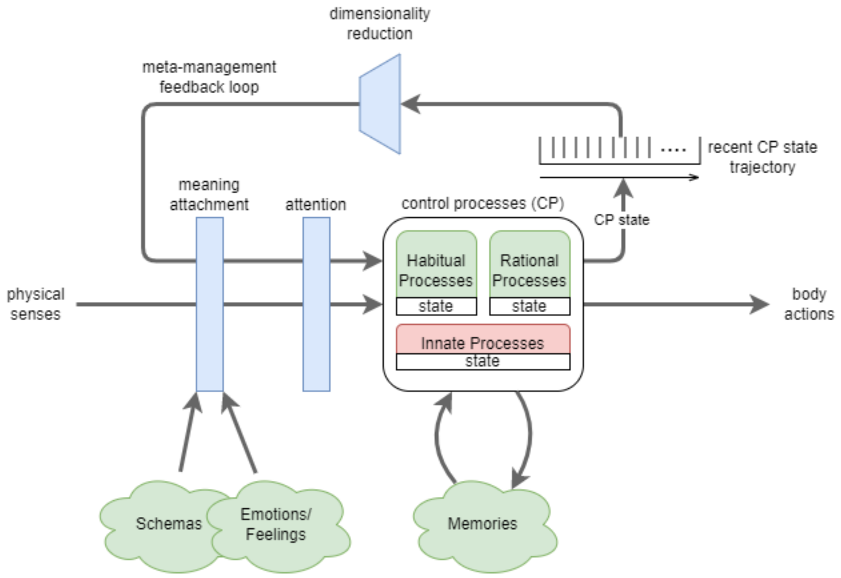
We are now in a position to state the core thesis of this treatise, which is this:

- Subjective experience is the result of the meta-management feedback loop, an inline meta-management architecture, and the cyclic processing of self-observational state interpreted from that feedback loop.

This part explores that core thesis in detail. It starts by summarising the processes underlying subjective experience within the human cognitive processes, and then attempts to define subjective experience in more general terms that might be applicable to arbitrary evolved and synthetic agents.

V.1 The Architecture of Subjective Experience

The following illustration pulls together the various processes described in earlier sections, creating a description of the major *functional* components involved in the construction of subjective experience within humans. The implication is not that the brain contains specific components and pathways as per this diagram, but that these broad functions and connectivities are present in some way:



- ***Functional processes underlying subjective experience in humans**. Habitual, rational, and innate control processes receive sensory input about the state of the body and the environment and produce body actions that effect change against the state of the body and environment. Control processes retain state, enabling multi-iteration processing. The current and recent trajectory of control process state (CP state) is captured and a dimensionality reduced high-level representation is made available as an additional sense to the control processes via the meta-management feedback loop. Environment, body, and cognitive schemas attach additional contextual and meaning information to sensory input, for example providing source labelling to the control processes. Emotions/feelings add an extra level of "importance" to some information. Attentional control mechanisms "choose" which sensory information is attended to.*

At its core, these functional processes take the externally focussd physical senses (sight, sound, smell, touch, taste) and the internally focused physical senses (proprioception, balance, etc.), and process them through the action of the innate, habitual and rational control processes, producing appropriate body actions. Schemas for body and environment attach additional contextual and meaning information to the raw senses, including "source labelling" - identifying whether the sense indicates an external or internal source. Emotions/feelings likewise attach additional meaning information to the raw senses. Emotions likely also a have more direct manipulative effect on the control processes, but that is not important for our discussion. Attention affects which information from those sense modalities becomes the focus of the control processes at any given moment.

Over the course of multi-iteration processing, the state and recent trajectory of the control processes are captured through some means and a dimensionality reduced representation is constructed. That is made available as an additional sense, a *cognitive sense*, via the meta-management feedback loop. That cognitive sense goes through the same processes for modelling, meaning attachment and source labelling as any other sense, including through the development of cognitive schemas.

The cognitive sense likely also encounters the same attentional control as for any other sense. The full extent to which attention attenuates the meta-management feedback loop cannot be clarified at this stage. Anecdotally there is definitely a way in which we can attend less or more to our own mental states, but conversely the arguments of the need for meta-management suggest that the data feed from the feedback loop should always be available. The answers are not important for an understanding of subjective experience. For now we will simply assume that handling of meta-management feedback has the same mixture of "unconscious" and "conscious" processing that we are familiar with for other things - to whatever extent that actually means.

Through the action of multi-iteration processing, this system is capable of entertaining trains of thought. The kinds of trains of thought, the topics that are possible to be thought about, and the duration of any given train of thought are all constrained by the computational capacity and representational qualities of the components of this system. In short, the representational complexity associated with any train of thought is limited by the understanding inherent within the system.

Memory enables past computational states to be recovered.

V.2 Subjective Experience States

I believe that the processes described above are entirely sufficient for the creation of all of our subjective experiences. These processes can be used to explain the specific nature of the content of subjective experience, including its "raw feels", and of the existence of subjective experience in the first place.

The meta-management feedback loop turns the cognitive state into another sense like any other, a *cognitive sense*. Any processing in relation to that sense is an act of higher-order processing - cognitive processing about cognitive processing. The outputs from such processing are higher-order representations - representations about prior cognitive processing. Those higher-order representation outputs are subsequently available as both first-order state for further first-order processing, and as a source of subsequent higher-order outcomes. Predominantly this cognitive sense will be used for straightforward meta-management purposes that elicit no specific nature beyond unconscious processing; eg: through learned habitual meta-management. It is also quite easy to see how all sorts of arbitrary thoughts about thoughts and thinking are possible from such an architecture.

With the powerful modelling capability of the brain, including the cognitive schema and with the help of source labelling, it is possible for the brain to develop models that capture various causal relations. One such causal relation is that the individual "observes" things. For example, through a long series of experiments during infancy, we develop a cause and effect model that when we see (observe) something through our visual sense, then that something exists and it can be interacted with. To put that in contrast, consider a hypothetical opposite - where our visual senses are predominated by random hallucinations - the same series of infantile experiments would lead us to ignore the visual sense because there is no relationship between the seen objects and any kinds of interactions that are possible. I must clarify that this modelled notion of "observe" is not a spoken-language understanding of such a word. It is a primitive modelling of a cause-effect relationship, and it associates that relationship to other primitive models about various kinds of actions. That primitive model only later becomes associated with a spoken word and then with its more culturally derived additional meanings.

Similar to the cause-effect model that we build from our visual sense, we can build a cause-effect model from our cognitive sense, thanks to the meta-management feedback loop. One such example is that the brain can develop a model that captures the causal relation between first-order experiences (perceptions, thoughts, etc.) and subsequent higher-order representations in relation to those experiences. Thus the brain can develop a notion of the fact that it "observes" its own cognitive state (again, in the primitive sense of that word). When the brain constructs a higher-order representation about a prior cognitive state and interprets that in relation to its cause-effect model of "observe", it produces a very special kind of higher-order thought, which I call a *higher-order observational state* (HOOS). The HOOS is the essence of subjective experience in its most basic form.

The sections that follow discuss in more detail how the HOOS produces the various phenomena of subjective experience. In order to aid that, a slightly more precise definition will be useful.

Given any prior cognitive state *x*, a HOOS is a primitive (ie: non-language, non-visual) representation of:

- "[I] [OBSERVING] [x]".

The "I" here captures a few things, constructed from the schema and meaning attachment processes:

- It is a statement (in a figurative sense) of location - ie: the place where the observation is occurring is within the bounds of my body.
- It is a statement of identity - ie: the thing doing the observing is me; it is not some other individual.
- It is a statement of source - ie: the observation came from "my" senses.

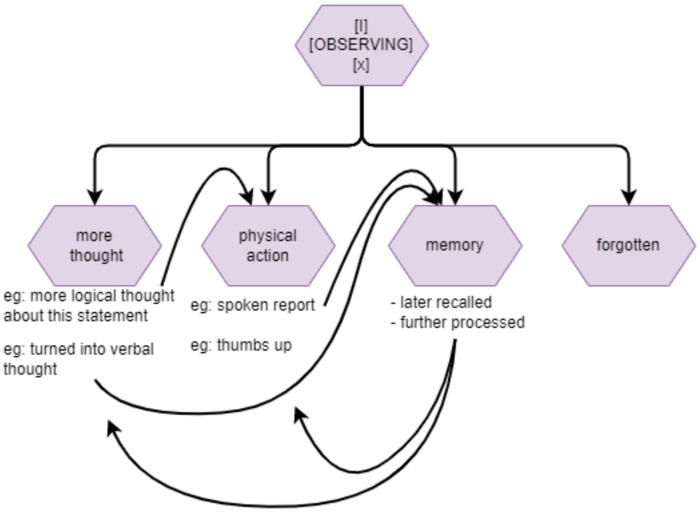
At this stage, this is not a worded "I" as in the form of worded thoughts. It is more of a raw feeling, a knowing.

The "observing" component here is also not a language reference, but simply a reference to that most primitive modelled cause-effect relationship discussed earlier. It is thus not just a statement that *x is*. Rather:

- It is an acknowledgement that a relationship is present: a relationship between *x* and something else (in this case "I").
- It is an identification of all the implied possibilities for future interactions (as understood through that primitive cause-effect model).

V.3 Interpretation of Subjective Experience States

With the construction of a HOOS, you do not yet have subjective experience. A HOOS is just a state. States don't *do* anything, unless they are interpreted. With a representation of HOOS, a few things may happen next. Illustrated in the following diagram.



- ***Options for consequent interpretation of a higher-order observational state (HOOS).*** *Various next steps are possible from a HOOS representing "[I] [OBSERVING] [x]": further thought stemming from the HOOS, physical actions that report the existence of the HOOS, "storage" to memory (particularly episodic memory), or discarding of the HOOS without any further processing.*

More thought. The HOOS may be further interpreted resulting in additional thought. For example, the wordless representation of "[I] [OBSERVING] [*x*]" may be turned into a verbal language representation "I am observing *x*" in the mind. Such further thought could also entail logical rationalisation about the fact that I am "aware" of what I observe, leading to more general statements about consciousness. Those more general statements could then be further acted upon, such as by constructing HOOS about them, producing physical action, or remembering them for later.

Physical action. The HOOS may lead to physical action, such as a spoken report or some other physical action (eg: a thumbs up) that acknowledges awareness of observing *x* [citation, recognition from non-verbal half of split brain patient].

Memory. The HOOS may be merely "saved" to episodic memory before attention switches to something else. Later, that same HOOS (or some approximation thereof [citation, recall is reconstructed]) may be recalled and further processed. Such further processing could be in the form of more thought or physical action, both of which may be further remembered and recalled.

Forgotten. The HOOS may be simply discarded, with no further processing nor memorisation of its existence.

The interpretations above begin a chain reaction, a train of thought, that convinces the individual of the existence of the HOOS and of its specific nature. For example, by way of its effect on subsequent processing, memory and subsequent recall, or physical behaviour that reports its existence. It is through that convincing that we state that we have just had subjective experience.

The outcomes of all of the above entail the *content* of subjective experience. Most importantly, this includes the specific property of the content that claims "I am experiencing this". So this also explains the *existence* of subjective experience - that it is processing of a state that claims "I am experiencing this". And why should such a state and associated processing occur? Because it is a direct result of the mechanisms underlying meta-management.

V.3.1 Rounding it Out

Some of the implications of what I am trying to say may not be obvious at first glance, so I shall now summarise by restating some of the conclusions so far.

The HOOS is effectively a state that represents that subjective experience has occurred. However, it is not subjective experience on its own; it is merely the content of the experience: a statement that the individual has observed whatever is the topic of the so called subjective experience.

The processing that follows is also not itself subjective experience; it is merely the interpretation of the HOOS representation, and has as its output another representation. In the brain, that interpretation occurs through the action of millions of individual neurons. There is no place to pinpoint as the thing that has subjective experience.

There is also no time to pinpoint as the moment of subjective experience. At any given moment in time, the only thing that exists of subjective experience is a representation that subjective experience occurred, plus the in-flight action of millions of individual computational units interpreting that representation.

The only evidence that we have for the supposed existence of subjective experience is the actions that follow the generation of a HOOS: further thought plus further representations of subjective experience, external actions such as verbal report with regard to our observations, and the subsequently processed memories of the former two.

So, while the HOOS captures the essence of subjective experience and in that sense it is subjective experience in its most basic form, it is only through the subsequent train of thought that we collate sufficient evidence of the existence of the HOOS in order for us to convince ourselves that we have just had subjective experience.

In conclusion, we need to shift our view from looking for a particular place or point in time when subjective experience occurs, to seeing it as duration over which various aspects of its content are processed. The exact point in time when the subjective experience starts cannot be identified, because it is gradually built up from earlier computational states. The point in time when a given subjective experience ends also cannot be identified as computation gradually shifts from being about that observation to doing things as a result of the observation.

V.4 Explaining Phenomenology of Subjective Experience

To show more clearly how the above creates subjective experience, let us examine some specific phenomena of subjective experience.

V.4.1 Homunculus

In chapter IV.1 I discussed briefly the old notion of a *homunculus*, a "little person" that sits in our heads and does all the observing. While that old notion has been thoroughly dismissed, it remains a very accurate description of what it "feels like" to be conscious. There is something specific about the way that we feel like we observe our perceptions and our thoughts as if the "we" doing the observation is somehow separate from the "we" doing the perceptions and thoughts.

With the understanding of a meta-management feedback loop and its associated cognitive sense, and with an understanding of the kinds of schema that the brain builds up, it is not so hard to understand how subjective experience "feels like" a homunculus. The cognitive "observation" cause-effect model that forms of the basis of a HOOS is part of the cognitive schema. That schema groups all of the cognitive actions into a single identity - a sort of cognitive organ. That identity is useful for the purpose of modelling those cause-effect relationships between states and and possible interactions. Thus the "source" of thought is labelled as that cognitive organ.

In principle there's no need for the brain to represent that cognitive organ as having a location in physical space. The brain also rarely encounters any sensory information that would enable it to locate that cognitive organ in space relative to the rest of the body. So to whatever extent the cognitive organ is located in space, it will have a vague and fuzzy boundary - something consistent with our own experience.

The apparent location of the homunculus within the bounds of our head is curious, but again I think it has a relatively simple explanation. The brain constructs body schema to identify the specific location of its senses. Many of our physical senses are hosted within our head (sight, sound, smell, taste), and so the schema naturally locates them as such. Although our vision is purposed to gain information about the world outside, our visual sense organs are located within our head. This affects interpretation of the perspective of our vision (things at the centre of our visual field are at about the height of our head, while things at the bottom of our visual field are below our heads). If the brain develops a body schema that identifies the majority of its senses as located within our heads, it's not a far stretch to do likewise for the cognitive sense.

V.4.2 Conscious and Unconscious Thought

We have some thought that is always unconscious, some that can be selectively accessed when attention so chooses, and some where we are always actively involved. This seems arbitrary. Why does some processing appear as conscious thought, while other processing does not? How does an understanding of HOOS help explain this?

Let us discuss entirely unconscious thought first. The answer simply is that there is no need for that form of processing to be represented by the meta-management feedback loop. Feedback will be optimised by evolution to contain the minimal set of information needed for effective meta-management of control processes. Anything else would serve to consume extra neural capacity without further benefit, and would simply be evolved away. There are two possible reasons why the processing does not need to be captured by the feedback loop. One is that the process is inherently stable (eg: inherently convergent), and thus simply does not to be meta-managed. A second possibility is that it is sufficiently meta-managed indirectly through meta-observation of other processing.

