sciendo This is an open access article licensed under the Creative Commons BY-NC-ND License.

Journal of Artificial General Intelligence 12(1) 26-70, 2021
DOI: 10.2478/jagi-2021-0002

Submitted 2020-08-14
Accepted 2021-04-12

The Synthesis and Decoding of Meaning

H. Georg Schulze SCHULZE@MSL.UBC.CA
Monte do Tojal
Caixa Postal 128
Hortinhas, Terena 7250-069
Portugal

Editor: James Marshall

Abstract

Thinking machines must be able to use language effectively in communication with humans. It requires from them the ability to generate meaning and transfer this meaning to a communicating partner. Machines must also be able to decode meaning communicated via language. This work is about meaning in the context of building an artificial general intelligent system. It starts with an analysis of the Turing test and some of the main approaches to explain meaning. It then considers the generation of meaning in the human mind and argues that meaning has a dual nature. The quantum component reflects the relationships between objects and the orthogonal quale component the value of these relationships to the self. Both components are necessary, simultaneously, for meaning to exist. This parallel existence permits the formulation of 'meaning coordinates' as ordered pairs of quantum and quale strengths. Meaning coordinates represent the contents of meaningful mental states. Spurred by a currently salient meaningful mental state in the speaker, language is used to induce a meaningful mental state in the hearer. Therefore, thinking machines must be able to produce and respond to meaningful mental states in ways similar to their functioning in humans. It is explained how quanta and qualia arise, how they generate meaningful mental states, how these states propagate to produce thought, how they are communicated and interpreted, and how they can be simulated to create thinking machines.

Keywords: meaning, meaning coordinates, quanta, qualia, meaningful mental states, hedonic states, language, thought, symbols, simulated physiologies

1. Introduction

This paper is ultimately about building an artificial general intelligent system. Because meaning is central to intelligence and because general intelligent systems will be expected to communicate with humans, this paper specifically addresses meaning in the framework of artificial intelligence (AI) systems.

The paper is organized as follows. I start from a criticism of the Turing Test (TT) and identify what I consider to be the source of the discontent engendered by its failure to lead to solid advances in artificial intelligence (AI). I then look at the extent that the main approaches to explain meaning succeed in doing so. This confirms the need to search for alternative approaches and it allows me to discern simultaneously directions to avoid and possible directions to pursue.

Because the overall aim is to provide a path forward in building an artificial general intelligent system, I next consider mental processes in humans to synthesize, drawing on diverse fields, a concept of meaning. I include the formation of meaningful mental states, the initiation and progression of rudimentary thought, the generation of symbols, the use of language to convey and interpret meaning and I also address the emergence of consciousness. In the following section, these ideas are applied to the problem of making machines that think. I discuss the types of algorithms and data needed and consider aspects of simulating thinking in machines before concluding the work. To maintain a more readable exposition and a clear focus, I elected to present supporting and auxiliary material, along with more extensive discussions, in a number of appendices rather than in the main text.

The concepts presented here are based on the notion that the emergence of meaning is dependent on causal functions consisting of connectionist brain structures with specific architectures. These structures arise from developmental functions that are partly shaped by evolutionary functions. At some level, one or more connectionist structures give rise to internal representational states, mental states, that permit symbol formation – symbols that can be manipulated within recurrent connections and that can be communicated to the external world by the musculature. These causal functions can be activated under untutored (natural or self-directed) conditions or under tutored (teaching) conditions. This structure is not intended to engage in the ongoing debate regarding representationalism (*e.g.*, Hutto and Myin, 20218), but is meant to provide some conceptual coherence to this work.

Though the approach is causal and the aim is to make machines that think, it is also not intended to assert that simulating mental states in machines would identically create such states in machines. Such a strong claim might require an extremely fine-grained correspondence between human mental states and mental states simulated *in silico* and it might require a highly detailed simulation of the entire brain (*e.g.*, Eth *et al.*, 2013). To make an analogy with principal component analysis, making a strong claim amounts to accounting for all of the eigenvectors contributing to the vast extent, if not all, of the variance in the system. The intention here is rather more modest – that simulating just a few principal components, but the critical ones, would be sufficient to build an artificial general intelligent system capable of generating and interpreting meaning digitally, with simulated mental states functioning in a manner analogous to those in humans and therefore enabling machines to virtually 'think'. Though adequately simulating mental states also seems to require some simulated physiology, replicating the entire brain might not be necessary.

2. The Fallacy of the Turing Test

The question of whether a machine can think is a captivating one. It arose because machine capability has been expanding, gradually encompassing many feats formerly accomplished only by humans. Addressing this question, Turing (1950) proposed that a machine could be deemed a thinking entity if, upon remote interaction with said machine, a human could not discern it to be a machine, but perceived it as another human. Thus, the machine use of language had to be indistinguishable from the human use of language. This requires from them the ability to generate and interpret meaning when communicating with a partner. In standardized form, this comparison between human- and machine-based use of language became known as the Turing Test. It has had wide application as an intelligence test of machines and, more importantly, a profound influence

on questions regarding the nature of mental processes such as thinking, intelligence, and consciousness (Muller and Ayesh, 2012).

Despite its impact, however, there is widespread and growing dissatisfaction with the TT (Fjelland, 2020; French, 2012; Muller and Ayesh, 2012). Many feel that research, in particular AI research, has not progressed in directions leading to fundamental understanding and long-term payoffs (*e.g.*, York and Swan, 2012). Instead, it resembles more the use of gimmickry and trickery – the exploit of purely technological processes rather than an embodiment of true insight (Bianchini and Bruni 2012; Greif, 2012). Several remedies are proposed to counter the perceived shortcomings of the TT. These seem to cluster along two directions: by expanding the sensory modalities of the machine and by expanding its behavioral capabilities. Thus, on one side, there is a push to extend the test to include human sensory capabilities beyond those needed for textual language processing, capabilities such as touch, vision, smell, hearing and so on. The impetus on the other side is to include in the test complex behaviors, other than engaging in discourse, like avoiding obstacles, recovering from falls, searching for objects, and so on (Muller and Ayesh, 2012; Linson *et al*., 2012; Zillich, 2012).

However, neither one direction nor the other, or their combination, may be effective. This is because, at a certain level of analysis, the disillusionment with the TT seems to stem from comparing the outcome of one set of processes, those in humans, to the outcome of another set of processes, those in machines. If the outcomes of these different sets of processes are similar enough, it is then assumed that the processes themselves must be similar. Moreover, there is the implication that the similarities may then extend to include still other processes and their outcomes. Such conclusions seem erroneous because different processes could conceivably produce similar results or outcomes, hence more comparisons of outcomes would not necessarily reveal whether the processes are similar or not.

2.1 Meaning Cannot Be Inferred from Process Outcomes

Searle (1980) gave a more granular explication of the problem of generalizing from an outcome to a process. In his influential Chinese room argument (CRA), an English-speaking human, unable to understand Chinese, is confined to a room with an instruction booklet in English. The room permits its occupant to receive written Chinese characters from the outside through a mail slot and to produce written Chinese characters that can be returned to the outside through another slot. The booklet gives instructions on how to respond to different Chinese characters. If given a booklet with suitably good instructions, the room will appear to understand Chinese by virtue of its appropriate responses to Chinese characters. Searle argues that, despite appearances, this is not the case. Likewise, it cannot be concluded that a machine is intelligent merely because it executes relevant computer code in a manner that makes it seem intelligent.

By virtue of differentiating between a process and its outcome, one could assert that adding the outcomes of more processes to the comparison will not help. As alluded to above, expanding the TT to include a wider variety of tests will not reveal whether these outcomes occur through (nearly) identical processes. Comparing more outcomes may limit the subset of processes capable of producing these outcomes, but it may not necessarily constrain the paths taken to these outcomes. Even when comparing an extremely large number of outcomes, the number of processes capable of producing them may not be necessarily limited to processes of the kind that generate the human mind, though such a threshold may exist. Thus, on one hand there is Turing's claim that when a machine obtains the same outcome than a human on a suitably characteristic task it is sufficient to label the machine as 'thinking'. On the other hand, there is Searle's

argument showing that the same outcomes could be obtained via different processes, with the processes used by machines being such that they could not be labeled as constituting 'thinking'. And these two views may not be reconcilable except possibly when approaching an unrealistically large number of outcomes to be compared.

2.2 Meaning Cannot Derive from Symbol Manipulation

Computation is an inherent concept in the TT (*e.g.*, Hodges 2012) and this focus on computation effectively emphasizes syntactical operations. Thus, the manipulation of symbols came to be perceived as a way to generate artificial intelligence with a further perception being that semantics emerge from syntax. Ford (2011), for example, analyses an argument by Rapaport (2006) claiming that syntax is sufficient for meaning to emerge. According to Ford, Rapaport asserts that formal operations on symbols are all that we have and that meaning must arise from these formal operations (*i.e.*, from syntactical operations). Thus, a system must be understood in terms of itself, because uninterpreted symbols are all that we ever get to start with. With the CRA, Searle cast doubt on the idea that formal operations or computations were sufficient to generate meaning by showing that the meaning of words existed separately from their manipulation. If an empty box was moved from A to B, it did not gain content by virtue of being moved: it still was empty when arriving at B. Likewise, manipulating symbols does not imbue them with meaning. This argument of Searle, that formal operations do not provide insight into the semantic content of the Chinese characters and can never do so, is defended by Ford by pointing out that the formal operations available to the person in the room do not constitute the requisite experiences to relate the Chinese characters to their accepted semantic content. The formal operations on Chinese characters executed by the person in the room may very well constitute experiences that become associated with, thus impart meaning to, those same Chinese characters, but not in a way that reflects their original semantic content. The analysis leads Ford to point out that we don't really know just how we get meaningful experiences. Are neurons/biological units necessary to generate meaning? Do we need to study the brain, as Ford (2011) claims, to understand how we get meaningful experiences?