For the rest of processing, meta-management is required and thus the meta-management feedback loop represents it, but whether we have subjective experience of those states is a factor of attention and learning. It has been observed that cognitive processes are harder to observe as we become proficient with them (Baars, 2021). This is easy to understand in relation to habitual vs rational thought, both with respect to first-order and meta-management processing. For many of us, the answer to the maths question "5 + 3" is recalled rather than worked out - the answer simply appears in our thought some time after we consider the question. At the time of learning such a relation, meta-management is heavily involved as part of the training mechanism. At that point various HOOS about the intermediate steps can be readily produced. Once learned sufficiently well however, the answer to "5 + 3" is simply recalled in a habitual way rather than actually being calculated. Any meta-management that might be involved too would be carried out in a habitual way that simply ensures that the previously proven best outcomes are repeated - and generation of a HOOS is likely not one of those well-repeated outcomes.

For more elaborate multi-iteration processing, and in particular where rational control processes are involved, thought appears to proceed in a more slow-paced sequential way with many intermediate states. It is exactly this kind of thought where I have argued that meta-management is important. More specifically, it is exactly this kind of thought where rational meta-management is required, and that is where HOOS are more likely to be formed.

In summary, subjective experience is the collection of certain processing iterations where aspects of the processing behaviour require meta-management, and the content information of that conscious thought is constrained to the information that is needed for the purpose of meta-management. So called “unconscious thought” is everything else that occurs within the brain.

V.4.3 Stream of Consciousness

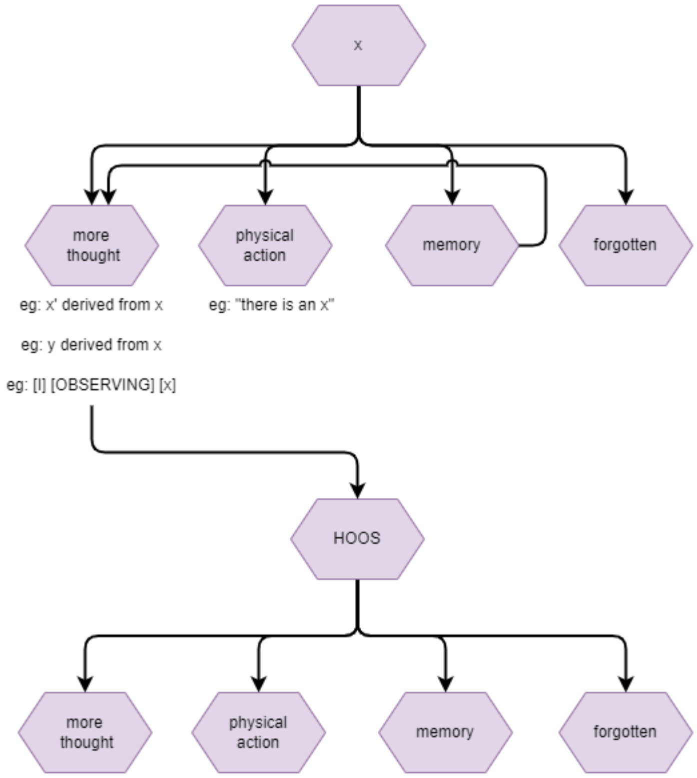
The reader may notice an implication in the statements about the continuity of HOOS and associated subjective experience. I have implied in several ways that HOOS are only sometimes constructed. This would imply that subjective experience only sometimes exists. But that is not our experience. It appears to be “always on” - a continuous stream of consciousness.

We recognise that there are times when we aren’t consciously aware of actions, such as when walking or driving along a familiar road, but that is usually because were lost in thought at the time. So even during those moments it as if we still had a continuously present, unbroken, stream of subjective experience.

HOOS and associated interpretations may occur in a number of different scenarios, and such states may be constructed through a number of different sequences of events. By examining this, we may be able to form some accurate statements about whether subjective experience can be said to be “always on”.

Let us start by considering some initial topic *x* that has no representation of subjective experience. For example, consider noticing an apple sitting on a table, and wondering about some unusual-shaped texturing in its surface.

From a processor state representing *x* a number of things may occur next, illustrated in the following diagram.



- ***Options for consequent interpretation of a representation of x.*** *Further thought stemming from x. Physical actions that relate to x in some way. “Storage” to memory, particularly episodic memory. Discarding of x before it has been “stored” to memory.*

More thought. The representation *x* may immediately be further processed (interpreted) in some way that produces new representations. Perhaps this results in further representations about different aspects of *x* (“that apple looks tasty”), or a deeper analysis of *x* in relation to the individual (“I saw that apple yesterday”). Such thought might gradually or rapidly “wander” from *x* to something else entirely. It might also involve converting the form of the representation or “enacting” it within the mind, for example as verbal language in the mind.

One of the kinds of thought possible here is to combine additional context with the existing representation. So, based on whatever attentional or motivational trigger that may cause it, a thought of *x* (“there is an apple”) may become a thought of “[I] [*x*]” (“there is an apple near me”), and then a thought of “[I] [OBSERVE] [*x*]”. This likely occurs via a series of intermediate steps that are hard to quantify without a deeper understanding of the representations actually employed within the brain.

Physical action. A representation of *x* may be further processed (interpreted) as requiring or enabling an external action. For example it may lead to the individual reaching for and taking the thing in question. It may lead to the individual speaking something about *x*: "oh look, an apple".

Episodic memory. The representation may be merely "stored" to memory, particularly to episodic memory, and recalled at a later time. When recalling memories we (usually) have some idea of where the memory came from, known as *meta-memory* (Dunlosky & Bjork, 2008; Shimamura, 2000; Fernandez-Duque, 2000; Benjamin et al, 1998; Metcalfe & Shimamura, 1994). If we saw an apple, we remember that it was ourselves who saw the apple. If however we were told about an apple by a friend, with that friend going to elaborate details about how it looked and where they found it, creating a detailed visual image within our imagination, we still know that the image we remember was described to us rather than directly observed through our eyes. This is the result of the labelling function. Our recalled representation includes labels attached at the time of original representation.

So, for any recalled *x*, there is a recalled label that identifies the original source of *x*. That is sufficient to construct further representations of many statements about *x* well after the event where *x* originally occurred. In particular, it enables construction of representations of new statements about *x* in relation to "I". For example, "I was told about *x*", "I saw *x*", "I observed *x*".

A HOOS about *x* does not need to be constructed at the time of the original observation of *x*. It is sufficient that *x* can be recalled, and that it is labelled with "I".

Forgotten. Attention may suddenly shift before *x* has had time to be "stored" into episodic memory, and it is forgotten. For some *x* that has the potential of leading to a HOOS, if it is entirely discarded before being "stored", there may be no evidence available for the absence of the ability to recall that event. So it is entirely possible that there are many moments in our life for which we never have subjective experience. Whether this kind of forgetting without evidence actually occurs in humans, and to what extent it occurs, may not be easy to measure.

Now, I have taken a certain amount of creative license here, making some assumptions about how the brain operates that could well be factually inaccurate. Nevertheless, I think it illustrates an important point of HOOS - that these states only occur because a particular sequence of processing steps have been taken that produced a HOOS, and that many other sequences exist that do not produce HOOS. Considering the finite computational and energy resources of the brain, there is every reason to expect that HOOS are only constructed under specific situations where there is something that triggered it - eg: a previously selected goal of observing one's own thoughts, or idle mind wondering looking for novel stimulation.

On the other hand, it appears to us that HOOS are always present. Firstly, this is because almost any given (recent) moment in time is captured in some way within memory, is able to be recalled, and carries with it sufficient contextual information for constructing a HOOS about that memory. Secondly, likely many moments in time are not remembered in such a way as to construct HOOS, but in the same way that they lack sufficient information to construct HOOS, they also lack sufficient information to identify that a HOOS should be present. So we never become alarmed that a HOOS is not present for a given moment, and we are only aware of those where HOOS are present. We are false-negative blind.

You can actually see this effect at play for yourself. If you recall the actions that you took earlier today you will probably perceive them as having a similar level of subjective experience as you do in the present moment. Likewise if you recall scenes from a holiday a year or more ago you will probably find a similar result. However notice that your recall of the holiday is of only a few scenes here and there, with gaps in between. Perhaps you remember being at a beach, but not how you got there from the hotel. You may be able to construct HOOS from the memories that you can recall, but not from those that you have forgotten. Now, try to recall things about your childhood - not visual scenes, but facts. Many of us have memories of our childhood that we don't actually remember from a first person point of view, but because our parents repeatedly told us. So the facts remain long after the first-person memory has been lost. For those memories, we cannot construct HOOS, unless it has been so instilled in us that we have created artificial first-person memories.

So, subjective experience is not "always on". It is not a continuous stream. It is sparsely constructed, and only at the times when it is needed or desired. But it can be constructed post-hoc from many of our memories, particularly for the more recent episodic memories, and so we trick ourselves into thinking that we had subjective experience at the time. We conclude, falsely, that we have a continuous stream of subjective experience because by sampling a few random data points we find that we can construct a HOOS about any of the situations where we expect to be able to construct HOOS.

Contrast all of the above to the meta-management feedback loop itself, the real source of HOOS. As discussed above, we don't have the empirical evidence to be certain about the continuity of the meta-management feedback, but in a first simplistic interpretation, we would expect that the meta-management feedback loop is indeed "always on". So even though HOOS are only sparsely constructed, any given thought is potentially available for construction of HOOS.

V.5 Conditions of Subjective Experience

One of the problems plaguing studies of consciousness is that we lack an understanding of the necessary and sufficient conditions for production of subjective experiences. Without that knowledge we are unable to be certain which things exhibit consciousness and which do not. For example, the Φ (phi) measure of consciousness used within Integrated Information Theory (IIT) (discussed in chapter I.3) provides a measure of the *degree* of consciousness in a system, if that system were to have consciousness. The measure of phi can also provide a *necessary* condition of consciousness -

that a value of phi below some minimum value would lead to the conclusion that the system is definitely not conscious. However we still lack *sufficient* conditions - what characteristics together are sufficient for consciousness to arise in a system?

I have so far focused on human subjective experience. The explanation I have proposed has been somewhat tailored towards that human experience. This means that the architecture described is not necessarily the only architecture capable of producing subjective experience. With that caveat in place, the following attempts to define a first draft of the *necessary* and *sufficient* conditions of subjective experience based on the analysis so far. The statements here are somewhat vague, but I believe they offer a good starting point for further refinement.

- **Access to computational state.** A feedback loop that provides a summary of CP state and feeds it back in at the level of a sense. In other words, the CP state sense should have the same capacity for applying modelling and meaning attachment as for any other sense. In a biological organism defined through evolutionary forces such a feedback loop would only evolve in the context where it meets an evolutionary need. Thus, when considering the conditions for evolution of such a feature, one must also consider that a certain level of environmental and behaviour complexity are required in order to lead to the evolution of multi-iteration processing and the need for access to CP state in order to maintain stability of that multi-iteration processing.
- **Processing of computational state.** The computational state supplied via the feedback loop must be processed in some meaningful way, such as for the purpose of meta-management.
- **Observational state.** The computational processes must be capable of taking the feedback loop and constructing a HOOS about at least some kinds of computational state.
- **Representational and processing capacity and qualities.** The representational capacity and qualities of the feedback loop must convey sufficient information to the computation that meaningful computations can be performed about that state. Likewise, the computational processes must be capable of producing new representations about that computational state, and those new representations must be made available via the computational state feedback loop. Together, repeated through multiple iterations, the representations and processing must be capable of retaining sufficient informational complexity without attenuation for long enough that a HOOS can be constructed, processed, and acted upon in some way that can be later referenced as evidence that it existed in the first place. There is a “weakest link” effect here - if any one step in that multi-iteration processing loop fails to sufficiently process or represent the self-referential or the observational aspect of the HOOS, then the phenomena of subjective experience collapses.
- **Modelling.** In order for a HOOS to be constructed, the computational processes must perform sufficient and sufficiently flexible modelling that they can a) model the existence of a cause-effect relationship involving the observation of computational states, b) identify specific instances of such observations, c) label a specific computational state as being an instance of an observation, and d) further process that label against its associated computational state.
- **Attention.** If the computational processes incorporate a mechanism of attentional focus, then that mechanism too must retain attention long enough that a HOOS can be constructed and processed in a meaningful way.
- **Self-schema.** It helps if the computational processes incorporate a schema of the self. However it is likely not necessary.

V.6 Degrees of Subjective Experience

There is a lot of room for variations within the statements above. In particular, there is a possibility of variations in *degree*. Representational capacities and qualities can be varied while still meeting the conditions. The degree of representational complexity and the specific nuanced qualities of those representations all affect the kinds of HOOS that can be constructed. The extent to which the individual has a concept of their own identity and the complexity and properties of that representation also influence the kinds of information that can be attached to a HOOS, and the kinds of interpretations that are possible from a HOOS.

Trains of thought are at constant risk of attenuation due to representational capacity limits or attentional focus shifts. The duration over which a train of thought is possible that covers the topic of a HOOS affects the depth and complexity of subsequent actions and processing. Where consequent processing produces new models about concepts of self-awareness, the ability to form those models and the depth of detail to which they can be constructed is affected by the duration of such trains of thought. A sub-human intelligence may observe its own computational state, but be unable to consider it beyond that mere observation.

Thus there are many possible variations of degree of subjective experience, and of degree of representational modelling that can be constructed as a result of that subjective experience.

Does this level of degree extend towards zero? In the information theoretic sense, can you have 1 bit of subjective experience? The answer is a firm no. The processes described above require a minimum level of representational and processing complexity, below which no HOOS will be constructed. Further theoretic and empirical work will be required to identify where that limit lies.

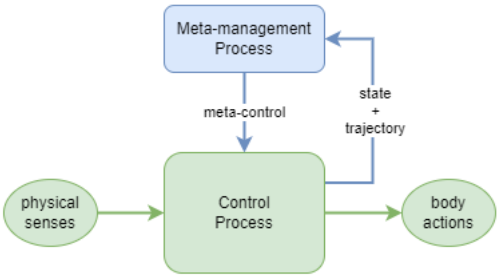
V.7 An example of an alternative consciousness

I shall conclude this part with a thought exercise to consider what it might like to be a consciousness with a slightly different brain architecture. This is largely for a bit of fun, but the particular example I shall use also highlights once again why I believe humans employ an inline meta-management architecture.

Consider your own human introspective experience. You can observe your physical senses of sight, sound, touch etc. Likewise you can observe your own thoughts, images that you imagine, words spoken that only you can hear. You can freely switch attention between any of these, and to some extent even choose to attend to multiple sensory modalities at the same time including both external physical senses and that of your cognitive sense. Importantly, you experience a high degree of detail about the external environment, particularly when you look at it.

As claimed earlier, that is all because of the inline meta-management architecture.

Now consider the independent meta-management architecture described in chapter III.6, and repeated here:



- ***Alternative meta-management architecture.*** *An independent meta-management architecture.*

The first-order control process only receives as input the physical senses. It holds state related to the processing of those senses and associated tasks, but does not have access to directly observe that state. Likewise, though the behaviour of the control process is affected by signals from the meta-management process, it has no direct computational access to those signals. The control process cannot form any conclusions about its own state.

For the meta-management process, its availability of access to information is the opposite as for the first-order control process. It takes as input the computational state of the control process and so has a clear and detailed "view" of that state. In contrast, it has no direct access to the external physical senses. The control process state likely incorporates some information about the external senses, but the representation held within that state is tuned to the needs of subsequent processing rather than for the purpose of representing current environmental and body state. For example, it may be focused more on conclusions drawn such as about the next action that should be taken. Thus the meta-management process has only indirect evidence about the state of the environment and body.

Where does the HOOS get constructed? The control process has no access to its own state, so necessarily if a HOOS is to be constructed it will be constructed within the meta-management process. Furthermore, it is reasonable to state for our thought experiment that the meta-management process has its own state, so that it can hold that HOOS once it has constructed it.

Oh, but if it has state, then its state trajectory needs to be meta-managed. We have already stated that a feedback-loop and specific meta-management processing is required for anything beyond simple reactive/habitual meta-management. So in order to avoid an infinite regress, our meta-management processor will need to have its own inline meta-management for the purpose of meta-meta-management of itself. Well, maybe that's true, maybe it's not, but this is our thought experiment so we can decide what we want. While we're at it, let's assume that this meta-management process has all the same capabilities of modelling and schema generation that we do.

What do we have now? We have a meta-management process that can observe and manage its own computational state in the same way that we humans can. It has an input sense that conveys information about something external to itself: the state of the first-order control process. The meta-management process has effector capabilities against that control process state, and so it can build cause-effect models encapsulating its ability to "observe" those states. Thus, it considers the control process state to be a part of "itself". However, its awareness of the external physical environment and of the body that contains all this processing is vague at best.

I imagine that the subjective experience of the meta-management process (for that is where the subjective experience occurs in this example), would be something akin to a pilot sitting in an advanced flight simulator but where the computer screen "windows" have been turned off and there's no pistons underneath the cabin creating the effect of motion. They can turn knobs and move joysticks, changing their internal state. They can see readouts on the altimeter and compass, indicating their current state. But they would struggle to grasp that any of that relates to anything that might occur "outside".

Part VI - The Intuitional Gap

VI.1 The Explanatory Gap and the Problem with Intuition

A common issue with materialistic explanations of consciousness is that they suffer from an *explanatory gap* between the properties that they are capable of explaining and the apparently inexplicable properties of conscious subjective experience (Levine, 1983; Van Gulick, 2022). The claim is that none of the physical processes understood today can explain the “raw feels” of consciousness. This “Hard Problem of Consciousness” (Chalmers, 1996) has been used to justify non-physical explanations such as *pan-psychism*, where consciousness is treated as a fundamental property of nature.

In practice, it is hard to clearly define what these properties are that cannot be explained by materialistic means. Terms like “raw feels” and “what it is like” don’t offer much detail. It seems that the best anyone can do is to provide a few examples and then hope that their readers recognise what they are attempting to refer to by considering their own experiences. That is a reasonable starting approach when one lacks a deeper scientific understanding. Generally, as scientific understanding of a topic improves, the use of examples gets replaced by definitions that become more precise over time. Unfortunately, that doesn’t seem to be occurring for the study of these non-physical properties of consciousness. We seem to completely lack any basic idea that could even get us started to explain these properties, let alone consciousness itself (Block, 2002).

If we lack any conception of how to handle all this, where do these ideas come from in the first place? Our intuition. We introspect our own minds, examining what we observe, forming ideas about what those things are that we observe, and then try to describe them to others. This all occurs within the nebulous thoughts that float around our heads, so we lack any means to measure those observations, to hold them up and compare them accurately against anything else. We have no way of proving that there is anything special about our internal observations. We just *think* that there is something special about them.

Now, we know that our minds have their limitations, that they do not always produce rational conclusions (Cushman, 2020). These ideas that we form about the special nature of subjective experience were constructed by these same irrational minds. If these ideas were formed by one irrational mind examining another, then we might be better off, but unfortunately that is not possible. So these ideas were formed by each irrational mind examining itself, using those same irrational processes to interpret the observations of questionable accuracy made of that irrational behaviour. We are not in a strong position.

It is from here that I provide my explanation. The explanatory gap is not a lack in our ability to explain the things that *are* through materialistic processes. It is an *intuitional gap*: the gap between what is known to exist and what our faulty intuition makes us believe exists.

To make this case, I shall now offer two arguments that support this conclusion.

VI.2 Argument By Elimination

The notion of a *philosophical zombie* (p-zombie for short) offers as a tool for examining some questions of consciousness (Chalmers, 1996). It works like this. Firstly, you imagine two individuals, one a real human being, and the other a zombie that walks and talks like a human but is claimed to be missing some important aspect of its experiential inner life. Then you use this hypothetical example to argue one way or another about the nature of those missing aspects of its inner life and to what extent they serve a purpose.

Let us use such a tool, building up from a simple example to a more complex one, and see what we find.

The first p-zombie is an anthropomorphised Turing-test passer. It looks, walks, talks like a normal human and in all other ways behaves just like a normal human. It is *behaviourally indistinguishable* from a normal human being - anyone attempting to determine whether this individual was a real human or a philosophical zombie would be unable to make a determination with any confidence better than 50%. And that is where the similarities stop. The zombie controls all of its behaviour through processes that do not produce subjective experience of any sort or degree. Perhaps for example the structure of its brain is such that it has no access to its own thoughts in any direct way. Where it produces the appropriate external behaviours with regards to thought, it is done via a kind of *blindsight*. Here I am making reference to experiments showing that some people with specific kinds of damage to their visual processing can still react to visual scenes, even though they don’t know that they do it (Trevethan and Sahraie, 2009; Weiskrantz et al, 1975). If we were to attempt to imagine the inner life of the zombie, it would be as if it was permanently in a dreamless sleep - and all of its behaviour was done while sleepwalking.

Such an example is not hard to imagine. For many centuries people believed that humans were entirely different from animals. Even when it was recognised that humans are also an animal, it was still believed that humans were the only animal to have consciousness [citations]. So the above is just an animal that looks like a human. Well, looks, walks, talks and behaves in all ways like a human. Perhaps that is stretching the analogy a little too far. How about an alien? Surely there are aliens that we may one day meet that have entirely different kinds of brains, employing entirely different kinds of structures and processing. Maybe one of those alien species is like a p-zombie - it controls all of its behaviour without the need to observe its own internal processing state. Mind you, it wouldn’t look or behave like a human, and if we’re trying to use this to gain an understanding of the “extra stuff” that is added with subjective experience then there’s just too many other things that are different here too. While we’re questioning the efficacy of our own hypothetical scenarios, it’s worth taking a second look at this blindsight-powered verbal report of thought. Does that really make any sense? Dennet calls these these experiments *intuition pumps* (Dennet, 1991), and the problem with them is that they seem to make sense on face value but break down when you look deeper. If our p-zombie can report on its own thought, and report on the fact that it has thoughts, and report on its thought about thought, just like any human could, how could it possibly do that without the same internal reporting mechanisms that are available to humans?

Let us give up on that example and move to something more advanced. Where the first p-zombie was behaviourally indistinguishable from a normal human, the second p-zombie is additionally *computationally indistinguishable*. To whatever extent that it might be theoretically possible here or anywhere in the future to examine all of the computational processes within the mind of this p-zombie, and to compare them to a normal human, no difference would be found. Furthermore, the p-zombie has introspective access to its own computational state. The representations it constructs about those computational states, and the interpretations that it is capable of making from those representations are indistinguishable from the kinds of representations and interpretations possible within a normal human. To the greatest extent this p-zombie *is* a normal human. The only limitation imposed on it is that its computational state can only be derived from materialistic physical processes that we understand today.

Now we get to ask two very interesting questions about our second p-zombie. Firstly, is it possible for a collection of entirely materialistic computational processes to produce the same kinds of computational results as a human? Secondly, what, if anything, does the human have that the p-zombie does not?

Many past issues with materialistic explanations of consciousness have been based on the claim that the answer to the first question is “no”. I believe that we are finally in the position to strongly affirm the answer “yes”. Some of the past struggles have been a failure to find any explanation for the functional purpose of consciousness. To that we have the explanation of multi-iteration processing, complexity, and meta-management. Other struggles have been due to an inability to find a plausible explanation for what kinds of states subjective experience is. For that we have cause-effect models of observation, HOOS representing that observations have taken place, and interpretation of those HOOS producing further representations that a HOOS has been experienced.

More generally, I believe that the mechanisms described within this treatise are capable of producing all of the computational states that we know and love and call subjective experience. This includes the “raw feels” of that subjective experience, which itself is just another representation and associated interpretation. Moreover, not only are these mechanisms *capable* of producing such computational states, but they are *likely* to produce such computational states, given sufficient modelling capabilities and mind-wondering characteristics that are typical of humans.

So, can a materialistic computational p-zombie construct the same computational states as humans? Absolutely.

With the first question answered in the affirmative, the second question reduces to asking whether there exists anything *beyond* computational state. Here we hit up firmly against the intuitional gap. For many, there absolutely must be something extra beyond computational state (Levine, 1983). Many claim that our mental experience cannot be reduced to computational state, that there is some part of thought that occurs *beyond* the confines of computations. I have already combatted that claim to the best of my ability while answering the first question.

There are also those who accept that materialistic computation is sufficient for control of our human bodies and mind states, but that there is still something extra. This leads to the curious case of epiphenomenalism, where subjective experience and consciousness in general are considered non-causal - having no impact on those computations (Robinson, 2019). This is easily discarded as absurd. If consciousness is non-causal then it should not influence my report that I am conscious. But I do report that I am conscious, which either means that my consciousness *caused* that, or that something else caused my report and that my consciousness is reduced to…well…something that I am not even aware of.

At this point this second question turns on itself. Does there exist anything beyond computational state? Why would there be? If all human behaviour, external and internal, can be explained through materialistic processes, what need could there possibly be that invokes “something extra”? By occam’s razor, I claim that there is none.

VI.3 Argument By Delusion

A common belief is that the evolved purpose of our physical senses and associated perceptions is to reveal the true state of the world to our logical processes, in order that we might respond to them accurately. Of course that purpose can never be attained as there will always be information about the world that we cannot obtain. For one thing, our computational and representational capacity is limited and thus could never, for example, represent all of the state of the world to the level of individual atoms. Secondly, our senses only ever catch a glimpse of the world, from the point of view of wherever we happen to be located at the time. So the common belief can be stated more clearly as: the evolved purpose of our perceptions is to reveal the true state of the world as accurately as possible, within the computational and representational limitations of our brain.

Another aspect to the process of perception is interpretation. The glimpses provided by our physical senses at any given moment can be augmented by prior gained knowledge about the world. Thus the raw senses are interpreted within the context of a learned *model* of the world. This augmentation serves to add further information to perception. Hopefully it also increases accuracy of perception at the same time, but that is not necessarily true. In 1781 Immanuel Kant argued extensively that our perceptions are biased by a priori assumptions and that these assumptions never accurately reflect reality (Kant, 1929).

An excellent case in point is that of perception of colour. We know now that colour simply does not exist in any way that is even similar to how we perceive it. What does exist, to the best of our scientific knowledge, are electromagnetic waves of various frequencies interacting with the nano-scale texture of objects surrounding us, being absorbed, refracted, reflected in complex ways, and leading to further electromagnetic waves of various frequencies making contact with the cones and rods in our eyes. The electromagnetic waves that happen to interact with our eyes form a complex *spectral power distribution*, a mixture of different wavelength frequencies with different amplitudes, not to mention the variations and possible effects of phase and polarization. In contrast, the photosensitive receptors in our eyes produce only three signals (for most people) for any given “spot” within the visual field: the amplitude of signals detected by the S, M, and L cones. These *sample* the received electromagnetic spectra over heavily overlapping regions of wavelength with response peaks that roughly correspond to what we ultimately perceive as blue, green-yellow, and yellow-red colours, respectively. By the time that we perceive these signals, however, our brains have interpreted those three signals into a single "colour" in a virtual colour space. That colour space has values that map directly onto single electromagnetic wavelengths, so-called spectral colours, such as blue (420nm), green (534nm), and red (564nm). It has non-spectral colours that exist only as a combination of multiple wavelengths, for example white (a mixture of roughly equal amplitudes of blue, green, and red wavelengths). Additionally, we have metamerism effects whereby we perceive the same colour for completely different spectral distributions. For example, while the perceptual colour yellow can be produced by a single wavelength, it can also be produced through a combination of blue, green, and red wavelengths (see the Wikipedia article on Metamerism for details).

More recently, Donald Hoffman extended this via the *User Interface Theory of Perception* (Hoffman, 2009, 2023; Hoffman et al, 2015). It claims that not only is it wrong to assume that our perceptions are accurate representations of reality but that it is an entirely wrong way of looking at the underlying purpose of perception. When looking at perception with an evolutionary perspective, we see that perception is in aid of survival, and so only those aspects of perception that increase the fitness of a species will be retained. Hoffman likens this to the way that a user interface is designed to help us get a particular task done while avoiding unnecessary details, and states "natural selection fosters perceptions that act as simplified user interfaces, expediting adaptive behaviour while shrouding the causal and structural complexity of the objective world" (Hoffman, 2009, p. 163).

This can be made even clearer if we look at this from an optimisation point of view. The purpose of the group of neurons that control our behaviour is to control that behaviour as effectively as possible while consuming as few resources as possible (energy and time). Simpler computations that perform almost as good as more complex computations win. The control neurons need to know about the state of the world, as generated by perceptual neurons. The purpose of the perceptual neurons is to construct representations of the world that are most effective in aiding the needs of the control neurons, while also using as few resources as possible. Simpler representations that lead to almost as good outcomes win over more complex representations. Furthermore, abstract representations that do not have direct one-to-one mappings to physical reality (eg: virtual colour spaces and metamerisms) but that aid survival win over representations that are more realistic but cost more.

Unfortunately, a system that evolves to maximise such a tenuous, abstract, and moving target as survival will on occasion misfire. This is exemplified in the effects of hallucinations and delusions. It is common enough for the brain to be thoroughly convinced of something which is clearly wrong to everyone else. Furthermore, even with the evidence clearly presented and accepted by the individual, they refuse to change their deluded conclusion. Hallucinations and delusions have one thing in common that extends the problem: the tools that one depends on in order test the accuracy of their perceptions are the same tools that are in some sense "broken", and thus they are unable to accurately inform the individual of the inaccuracy of their perception.

In the famous case of the mathematician John Nash, who experienced hallucinations, voices, and irrational thoughts for much of his career, he was able to "intellectually reject" his hallucinations (see Wikipedia article. This is not particularly hard for visual and audial hallucinations - you can step forward and check that the visual hallucination can be interacted with in the normal way, or ask someone for a second opinion. It is harder, but also possible, for one to identify mental hallucinations and irrational thoughts through similar means. In general it requires a stable ("sane") reference point - being able to compare ones own thoughts to that of others. It is much harder still if are no sane reference points because everyone around you is also delusional in the same way.

The notion that our subjective experience carries any extra "raw feels" beyond computational representation and interpretation, and our intuition that materialistic processes cannot explain those raw feels, is a delusion. A collective delusion. A delusion that stems quite naturally from the fact that our perceptions are optimised for fitness not accuracy and from the specific architecture of our brain.

How else could our perception be of our internal computational state? What other kinds of "raw feels" might it be possible for one to experience. Could there be any computational way of observing our own computational state without "raw feels"? One aspect of "raw feels" is that we observe our inner state at a certain level within its processing - not at the true raw sensory inputs, but after a significant amount of interpretational processing has already occurred; and not every consequent processing step, but just at certain sampling points afterwards. The lack of observational access to all those intermediate computations creates an impression that certain things "just are". What would it be like if we could observe the electrochemical interactions within each and every rod and cone receptor in the eye, observe each resultant electrical impulse as it leaves the receptor and travels down the optic nerve, watch as millions of such signals reach the visual processing centers of the brain [citation, number of synapses in optic nerve], and continue to observe each and every neuronal computation that ultimately results in the brain forming a representation of an apple on a table followed by an abstract thought about that visual image? Firstly, any meaning that we form from our observations is driven by the models that we construct through experience, and the specific structure and content of those models is driven by the information available at the times of our experiences. So the mental models that we form about the concept of "observe" would be completely different if we had such detailed access. For example, our concept of "observe" would probably be more like a concept of "process". Secondly, unlike the joyful ride that is the 2014 Hollywood movie "Lucy", evolution would never produce an individual that had such access to their internal processes. Ignoring the obvious issue with infinite regress on the number of neurons required to process all that information, there is simply no need to observe the internal operations of the brain processes to such an extent. Evolution creates a "user interface" to our computational state that is as simple as possible, and at the most suitable level of abstraction, in order to meet the needs that they serve.