2.3 Meaning Cannot Derive from Symbol Grounding

At this point it should be clear that a related problem arises – known as the symbol grounding problem (SGP). The SGP addresses the connection between manipulated symbols and their referents (Harnad, 1990). How is the word 'tree', or the Chinese character for tree, connected with a tree, that is, with a real world object? As argued by Ford, the formal operations on symbols, *i.e.*, manipulating Chinese characters, does not give semantic content to them and meaning cannot arise from manipulating them. The SGP addresses the fact that the person in the room cannot know what the Chinese characters refer to in the outside world, and hence cannot meaningfully interpret sequences of Chinese characters. Although the person in the room may begin to form internal representations of certain Chinese characters by virtue of his experiences with them, these internal representations are still not linked with their accepted external referents. Thus, the SGP can be seen as the problem of providing an artificial agent with the ability to link the symbols that it manipulates to the symbols' corresponding referents in the external world. Although many researchers have come to view a solution to the SGP as a means to attach meaning to symbols, Rodríguez *et al*. (2012) argue that Harnad never intended to show that, by grounding its symbols, an artificial agent can grasp their meanings. Thus, being able to pick out the external referents corresponding to its manipulated symbols does not give the machine

meaningful internal states (Rodríguez *et al*., 2012). From this follows the important conclusion that symbol grounding does not generate meaning. In my view, it also suggests that semantic content, at least by itself, does not constitute meaning.

Furthermore, even with symbol grounding, another problem arises. Because each symbol is grounded to its respective referent, irrespective of sequence, 'man bites dog' and 'dog bites man' will have the same meaning. Thus two new symbols need to be created to differentiate between these different meanings or the meanings have to be conveyed through the way in which the symbols are sequenced, leading us back to symbol manipulation. When symbol manipulation is used to construct superordinate bindings according to consistent rules of manipulation, a particular sequence of symbols essentially functions as a new symbol. Symbol manipulation, fundamentally, creates new symbols, not meaning.

2.4 Meaning Cannot Derive from Social Interactions

Due to the elusive nature of meaning, the search for it has also moved to the field of social interactions. Indeed, both the problems of symbol grounding and semantic meaning surface whenever two or more parties need to communicate. Thus, if meaning is definable in terms of meaningful mental states that, in turn, are linked to words, how can one agent exchange thoughts with another agent who is likely to have rather different meaningful mental states? Hence, for agents to communicate, we must assume that the same words link to the same meaningful mental states in those agents able to communicate. But this should make misunderstanding in communication improbable when, in fact, it is a common occurrence (Arrighi and Ferrario, 2005). Furthermore, because agents have different histories, they would have had different experiences with the real world leading, not only to different meaningful mental states, but also to different, but possibly overlapping ways, in which the words that they use relate to their external referents. For example, one agent may only have experience with conifers while another has only experience with deciduous trees, thus their 'tree' symbols would be grounded in an overlapping, but non-identical, manner. Consequently, agents may interpret words or sets of words such as expressions, differently. These difficulties have been approached by translating the differing mental concepts of different agents into a created framework that is then shared to permit communication or by retaining the different concepts of each agent but allowing them to find a compromise between meanings that are at variance. In the latter case, meaning coordination or meaning negotiation (MN) can be used. With meaning coordination agents attempt to find mappings between the meanings of expressions and with MN agents try to reach an agreement about the meaning of a word or a concept when mapping is not possible (Arrighi and Ferrario, 2005; Burato *et al.*, 2012). AI scholars have dealt with the problem of MN by using argumentation schemes, beliefs merging and information fusion operators, and ontology alignment, but the proposed approaches depend upon the number of participants (Burato *et al.*, 2012). Regardless, though, of how the issue of meaning in communication is addressed, the reader will realize that some meaning must exist before agents will attempt to communicate them and this presupposes the existence of meaningful mental states. Put differently, some common understanding has to exist *a priori* – before any negotiation can occur.

3. Mental Processes in Humans

The fundamental problem with the TT then is that different algorithms can produce the same results. Thus, one cannot infer from similar results that similar algorithms were used to produce

them. Therefore, the TT cannot really tell us, unambiguously, whether machines can think or not. Furthermore, the TT depends on symbol manipulation and symbol manipulation does not seem capable of generating meaning - neither syntactical operations, nor operations to imbue symbols with semantic content, nor symbol operations in social contexts succeed in generating meaning. And if meaning cannot be generated by a machine, how can it be said to be able to "think"? A better test is needed and a better way to generate machines that can think.

We are forced, therefore, to turn our attention to understanding the processes that underpin human thinking and to implement them algorithmically. In this context, two related problems arise. First, the relevant data are accessible to the human mind in a different form compared to how the same data are accessible to machines, thus even if the same algorithms could be implemented in machines, the outcomes may not be informative. It will be necessary to analyze the types and nature of the data on which the algorithms of the human mind operate and to understand how these data are accessed. Second, it is possible that the algorithms implemented by the human mind cannot be implemented by machines because their substrates or computational platforms do not allow for it. Therefore, simulating in machines essential aspects of the human computational platform might be called for.

What, then, are the human mental processes that need to be understood and simulated to generate machines that can think? Under the Representational Theory of Mind (Pitt, 2020), mental processes such as thinking are viewed as sequences of mental states. Because mental states are intentional states, they are sequences of mental states about something. But are they meaningful or do they need to be? I want to point out here that, if meaning is to be communicated, it must arise within the communicating agent and this presupposes the existence of meaningful mental states in that agent. Thus, a machine does not only need to produce a sequence of mental states, but this sequence of mental states needs to be a sequence of meaningful mental states or result in a meaningful mental state that can be communicated. These considerations suggest that to understand the human mental processes necessary to produce thinking makes it essential to consider the nature of meaning itself, how meaningful mental states arise, how they are sequenced, and how they are communicated.

3.1 Meaningful Mental States

I consider all meaningful mental states to arise from conscious experiences, the latter addressed further below. These meaningful mental states are generated through developmental functions, the types and specifics of which are, in turn, partly shaped by evolutionary functions. A discussion of developmental and evolutionary functions as causal theories of mental content, and content determination in representational theories of mind, can be found in Adams and Aizawa (2017) and Pitt (2020), respectively. These causal functions can be activated under unstructured (natural or self-directed) conditions or under structured (teaching) conditions (*cf.* Appendix A).

Thus, a conscious experience of X generates a meaningful mental state x through such a causal function. The ultimate function of x is to enhance the survival and fitness of its bearer by permitting more advantageous behavioral responses to X. Therefore, the 'accuracy' of x is important as it will ultimately have a fitness benefit. The accuracy of x can be finely honed by the trade-offs present in the acquisition environment. Thus, a conscious experience of X generates a meaningful mental state x and its accuracy is shaped by the benefits of x well-functioning under conditions of commission and omission. For example, if X were a predator, an x of good quality would allow evasion or escape from the predator. It would also avoid the unneeded energy expenditure incurred when fleeing something innocuous that was mistaken for the predator.

Likewise, it would avoid freezing, and thus avoid missing potential opportunities to engage in beneficial activities such as grazing, grooming, or procreation. Therefore, the generation of x is optimal when it develops in response to, and only in response to, X. And, because it does not emerge full-blown, but is a process of adjustment or development, errors occur, at least initially.

A primary meaningful mental state is an original meaningful mental state as it evolves in the course of development from birth onward, rather than a derivative mental state or a mental state triggered by other mental states due to interactions, iterations, recursions, inductions, referencing, embedment or other higher-order processes. All other meaningful mental states are derivatives of primary meaningful mental states and therefore possess the same nature as primary meaningful mental states, but with their attributes in modified form.

A primary meaningful mental state is a state that is generated by a conscious experience that includes a specific 'formative object' or a specific 'formative action'. The specific formative object or formative action is the most salient object or action present in the experience. Henceforth, I shall use "formative object" to refer to either a formative object or a formative action, or both when they occur together, depending on the context. More specifically, a primary meaningful mental state is a state deriving from the perception of the physical attributes of the formative object being experienced and from the feelings that arise in how the formative object is being experienced. It forms in the process of experiencing something for the first time and is modified and elaborated upon by subsequent similar experiences and by the simultaneous or independent recall of these experiences. Recalling an experience induces the related meaningful mental state, but in modified form, for example, with reduced or augmented intensity.

The formative object has at least one physical dimension, such as intensity, length, color, weight, and speed, that can be registered or perceived by the sensory systems when engaging with it and that is quantifiable in some ordinary sense. Furthermore, engagement with the formative object always provokes a feeling. Though feelings might be construed as having various attributes, they have at least a valence. A valence itself has the attributes of being positive or negative (*i.e.*, agreeable or disagreeable, respectively) and of an intensity of positivity or negativity. Perception of the physical attributes of the formative object allows the agent to orient itself physically with respect to the formative object – what it is, where it is, how it is. I consider these essentially as 'cognitive' aspects and they pertain in the main to the agent's external world (*i.e.*, objective conditions or states). Valences reveal how engagement with the same formative object affects the state of the agent – beneficially or adversely. I consider these essentially as 'affective' aspects and they pertain in the main to the agent's internal world (*i.e.*, subjective conditions or states). Both of these aspects always occur together, but with attributes that can vary independently to reflect realities in the environment of the agent and of the agent itself. This manner, of linking the state of the external world to the state of the self through valences, permits an agent to navigate the external world advantageously.

The specific manner of primary meaningful mental state formation is conceived on connectionist principles (*cf.* Appendix A) that, at some level, give rise to representational states that permit the production of discrete words such as "x". Pitt (2020) provides a treatment of computational theories of mind that includes the classical and connectionist views. A meaningful mental state x, therefore, is conceptualized as a certain pattern of activation in an assembly of nodes of a given architecture of weighted connections produced in response to the presence of X. The exact same assembly of nodes will give rise to a different pattern of activation of these nodes in response to the presence of Y, thus producing the meaningful mental state y, different from x. The presence of X results in input from the sensory organs being distributed simultaneously via two parallel pathways to the input nodes of the meaningful mental state-producing assembly of

nodes. The first of these involves sensory and cognitive processing while the second involves affective and emotional processing that is dependent on a combination of the prevailing internal state and the internal state induced by X. This system can be considered a simplified description of sensory input that is distributed by the thalamus to cortical areas for cognitive processing and limbic areas for affective processing before subsequent integration of such processed cognitive and affective components in a different neural assembly such as the cingulate cortex. (Vertes *et al.*, 2015a,b; Rolls, 2019; *cf.* also Appendix D) The generation of x is therefore based on the presence of both types of input as presented to the input nodes of the assembly. Without the presence of both types of input, a mental state ⋇ might arise, but it would not be a meaningful mental state.

On this view, a primary meaningful mental state is a mental object that is about the formative object and an associated affective experience of engagement with the formative object. Clearly then, primary meaningful mental states have both intentional and phenomenal character, as do their derivatives by virtue of being later states that have integrated at least some aspects of the earlier primary meaningful mental states as described further below.

Support for this view of the joint presence of two orthogonal dimensions of the formative object, cognitive and affective, derives from a non-associative learning process called *habituation*. Here, a new stimulus initially elicits a response that diminishes with repeated presentation, for example, the startle response following the presentation of a loud noise (Kupfermann, 1991). Currently prominent theories of habituation all have an arousal system (Thompson, 2009). Thus, the presence of a novel stimulus is noticed via sensory systems that respond to its physical dimensions and provokes arousal and awareness of it. Though completely novel stimuli provoke some disagreeable arousal, novel stimuli that can be interpreted in the context of an existing schema (Piaget, 1977) might be accompanied by a more defined positive or negative character. Furthermore, the *mere exposure effect* also has an affective component as previous exposure to novel, but neutral, stimuli leads to the preference of presented stimuli compared to stimuli not presented (Zajonc, 1968). The preference component of exposure arises when stimuli are presented for a period too short to permit recognition. Thus, when stimuli are presented for only 5 ms, a stimulus preference already becomes evident while there is no effect on recognition yet (Bornstein and D'Agostino, 1992). This is consistent with habituation, but suggests that, upon habituation, there is not only a reduction in disagreeable arousal provoked by a novel neutral stimulus, but that preference to the novel stimulus arises in the course of habituation. If the neutral stimulus is completely unfamiliar, the valence tends to be negative, if it is somewhat familiar, the valence is positive. However, the point here is that all stimuli, even neutral ones—those that do not already have a valence—are 'classified' or labeled with an incipient valence. Hence, in the normal case, a positive or negative valence, imperceptible or intense or in-between, always accompanies perception of any stimulus.

In the absence of affect meaninglessness arises. Feelings that life is empty and meaningless is a common experience of persons with affective disorders (Nichols *et al*., 2021; Prigerson *et al*., 2021; Tolentino and Schmidt, 2018; Vanhooren, 2019). Furthermore, The Diagnostic and Statistical Manual of Mental Disorders-5 requires the presence of one of two major criteria, depressed mood or anhedonia, for a diagnosis of major depressive disorder with the presence of anhedonia shown to be indicative of a more severe case (Tolentino and Schmidt, 2018). Similarly, the presence of emotional numbness contributes to the diagnosis of prolonged grief disorder (Prigerson *et al*., 2021). Thus, the affective mental disorders forcefully bear out that meaninglessness is experienced in the presence of anhedonia (clinically, a loss of interest or

pleasure) and the presence of emotional numbness (clinically, the absence or marked reduction of emotional experience).

Detailed treatments in support of the view of the joint presence of two orthogonal dimensions of the formative object, cognitive and affective, can be found in Appendix B (Habituation and Arousal), Appendix C (Homeostatic Mechanisms and Hedonic States) and Appendix D (Brain Structures Subserving Meaningful Mental States).

3.2 Symbols

The pattern of activation at the output nodes of this assembly gives rise to a representational state that approximates a classical discrete representational state. The representational state is therefore a sub-state of the meaningful mental state with which it is associated and it is postulated that its generation indicates that the meaningful mental state with which it is associated has been fully formed or 'completed'. This completion makes possible a transition to the next meaningful mental state in a thought sequence.

The formation of a representational state can therefore be conceived of as the formation of an internal 'symbol' $\boldsymbol{x}$. When, through the activation of further neural architecture, a set of musculature is activated to produce a discrete word such as "x", the external symbol x is generated. Hence the formative object X constitutes the referent of the symbols $\boldsymbol{x}$ and x, but can only become the referent of a given symbol after the symbol has been created. This amounts to an organic form of symbol grounding. These ideas are schematically shown in Figure 1.

The cognitive part of the relationship that arises between the formative object and the primary meaningful mental state along pathway A (*cf.* Figure 1) is defined as a 'quantum' and can be given an intensity, strength, probability, and so on. The affective part of the relationship that forms between the formative object and the primary meaningful mental state along pathway B is defined as a 'quale' and can be also be given an intensity, strength, probability, and so on. One can conceive of the quanta and qualia as traces being laid down in memory where memory must be understood in terms of the strength of the synaptic connections in the pathways of the related neural assemblies. Conceptually, a quantum pertains to the strength of the cognitive aspect of the relationship between a meaningful mental state and its formative object and, by extension, between a symbol and its referent and a quale to the strength of the affective aspect of those relationships. I now define that pair of values, a quantum and a quale, associated with a given meaningful mental state, as a *meaning coordinate*, analogous to a Cartesian coordinate.

The strengths of these relationships' cognitive aspects, thus the intensities of the quanta, increase with the frequency and intensity of occurrences and gradually decrease in the absence of occurrences. The strengths of these memory traces, and their changes, can very between individuals and be affected by various factors including ageing, injury, and disease. Likewise, the strengths of the affective aspects of these relationships, thus the intensities of the qualia, are modified. Because the experiences of the formative object might vary, the generated feelings also vary, causing the relationship strengths to be augmented or diminished by virtue of a 'summation' of previous and newly generated valences depending on whether a positive or negative valence was generated. They also gradually decay in absolute value in the absence of occurrences. The waxing and waning of these strengths constitute, respectively, types of learning and forgetting, and the process is analogous to an infinite impulse response filter where a fraction of the new association strength is added to the sum of previous strengths. These modifications of memory traces never stop, thus learning and forgetting are lifelong. By implication, the cognitive and affective aspects of the conceptual relationship between a symbol and its referent likewise

change, thus symbol grounding is not seen as static. Symbol grounding should furthermore be considered to imply both cognitive and affective grounding.

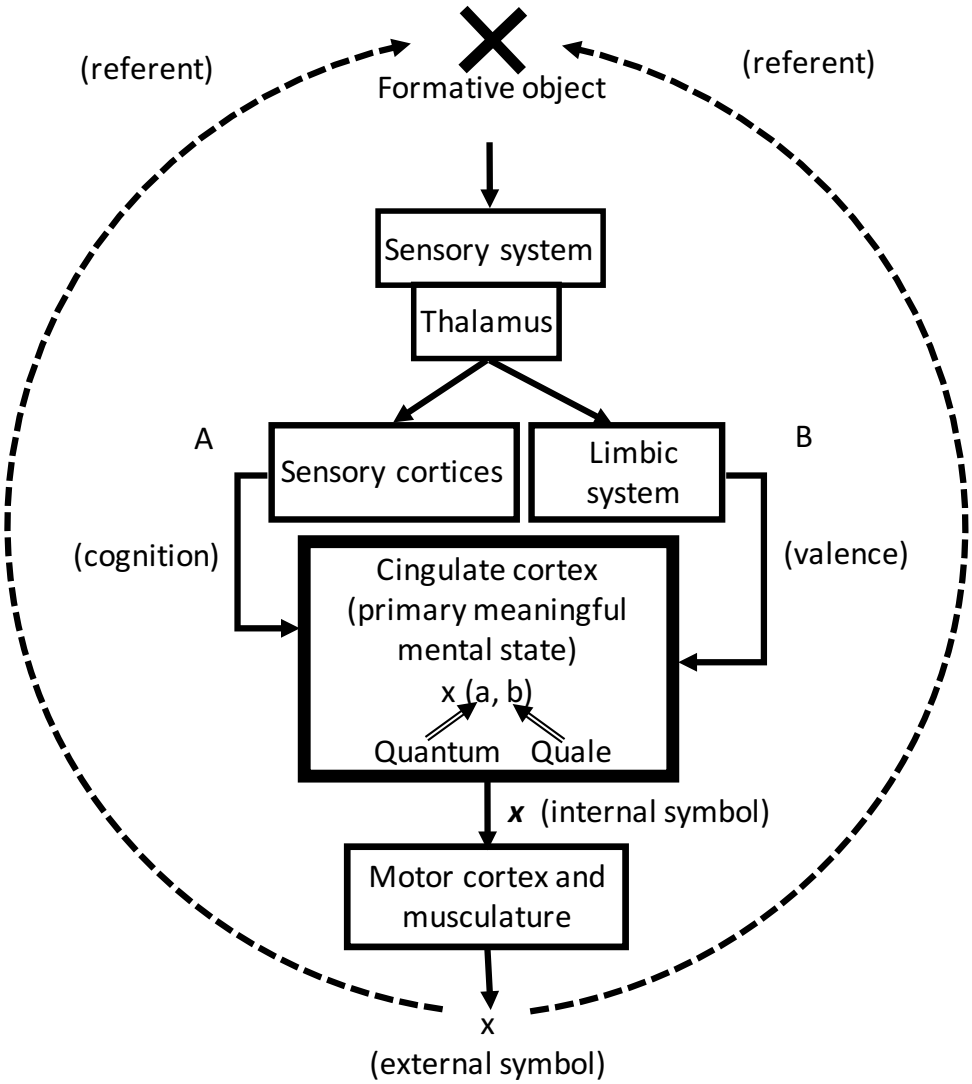

Figure 1
Schema of the formation of a primary meaningful mental state, x(a,b), from the presence of a formative object X. Simultaneous but separate cognitive, A, and emotional, B, processing pathways contribute to the cognitive, a, and affective, b, aspects of the primary meaningful mental state x. The cognitive aspect is called a 'quantum' and that of the affective aspect a 'quale' and together they form a meaning coordinate. The intensities of the quantum and quale can vary independently. Also shown are putative brain structures subserving the generation of primary meaningful mental states, leading to the formation of representational states that constitute internal and external symbols. Potential connections from other brain structures contributing other effects are not excluded. Conceptually, the dashed arrows indicate that the quantum represents the cognitive grounding of a symbol to its referent and the quale represents the affective grounding of the same symbol to its referent.