What about other aspects of "raw feels"? I have so far been unable to find a list of the apparent properties of "raw feels", so I find it very hard to be more specific at this point. Until such a list is identified, a more general statement will have to suffice. The "delusion of raw feels" and the specific properties of that delusion are due to the compounding effect of all of the computational processes within the human brain. Not the least of those processes and their effects include:

- cause-effect modelling that constructs concepts of “I” as different from other individuals, and the specific properties of the constructed models
- cause-effect modelling that constructs concepts of “observe” from a) the fact that I can obtain information about the external world independent of others, b) the fact that I can obtain information about the state of my own computational processing that others cannot obtain, and c) that in both cases that information can be useful to me.
- meaning attachment that associates “I” with “observe”
- feelings/emotions that serve other utilitarian functions through the action of attaching further contextual information to the combined concept of “I observe”.

So we observe our own thoughts, identify them as our own, and feel strongly about that identification and all that is connoted by it. Alternative architectures would construct different concepts and produce different effects, but for humans the particular concepts and effects that we produce in relation to “raw feels” are a foregone conclusion driven by our evolution and the societies in which we live.

Part VII - Predictions and Conclusions

In this final part I shall take a little time to draw out some of the consequences of my theory of subjective experience. There are a few reasons for doing so. Firstly as a place to fill out some of the details of explanations presented earlier that otherwise would have made them longer. Secondly as a chance to deal with some of the likely rebuttals. And lastly as an opportunity to be a little more speculative.

VII.1 Convergence of Meta-Management

In developing this theory I have been palpably aware of concerns that what I am proposing makes no sense because such a system could never be stable enough to operate effectively. I have made the claim that habitual and rational systems learn off each other. That alone can produce unstable results. Indeed, I have made the claim that they are so unstable that they need a meta-management process in order make them stable. But then I’ve twisted that assumption on its head by claiming that somehow those very same unstable systems can meta-manage themselves. Without sound empirical results what basis do I have to believe that this could possibly work? And if it’s not a feasible architecture, then my entire argument for the basis of subjective experience may be in question.

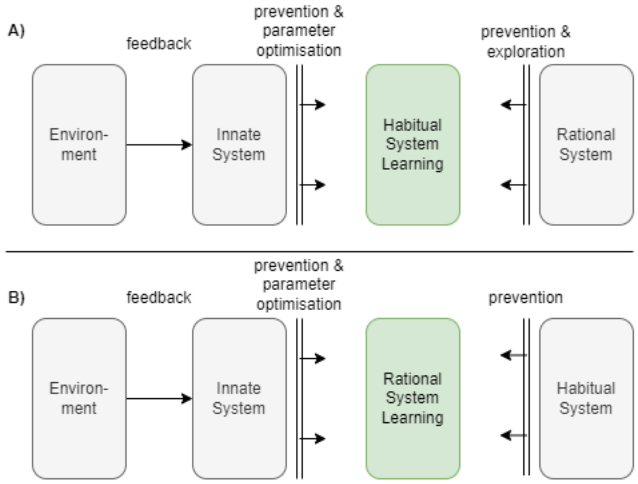
I have attempted to keep this treatise focused on the theoretical for two reasons. One is that I wanted to provide food for thought to those who claim that subjective experience cannot be explained through materialistic means, by showing that it is indeed possible. To that end I only needed to present a *plausible* materialistic explanation. The detail that I have included hopefully provides a reference point for others to pick up from. The second reason is that I only have so much time in the day. It takes long enough to generate and write about a plausible theory. It takes longer still to do all the empirical research needed to provide data proof behind such a theory. Thus my original intent was not to provide a watertight solution for every question, and I certainly did not intend to prove how such a system could attain stability.

So, perhaps it is more to alleviate my own discomfort than that of anyone else that I shall now offer a few suggestions about how such a system could attain stability.

For this endeavour I shall once again borrow from the AI literature, in particular the concept of *convergence*. In AI research, a learning system is “trained” through repeated trials. After each trial a measure of the system’s behaviour is produced relative to the objective that the AI researcher has imposed upon the system. So, for example, if the system is being trained to navigate a maze environment, the objective measure may record how many steps are required in order to complete the maze with the intention that the system requires fewer and fewer steps as it learns. The system is typically initialised with random parameters, resulting in wildly chaotic initial behaviour. The system is said to *converge* from that chaotic behaviour towards behaviour that is a) more stable (ie: consistent behaviour for similar problems), and b) with higher utility (ie: it is more successful in meeting the objectives and does so more efficiently).

In a biological setting, the habitual and rational systems need to converge likewise. The difficulty is that aside from some basic principles like “stay alive”, the objective measure is not known a priori. Rather, various factors play together to constrain behaviour and those constraints become refined over time as the individual develops and learns. As the individual begins to understand the world around them, including the social world, knowledge of what is good and bad behaviour further refines and narrows what is incorporated into the ever changing model of objective measure. Thus the ever refined constraints impose ever narrower boundaries on ideal behaviour, which I refer to as *narrowing convergent boundaries*. The narrowing convergent boundaries are created and refined by constant interactions between the innate system, habitual system, the rational system, and pressures from the external environment.

Panel A in the diagram below illustrates how the external environment, the innate system, and the rational system may act to apply narrowing convergent boundaries against the learning within the habitual system. The individual interacts with the environment, firstly in an experimental way. Feedback signals received from the environment and interpreted by the innate system apply an after-the-fact convergent pressure. This presents in the form of negative and positive rewards and in the form of prediction error. When the rational system is operating near its optimum it provides a before-the-fact convergent pressure. This presents in the form of prevention (usually to avoid a catastrophic error occurring) and in the form of encouraging exploration. For example, the individual may have developed a habit of walking a particular route through the forest to the nearest watering hole, but the rational system will prevent that habituated path because it remembers that a tree had fallen and blocked the path at a location that is currently out of sight.



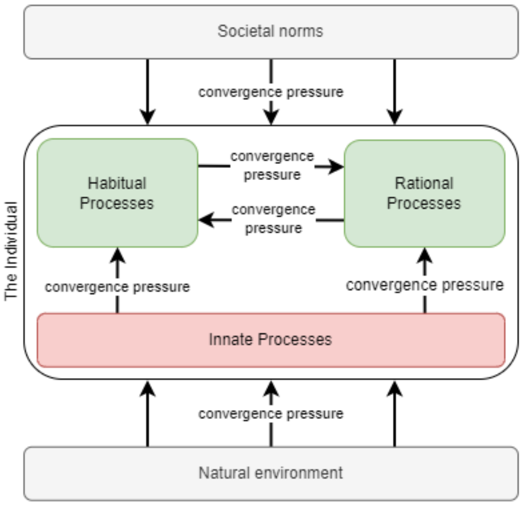
- ***Narrowing convergent boundaries.*** *A: behaviour of the habitual system is trained through narrowing convergent boundaries created by the environment and innate system on one hand, and by the rational system on the other. B: behaviour of the rational system is trained through narrowing convergent boundaries created by the environment and innate system on the one hand, and by the habitual system on the other.*

When the rational system is not operating near its optimum, the environment and habitual system act to apply a narrowing convergent boundary, illustrated in Panel B. The external environment applies after-the-fact feedback in the form of positive and negative rewards, interpreted by the innate system, which the rational system can use for future improvement through both rational adaptation and through whatever underlying learning mechanisms it operates on. As actions become more strongly habituated, it becomes harder for the rational system to counteract them. Thus the habitual system also applies a before-the-fact prevention against the most catastrophic errors.

The modalities of environmental pressures requires some further discussion. Discussion in chapter IV.3 already mentioned pure positive and negative rewards such as hunger, satiation and pain. It also mentioned learned higher-order feedback mechanisms that layer on top of those innate learning mechanisms.

One further learning layer exists on top. With such advanced and flexible learning and meta-management comes a high degree of freedom. For example, many different modes of behaviour are potentially equally effective. Most negative short-term outcomes can be adapted to, so that the total outcomes is net positive. This can lead to a high degree of instability - if so many modes of behaviour are equality advantageous, there is little learning pressure to maintain a consistent mode of behaviour. In turn, that can lead to constant re-learning within the habitual system. Ultimately, the individual may converge extremely slowly towards its most optimum set of behaviours. It has been suggested that society plays an important role here (Damasio, 2006).

Effectively, society is a collective behavioural exploration and learning system that spans a longer time scale than that of an individual. There are many possible systems of societal norms, and the societies that thrive will tend to have produced systems of societal norms that benefit the society and the individuals within it and those systems will tend to remain relatively stable over time. Through constant feedback from adults and peers, a naïve individual's higher order processes are trained to produce behaviours that are consistent with the norms of the society in which they live. That societal pressure thus provides a significant portion of the narrowing convergent boundary exerted by the environment on the individual's learning, particularly for its rational systems including meta-management.



- ***Convergence pressures operating against the control systems of an individual.*** *A human individual has many different pressures that create the narrowing convergent boundaries in which they develop. At one extreme this includes primitive environmental factors like finding food and avoidance of dangerous predators. At the other extreme this includes complex and nuanced societal norms. The innate, habitual, and rational processes interact to interpret and model those external pressures, internalising them so that the individual's own brain can provide its own objective measure.*

VII.2 On the Causality of Consciousness

Many of the debates around the causal nature of consciousness are in the form of deep metaphysical ontological questions about mind-states, physical-states, and the kinds of relationships between them (Yablo, 2002; Kim, 2002; Robb et al 2023; Robinson & William, 2019). Those debates arise even assuming a materialistic basis for those mind-states. I have neither the skill nor the inclination to attempt a response to those philosophical questions. I merely hope that those more philosophically skilled will find my scientific explanations as a useful grounding for their metaphysical arguments - hopefully in support of my own conclusions.

To the more practical scientific questions of causation is what I shall devote some time here.

In now famous work by Libet, it was found that electrophysiological readiness potentials (RPs) that reliably indicated the beginnings of cerebral function leading to a decision to act occurred between 350ms to 400ms prior to the individual being consciously aware of that decision (Libet, 1985 & 2004). This result suggests that conscious awareness of a decision is an "after the fact" event, and worse yet that it is epiphenomenal - having no causal effect. Libet softened the blow a little in that same work in the discovery that subjects could "veto" the motor performance if a counter-decision is made (supposedly consciously) within 150ms after the conscious recognition of the first decision. That work is further backed up by observations of causal chains of brain activity that occur over several seconds leading up to a conscious decision, and that the specific patterns of activity can be used by an observer to predict the selection given a predefined set of available choices (Haynes, 2013; Soon et al 2008).

Other experiments have found that people can judge themselves to have at least partial control of a movement that in fact they lacked control of (Wegner and Wheatley, 1999; Wegner, 2002; Linser and Goschke, 2007). And to the more abstract examples, it has been found that people will sometimes confabulate their reasons for judgments - unknowingly giving false reasons for their conclusions, when the experimenter knows that their judgements were made for other reasons (Nisbett and Wilson, 1977a & 1977b).

I can see two questions of interest here with respect to consciousness and causality, stated as such:

1. Is subjective experience the *cause* of its intent?
2. Does the act of having a subjective experience have any effect on the individual's subsequent state?

Both of these questions need some clarification before I continue.

Recall from chapter I.3 that experiential *intent* does not take its usual English meaning, but rather is best thought of as the target or focus of the experience. This is clearest in the case of thought: is the subjective experience of thought the cause of that thought, or does it occur afterwards? Replace "thought" with "decision" and you have the questions posed by Libet.

In order to provide some clarity to the second question, I shall rephrase it with a slightly different meaning and it will be to this meaning that I attempt to provide an answer: is it ever the case that the state of having a subjective experience has causal effect on subsequent computations?

The empirical results that we have from the works of Libet and others is a good start to begin to answer these questions, but their interpretation is still speculative in nature because we lack a clear understanding of the functional purpose of consciousness and its underlying mechanisms. Perhaps we can now add something more concrete to the debate by examining the effects of the claims of this treatise.

Firstly, as claimed, subjective experience is created from a sequence of computations. There is not just one kind of computational sequence that we need to consider. I shall consider three broad kinds of sequence:

1. sequences of events that for some meaningful duration were controlled entirely through habitual processes with minimal or no use of the meta-management feedback loop
2. sequences of events that produce HOOS
3. sequences of events that employ significant meta-management processing, but form no HOOS

Clearly in the first case subjective experience is not causal as there is no subjective experience in the situation, no HOOS is constructed and interpreted. The events may be later recalled, and a HOOS constructed from them, and at that point there will be a subjective experience of the memory of the events. But at the time of the events in question, there was is subjective experience.

In the second case, a HOOS is constructed. A HOOS is always *about* something, and that something always occurs *prior* to the construction of the HOOS. In the overly simplistic case of a HOOS about seeing an apple, clearly the HOOS is in no way causal in terms of the seeing of the apple. Likewise, a HOOS about a decision to eat that apple is also non-causal in terms of that particular decision.

But this does not mean that HOOS are non-causal in absolute terms. On the contrary, HOOS may be further interpreted. Whatever result is produced by that interpretation is *caused* by the HOOS (in conjunction with the specific nature of the interpretation process executed). That result could be the start of a lengthy thought process about the nature of self-awareness, it could be a body action reporting that you observe whatever the target of the HOOS was, it could be a decision to eat an apple in order to produce HOOS about eating apples, or it could be any number of other more subtle effects, all of which would be caused by the presence of that original HOOS.

What happens if a HOOS is constructed but not interpreted? HOOS are just representational states after all. A HOOS is entirely non-causal with regards to any computational results that are produced at that moment if it is constructed and merely “stored” to memory before attention switches to something else. But it still *causes* something - the particular episodic memory of that HOOS. If that same HOOS was constructed and then immediately discarded without memory storage, then the answer is a little less clear. Did that particular HOOS *cause* the subsequent processing outcome that concluded that attention should be shifted elsewhere? If yes, then the HOOS caused something. On the other hand, if the HOOS was discarded due to some other attentional control process that was entirely unaffected by whatever state happened to already be present, then we can say that the HOOS *caused* nothing.

Another nuance comes from the fact that processing in the human brain is not so step-wise as it is in von Neumann computers. The events entailing seeing an apple and then deciding to eat the apple do not occur as two discrete processing steps. Rather, recall from chapter I.4 that processing outcomes are reached through a continuous motion of waves of neural signalling spanning many regions of brain with multiple repeated and varied interactions. If anything, the conclusion “eat the apple” slowly resolves from a vague proto-decision until it is clearly defined enough that actions can be initiated from it. And even then the proto-decision may still resolve further while action is already in play, perhaps to further refine the action plan. Likewise, the computational state that holds the representation of that proto-decision is not a discrete one-thing-holder like in von Neumann computers. It is a mass of neurons that represents through whatever fluffy organic mess of a means that evolution has stumbled upon. We have every reason to suspect that it is quite capable of holding multiple proto-anythings at the same time, including a proto-decision-to-eat-apple and a proto-HOOS. Ultimately the decision to eat the apple may have been caused in some way by the HOOS that was being constructed at the same time. That is not to say that a HOOS would always be constructed in parallel, but that it could be.