3.3 Derivative Meaningful Mental States

When commerce occurs in close temporal proximity with two objects that have become formative objects due to the fact that they have generated, independently, primary meaningful mental states, their respective meaningful mental states become linked by a memory trace. As for quanta and qualia traces, a linking trace is also understood in terms of the strengths of the synaptic connections in the relevant neural pathways. The strengths of synaptic connections determine, in turn, a probability with which a current meaningful mental state can link to a subsequent one. Furthermore, the linking memory trace also has a dual character and constitutes a linking meaningful mental state that is a derivative meaningful mental state. Its quantum intensity also increases with the salience (*e.g.*, frequency or intensity) of joint or close temporal occurrences of the formative objects and gradually decreases in the absence of occurrences. This quantum strength may be given a positive value if x occurs before y and a negative value if y occurs first. Its quale strength derives from the affective states induced by their joint occurrences. These strengths wax and wane in a manner similar to that previously described. A quantum or quale strength that exceeds a certain threshold, possibly varying but high, causes the formation of a secondary—also derivative—meaningful mental state from the linking meaningful mental state. The formation of a secondary meaningful mental state occurs through a psychological process known as *chunking* or *binding* and the sign of the quantum strength, indicating temporal information, is supported by the forward bias of binding (*cf.* Appendix E). A linking meaningful mental state remains in existence after the formation of a secondary meaningful mental state even though it can no longer produce a secondary meaningful mental state. Once in existence, any primary or secondary meaningful mental state can attain, through binding, a discrete association with any other such meaningful mental state and such an association can also become a secondary meaningful mental state. This is illustrated in Figure 2.

It is postulated that the richness of experience derives from multiple linking states and associations between meaningful mental states. It is also implied that new symbols, one internal and one external, related to the new meaningful mental state, arise in the process. Symbols that form in response to secondary meaningful mental states are likely to be single symbols like ***w*** (*e.g.*, 'welt') or simple compound symbols like ***xy*** (*e.g.*, 'mosquito bite'). Symbols can also occasionally form in response to linking meaningful mental states and are likely to be complex descriptive symbols, for example, ***x in the y*** (*e.g.*, 'mosquito in the air'). It relates to the previous contention that symbol manipulation produces new symbols, not meaning, because the meaningful mental states emerge before they become represented by symbols.

Connectionist models can also be trained to produce a specific output pattern, that is, a symbol, when presented with any input pattern, that is, a formative object, from a category of similar input patterns. This requires the repeated presentation of a variety of examples from the category used and the requirement to produce a given output pattern (*cf.* Appendix A). When the neural assembly is adequately trained, the meaningful mental state that obtains is therefore an amalgam or derivative of mature and immature primary meaningful mental states. This is because the presentation of category members might include members (*i.e.*, formative objects) not previously encountered for which no primary meaningful mental state exists. The presentation might trigger the incipient formation of a primary meaningful mental state for each of the unknown category members. Known category members, for example, a specific dog, will already have generated primary meaningful mental states. On this view, the symbols that arise (internal and external) are then symbols pertaining to a class or category of similar objects. Every derivative meaningful mental state so generated also has its own quantum and quale with initial

strengths being the mean of the strengths of existing members (*i.e.*, other meaningful mental states) of that class.

Formative object → Symbol ↓	Mosquito	Air	Bite	Mosquito bite
"Mosquito"	(0.9, -0.7)	(0.2, -0.70)	(0.75, -0.8)	(0.35, -0.8)*
"Air"	(0.2, -0.70)	(0.91, 0.9)	(0.1, -.04)	(0.0, -0.2)*
"Bite"	(-0.75, -0.8)	(0.1, -.04)	(0.87, -0.95)	(0.4, -0.5)*
"Mosquito bite"	(-0.35, -0.8)*	(0.0, -0.2)*	(-0.4, -0.5)*	(0.4 , -0.5)*

chunking

Figure 2
Primary and derivative meaningful mental states are shown in a grid-like fashion with all their associated meaning coordinates. Linking meaningful mental states form through linkages based on memory traces of existing meaningful mental states and subsequent consolidation through binding causes secondary meaningful mental state formation. Primary meaningful mental states are shown in light gray. Linking mental states are shown in dark gray. Secondary mental states are shown patterned. A fixed quantum binding threshold of 0.75 is assumed. The meaning coordinates of the meaningful mental states are shown in brackets as ordered pairs with putative values; the asterisk, *, denotes initial values (just after formation). For quanta, their temporal order is shown in row → column order. Thus, the positive quantum in the box 'Mosquito, Bite (0.75, -0.8)' indicates that 'Mosquito' temporally precedes 'Bite' while the negative quantum in 'Bite, Mosquito (-0.75, -0.8)' indicates that 'Bite' follows 'Mosquito'. If there is no difference in sign, there is no temporal effect. For qualia, the sign indicates whether the affect is positive or negative, thus both the abovementioned combinations are associated with negative affect.

3.4 Thought

The existence of meaningful mental states permits them to be sequenced based on their linkages in temporary ways, thus in a manner different from binding. Such temporary sequences constitute primordial thoughts that are further elaborated by specialized neural architectures. In principle, a thought is a succession of meaningful mental states that proceed based on the nature of either the quanta or the qualia of the current and succeeding meaningful mental states. Though thought progression occurs via the links represented by linking states, the normal 'contents' of a thought are primary and secondary meaningful mental states. The series of internal symbols provoked by such a succession of meaningful mental states produces the internal expression of a crude thought. In the more mature person, crude thoughts are modified by the rules of grammar and potentially other processes (*e.g.*, logic, prediction), but in the young child less so or not at all. Though an internal stimulus such as a perceived occurring internal state can initiate a thought process, the most common triggers are likely to be an external stimulus, for example, the presence of a supervisor giving instructions or the occurrence of an external object or event such as a thunderclap.

Due to forgetting, the absolute values of all quanta and qualia will diminish by some small amount, not necessarily all by the same amount. In contrast, due to learning, the presence of the

stimulus X will cause adjustments of the associated absolute values of the quantum and quale, often augmenting them. This might cause its quantum or quale to exceed in absolute value all others and so initiate a thought process. Once initiated, progression can depend on either quanta or qualia or a combination of both as shown in Figure 3.

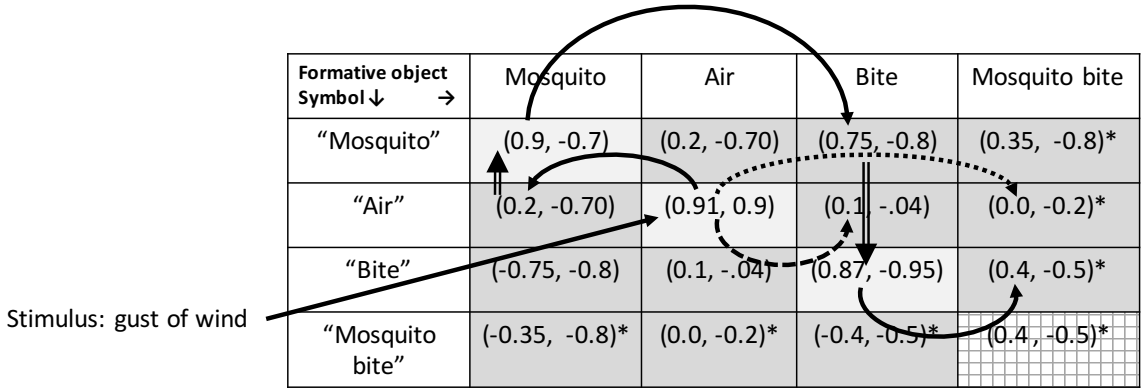

Figure 3
Illustration of thought progression via a series of related meaningful mental states started by an external stimulus. The stimulus, such as a gust of wind, caused an augmentation of the quantum and quale of the associated meaningful mental state 'Air'. This resulted in it having the largest quantum or quale value in the array of meaningful mental states and for the thought to originate there. This meaningful mental state progresses to the next one by finding that meaningful mental state that is currently linked most strongly to it. Thus, along the row containing 'Air', the column 'Mosquito' has the largest quantum link with it (solid arrow) and the meaningful mental state 'Mosquito' becomes activated (double lined arrow). From 'Mosquito' the following meaningful mental state is activated through a similar progression. Progressions can proceed based on quanta, qualia, or a combination of both. Further details can be found in the main text.

In the case of quanta, the succeeding meaningful mental state is that meaningful mental state that is most strongly associated with the current meaningful mental state other than the one immediately preceding it. With reference to Figure 3, for example, the occurrence of a gust of wind might augment the quantum associated with the meaningful mental state represented by the symbol 'Air' and so activate this meaningful mental state. Considering now all quanta along the row in which 'Air' is located, the link with the newly formed derivative meaningful mental state 'Mosquito bite' is weak (dotted arrow), that with 'Bite' is stronger (dashed arrow), and the strongest occurs in the column 'Mosquito' (solid arrow). Thus, after 'Air', the meaningful mental state 'Mosquito' is activated (double lined arrow). Considering now all quanta along the row in which 'Mosquito' is located, and proceeding as before, the meaningful mental state 'Bite' becomes activated (double lined arrow). This leads to the row containing 'Bite' and within that row, to the column 'Mosquito bite'. In the case of qualia, similar 'rules of progression' apply, but based on the absolute values of the relevant qualia. Initiated with 'Air', this leads to 'Mosquito' and then to either 'Bite' or directly to 'Mosquito bite'. Progression can also occur in a mixed fashion, depending on whether the strongest link is via quantum or quale. The sequence of progression, therefore the thought, might differ depending on the type of progression used. A thought is hypothesized terminated when thought progression is interrupted by an external or internal stimulus, in which case a new thought might commence, or when the needed energy

supply to the neural assembly is reduced or its activation is inhibited, such as by the onset of sleep.

Though 'rules of progression' of meaningful mental states in the thought process are alluded to, they are not formalized at this stage and will have to be elaborated upon in future work. However, a few observations might be relevant. (i) Normal progression from one meaningful mental state to the next excludes the meaningful mental state immediately preceding the current one if the current one is not the initial state. It is known that neurons are refractory for some period immediately after activation and that such refractory periods could substantially affect the activation patterns of neural assemblies (*e.g.*, Stein, 1965; Rolls, 1971; Cáceres and Perthame, 2014). Excluding in a thought process a reversion to the meaningful mental state immediately preceding the current one can thus broadly be justified on the basis of a variable refractory period for meaningful mental states. (ii) Though progression is here illustrated with the use of the strongest links, one might also consider progressions that are based on the biggest changes occurring in links. (iii) Furthermore, universal or near-universal associations, such as those related to breathing, might be disregarded either because augmentation is inhibited in the manner of habituation or because they are likely to be fraught with progression ambiguities. (iv) Succession of meaningful mental states primarily by quanta might, in general, have a more rational character while succession primarily by qualia might have a more emotional character. If this is borne out by experiment, two artificial agents, therefore, could be made to have different personality traits due to providing them with such different habitual patterns of thought. (v) It is hypothesized that succession by qualia is dominant in the young child. Eventually, more complex succession procedures develop, as the ability to take past thoughts and their consequences into consideration arise, and some shifting towards progression based on quanta occurs. (vi) Though some meaningful mental states are more labile and some are more stable, all meaningful mental states are dynamic in nature due to the constant adjustments of their meaning coordinates. They must reflect the constantly changing relationships between the state of an agent and the state of the environment with which it is interacting. Therefore, thoughts are not static entities, but have dynamic character. Even though the same thought might arise on different occasions, it is unlikely that these same thoughts will involve identical quanta and qualia. Thus, they might consist of a sequence of meaningful mental states that give rise to the same sequence of symbols, but the meaning coordinates of one or more of these meaningful mental states are likely to differ.

3.5 Communication and Language

In principle, when two individuals deploy the same symbol derived from the same formative object, that symbol is grounded to the same referent and they can communicate. This communication contains information, in the form of external symbols, but no meaning. Meaning is reconstructed by the receiver of the communication.

Specifically, a thought process, as a series of meaningful mental states, produces a series of internal symbols as explained above. When these symbols are also expressed externally in order to communicate, a message is produced that can be directed at a recipient. The symbols contained in the message, however, do not contain their associated quanta and qualia – they do not contain meaning coordinates. Therefore, they are devoid of meaning. If these symbols are recognized by the recipient—thus grounded to the same referents as they are in the producer—their perception provokes in the recipient those meaningful mental states associated with these symbols. Technically, therefore, meaning is not 'synthesized' when producing a message. Neither is, or can be, meaning 'decoded' from such a message because it contains no meaning.

On the above view, the success of communication depends on the extent that an intended sequence of meaningful mental states is induced in the recipient of the message. In the most basic case, a producer of a message might intend for the recipient to understand the meaningful mental states of the producer. The success of communication will then depend on the extent that the meaningful mental states of the producer overlap with the meaningful mental states evoked in the recipient and the natures of their overlapping and non-overlapping parts. The non-overlapping parts give rise to phenomena of the sort "gained in translation" or "lost in translation". When all the attributes of the intended and induced meaningful mental states overlap precisely, perfect communication ensues. It follows now that, to effect communication, supervised learning becomes essential (*cf*. Appendix A). It ensures that, in those individuals that need to communicate with one another, the same symbols are generated in response to a given formative object. Furthermore, these symbols would also activate the same meaningful mental states in these individuals as the formative object with which they are paired during training. Individuals can then refer to the same referent with identical symbols. However, supervised learning also admits intentional as well as inadvertent misrepresentations.

Perfect communication between humans is probability-wise negligible because the intended and induced meaningful mental states will vary to a lesser or greater extent for two major reasons. First, because different individuals have varying experiences, their symbols are rarely derived from identical formative objects, and, even if they were, these formative objects are unlikely to have been experienced in identical ways. Second, as explained above, thoughts are dynamic in character due to the constant adjustments of the meaning coordinates of meaningful mental states. Thus, even if the symbols used in communication between humans are derived from identical formative objects experienced in identical ways, it is unlikely that the sender of a message will be able to evoke meaning coordinates identical to the ones being experienced in the receiver of that message. In the absence of perfect communication, 'good enough' communication, dependent on sufficient overlap between intended and induced meaningful mental states, must then suffice. In artificial agents, where access to meaning coordinates is explicit, the failure of communication and the nature of this failure can be determined based on the differences between the respective sequences of meaning coordinates – the message-related meaning coordinates of the producer and the corresponding ones arising in the receiver.

When, in the majority of cases, identical symbols used by two individuals are grounded in the same referents, they can communicate in the same 'language'. When, for the same referent, the symbols generally differ, their languages differ. In the latter case, communication requires an extra step. In a very general sense, the symbol representing the intended meaningful mental state in one language must be matched to the symbol representing the same intended meaningful mental state in the other language via connections to their common referent. For various reasons, this matching will not be perfect. The matched symbol representing the intended meaningful mental state now has additional, thus augmented, overlapping and non-overlapping parts with the induced meaningful mental state. This brings with it an increased potential for misunderstanding and the need for additional error mitigation and error tolerance schemes. Appendix F further explicates the communication process delineated here.

3.6 Consciousness and Self-Consciousness

The quest for meaning confronts one rather fiercely with the problem of consciousness, and when pushed further, self-consciousness. This entanglement emerges immediately upon asking whether

an agent can have meaningful mental states without consciousness and whether it can have consciousness without meaningful mental states.

The problem is further compounded by the seeming existence of different types of consciousness. Block (2002) differentiates between phenomenal consciousness (he calls P-consciousness), access consciousness (he calls A-consciousness), self-consciousness and monitoring-consciousness. P-consciousness relates to the experience of receiving sensory inputs and processing them, is subjective in nature, and distinct from cognitive, intentional, and functional properties. A-consciousness is the information processing complement of P-consciousness. Relying on having access to and exploiting the right computational structures, it prods with—ready for use—informational aspects of the sensory inputs.

On the face of it, when trying to answer the question of whether meaningful mental states depend on one or another type of consciousness, a certain parallelism is observed between them. Considering that P-consciousness and A-consciousness are complementary and that meaningful mental states require the joint presence of quanta and qualia, it would seem that P-consciousness can be mapped onto or related to qualia and, likewise, A-consciousness to quanta. However, my reading of Block (2002) suggests that P-consciousness might be interpreted as a low-level consciousness, it being mostly related to sensory states while A-consciousness might be interpreted as a higher-level consciousness, it being mostly related to non-sensory cognition and information broadcasting. Thus, they seem hierarchical, while the quanta and qualia of meaningful mental states are not hierarchical – they are present jointly at all levels of meaningful mental states starting with the primary ones. Hence, the concepts of P-consciousness and A-consciousness appear nonconforming with meaningful mental states as exposed here.

An insight about consciousness might be gained by asking what the purpose of consciousness is. Here, considering homeostatic mechanisms (*cf.* Appendix C) could be helpful. Homeostatic mechanisms regulate many important physiological variables, such as body temperature, in an automatic manner that is generally below awareness. When unable to manage, physiological stress occurs and an awareness of the problem needs to arise to recruit a supporting domain general behavior – for example, donning a sweater. Recruitment occurs once the level of stress exceeds a certain, possibly varying, threshold, causing arousal via the locus coeruleus and generation of valences in the limbic system. Valences are the positive and negative hedonic states that reflect whether increasing stress is present (negative hedonic states) or whether stress is abating (positive hedonic states). This is one way in which qualia emerge. Valences serve to initiate, maintain, and eventually terminate the necessary behaviors. When several homeostatic mechanisms are experiencing regulation problems (*e.g.*, cold, hunger), attention is required to deal with the most critical regulation problem first.

The preceding description suggests that the induction of arousal (see also Appendices B, G) is a prime candidate for explaining consciousness. Arousal modulates the functioning of brain circuits and readies them for activity (Aston-Jones and Cohen, 2005); it 'switches on' certain brain circuits. After activation, this initially 'empty' conscious state can then be populated with content. For example, due to the state of arousal, awareness of sensory stimuli and feelings can occur (pathways A and B in Figure 1) and meaningful mental states arise along with phenomenal consciousness. I shall call the latter f-consciousness to differentiate it from Block's (2002) P-consciousness. Because arousal is usually provoked by something, there is an immediate population of the empty state with the arousing cause; thus an empty conscious state is unlikely to exist in the normally functioning brain.

Self-consciousness (s-consciousness) is a self-referential f-consciousness, it requires the ability to form self-referential meaningful mental states, along with the ability to perform

operations on them. The existence of two brain hemispheres, two near-identical functional 'brains' lodged in the same skull and connected to the same body and traveling through the world together, permits one hemisphere to objectively refer to the other as "I". It is used, along with mirror neurons, in Appendix G to elaborate the concept of s-consciousness.

4. Making Machines that Think

Meaningful mental states arise as a way to relate the objective world, including the body objectively experienced, to the subjective self—that self that experiences qualia. In humans, these relationships have a specific structure. This structure arises because humans have a certain anatomy with a certain physiology, different from those of other living things. Thus, a human is exposed to objective things through sensory systems that are largely like those of other humans and respond to them with a brain that is structured like, and operate in ways similar to, those of other humans. Humans develop in similar ways graded patterns of relationships between the objective and the subjective. Therefore, metaphorically speaking, humans are 'in tune', much as the instruments of an orchestra calibrated to concert pitch. And like the instruments of an orchestra, they can make music together – they can communicate.

Consequently, to generate and interpret meaning digitally, digital agents must have digital meaningful mental states themselves that function in a manner analogous to those in humans, that is, they must have cognitive and valence dimensions. Hence it will be necessary to simulate the formation of meaning coordinates to make machines that think. To accomplish this, machines need an embodiment – they must be given digital 'physiologies' that simulate those of humans (Kremelberg, 2019; Fjelland, 2020). This has two important implications. The first is that replicating the brain (Eckersley and Sandberg, 2013; Eth *et al.*, 2013) may not be necessary and the second is that simulating only key physiological processes, including central nervous system ones, might be good enough. For example, on the face of it, it seems unnecessary to simulate central neural control of blood pressure or many aspects related to the activation and control of the immune system.

A machine that can be programmed to have an awareness of things and to experience associated valences will have a simulated f-consciousness. It will then generate meaningful mental states without explicitly being aware that it is doing so and without the ability to refer to itself as experiencing f-consciousness. But is this sufficient? On some level, it probably is. Depending on the objective, the simulation of an array of mental states, thus implying a conscious state, could be enough to provide considerable usefulness. However, a fuller simulation of meaningful mental states as they occur in humans will also need the simulation of self-consciousness.

If digital agents can be constructed in such a manner, a basic framework of meaning can be generated in digital agents that is compatible with that in humans. Success in these simulation efforts will permit the exchange of meaning between agents that generate meaning naturally and those that do so synthetically.