One last nuance considers that the only way that we can know that a HOOS was constructed is through its effects. A HOOS that is entirely discarded before producing any effect is merely a representational state that was never interpreted. At the very least it must be “stored” into episodic memory so that it can be later recalled and reacted to. If that does not occur, then no subsequent computational processes are affected by the HOOS. From a computational point of view, the question of whether such a discarded HOOS was causal collapses to a tautology. But that is not the end of it. The construction of the HOOS used energy resources, resulting in changes to blood flow. An experimenter examining the activity of the individual’s brain via fRMI would see, ever so briefly, those associated blood flow changes. If, perhaps in the future, we had some advanced scanning technology that reveals the precise state of individual neurons, we would see even more clearly the brief existence of that HOOS within the mind of the individual. And all the while, the individual would claim that no such HOOS existed at that time, because they have no computational evidence that it existed.

There is one last kind of computational sequence to consider, where significant meta-management processing is employed but forms no HOOS. I have included this case last because it quickly becomes convoluted by questions around how we map this new mechanistic understanding onto our more nebulous conception of subjective experience. A hint of the issue has already appeared in the preceding paragraph. In many respects, this is a definitional problem and thus heavily affected by opinion, but let us now attempt to examine this to some extent. Meta-management processes, when they occur, are causal of what comes next (at least to every extent worth considering at this point). But meta-management processes do not always produce HOOS, so we likely would not choose to pin the definition of subjective experience to the action of meta-management processes per se. Additionally, the construction of HOOS do not require that meta-management actions be undertaken at the time. HOOS are constructed not from the interpretive processes that is meta-management, but rather from any processing that incorporates sensory data supplied by the feedback loop. We can also say that sometimes that sensory data provided by the feedback loop is processed without any kind of subsequent processing that recognises that the feedback loop was used. In other words, the computational state can be observed (via the feedback loop) without a HOOS being constructed about that observation. So we can say that a HOOS is just a recognition that an observation took place, but that the observation still took place regardless of whether a HOOS is constructed about it.

Now, that observation was indeed causal. By my definition, the act of computationally processing the signal from the feedback loop *is* the observation, and the computational process always produces an outcome (ignoring the moment when the energy levels required to convey electrical spikes across synaptic junctions drop towards zero as the individual dies), and thus the observation is always causal with respect to subsequent events. In contrast, the observation may *not* be causal with respect to the thing being observed - because the thing being observed is the dimensionality reduced representation of the computational state of the immediately prior iteration. However, that assumes a von Neumann style architecture where capturing of computational state occurs as a separate step after the fact. The observation and the thing being observed could resolve together as per the same argument for HOOS.

So, a HOOS is a state that represents a prior causal state, certainly at least with respect to subsequent events. What is subjective experience? As best as I can say here, it is an awareness of our own computational states. Coincidentally, that is what a HOOS is. But a HOOS is just a representation; it has to be interpreted to have any meaning or effect. Well, arguably the HOOS *is* the interpretation, of the original observation; the original observation that we've said is causal of subsequent events, while not being casual of the thing being observed. It is not clear whether to pick the observation or the HOOS of that observation as subjective experience.

In summary, for many cases observations and their HOOS are not the cause of the things that they are about, but rather are produced after that thing occurred. However, it is entirely plausible that sometimes observations occur and HOOS are constructed in parallel with the thing that they are about (over the course of gradual resolution) and as such contribute to the causal history of that thing. By definition, observations always cause subsequent events, but do not always cause the construction of HOOS. HOOS often cause subsequent processing outcomes, which themselves may or may not produce further observations and HOOS. It is possible for HOOS to be constructed that have no computational effect, both with respect to the thing that they are about, and with respect to any subsequent processing.

If we pick (somewhat arbitrarily) that the construction of a HOOS is the closest concrete thing we can attribute to subjective experience, then we can say that subjective experience is a mixture of causal and non-causal natures in the exact ways as described in the prior paragraph, except for the last sentence.

With respect to the last sentence, whether you can have entirely non-causal subjective experiences is another example of the problem of defining the mapping from our concept of subjective experience to the concrete mechanisms surrounding observation and HOOS. I would argue that, by definition, subjective experience only exists if the individual is aware that it existed, otherwise it was not subjective experience in the first place. And I would argue that this awareness occurs in the form of the underlying HOOS having a causal effect on some further result. But this is not the only way of defining subjective experience and not the only way to map from it to the mechanisms explained here. So I will leave any further discussion of whether there exists entirely non-causal subjective experience as a matter of definitional opinion and metaphysical debate.

VII.3 Comparison to other theories

The theory as presented bears similarities to a number of existing theories of consciousness. I shall briefly mention some of those and examine how the various theories fit together.

While the high-level meta-cognitive aspect of consciousness has been long suspected, few works address the underlying meta-management mechanisms of meta-cognition. Sloman (1978) applied abstract analysis of concepts associated with intelligence such as: notice, alert, interested, puzzled, surprised, understand, and others. They identified the importance of "administrative processes" in "intelligent" deliberative systems in order to support flexibility and creativity, but only in a later revised version of the book did they begin to use the term "meta-management". They highlighted the need for communication between different sub-systems of the brain, and suggested that those administrative processes are required in order to coordinate behaviour between those sub-systems. They went on to describe several specific systems that might be involved, including modelling of the environment, storage of facts about that environment and about the individual, motivational modelling, modelling of actions against purpose, tracking of current behaviour, and retrospective analysis of past behaviour. Beaudoin (1994) identified the importance of understanding those as "meta-" processes and identified several varieties, grouped under the umbrella term "meta-management": meta-goals, meta-planning, meta-procedure, meta-process. Beaudoin proposed a symbol-based meta-management system suitable for use within artificial autonomous agents. Sloman (2008) extended that research to the human sphere and identified a hierarchical behaviour control system involving three layers: i) reactive processes, ii) deliberative processes, and iii) meta-management processes. They also proposed various additional mechanisms that could influence the primary layer, such as "alarms" which mimic the effect of emotion to change our focus when a threat is identified.

The meta-management needs of intelligence is key to understanding how intelligence can function. So understanding what those meta-management are is of utmost importance. I have attempted to collate the many works in this area in order to identify the range of high-level meta-cognitive and low-level meta-management needs. In order to form the basis of a theory of *human* intelligence I have focused on connectionist architectures and formed my theory around those, extending the works of Beaudoin and Sloman to the connectionist domain. With the advances in deep neural networks today, it should be possible to build and train an artificial neural network based on my theory.

A large part of the theory presented involves higher-order thoughts (HOTs). The definition of a higher-order observational state (HOOS) is almost identical to the definition of a HOT put forward by Rosenthal (1997): a HOT is a particular kind of thought about a first-order mental state such that its contents represents that the individual is in that mental state, and that the HOT was formed through direct internal observation rather than through indirect inference. However, the constraint on HOT contents is not obvious from the term alone and the HOT term is now used by many other researchers who may have subtly different interpretations, so I chose to coin a term that made more direct reference to the observational aspect. Another motivation for coining a new term is that I wish to get away from making a distinction between perceptions, thoughts, and other kinds of experiences. The brain has states. Some represent thoughts, some perceptions, some others, some mixtures of different kinds. We should not be prematurely identifying higher-order states as being only one particular kind.

Rosenthal's variation of HOT theory (1997, 2004) proposes that all consciously experienced first-order mental states must have a HOT generated about them at the time in order to be consciously experienced. Detractors bemoan the inefficiency in the implied constant production of HOTs. Carruthers (1996, 2000, 2005) proposes instead that conscious mental states have a disposition towards the production of HOTs, and only need those HOTs to be produced on the rare occasion when one chooses to consciously introspect their thoughts. The *actualist* HOT theory of Rosenthal more closely aligns to our experience of a continuous stream of consciousness. The *dispositional* theory of Carruthers is more appropriate from a resource efficiency point of view. My own theory is closer to that of Carruthers. I have argued that HOTs/HOOS are only generated at the time that we choose to introspect our mental state. I have claimed that we are tricked into believing in a continuous stream of consciousness through the action of memory and source labelling.

Global Workspace Theory (GWT) (Baars, 1988 & 2021; Baars and Franklin, 2007) proposes that consciousness is due to the global broadcast of information. While it is clear that the content of consciousness contains information from many disparate brain processes and thus a global broadcast seems a likely source of that information, GWT does not explain *how* that global broadcast creates the phenomenological aspects of subjective experience. My own theory is entirely compatible with GWT and adds that missing explanation, by explaining how higher-order information about our mental state creates the phenomenological experience. The global broadcast of GWT may be a first-order process that broadcasts first-order state, such as in Global Neuronal Workspace Theory (Dehaene, Sergent, and Changeux, 2003), and this may occur in parallel to the higher-order representation supplied by the meta-management feedback loop. Alternatively, perhaps the global broadcast and the meta-management feedback loop are the same thing.

Lastly, I have provided context to Integrated Information Theory (IIT) (Tononi and Sporns, 2003; Tononi, 2004, 2008 & 2014). IIT applies a systems theory to information flow within a complex computational system and devises an objective measure of the degree of consciousness in that system. Detractors have long been dissatisfied because of a lack of a way to identify which of those systems are conscious in the first place. IIT measures the information flow in arbitrary connectionist architectures without an understanding of what that information represents. The theory as presented here operates against a higher-level conceptualisation of the processes in the brain and identifies specific kinds of representations and processing. It is only at that level that we can identify the necessary and sufficient conditions for consciousness, and I have provided an initial sketch of those conditions. At that higher level I have also identified that consciousness occurs by *degrees*. IIT provides the objective measure of degrees. My own theory explains how that produces subjective experience.

VII.4 Summary

We cannot understand consciousness on its own. We must consider it in the context of other processes occurring within the brain. So to understand consciousness we must look back at the problems faced during our evolution, and the mechanisms that evolved to resolve those problems. As our ancestors evolved advanced forms of computational processing, one significant problem that arose was the chaotic nature inherent within flexible deliberative thought and something was needed to reign in that chaos. The solution was simple and elegant: that the very same adaptive and chaotic deliberative process could learn to gain control over itself; all it needed was a feedback loop so that it could *observe* itself. Thus self meta-management evolved, and consciousness with it.

Meta-management enables the generation of higher-order thoughts. Schemata and source labelling enable the attachment of additional contextual information to both first-order and higher-order thought. One such form of higher-order thought generates a higher-order observational state (HOOS), which represents that the individual has observed their own internal state. Subjective experience is the result of the generation and subsequent processing of a HOOS. It occurs only with a level of detail and for a duration that is proportional to the extent of brain resources that are spent on the train of thought that may or may not ensue.

The phenomena of subjective experience itself does not meet any evolutionary need that anyone has been able to identify to date; but its underlying meta-management mechanisms do. Thus subjective experience is an emergent side-effect. It has causal power, like any other processing, that affects subsequent thought, so it is not epiphenomenal; but it likely has minimal to no effect on the particular perception or thought that is the *intent* of any given subjective experience.

The explanation given is entirely materialistic. No additional "something special" is required in order to explain the phenomenological aspects of subjective experience. But the explanation given is theoretical, and has many aspects that need empirical evidence. So this is not a *proof* that consciousness is explained through entirely materialistic means. However, by presenting a *plausible* materialistic explanation, it does prove that it is *plausible* for consciousness to have an entirely materialistic explanation.

Some outstanding questions remain. Firstly there are clear empirical questions regarding the meta-management architectures proposed. What are their relative merits in different systems? Does the inline meta-management architecture actually match up with the real architecture of the human brain? But a more puzzling question relates to the apparent prevalence of the generation of HOOS. While generation of HOOS is clearly just one example of the many possible forms of output from meta-management, its functional benefit to the individual is less clear. So why would it be generated with such regularity?

Issues of HOOS aside, the theory presented here is more than just a theory of subjective experience. I started this treatise by explaining that I believed that consciousness could only be understood in the context of meta-management and general intelligence. I then proceeded to spend the bulk of discussion on topics related to meta-management. But meta-management is only needed because of general intelligence. Thus this theory is also about the mechanisms underlying general intelligence, and I believe that more discoveries will follow from it.

References

- Armstrong, D. (1994). A Materialist Theory of the Mind. Routledge. (Originally published 1968) https://doi.org/10.4324/9780203003237
- Armstrong, D., & Malcolm. N (1984). Consciousness and Causality: A Debate on the Nature of Mind. Blackwell.
- Armstrong, S., Leike, J., Orseau, L., & Legg, S. (2020). Pitfalls of Learning a Reward Function Online. Proceedings of the International Joint Conferences on Artificial Intelligence (IJCAI-20). https://doi.org/10.24963/ijcai.2020/221
- Atasoy, S., Donnelly, I., & Pearson, J. (2016). Human brain networks function in connectome-specific harmonic waves. Nature Communications, 7, 10340. https://doi.org/10.1038/ncomms10340
- Atasoy, S., Deco, G., Kringelbach, M. L., & Pearson, J. (2018). Harmonic brain modes: a unifying framework for linking space and time in brain dynamics. Neuroscientist, 24, 277–293. https://doi.org/10.1177/1073858417728032
- Baars, B. J. (1988). A cognitive theory of consciousness. Cambridge University Press.
- Baars, B. J., & Franklin, S. (2007). An architectural model of conscious and unconscious brain functions: Global Workspace Theory and IDA. Neural Networks, 20(9), 955–961. https://psycnet.apa.org/doi/10.1016/j.neunet.2007.09.013
- Baars, B. J. (2021). On Consciousness: Science & Subjectivity - Updated Works on Global Workspace Theory. Nautilus Press.
- Bahrami, B., Olsen, K., Bang, D., Roepstorff, A., Rees, G., & Frith, C. (2012). What failure in collective decision-making tells us about metacognition. Philosophical Transactions of the Royal Society B: Biological Sciences, 367, 1350–1365. https://doi.org/10.1098/rstb.2011.0420
- Bayne, T., & Pacherie, E (2007). Narrators and comparators: the architecture of agentive self-awareness [Abstract from Springer]. Synthese, 159, 475–491. https://doi.org/10.1007/s11229-007-9239-9
- Beaudoin, L. (1994). Goal processing in autonomous agents [PhD thesis, The University of Birmingham]. https://citeseerx.ist.psu.edu/document?doi=382135c4379c08253810ef8f5823c469af6b69df
- Benjamin, A.S., Bjork, R.A., & Schwartz, B.L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. Journal of Experimental Psychology: General, 127(1), 55-68. https://psycnet.apa.org/doi/10.1037/0096-3445.127.1.55
- Bermúdez, J. L. (2005). The Phenomenology of Bodily Awareness [Abstract from Oxford Academic]. In D. W. Smith, & A. L. Thomasson (Eds.), Phenomenology and Philosophy of Mind (pp. 295-316), https://doi.org/10.1093/acprof:oso/9780199272457.003.0015
- Bernacer, J., & Murillo, J.I. (2014). The Aristotelian conception of habit and its contribution to human neuroscience. Frontiers in Human Neuroscience, 8, 883. https://doi.org/10.3389/fnhum.2014.00883
- Birch, J., Ginsburg, S., & Jablonka, E. (2020). Unlimited associative learning and the origins of consciousness: a primer and some predictions. Biology & Philosophy, 35, 56. https://doi.org/10.1007/s10539-020-09772-0
- Block, N. (1995). On a confusion about a function of consciousness. Behavioral and Brain Sciences, 18(2), 227–247. https://doi.org/10.1017/S0140525X00038188 (Full Text)
- Block, N. (2002). The Harder Problem of Consciousness. The Journal of Philosophy, 99(8), 391–425. https://doi.org/10.2307/3655621
- Booth, A. D. (1951). A Signed Binary Multiplication Technique. The Quarterly Journal of Mechanics and Applied Mathematics, 4(2), 236–240. https://doi.org/10.1093/qjmam/4.2.236
- Borţun, D., & Purcarea, V. L. (2013). Marketing and semiotic approach on communication. Consequences on knowledge of target-audiences. Journal of Medicine and Life, 6(1), 103–108. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3632356/
- Bostrom, N. (2003). Ethical issues in advanced artificial intelligence. Science Fiction and Philosophy: From Time Travel to Superintelligence, 277–284.
- Bostrom, N. (2014). Superintelligence: Paths, dangers, strategies. Oxford University Press.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. Neural Information Processing Systems (NeurIPS), 33, 1877–1901. (Full Text)

- Buckley, C. L., Kim, C. S., McGregor, S., & Seth, A. K. (2017). The free energy principle for action and perception: A mathematical review. Journal of Mathematical Psychology, 81, 55–79. https://doi.org/10.1016/j.jmp.2017.09.004
- Burnum J. F. (1993). Medical diagnosis through semiotics. Giving meaning to the sign. Annals of Internal Medicine, 119(9), 939–943. https://doi.org/10.7326/0003-4819-119-9-199311010-00012
- Carruthers, P. (1996). Language, Thought and Consciousness. Cambridge University Press.
- Carruthers, P. (2000). Phenomenal Consciousness: a naturalistic theory. Cambridge University Press.
- Carruthers, P. (2005). Consciousness: essays from a higher-order perspective. Oxford University Press.
- Carruthers, P., & Gennaro, R. (2020). Higher-Order Theories of Consciousness. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Fall 2020 Edition). https://plato.stanford.edu/archives/fall2020/entries/consciousness-higher/
- Carruthers, P., & Williams, D.M. (2022). Model-free metacognition. Cognition, 225, 105117, https://doi.org/10.1016/j.cognition.2022.105117.
- Carter, R., Aldridge, S., Page, M., & Parker, S. (2019). The brain brain book: An illustrated guide to its structure, function, and disorders (3rd ed.). DK Publishing.
- Chalmers, D. J. (1996). The Conscious Mind. Oxford University Press.
- Chalmers, D. J. (2013). Panpsychism and Panprotopsychism. Amherst Lecture in Philosophy, 8, 1-35. (Full Text)
- Chappell, J., & Sloman, A. (2007). Natural and artificial meta-configured altricial information-processing systems. The International Journal of Unconventional Computing. 3(3), 211-239.
- Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behavioral and Brain Sciences, 36, 181–204. https://doi.org/10.1017/s0140525x12000477
- Clark, A. (2019). Beyond desire? Agency, choice, and the predictive mind. Australasian Journal of Philosophy, 9, 1–15. https://doi.org/10.1080/00048402.2019.1602661
- Clark, A., & Karmiloff-Smith, A. (1993). The cognizer’s innards: a psychological and philosophical perspective on the development of thought. Mind and Language, 8, 487–519. https://doi.org/10.1111/j.1468-0017.1993.tb00299.x
- Cleeremans, A. (2007). Consciousness: the radical plasticity thesis. Editor(s): Rahul Banerjee, Bikas K. Chakrabarti. Progress in Brain Research, 168, 19-33. Elsevier. https://doi.org/10.1016/S0079-6123(07)68003-0.
- Cleeremans, A., Achoui, D., Beauny, A., Keuninckx, L., Martin, J-R., Muñoz-Moldes, S., Vuillaume, L., & de Heering, A. (2020). Learning to Be Conscious. Trends in Cognitive Sciences, 24(2), 112-123, https://doi.org/10.1016/j.tics.2019.11.011.
- Cleeremans, A., Achoui, D., Beauny, A., Keuninckx, L., Martin, J-R., Muñoz-Moldes, S., Vuillaume, L., & de Heering, A. (2021). Do You Need to Be Conscious to Learn to Be Conscious? Trends in Cognitive Sciences, 25(1), 9-11. https://doi.org/10.1016/j.tics.2020.10.002
- Cleeremans, A., Timmermans, B., Pasquali, A. (2007). Consciousness and metarepresentation: A computational sketch. Neural Networks, 20(9), 1032-1039, https://doi.org/10.1016/j.neunet.2007.09.011.
- Colas, C., Fournier, P., Sigaud, O., Chetouani, M., & Oudeyer, P-Y. (2019). CURIOUS: Intrinsically Motivated Modular Multi-Goal Reinforcement Learning. Proceedings of the 36th International Conference on Machine Learning (PMLR), 97, 1331-1340. https://proceedings.mlr.press/v97/colas19a.html.
- Crane, T. (2009). Intentionalism. In B. McLaughlin, & A. Beckermann (Eds.), The Oxford Handbook to the Philosophy of Mind (pp. 474–93). Oxford University Press. (Abstract)
- Cushman, F. (2020). Rationalization is rational. Behavioral and Brain Sciences, 43, E28. https://doi.org/10.1017/S0140525X19001730
- Cushman, F., & Morris, A. (2015). Habitual control of goal selection in humans. Proceedings of the National Academy of Sciences (PNAS), 112(45), 13817–13822. https://doi.org/10.1073/pnas.1506367112
- Damasio, A. R. (2006). Descartes’ Error. Vantage. (Original work published 1994)
- Daw, N.D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioural control. Nature Neuroscience, 8, 1704–1711. https://doi.org/10.1038/nn1560
- Dayan, P. (2008). The role of value systems in decision making. In C. Engel, & W. Singer (Eds.), Better Than Conscious? Implications for Performance and Institutional Analysis (pp. 51-70). MIT press. (Full Text)
- de Vignemont, F. (2020). Bodily Awareness. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Fall 2020 Edition). https://plato.stanford.edu/archives/fall2020/entries/bodily-awareness/.

- de Wit, S., Corlett, P.R., Aitken, M.R., Dickinson, A., & Fletcher, P.C. (2009). Differential Engagement of the Ventromedial Prefrontal Cortex by Goal-Directed and Habitual Behavior toward Food Pictures in Humans. Journal of Neuroscience, 29(36), 11330-11338. https://doi.org/10.1523/JNEUROSCI.1639-09.2009
- Deco, G., Jirsa, V. K., Robinson, P. A., Breakspear, M., & Friston, K. (2008). The dynamic brain: from spiking neurons to neural masses and cortical fields. PLOS Computational Biology, 4(8), e1000092. https://doi.org/10.1371/journal.pcbi.1000092
- Dehaene, S. (2014). Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts. Penguin Books.
- Dehaene, S., Sergent, C. & Changeux, J. P. (2003). A neuronal network model linking subjective reports and objective physiological data during conscious perception. The Proceedings of the National Academy of Sciences (PNAS), 100, 8520-5. https://doi.org/10.1073/pnas.1332574100
- Dennett, D. C. (1991). Consciousness explained (P. Weiner, Illustrator). Little, Brown and Co.
- Descartes, R. (1911). The Principles of Philosophy (E. Haldane & G. Ross, Trans.). Cambridge University Press. (Original work published 1644)
- Dolan, R.J., & Dayan, P. (2013). Goals and habits in the brain. Neuron, 80, 312–325. https://doi.org/10.1016/j.neuron.2013.09.007
- Douskos, C. (2018). Deliberation and Automaticity in Habitual Acts. Ethics in Progress, 9(1), 25–43. https://doi.org/10.14746/eip.2018.1.2
- Dunlosky, J., & Bjork, R.A. (2008). Handbook of metamemory and memory (1st ed.). Psychology Press.
- Earl, B. (2014). The biological function of consciousness. Frontiers in Psychology, 5, 697. https://doi.org/10.3389/fpsyg.2014.00697
- Edelman, G. M. (1987). Neural Darwinism: The Theory of Neuronal Group Selection. Basic Books, Inc.
- Edelman, G. M. (2003). Naturalizing consciousness: a theoretical framework. The Proceedings of the National Academy of Sciences (PNAS), 100, 5520-4. https://doi.org/10.1073/pnas.0931349100
- Edelman, G. M., Gally, J. A., & Baars, B. J. (2011). Biology of consciousness. Frontiers in Psychology, 2, 4. https://doi.org/10.3389/fpsyg.2011.00004 (Full Text)
- Edelman, G. M., & Tononi, G. (2000). A universe of consciousness: how matter becomes imagination. New York, NY: Basic Books.
- Erulkar, S.D., & Lentz, T.L. (2023). Nervous system. Encyclopedia Britannica. https://www.britannica.com/science/nervous-system
- Eysenbach, B., Gupta, A., Ibarz, J., & Levine, S. (2019). Diversity is All You Need: Learning Skills without a Reward Function. ArXiv, 1802.06070. https://doi.org/10.48550/arXiv.1802.06070
- Favela, L. H. (2020). Dynamical systems theory in cognitive science and neuroscience. Philosophy Compass, 15, e12695. https://doi.org/10.1111/phc3.12695
- Fernandez Cruz, A.L., Arango-Muoz, S., Volz, K.G. (2016). Oops, scratch that! Monitoring ones own errors during mental calculation. Cognition, 146, 110-120. https://doi.org/10.1016/j.cognition.2015.09.005
- Fernandez-Duque, D., Baird, J.A., & Posner, M. I. (2000). Executive attention and metacognitive regulation. Consciousness and Cognition, 9, 288–307. https://doi.org/10.1006/ccog.2000.0447
- Fleming, S.M., Dolan, R.J., & Frith, C.D. (2012a). Metacognition: computation, biology and function. Philosophical Transactions of the Royal Society B: Biological Sciences, 367, 1280–1286. http://doi.org/10.1098/rstb.2012.0021
- Fleming, S.M., Huijgen, J., & Dolan, R.J. (2012b). Prefrontal contributions to metacognition in perceptual decision making. The Journal of Neuroscience, 32(18), 6117-6125. https://doi.org/10.1523/JNEUROSCI.6489-11.2012
- Fleming, S.M., & Lau, H.C. (2014). How to measure metacognition. Frontiers in Human Neuroscience, 8, 443. https://doi.org/10.3389/fnhum.2014.00443
- Fodor, J. (1975). The Language of Thought. Thomas Y. Crowell.
- Fodor, J. (1981). Representations. MIT Press.
- Fodor, J. (1987). Psychosemantics. MIT Press.
- Fodor, J. (1990). A Theory of Content and Other Essays. MIT Press.
- Fodor, J. (1994). The Elm and the Expert. MIT Press.
- Fodor, J. (2008). Lot 2: The Language of Thought Revisited. Clarendon Press.
- Fonseca-Azevedo, K., & Herculano-Houzel, S. (2012). Metabolic constraint imposes tradeoff between body size and number of brain neurons in human evolution. Proceedings of the National Academy of Sciences (PNAS), 109(45), 18571-18576, https://doi.org/10.1073/pnas.1206390109

- Franklin, S. & Graesser, A. (1999). A software agent model of consciousness. Consciousness and Cognition, 8, 285-301. https://doi.org/10.1006/ccog.1999.0391
- Friston, K. J. (2005). A theory of cortical responses. Philosophical Transactions of the Royal Society B: Biological Sciences, 360(1456), 815–836. https://doi.org/10.1098/rstb.2005.1622
- Friston, K. J. (2008). Hierarchical models in the brain. PLOS Computational Biology, 4(11). https://doi.org/10.1371/journal.pcbi.1000211
- Friston, K. J. (2010). The free-energy principle: a unified brain theory? Nature Reviews Neuroscience, 11, 127–138. https://doi.org/10.1038/nrn2787
- Friston, K. J. (2019). A free energy principle for a particular physics. ArXiv, 1906.10184. https://doi.org/10.48550/arXiv.1906.10184.
- Friston, K. J., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active inference: a process theory. Neural Computation, 29, 1–49. https://doi.org/10.1162/NECO_a_00912
- Friston, K. J., Kahan, J., Razi, A., Stephan, K. E., & Sporns, O. (2014). On nodes and modes in resting state fMRI. NeuroImage, 99, 533–547. https://doi.org/10.1016/j.neuroimage.2014.05.056
- Friston, K. J., Kilner, J., & Harrison, L. (2006). A free energy principle for the brain. Journal of Physiology-Paris, 100(1-3), 70–87. https://doi.org/10.1016/j.jphysparis.2006.10.001
- Frith, C.D. (2008). Social cognition. Philosophical Transactions of the Royal Society B: Biological Sciences, 363, 2033–2039. https://doi.org/10.1098/rstb.2008.0005
- Gans, J. S. (2017). Self-Regulating Artificial General Intelligence. ArXiv, 1711.04309. https://doi.org/10.48550/arXiv.1711.04309
- Gennaro, R. (1996). Consciousness and Self-Consciousness. John Benjamins.
- Gennaro, R. (2012). The Consciousness Paradox: Consciousness, Concepts, and Higher-Order Thoughts. MIT press.
- Gęsiarz, F., & Crockett, M.J. (2015). Goal-directed, habitual and Pavlovian prosocial behavior. Frontiers in Behavioral Neuroscience, 9, 135. https://doi.org/10.3389/fnbeh.2015.00135
- Ghosh, D., Gupta, A., & Levine, S. (2019). Learning Actionable Representations with Goal-Conditioned Policies. ArXiv, 1811.07819. https://doi.org/10.48550/arXiv.1811.07819
- Gleitman, H., Fridlund, A. J., & Reisberg, D. (2004). Psychology (6th ed.). Norton.
- Glorot, X., Bordes, A. & Bengio, Y. (2010). Deep Sparse Rectifier Neural Networks. Proceedings of the 14th International Conference on Artificial Intelligence and Statistics, Journal of Machine Learning Research, 15, 315-323. https://proceedings.mlr.press/v15/glorot11a.html
- Godfrey-Smith, P. (2016). Other Minds: The octopus and the evolution of intelligent life. Farrar, Straus and Giroux.
- Goff, P., Seager, W., & Allen-Hermanson, S. (2022). Panpsychism. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Summer 2022 Edition). https://plato.stanford.edu/archives/sum2022/entries/panpsychism/
- Graziano, M.S.A., & Kastner, S. (2011). Human consciousness and its relationship to social neuroscience: a novel hypothesis. Cognitive Neuroscience, 2, 98–113. https://doi.org/10.1080/17588928.2011.565121
- Graziano, M.S.A. (2017). The Attention Schema Theory: A Foundation for Engineering Artificial Consciousness. Frontiers in Robotics and AI, 4, 60. https://doi.org/10.3389/frobt.2017.00060.
- Guyer, P., & Horstmann, R-P. (2023). Idealism. In E. N. Zalta, & U. Nodelman (Eds.), The Stanford Encyclopedia of Philosophy (Spring 2023 Edition). https://plato.stanford.edu/archives/spr2023/entries/idealism/
- Hameroff S. (2021). 'Orch OR' is the most complete, and most easily falsifiable theory of consciousness. Cognitive Neuroscience, 12(2), 74–76. https://doi.org/10.1080/17588928.2020.1839037
- Hameroff, S., & Penrose, R. (2014). Consciousness in the universe: A review of the 'Orch OR' theory. Physics of Life Reviews, 11(1), 39–78. https://doi.org/10.1016/j.plrev.2013.08.002
- Hameroff, S., & Penrose, R. (1996b). Conscious events as orchestrated space–time selections. Journal of Consciousness Studies, 3(1), 36–53. (Full Text)
- Hameroff, S., & Penrose, R. (1996a). Orchestrated reduction of quantum coherence in brain microtubules: A model for consciousness. Mathematics and Computers in Simulation, 40(3–4), 453–480. https://doi.org/10.1016/0378-4754(96)80476-9
- Haynes, J-D. (2013). Beyond Libet: Long-Term Prediction of Free Choices from Neuroimaging Signals. In A. Clark, J. Kiverstein, & T. Vierkant (Eds.), Decomposing the Will (pp. 60-72). Oxford University Press.