4.1 Algorithms

What then, are the algorithms needed to simulate human mental processes and how can they be implemented – even if by approximation? Below, I briefly discuss the algorithms that I perceive to be necessary to create thinking machines. They are enumerated in an order that might facilitate effective implementation.

- First, there are sensory systems providing input from the external world and the physical body (*cf.* Figure 1). The algorithms needed here depend much on the type of artificial agent one is dealing with. Unless one is dealing with a robot, equipped with an array of sensory systems to emulate those that humans have—touch, taste, smell, sight, and hearing—and an ability to execute behaviors, simplifications will have to be made to compensate for the missing attributes. For example, an artificial agent that is solely computer based, might have sight and hearing and, perhaps, by implication touch (due to temperature monitoring of the hardware). Thus input from other senses such as smell and taste and the execution of behaviors need to be derived vicariously – perhaps from pre-training with text where words such as "smell" could be construed as activating an olfactory system. Alternatively, synthetic senses and a simulated environment, within which they function, could be created.
- Second, sensory input needs to provoke cognition and generate a salience or prominence for that input (*cf.* Figure 1, pathway A). This salience can be based on the intensity, negative entropy, familiarity or another suitable aspect of the input and it produces a quantum for that input.
- Third, sensory inputs need to cause arousal and generate valences (*cf.* Figure 1, pathway B). A general description of arousal and the emergence and role of valences can be found in Appendix C and detailed suggestions on how to implement valence generation can be found in Appendix H. The valence generated in response to a given input produces a quale for that input.
- Fourth, the quantum and quale generated in response to a given input are combined into a meaning coordinate for that input (*cf.* Figure 1). The meaning coordinate is then assigned to a structured memory (*cf.* Figure 2). If a meaning coordinate for that input already exists, it is modified with the new quantum and quale according to some chosen learning rate. All other meaning coordinates decay according to a chosen decay rate. The generation or modification of a meaning coordinate represents the activation of a meaningful mental state and hence phenomenal consciousness in the artificial agent.
- Fifth, linking and binding based on the frequency of joint occurrences (*cf.* Figure 2 and Appendix E) are implemented to generate derivative meaningful mental states – linking meaningful mental states and secondary meaningful mental states. A threshold is set to determine when a linking state precipitates a secondary state; though linking states persist, they can only generate a secondary state once.
- Sixth, supervised machine learning is needed for symbol formation (*cf.* Figures 1 and 2). Within the social context of the artificial agent, there will be a number of symbols frequently used. Starting with a core set of symbols, and gradually expanding that set, an artificial agent is trained to associate a given input (*i.e.*, a given formative object), and its meaning coordinate, with the commonly used symbol for that input. If the input is new, a new meaning coordinate is formed and the agent learns about that formative object and its 'name'. If the input has already been encountered, a meaning coordinate should exist and be in the process of being modified; the agent now learns only the name of the previously encountered formative object. After training, the presentation of a formative object as input should activate its associated meaning coordinate and that should then activate as internal and external symbols the training symbol used for that formative object. The presentation of a given training symbol should do the reverse by activating the meaning coordinate associated with the given formative object (*i.e.*, activating the meaning coordinate of the referent of the symbol). Eventually, through training, symbols

would be assigned to the vast majority of meaning coordinates; an advanced algorithm might incorporate means of self-determined (*i.e.*, unsupervised) symbol generation to accommodate the remainder.

- In the seventh position is the simulation of thought processes as explained earlier (*cf.* Figure 3).
- Eighth, incipient communication. Here a difficulty arises because the trigger for engaging in communication has not been determined. However, the discussion below of autonomy and how behaviors are activated is relevant here. Regardless, what could be communicated is the most recent train of thought, but, at a minimum, the current meaningful mental state. The current meaningful mental state or the most recent thought provides the core set of symbols to be used in communication. They will require algorithms to effect proper symbol sequencing according to the rules of grammar. Algorithms are also needed for desequencing received communication and evocation of the relevant sequence of meaningful mental states in the recipient of a message. Evocation of a meaningful mental state might cause cognitive and valence changes to other meaningful mental states according to some predetermined rules of change. This, in turn, affects meaning coordinates (see the fourth point above).
- Ninth, self-consciousness. Basic ideas for simulation are provided in Appendix G. I suggest to create two parallel systems, to simulate the two hemispheres, one of which would be assigned as the dominant hemisphere. Starting with the output from the thalamus (*cf.* Figure 1), a pathway A and a pathway B, up to and including the formation of internal symbols, are created for the dominant hemisphere and pathways A' and B' for the other hemisphere. The dominant hemisphere alone is extended to produce external symbols. Subtle differences are needed between them, such as learning rates, decay rates, and perhaps rules of thought progression. Furthermore, both hemispheres are enabled to communicate with each other according to methods implemented above for incipient communication. This would permit and effect internal dialogue, though parameters for initiating and terminating internal dialogue would need to be established.
- Finally, autonomy can be implemented. Homeostatic mechanisms and how they activate behaviors through the generation of valences are relevant here (*cf.* Appendix C). Valences provide the key to imparting true autonomy to synthetic agents due to their behavioral guiding functions. They determine whether an agent should cease (displeasure) or continue (pleasure) a behavior. The strength of a given valence relative to those of other valences determines what the agent will pay attention to, and in what temporal order. Therefore, by simulating such valence-generating physiological processes (*cf.* Appendix H), agents can be imbued with a degree of autonomy – they will act based on their current array of valences with the strongest valence most likely determining the course and duration of action. This also relates directly to reinforcement learning (*cf.* Appendix C). It is reasonable to suggest that communication too is initiated, in principle, as a behavior provoked by strong valences.

4.2 Data Types

In humans, valences also reflect whether increasing stress is present or whether stress is abating. Once stress emerges, it causes arousal and enters awareness as a hedonic state, that is, qualia emerge. But this implies that consciousness and meaning are dependent on the particular stresses that occur in human beings and these must necessarily be different from those that occur in

machines. Thirst, a form of stress that occurs in humans due to inadequate water intake, does not occur in machines. And the type of stresses that occur in machines, do not occur in humans. Paraphrasing Fjelland (2020), the real problem with the TT is that computers are not embodied in this world the way humans are. However, by understanding the fundamental mechanisms by which stresses in human systems are transduced into qualia, the processes can be simulated *in silico* without having to replicate the entire brain, an issue considered probable (Eth *et al.*, 2013) and potentially dangerous (Eckersley and Sandberg, 2013). Thus, human-like stress systems can be grafted, in a manner of speaking, onto machines (and possibly onto their native stress systems, for example, related to temperature).

Fundamentally, quantum-quale pairs relate causes with stresses. By defining meaning coordinates as quantum-quale pairs they simulate the human data types used when meaningful mental states occur. Though they hint at complex numbers, it is as yet unclear whether a similar meaning coordinate arithmetic can be formulated that might incorporate rules of derivative mental state formation and rules of thought progression. Furthermore, the symbols used in communication between humans derive from, and activate, meaningful mental states. To communicate with humans, digital agents need to generate meaningful mental states similar to that state that would give rise to a given symbol or to the meaningful mental state that the use of that symbol would provoke in human beings.

Data types used by thinking machines must therefore always include meaning coordinates – a data type with a two-dimensional character.

4.3 Simulations

Is it then necessary or desirable to add more components to the TT? Would not a more efficient route be to simulate those very processes that occur in humans to produce the desired outcomes? More specifically, could a machine be made to simulate a select set of characteristic human mental processes? And, if so, would it constitute thinking? Put differently, could a machine produce virtual thinking; and would that be largely satisfactory? Or should the appropriate focus instead be to create good human/machine interfaces, rather than to create more human-like thinkers? The pressures for human indistinguishability (Linson *et al.*, 2012; Zillich, 2012) may then be alleviated somewhat. I must admit, though, that creating good human/machine interfaces may resurrect the indistinguishability problem, but perhaps with a different emphasis. Machines clearly can do very useful things and the now ubiquitous and extensive human/machine and machine/machine connections demand increased cooperation and coordination between them (Burato *et al.*, 2011). Creating a seamless human/machine interface could immensely amplify their utility and that of the networks wherein they are embedded. Consequently, one could hold the opinion that, if machines can think, they should be allowed to think the way machines do but be able to communicate well with humans.

Regardless, whether one aims to pursue a good human/machine interface or machine thinking indistinguishable from human thinking at some specified level of discrimination, a common requirement is this: a machine has to be able to communicate with humans. Thus, it has to be able to synthesize meaning and deduce meaning. Hence, the task is instantaneously saddled with an anthropocentric burden that leads inexorably back to the need to simulate relevant human mental processes in machines.

But, to simulate these processes, they must be understood. And it is my perception that they are not understood well enough to be simulated. The lack of an understanding of the mental processes underlying the generation and deduction of meaning is one of the main reasons for the

AI dependence on mimicry. The other main reason is that the nature of a datum accessed *in vivo* is different from the nature of the same datum accessed *in silico*. Adding mountains of data and highly sophisticated techniques to process them will not improve the likelihood of producing machine thinking. The focus should not be on more data combined with sophisticated techniques, with the hope that, miraculously, some mental process will emerge from this 'mud'. More can be accomplished, I believe, with fewer data and data in the appropriate form, with the right mechanisms, and with an adequate conceptual understanding of the processes. The aim is to pursue a solid solution, rather than a smart solution (Hausser, 2014). Therefore, the emphasis here is on the basic nature and architecture of meaningful mental processes, the basic nature of the data that they operate on, and how to implement them computationally.

5. Conclusions

Recall that the implication of the TT was that thinking was dependent on performing the right computations. Therefore, devising a suitable algorithm to get a machine to perform the proper computations was all that was needed to generate thinking machines. The search for the right algorithms rested heavily on comparing the algorithms' machine-implemented computational outcomes to human 'computational' outcomes. In the following discussion, it was pointed out that obtaining the same outcomes did not mean that the same processes were executed. Furthermore, using the same processes in machines and humans but on data that differed would produce different outcomes and it would also not reveal whether machines were capable of thinking. Thus, one arrives at the conclusion that the creation of thinking machines requires the generation of human-like meaningful mental processes in machines via the simulation of key processes that occur in humans as well as the simulation of data used by those meaningful mental processes. The processes simulated in machines must then operate on the appropriately simulated data.

In this work, I addressed what I saw as the shortcomings of the Turing test and related attempts to investigate meaning. It led me to focus on a more fundamental analysis of meaning and its generation. This allowed me to provide an outline of what I perceived to be the basic nature of meaning, the architecture and operation of mental processes required to produce meaning and the algorithms needed to implement them. I have also discussed the nature of the data that these mental processes operate on. Taken together, it seems indeed possible to generate thinking machines and with this work I hope to provide a sketch of the way forward.

Acknowledgements

I must acknowledge Lawrence Ward, for introducing me to the topic many years ago, and Robin Turner, for interesting discussions about it. I also want to thank the anonymous reviewers whose valuable comments helped me to improve this work.

Appendix A. Supervised and Unsupervised Connectionist Models

Meaningful mental state formation is based on connectionist principles that are instantiated by connectionist models, also known as artificial neural networks, with specific architectures. Briefly, connectionist models consist of simple computational units or nodes—artificial neurons–

–that are connected in layers. Two of these layers interface with the surrounding world of the network. The first receives input from the surrounding world and the last provides an output to this world. In between are one or more "hidden" layers, the nodes of which perform individual computations on the inputs that they receive, such that collectively they transform a given network input into a network output. Though there are various rules of computation, the most basic one involves the summation of inputs to a node and the transformation of the resulting sum with a step transfer function to produce the output from that node. Nodes in one layer are connected to nodes in one or more other layers in some specific manner that constitutes the network architecture. The connections between nodes have strengths (or "weights") that modify the inputs to the nodes. The rules of computation and the network architecture are generally fixed, but the weights can be varied, and together all these elements account for the network's performance. An accessible introduction to these concepts can be found in Goldwag and Wang (2019).

Because the weights are an important determinant of network performance, but the weights needed for good network performance are unknown, the network is given small weights initially. Upon repeated sequential presentation of all the input data, these weights are then gradually adjusted in a 'learning' process according to certain learning rules. Though most learning is based on stationary data, important advances are also being made with lifelong learning to accommodate real-world data that are non-stationary and temporally correlated (Parisi *et al.*, 2019).

Learning methods fall into one of two broad categories – supervised and unsupervised learning. In the former case, input data are paired with the desired output, for example, when the input image of a dog is presented, the neural network is required to activate an output node that represents "dog" and a different output node that represents "cat" when a cat image is presented. Using a variety of dog and cat images, and terminating training before overtraining occurs (for example, with the use of training, validation, and test image sets), a network can identify with high accuracy a dog or cat image, even if it has not been exposed to that image before. In unsupervised learning, the learning rules permit the automatic partitioning of input images into dog and cat images, that is, without requiring the images to be labeled as "dog" or "cat". Hybrid approaches aim to take advantage of combined supervised and unsupervised methods (*e.g.*, Van Engelen and Hoos, 2020).

Appendix B. Habituation and Arousal

The key contention in this work is that meaning has two orthogonal dimensions. Similar to the central postulate of Schachter's (1964) cognition-arousal theory, that an emotion is a function of cognition and arousal, meaning is considered to be a function of cognition and affect (while emotions might be considered as derivative affective states).

To justify this view of meaning, it is instructive to consider the non-associative learning process called *habituation*. Any sensory stimulus that is highly novel and/or intense can induce a startle response. The stimulus may be important to the well-being of the animal (*i.e.*, salient) and the startle response is often critical to the survival of an animal in its normal environment (Faingold *et al.*, 2014; Ruiz-Salas and De la Casa, 2020). Habituation is the decrease in a behavioral response, such as the startle response, that occurs when an initially novel, repeatedly presented, stimulus turns out to be non-noxious. Thus the stimulus becomes familiar, is now perceived as meaningless, and can be safely ignored in an environment of constant sensory

stimulation (Leader, 2016). This ability to distinguish novel from familiar stimuli, even following a single exposure to a stimulus, and to ignore meaningless ones allows the brain to rapidly encode significant events and is essential for the efficient functioning and survival of the organism (Bonzanni *et al.*, 2019; Leader, 2016; Faingold *et al.*, 2014).

Importantly, the startle magnitude varies depending on the affective state of the animal during the presentation of the startle-inducing stimulus. It is reduced in the presence of positive affect and enhanced in the presence of negative affect (Ruiz-Salas and De la Casa, 2020). I consider the startle response to be affectively negative, thus in the presence of a pre-existing positive affect, the summation of affects produces a reduced negative affect and a reduced startle response. The opposite happens in the presence of a pre-existing negative affect. Because affective conditions always exist in some form, it is likely that the startle response also occurs for low intensity novel stimuli, but that it can be masked by existing positive affect and thus require a normally more intense stimulus to be provoked.

Though habituation seems to occur to non-noxious stimuli perceived as meaningless, I deem a more nuanced situation to exist. This is deduced from Zajonc's (1968) *mere exposure effect* where the mere exposure to an unfamiliar non-noxious stimulus generates an eventual preference for that stimulus compared to an unfamiliar non-noxious stimulus to which no exposure occurred. It seems reasonable to argue that preference in this context involves some form of positive affect. This leaves us in the position where it must be concluded that all stimuli are associated with some form of affect. Unfamiliar non-noxious or neutral stimuli generate an initial negative affect that attain a more positive character as they become familiar through repeated exposure in the habituation process. Noxious and pleasant stimuli are already associated, *a fortiori*, with affect.

The startle response and response habituation are so critical that they already exist prenatally (Leader, 2016). Habituation in the fetus implies that some form of 'recognition' of the stimulus already occurs, in changing from 'foreign' to 'familiar', and the startle response implies that some form of affect is already present. Therefore, the initial startle response to a completely novel stimulus has *ab initio* an affective component and a cognitive component. Clearly, the requisite brain structures must permit this and in subsequent development these structures, as well as the adaptations that they permit, are expanded and elaborated upon. Thus, throughout life, stimuli—the formative objects—simultaneously involve cognition and affect and jointly these imbue them with meaning.

Finally, currently prominent theories of habituation all have an arousal system (Thompson, 2009). It is taken to be the same arousal system discussed in the context of homeostatic systems (*cf.* Appendix C). The habituation arousal system is also considered the same system involved in producing consciousness (*cf.* Appendix G). Thus a novel stimulus involving a startle response is considered to provoke arousal, thus causing the attendant cognitive and affective processes to occur in a conscious state.

Appendix C. Homeostatic Mechanisms and Hedonic States

Hedonic states are critical to the understanding of meaning because they constitute qualia without which meaning cannot exist. Besides being an essential component of meaning, hedonic states serve three important functions: (i) they motivate behavior, (ii) they guide behavior and (iii) they promote learning. An important implication that arises from these three functions is that they can provide an organism or agent with autonomy or self-directedness in the sense that the activation and execution of behaviors arise from internal processes. These internal processes are primarily

homeostatic mechanisms that serve to maintain physiological functioning and integrity and they generate hedonic states to recruit assisting behaviors when necessary (Damasio 1994; Schulze, 1995). Homeostatic concepts are finding increasing AI applications by simulating artificial agents with intrinsic motivation and reinforcement learning-dependent adaptive behaviors (Andersson *et al.*, 2019; McCall *et al.*, 2020; Yoshida, 2020). A brief introduction to homeostatic systems is given here followed by a treatment of the generation and functioning of hedonic states.

Homeostatic Mechanisms

A homeostatic mechanism (HM) is a control system that regulates a variable and a schematic diagram of a HM is shown in Figure C1. A well-known example is a thermostat that regulates the temperature in a room. The level at which the temperature is set to be regulated is provided by the set point (SP) and the current temperature of the room, the current point (CP), is provided by a sensor in the room. The SP and CP are fed into a controller for comparison. If the CP deviates from the SP a regulation error, E, exists that is defined as follows when expressed as a percentage:

$$E = 100\times\left(\frac{CP-SP}{SP}\right) \qquad \text{(C1).}$$

Thus, if the regulation error is negative, the temperature in the room is too low and the controller sends an activation signal to an effector to let hot air in. If the regulation error is positive, the temperature in the room is too high and another effector is likewise activated to let cold air in. In practical applications, however, the temperature is typically allowed to fluctuate a little above and below the SP before activating any of the effectors. The dashed lines above and below the upper border of the shaded portion of the room represent the upper and lower limits between which the temperature is allowed to fluctuate before activating the alarm. It is a visual representation of a room as a container with hot and cold air inlets and a heat sink/outlet. The fraction of the container that is shaded shows the current amount of heat in the container as a proportion of the maximum possible that the container can hold. Depending on outside conditions, heat is inadvertently lost or gained through a vent/window. That is, heat 'leaks' into or out of the room, causing its temperature to increase or decrease in undesirable ways. To counter such changes, the controller either lets cold or hot air into the room to bring the temperature back to a desirable level indicated by the SP. A SP must be selected that is between the minimum and the maximum that the room can hold. The figure shows the current amount of heat in the room corresponding to a SP at 75 % of capacity. Normally, once the SP is specified, the regulation of room temperature proceeds automatically without further intervention being necessary.

In rare cases, however, such as on extremely cold days, the HM is unable to regulate the temperature and additional action, that cannot be provided by the HM, is required to help maintain the room temperature at the desired level. Imagine, for the sake of illustration, that an alarm is then activated to have the occupant of the room close a vent or window to prevent heat from escaping.

Hedonic Sates

Homeostatic mechanisms are essential for regulating body systems to permit survival. We now apply the HM concept to body temperature regulation. Body temperature is regulated accurately around a temperature SP of 36.6 °C and within a very close range of about 0.5 °C to either side, even when outside temperatures may vary considerably. When the body temperature CP is too high, the body is overheating and sweating is induced. When the CP is too low, it is cooling too

much and shivering is triggered. As in the case of the room, body temperature is normally regulated without the need for intervention. However, if neither of these two automatic corrective processes is capable of restoring body temperature to the proper level, *i.e.*, bringing the CP to (near) the same value as the SP, a behavior *must* be activated to help restore body temperature, otherwise survival becomes at risk.