- Henshilwood, C.S., Sealy, J.C., Yates, R., Cruz-Uribe, K., Goldberg, P., Grine, F.E., Klein, R.G., Poggenpoel, C., van Niekerk, K., & Watts, I. (2001). Blombos Cave, Southern Cape, South Africa: Preliminary Report on the 1992–1999 Excavations of the Middle Stone Age Level. Journal of Archaeological Science, 28(4), 421-448. https://doi.org/10.1006/jasc.2000.0638
- Herculano-Houzel, S. (2012). The remarkable, yet not extraordinary, human brain as a scaled-up primate brain and its associated cost. Proceedings of the National Academy of Sciences (PNAS), 109(Supplement 1), 10661-10668. https://doi.org/10.1073/pnas.1201895109
- Hochreiter, S., Bengio, Y., Frasconi, P. & Schmidhuber, J. (2001). Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. In S. C. Kremer & J. F. Kolen (Eds.), A Field Guide to Dynamical Recurrent Neural Networks (pp. 237-243). IEEE Press. https://doi.org/10.1109/9780470544037.ch14.
- Hoffman, D. D. (2009). The Interface Theory of Perception: Natural Selection Drives True Perception to Swift Extinction. In S. Dickinson, A. Leonardis, B. Schiele, & M. Tarr (Eds.), Object Categorization: Computer and Human Vision Perspectives (pp. 148-166). Cambridge University Press. https://doi.org/10.1017/CBO9780511635465.009
- Hoffman, D. D. (2023). The Interface Theory of Perception. In J. T. Wixted, & J. Serences (Eds.), Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience. Wiley. https://doi.org/10.1002/9781119170174.epcn216
- Hoffman, D. D., Singh, M. & Prakash, C (2015). The Interface Theory of Perception. Psychonometric Bulletin & Review, 22, 1480–1506. https://doi.org/10.3758/s13423-015-0890-8 (Full Text)
- Humphrey, N. (2002). The uses of consciousness. In N. Humphrey (Ed.), The mind made flesh: Essays from the frontiers of evolution and psychology (pp. 65–85). Oxford University Press.
- Hunt, T., & Schooler, J. W. (2019). The Easy Part of the Hard Problem: A Resonance Theory of Consciousness. Frontiers in Human Neuroscience, 13, 378. https://doi.org/10.3389/fnhum.2019.00378
- Johnson, M. K., Hashtroudi, S., & Lindsay D. S. (1993). Source Monitoring [Abstract from APA PsycNet]. Psychological Bulletin, 114(1), 3–28. https://psycnet.apa.org/doi/10.1037/0033-2909.114.1.3
- Kahneman, D. (2011). Thinking, fast and slow. Farrar, Straus and Giroux.
- Kanai, R., Chang, A., Yu, Y., de Abril, I. M, Biehl, M, & Guttenberg, N. (2019). Information generation as a functional basis of consciousness. Neuroscience of Consciousness, 2019(1), niz016. https://doi.org/10.1093/nc/niz016
- Kant, I. (1929). Critique Of Pure Reason (N. K. Smith, Trans.). Macmillan. (Original work published 1781)
- Kaplan, R., & Friston, K. J (2018). Planning and navigation as active inference. Biol. Cybern., 112, 323–343. https://doi.org/10.1007/s00422-018-0753-2
- Karmiloff-Smith, A. (1992). Beyond Modularity: A Developmental Perspective on Cognitive Science. MIT Press
- Karmiloff-Smith, A. (1994). Précis of Beyond modularity: A developmental perspective on cognitive science. Behavioral and Brain Sciences, 17(4), 693-707. https://doi.org/10.1017/S0140525X00036621
- Kilner, J. M., Friston, K. J., & Frith, C. D. (2007). Predictive coding: an account of the mirror neuron system. Cognitive Processing, 8(3), 159–166. https://doi.org/10.1007/s10339-007-0170-2
- Kim, J. (2002). The Many Problems of Mental Causation (Excerpt). In Chalmers, D. (Ed.), Philosophy of Mind: Classical and Contemporary Readings (170-179). Oxford University Press.
- Koriat A. (2007). Metacognition and consciousness. In P. D. Zelazo, M. Moscovitch, & E. Thompson (Eds.), The Cambridge Handbook of Consciousness (pp. 289–325). Cambridge University Press.
- Kotchoubey, B. (2018). Human Consciousness: Where Is It From and What Is It for. Frontiers in Psychology, 9, 567. https://doi.org/10.3389/fpsyg.2018.00567
- Kriegel, U. (2003). Consciousness as intransitive self-consciousness: two views and an argument. Canadian Journal of Philosophy, 33(1), 103–132. https://doi.org/10.1080/00455091.2003.10716537
- Kriegel, U. (2004). The functional role of consciousness: a phenomenological approach. Phenomenology and the Cognitive Sciences, 3, 171–93. https://doi.org/10.1023/B:PHEN.0000040833.23356.6a
- Kriegel, U. (2006). The same-order monitoring theory of consciousness. In U. Kriegel & K. Williford (Eds.), Self-Representational Approaches to Consciousness (pp. 143-170). MIT Press.
- Kriegel, U. (2009). Subjective Consciousness. Oxford University Press.
- Kriegesgorte, K. (2015). Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. Annual Review of Vision Science, 1, 417–446. https://doi.org/10.1146/annurev-vision-082114-035447

- Krishevsky, A., Sutskever, I., & Hinton, G. (2012). ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems (NeurIPS), 25, 1097–1105. (Full Text

- Krueger, P. M, & Griffiths, T. (2018). Shaping Model-Free Habits with Model-Based Goals. Cognitive Science. (SemanticScholar) (Full Text)

- Krüger, J., & Aiple, F. (1988). Multimicroelectrode investigation of monkey striate cortex: spike train correlations in the infragranular layers [Abstract from PubMed]. Journal of Neurophysiology, 60(2), 798–828. https://doi.org/10.1152/jn.1988.60.2.798

- Lazaridis, A., Fachantidis, A., & Vlahavas, I. (2020). Deep Reinforcement Learning: A State-of-the-Art Walkthrough. Journal of Artificial Intelligence Research, 69, 1421-1471. https://doi.org/10.1613/jair.1.12412

- Levine, J. (1983). Materialism and qualia: The explanatory gap. Pacific Philosophical Quarterly, 64, 354-361. https://doi.org/10.1111/j.1468-0114.1983.tb00207.x (Full Text)

- Libet, B. (1985). Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action. Behavioral and Brain Sciences, 8(4), 529–566. https://psycnet.apa.org/doi/10.1017/S0140525X00044903

- Libet, B. (2004). Mind Time: The Temporal Factor in Consciousness. Harvard University Press.

- Linser, K., & Goschke, T. (2007). Unconscious modulation of the conscious experience of voluntary control. Cognition, 104, 459–475. https://doi.org/10.1016/j.cognition.2006.07.009

- Liao, S., & Gendler, T (2020). Imagination. In E. N. Zalta (ed.), The Stanford Encyclopedia of Philosophy (Summer 2020 Edition). https://plato.stanford.edu/archives/sum2020/entries/imagination/

- Livingstone, M. S., & Hubel, D. H. (1987). Psychophysical Evidence for Separate Channels for the Perception of Form, Color, Movement, and Depth. Journal of Neuroscience, 7, 236-368. https://doi.org/10.1523/jneurosci.07-11-03416.1987

- Llinás, R., Ribary, U., Contreras, D., & Pedroarena, C. (1998). The neuronal basis for consciousness. Philosophical Transactions of the Royal Society B: Biological Sciences, 353, 1841-1849. https://doi.org/10.1098%2Frstb.1998.0336

- Luo, T. Z., & Maunsell, J. H. R. (2019). Attention can be subdivided into neurobiological components corresponding to distinct behavioral effects. Proceedings of the National Academy of Sciences (PNAS), 116 (52), 26187-26194. https://doi.org/10.1073/pnas.1902286116

- Lycan, W. (1996). Consciousness and Experience. MIT Press.

- Lycan, W. (2004). The superiority of HOP to HOT. In R. J. Gennaro (Ed.), Higher-Order Theories of Consciousness: An Anthology (pp. 93–114). John Bejamins. https://doi.org/10.1075/aicr.56.07lyc

- Ma, W. J. (2019). Bayesian Decision Models: A Primer. Neuron, 104, 164–175. https://doi.org/10.1016/j.neuron.2019.09.037

- Mandrigin, A (2021). The where of bodily awareness. Synthese, 198, 1887–1903. https://doi.org/10.1007/s11229-019-02171-3

- Marcus, G. (2001). The Algebraic Mind. MIT Press

- Marković, D., Goschke, T., & Kiebel, S.J. (2021). Meta-control of the exploration-exploitation dilemma emerges from probabilistic inference over a hierarchy of time scales. Cognitive, Affective & Behavioral Neuroscience, 21(3), 509–533. https://doi.org/10.3758/s13415-020-00837-x

- McCarthy, J. (2008). The well-designed child. Artificial Intelligence, 172(18), 2003-2014. https://doi.org/10.1016/j.artint.2008.10.001

- McCulloch, W., & Pitts, W. (1943). A Logical Calculus of the Ideas Immanent in Nervous Activity. Bulletin of Mathematical Biophysics, 7, 115–133. https://doi.org/10.1007/BF02478259

- McGinn, C. (1989). Can We Solve the Mind-Body Problem? Mind, 98(391), 349–366. https://doi.org/10.1093/mind/XCVIII.391.349

- McGinn, C. (1991). The Problem of Consciousness: Essays Toward a Resolution [Abstract from PhilPapers]. Blackwell. https://doi.org/10.2307/2186044

- Metcalfe, J., & Shimamura, A.P. (Eds.) (1994). Metacognition: Knowing about knowing. MIT press.

- Miller P. (2016). Dynamical systems, attractors, and neural circuits. F1000Research, 5, F1000 Faculty Rev-992. https://doi.org/10.12688/f1000research.7698.1

- Millidge, B., Seth, A., & Buckley, C. (2021). Predictive Coding: a Theoretical and Experimental Review. Computer Science. https://doi.org/10.48550/arXiv.2107.12979 (Full Text)

- Mitchell, K. J., & Johnson, M. K. (2009). Source monitoring 15 years later: what have we learned from fMRI about the neural mechanisms of source memory? Psychological Bulletin, 135(4), 638–677. https://doi.org/10.1037/a0015849

- Mitchell, K. J., & Johnson, M. K. (2000). Source monitoring: Attributing mental experiences. In E. Tulving & F. I. M. Craik (Eds.), The Oxford handbook of memory (pp. 179-195). Oxford University Press.

- Morasso, P., Casadio, M., Mohan, V., Rea, F., & Zenzeri, J. (2015). Revisiting the body-schema concept in the context of whole-body postural-focal dynamics. Frontiers in Human Neuroscience, 9, 83. https://doi.org/10.3389/fnhum.2015.00083.

- Morsella, E. (2005). The Function of Phenomenal States: Supramodular Interaction Theory. Psychological Review, 112(1), 1000–21. https://doi.org/10.1037/0033-295x.112.4.1000

- Mountcastle, V. B. (1997). The Columnar Organization of the Neocortex. Brain, 120(4), 701-722. http://doi.org/10.1093/brain/120.4.701 (Full Text)

- Mumford, D. (1991). On the computational architecture of neocortex. Biological Cybernetics, 65, 135–145. https://doi.org/10.1007/bf00202389

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

- Nair, A., Pong, V., Dalal, M., Bahl, S., Lin, S., & Levine, S. (2018). Visual Reinforcement Learning with Imagined Goals. Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18). (Full Text)

- Nelson T.O. (1996). Consciousness and metacognition. American Psychologist, 51(1), 102–116. http://doi.org/10.1037/0003-066X.51.2.102

- Nessa J. (1996). About signs and symptoms: can semiotics expand the view of clinical medicine? Theoretical Medicine, 17(4), 363–377. https://doi.org/10.1007/BF00489681

- Nisbett, R. E., & Wilson, T. D. (1977a). Telling More Than We Can Know: Verbal Reports on Mental Processes. Psychological Review, 84(3), 231–259. https://doi.org/10.1037/0033-295X.84.3.231

- Nisbett, R. E. & Wilson, T. D. (1977b). The halo effect: evidence for unconscious alteration of judgments. Journal of Personality and Social Psychology, 35(4), 250–256. https://psycnet.apa.org/doi/10.1037/0022-3514.35.4.250

- Norman, E. (2020). Why Metacognition Is Not Always Helpful. Frontiers in Psychology, 11, 1537. https://doi.org/10.3389/fpsyg.2020.01537

- Oizumi, M., Albantakis, L., & Tononi, G. (2014). From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0. PLOS Computational Biology, 10(5), e1003588. https://doi.org/10.1371/journal.pcbi.1003588

- OpenAI, O., Plappert, M., Sampedro, R., Xu, T., Akkaya, I., Kosaraju, V., Welinder, P., D'Sa, R., Petron, A., Pinto, H.P., Paino, A., Noh, H., Weng, L., Yuan, Q., Chu, C., & Zaremba, W. (2021). Asymmetric self-play for automatic goal discovery in robotic manipulation. ArXiv, 2101.04882. https://doi.org/10.48550/arXiv.2101.04882

- Overgaard, M., & Kirkeby-Hinrup, A. (2021). Is Learning the Cognitive Basis of Consciousness? The Moral Implications of SOMA. Trends in Cognitive Sciences, 25(1), 8-9. https://doi.org/10.1016/j.tics.2020.08.004

- Paiva, A. R. C., Park, I., & Príncipe, J. C. (2010). Inner Products for Representation and Learning in the Spike Train Domain [Abstract from ScienceDirect]. In K. G. Oweiss (Ed.), Statistical Signal Processing for Neuroscience and Neurotechnology (pp. 265-309). Academic Press. https://doi.org/10.1016/B978-0-12-375027-3.00008-9

- Paris, S.G., & Winograd, P. (1990). How metacognition can promote academic learning and instruction. In B. F Jones & L. Idol (Eds.), Dimensions of thinking and cognitive instruction (pp. 15-51). Lawrence Erlbaum Associates, Inc.

- Pasquali, A., Timmermans, B., & Cleeremans, A. (2010). Know thyself: metacognitive networks and measures of consciousness. Cognition, 117(2), 182–190. https://doi.org/10.1016/j.cognition.2010.08.010

- Pattee, H. H. (2007). The Necessity of Biosemiotics: Matter-symbol Complementarity. In M. Barbieri (Ed.), Introduction to Biosemiotics (pp. 115-132). Springer.

- Pattee, H. H. (2021). Symbol Grounding Precedes Interpretation. Biosemiotics, 14(3), 561–568. https://doi.org/10.1007/S12304-021-09458-4

- Paul, E.J., Smith, J.D., Valentin, V.V., Turner, B.O., Barbey, A.K., & Ashby, F.G. (2015). Neural networks underlying the metacognitive uncertainty response. Cortex, 71, 306-322. https://doi.org/10.1016/j.cortex.2015.07.028

- Paulin, M.G., & Cahill-Lane, J. (2021), Events in Early Nervous System Evolution. Topics in Cognitive Science, 13, 25-44. https://doi.org/10.1111/tops.12461

- Peirce, C. S. (1982). The Writings of Charles S. Peirce: A Chronological Edition (Volumes 1–8). Indiana University Press.

- Peters, F. (2010). Consciousness as recursive, spatiotemporal self-location. Psychological Research, 74, 407–421. https://doi.org/10.1007/s00426-009-0258-7

- Picard, R. (1997). Affective Computing. MIT Press.

- Pierson, L. M., & Trout, M. (2017). What is consciousness for? New Ideas in Psychology, 47, 62-71. https://doi.org/10.1016/j.newideapsych.2017.05.004.

- Postle B. R. (2016). How does the brain keep information "in mind"? Current Directions in Psychological Science, 25(3), 151–156. https://doi.org/10.1177/0963721416643063.

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

- Rahimian, S. (2021). Consciousness in Solitude: Is Social Interaction Really a Necessary Condition? Frontiers in Psychology, 12, 630922. https://doi.org/10.3389/fpsyg.2021.630922

- Rao, R. P. & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. Nature Neuroscience, 2, 79–87. https://doi.org/10.1038/4580

- Rigoli, F., Pezzulo, G., Dolan, R., & Friston, K. (2017). A Goal-Directed Bayesian Framework for Categorization. Frontiers in Psychology, 8, 408. https://doi.org/10.3389/fpsyg.2017.00408

- Rescorla, M. (2019). A Realist Perspective on Bayesian Cognitive Science. In A. Nes & T. Chan (Eds.), Inference and Consciousness (pp. 40-73). Routledge. http://doi.org/10.4324/9781315150703-3

- Rescorla, M. (2020). The Computational Theory of Mind. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Fall 2020 Edition). https://plato.stanford.edu/archives/fall2020/entries/computational-mind/

- Robb, David, John Heil, & Sophie Gibb (2023). Mental Causation. In E. N. Zalta & U. Nodelman (Eeds.), The Stanford Encyclopedia of Philosophy (Spring 2023 Edition). https://plato.stanford.edu/archives/spr2023/entries/mental-causation/

- Robinson, W. (2019). Epiphenomenalism. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Summer 2019 Edition). https://plato.stanford.edu/archives/sum2019/entries/epiphenomenalism/

- Rolls, E. T. (2004). A higher order syntactic thought (HOST) theory of consciousness. In R. J. Gennaro (Ed.), Higher-order theories of consciousness (pp. 137–172). John Benjamins.

- Rolls, E. T. (2005). Emotion explained. Oxford University Press.

- Rosenthal, D. M. (1986). Two concepts of consciousness. Philosophical Studies, 49, 329–359.

- Rosenthal, D. M. (1993). Thinking that one thinks. In M. Davies & G. W. Humphreys (Eds.), Consciousness: Psychological and philosophical essays (pp. 197–223). Blackwell Publishing. (Full Text)

- Rosenthal, D. M. (1997). A Theory of Consciousness. In N. Block, O. Flanagan, & G. Güzeldere (Eds.), The Nature of Consciousness: Philosophical Debates (pp. 729-753). MIT Press/Bradford Books. (Full Text)

- Rosenthal, D. M. (2004). Varieties of higher-order theory. In R. Gennaro (Ed.), Higher-Order Theories of Consciousness: An Anthology (pp. 17-44). John Benjamins. http://doi.org/10.1075/aicr.56.04ros

- Rosenthal, D. M. (2005). Consciousness and Mind. Oxford University Press.

- Rosenthal, D. M. (2008). Consciousness and Its Function. Neuropsychologia, 46(3), 829-40. http://doi.org/10.1016/j.neuropsychologia.2007.11.012

- Rosenthal, D. M. (2012). Higher-order awareness, misrepresentation and function. Philosophical Transactions of the Royal Society B: Biological Sciences, 367, 1424–1438. http://doi.org/10.1098/rstb.2011.0353

- 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. Frontiers in Artificial Intelligence, 3, 30. https://doi.org/10.3389/frai.2020.00030

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

- Scarantino, A., & de Sousa, R (2021). Emotion. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Summer 2021 Edition). https://plato.stanford.edu/archives/sum2021/entries/emotion/

- Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural Networks, 61, 85-117. https://doi.org/10.1016/j.neunet.2014.09.003.

- Schrittwieser, J., Antonoglou, I., Hubert, T. et al (2020). Mastering Atari, Go, chess and shogi by planning with a learned model. Nature, 588, 604–609. https://doi.org/10.1038/s41586-020-03051-4

- Schroeder, T. (2020). Desire. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Summer 2020 Edition). https://plato.stanford.edu/archives/sum2020/entries/desire/

- Searle, J.R. (1990). Who is computing with the brain? Behavioral and Brain Sciences, 13(4), 632-64. https://doi.org/10.1017/s0140525x00080663

- Serin, Y., & Tek, N. A. (2019). Effect of Circadian Rhythm on Metabolic Processes and the Regulation of Energy Balance. Ann Nutr Metab, 74(4), 322–330. https://doi.org/10.1159/000500071
- Seth, A.K., Bayne, T. (2022). Theories of consciousness. Nature Reviews Neuroscience. https://doi.org/10.1038/s41583-022-00587-4
- Shanahan, M. (2005). A cognitive architecture that combines internal simulation with a global workspace. Consciousness and Cognition, 15(2), 433-49. https://doi.org/10.1016/j.concog.2005.11.005
- Shanahan, M. (2008). A spiking neuron model of cortical broadcast and competition. Consciousness and Cognition, 17(1), 288-303. https://doi.org/10.1016/j.concog.2006.12.005
- Shapiro, L.A. (2013). Dynamics and Cognition. Minds and Machines, 23(3), 353-375. https://doi.org/10.1002/wcs.1200
- Shimamura, A.P. (2000). Toward a Cognitive Neuroscience of Metacognition. Consciousness and Cognition, 9(2), 313-323. https://doi.org/10.1006/ccog.2000.0450.
- Siegel, S. (2021). The Contents of Perception. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Fall 2021 Edition). https://plato.stanford.edu/archives/fall2021/entries/perception-contents/
- Siewert, C. (2004). Is Experience Transparent? Philosophical Studies, 117(1/2), 15–41. https://doi.org/10.1023%2FB%3APHIL.0000014523.89489.59
- Siewert, C. (2022). Consciousness and Intentionality. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Summer 2022 Edition). https://plato.stanford.edu/archives/sum2022/entries/consciousness-intentionality/
- Sloman, A. (1978). The Computer Revolution in Philosophy (Revised 2018). Harvester Press (and Humanities Press).
- Sloman, A. (1998). Damasio, Descartes, alarms and meta-management. SMC'98 Conference Proceedings. IEEE International Conference on Systems, Man, and Cybernetics, 3, 2652-2657. https://doi.org/10.1109/ICSMC.1998.725060
- Sloman, A. (2001). Beyond Shallow Models of Emotion. Cognitive Processing, 2(1), 177–198. (Full Text)
- Sloman, A. (2008). Varieties of Metacognition in Natural and Artificial Systems. In M. T. Cox & A. Raja (Eds.), Metareasoning: Thinking about Thinking. MIT Press. https://doi.org/10.7551/mitpress/9780262014809.003.0020
- Smart, J. J. C. (2022). The Mind/Brain Identity Theory. In E N. Zalta & U. Nodelman (Eds.), The Stanford Encyclopedia of Philosophy (Winter 2022 Edition). https://plato.stanford.edu/archives/win2022/entries/mind-identity/
- Smith, D. W. (2018). Phenomenology. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Summer 2018 Edition). https://plato.stanford.edu/archives/sum2018/entries/phenomenology/
- Smithies, D., & Weiss, J. (2019). Affective Experience, Desire, and Reasons for Action. Analytic Philosophy, 60(1), 27-54. https://doi.org/10.1111/phib.12144
- Smolensky, P. (1988). On the Proper Treatment of Connectionism. Behavioral and Brain Sciences, 11(1), 1–74. https://doi.org/10.1017/S0140525X00052432
- Snodgrass, M., Kalaida, N., & Winer E. S. (2009). Access is mainly a second-order process: SDT models whether phenomenally (first-order) conscious states are accessed by reflectively (second-order) conscious processes. Consciousness and Cognition, 18, 561–564; discussion 565–567. https://doi.org/10.1016/j.concog.2009.01.003
- Snow, N.E. (2006). Habitual Virtuous Actions and Automaticity. Ethical Theory and Moral Practice, 9, 545–561. https://doi.org/10.1007/s10677-006-9035-5
- Soon, C., Brass, M., Heinze, H-J., & Haynes, J-D. (2008). Unconscious determinants of free decisions in the human brain. Nature Neuroscience, 11, 543–545. https://doi.org/10.1038/nn.2112
- Spillmann, L., Dresp-Langley, B., & Tseng, C.-h. (2015). Beyond the classical receptive field: The effect of contextual stimuli. Journal of Vision, 15(9), 7. https://doi.org/10.1167/15.9.7.
- Stoljar, D. (2023). Physicalism. In E. N. Zalta & U. Nodelman (Eds.), The Stanford Encyclopedia of Philosophy (Summer 2023 Edition). https://plato.stanford.edu/archives/sum2023/entries/physicalism/
- Sutton, R. S., & Barto, A. G. (1998). Reinforcement Learning: An Introduction. MIT Press.
- Sutton, R. S., & Barto, A. G. (2018). Reinforcement Learning: An Introduction (2nd ed.). MIT press.
- Timmermans, B., Schilbach, L., Pasquali, A., & Cleeremans, A. (2012). Higher order thoughts in action: consciousness as a unconscious re-description process. Philosophical Transactions of the Royal Society B: Biological Sciences, 367, 1412–1423. https://doi.org/10.1098%2Frstb.2011.0421

- Tononi, G. (2004). An information integration theory of consciousness. BMC Neuroscience, 5, 42. https://doi.org/10.1186/1471-2202-5-42
- Tononi, G. (2008). Consciousness as integrated information: A provisional manifesto. Biological Bulletin, 215, 216-42. https://doi.org/10.2307/25470707
- Tononi, G., & Edelman, G. M. (1998). Consciousness and complexity. Science, 282, 1846-51. https://doi.org/10.1126/science.282.5395.1846
- Tononi, G., Sporns, O. (2003). Measuring information integration. BMC Neuroscience, 4, 31. https://doi.org/10.1186%2F1471-2202-4-31
- Tononi, G., Sporns, O., & Edelman, G. M. (1994). A measure for brain complexity: relating functional segregation and integration in the nervous system. The Proceedings of the National Academy of Sciences (PNAS), 91, 5033-7. https://doi.org/10.1073/pnas.91.11.5033
- Trevethan, C. T., & Sahraie, A. (2009). Blindsight. In William P. Banks (Ed.), Encyclopedia of Consciousness (pp. 107-122). Academic Press. https://doi.org/10.1016/B978-012373873-8.00012-8
- Tsakiris, M., & Haggard, P. (2005). Experimenting with the acting self. Cognitive Neuropsychology, 22(3-4), 387-407. https://doi.org/10.1080/02643290442000158
- Tulving, E. (1987). Multiple memory systems and consciousness. Human Neurobiology, 6(2), 67–80.
- Turova, T., & Rolls, E. T. (2019). Analysis of Biased Competition and Cooperation for Attention in the Cerebral Cortex. Frontiers in Computational Neuroscience, 13:51. https://doi.org/10.3389/fncom.2019.00051
- Turvey, B.E., & Crowder, S. (2017). Investigators and the Scientific Method. In B. E. Turvey, & S. Crowder (Eds.), Forensic Investigations (pp. 67-90). Academic Press. https://doi.org/10.1016/B978-0-12-800680-1.00004-3.
- Tye, M. (1996). The function of consciousness. Noûs, 30(3), 287–305. https://doi.org/10.2307/2216271
- Tye, M. (2021). Qualia. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Fall 2021 Edition). https://plato.stanford.edu/archives/fall2021/entries/qualia/
- Valdez, P. (2019). Circadian Rhythms in Attention. The Yale Journal of Biology and Medicine, 92(1), 81–92. (Full Text)
- van Es, D.M., & Knapen, T. (2019). Implicit and explicit learning in reactive and voluntary saccade adaptation. PLOS ONE, 14(1), e0203248. https://doi.org/10.1371/journal.pone.0203248
- Van Gulick, R. (1992). Nonreductive Materialism and the Nature of Intertheoretical Constraint. In A. Beckermann, H. Flohr, & J. Kim (Eds.), Emergence or Reduction?: Essays on the Prospects of Nonreductive Physicalism (pp. 157-179). De Gruyter. https://doi.org/10.1515/9783110870084.157
- Van Gulick, R. (2022). Consciousness. In E. N. Zalta, & U. Nodelman (Eds.), The Stanford Encyclopedia of Philosophy (Winter 2022 Edition). https://plato.stanford.edu/archives/win2022/entries/consciousness/
- Vierra, N. C., O'Dwyer, S. C., Matsumoto, C., Santana, L. F., & Trimmer, J. S. (2021). Regulation of neuronal excitation–transcription coupling by Kv2.1-induced clustering of somatic L-type Ca2+ channels at ER-PM junctions. Proceedings of the National Academy of Sciences, 118(46), e2110094118. https://doi.org/10.1073/pnas.2110094118
- Von Neumann, J. (1966). Theory of Self-reproducing Automata. Edited and completed by A. W. Burks, University of Illinois Press, Urbana and London (pp. 74–87 and pp. 121–123). (Full Text)
- Xiao, L., Bahri, Y., Sohl-Dickstein, J., Schoenholz, S. S. & Pennington, J. (2018). Dynamical Isometry and a Mean Field Theory of CNNs: How to Train 10,000-Layer Vanilla Convolutional Neural Networks. Proceedings of the 35th International Conference on Machine Learning, PMLR, 80, 5389-5398. (Full Text)
- Yablo, S. (2002). The Many Problems of Mental Causation (Excerpt). In D. Chalmers, D. (Ed.), Philosophy of Mind: Classical and Contemporary Readings. Oxford University Press (pp. 179-197).
- Walsh, K. S., McGovern, D. P., Clark, A., & O'Connell, R. G. (2020). Evaluating the neurophysiological evidence for predictive processing as a model of perception. Annals of the New York Academy of Sciences, 1464(1), 242–268. https://doi.org/10.1111/nyas.14321
- Wang, T., Bao, X., Clavera, I., Hoang, J., Wen, Y., Langlois, E., Zhang, S., Zhang, G., Abbeel, P., & Ba, J. (2019). Benchmarking Model-Based Reinforcement Learning. ArXiv, 1907.02057. https://doi.org/10.48550/arXiv.1907.02057
- Wayman, E. (2023). When Did the Human Mind Evolve to What It is Today? Smithsonian Magazine. https://www.smithsonianmag.com/science-nature/when-did-the-human-mind-evolve-to-what-it-is-today-140507905/
- Webb, T.W., & Graziano, M.S.A. (2015). The attention schema theory: a mechanistic account of subjective awareness. Frontiers in Psychology, 6, 500. https://doi.org/10.3389/fpsyg.2015.00500.

- Weiskrantz, L., Warrington, E., Sanders, M., & Marshall, J. (1975). Visual Capacity in the Hemianopic Field Following a Restricted Occipital Ablation. Brain: A Journal of Neurology. 97:709-28. https://doi.org/10.1093/brain/97.1.709
- Whitmarsh, S., Oostenveld, R., Almeida, R., & Lundqvist, D. (2017). Metacognition of attention during tactile discrimination. NeuroImage, 147, 121-129. https://doi.org/10.1016/j.neuroimage.2016.11.070
- Winkielman, P., & Schooler, J. (2012). Consciousness, metacognition, and the unconscious. In S.T. Fiske, & C.N. Macrae (Eds.), The SAGE Handbook of Social Cognition. SAGE Publications Ltd. http://doi.org/10.4135/9781446247631.n4
- Wegner, D. M., & Wheatley, T. (1999). Apparent mental causation: sources of the experience of will. American Psychologist, 54(7), 480–492. https://doi.org/10.1037//0003-066x.54.7.480 (Full Text)
- Wegner, D. M. (2002). The Illusion of Conscious Will. MIT Press/Bradford.
- Wegner, D. M. (2003). The mind’s best trick: How we experience conscious will. Trends in Cognitive Sciences, 7(2), 65–69. https://doi.org/10.1016/S1364-6613(03)00002-0
- Zeki, S. M., & Shipp, S. (1988). The Functional Logic of Cortical Connections. Nature, 335, 311-317. https://doi.org/10.1038/335311a0