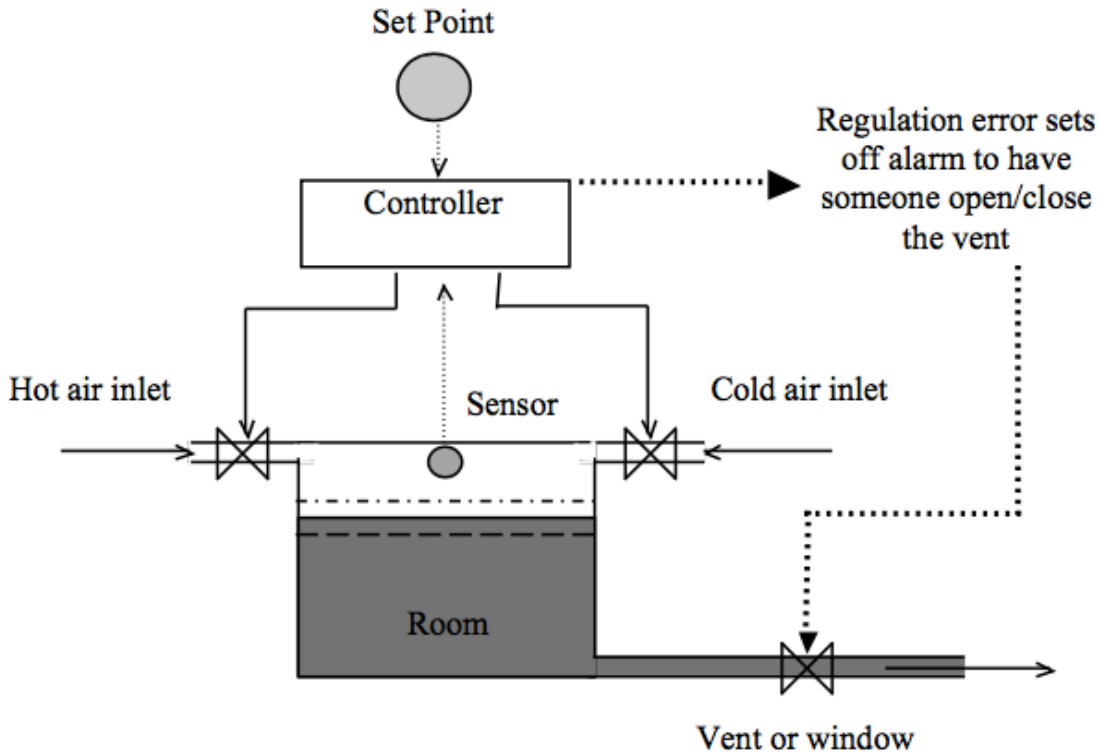

Figure C1
The figure shows a basic homeostatic mechanism regulating the temperature of a room by letting either hot or cold air in to bring the room temperature to the desired level. The desired level is specified by the set point. A sensor detects the current room temperature, referred to as the 'current point'. The controller compares the current point to the set point and activates the hot air inlet effector if the current point is lower or the cold air inlet effector if it is higher. If effector activation fails to restore the room temperature to the set point level and it rises above or drops below certain thresholds (the dashed lines near the upper boundary of the shaded area), an alarm is sounded to recruit external action to help maintain homeostasis. Arrows indicate direction of information or mass-energy flow.

Because regulation errors arise from conditions of physiological stress that require behavioral intervention, arousal is required to make the organism aware of the need to take action. Thus, once the level of stress exceeds a certain, possibly varying, threshold, arousal via the locus coeruleus occurs. The organism now becomes aware of the consequences of the regulation error and can take remedial action. Though this arousal is taken to occur through the same arousal system involved in the startle response (*cf.* Appendix B), it does not necessarily cause one.

Furthermore, due to a regulation error, a hedonic state is generated by the relevant neural circuitry. When the absolute value of a regulation error increases, a negative hedonic state or

displeasure is generated. When the absolute value of a regulation error decreases, a positive hedonic state or pleasure is generated. This implies that a large negative hedonic state can abruptly change into a similarly large positive hedonic state depending on how the regulation error is changing. It is important to realize that, as the executed behavior succeeds in restoring balance, less pleasure is received from repeating the behavior. If the behavior is nevertheless repeated, it will begin to upset the desired balance, but in the opposite direction, and the behavior will induce progressively more distress until it is suspended.

Hedonic states need not relate in a linear manner to regulation error, but are likely to be non-linear as shown in Figure C2 for illustrative purposes. For example, when blood glucose drops below the SP for blood glucose such that a -2 % regulation error occurs, a negative hedonic state with intensity about 40 % of the maximum possible intensity is generated as shown in Figure C2(a). This hedonic state, in conjunction with the sensations present when the body is in this condition, generates the feeling of hunger. In contrast, if the regulation error were +2 %, a negative hedonic state is also generated, but, because the sensations normally are different due to the ingestion of food, a different feeling results. This feeling is one of oversatiation and labeled "Stuffed" in Figure C2 for convenience. Similar arguments can be made about thirst, but note that excess water can be eliminated via increased urine production, so that the ingestion of too much water may not generate as strong a negative hedonic state as a corresponding shortage of water.

The functional forms of hedonic states may vary between persons; thus they may contribute to phenomena such as temperament. For example, a person that had hedonic states with modest slopes, as those depicted in Figure C2(a), will appear to respond to hunger and thirst by consuming larger portions, and more infrequently, than a person that had hedonic states with steeper slopes like those in Figure C2(b). Consequently, the first one will appear to be more laid-back or relaxed and the latter one more high-strung or tense.

The implication here is that different functional forms will be necessary to simulate hedonic state generation induced by different HMs, but, within the family of functions utilized for a given HM, different parameter settings may provide useful ways to generate aspects of 'personality'. This approach, therefore, can be used to generate diversity amongst digital agents in such a manner that they will not only have varying synthetic experiences of qualia, but they will also appear to respond qualitatively differently to identical stimuli. Of course, differences between digital agents will also occur on the cognitive or quantal level by virtue of endowing them with different processing abilities and providing them with different experiences, but the point here is that, even when given identical cognitive abilities and experiences, different 'physiologies' (HMs and their respective hedonic states) will ensure individual variations within 'families' of similar behavioral responses.

Learning and Reinforcement

Hedonic states themselves provide valence and direction due to having intensity and being negative or positive. However, hedonic states need to be transformed into drive states to produce action. Their intensities are converted into drive states with the strength of the drive being proportional to the absolute value of the intensity of the hedonic state.

In the naïve person or animal, learning-informed modification of the drive state is non-existent, thus a pure feed-through of the drive state would be expected. Furthermore, except with reflexes and hard-wired behaviors, the behavioral responses to drive states would be somewhat random and commensurate in vigor only with the strength of the drive. In the experienced person or animal, learning leads to both a modification of the drive state intensity and an appropriate

response selection from the behavioral repertoire. Thus, the drive state depends on at least two types of inputs, one being hedonic state-related, the other being learning-related.

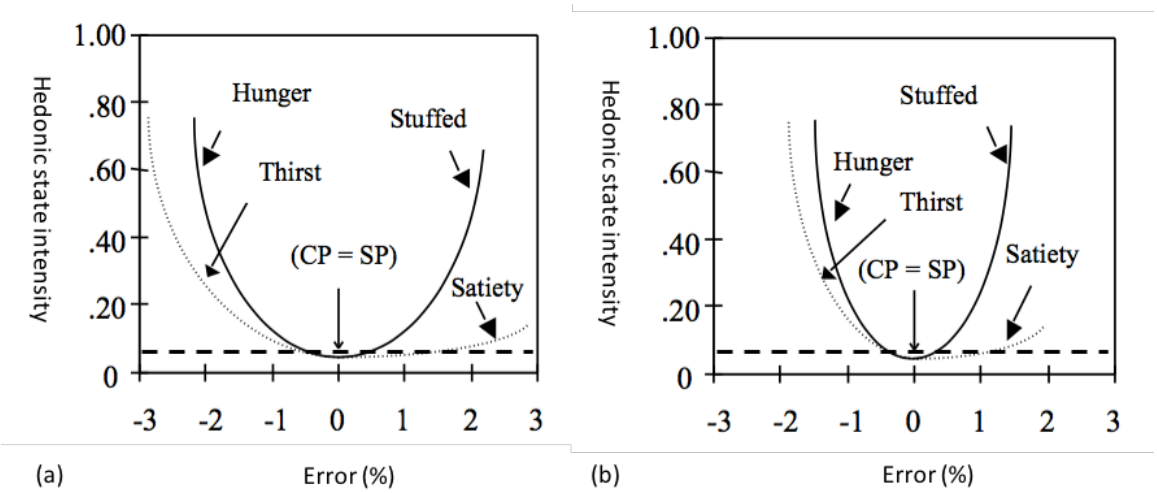

Figure C2

Hedonic state curves show how the intensity of hedonic states change as a function of physiological regulation error. When the current point moves away from (toward) the set point a negative (positive) hedonic state is generated. If the current point is at the set point, the hedonic state is neither positive nor negative, but neutral. Hedonic state curves, shown by way of illustration for hunger and thirst, respectively, suggest that they do not need to be identical or symmetrical around the set point. Hedonic state curves can also differ from person to person. In (a),'flatter' hedonic state curves are shown; therefore, the hedonic state intensities change more slowly than in (b). A person with flatter curves will react more slowly to hunger and thirst and this person will appear to be more relaxed. In contrast, the person with 'steeper' hedonic state curves will be more reactive and appear more high-strung. Dashed lines indicate putative thresholds above which arousal cause hedonic states to enter awareness.

Learning is reinforced by pleasure and pain that act as rewards and punishments, respectively. This process was called 'operant conditioning' by B. F. Skinner and leads to 'operant behavior' – behavior controlled by its consequences (Staddon and Cerutti, 2003). Indeed, learning improves phenomenally when occurring in the context of hedonic consequences (*e.g.*, Sgro *et al.*, 1967; Wagner, 1961; Wilkenfield *et al.*, 1992) and reinforcement learning algorithms have been successfully implemented in machine learning applications (*e.g.*, Ertel, 2017; Sutton and Barto, 1998; Szepesvári, 2010). The consequence is that learning results in behaviors that are highly effective and efficient and provide improved long-run homeostasis and thus enhanced survival and fitness. Importantly, because the reinforcer constitutes a quale, the combination of quale and quantum into meaning occurs and promotes learning – something that has meaning will be acquired much faster than something that has no meaning. One might call the first 'learning' and the last 'memorization'.

Independent Agency

A robot could be made and programmed to perform certain tasks when given instructions to do so. But why would it *want* to place a green cone on a red cube? Even though it is capable of

recognizing the objects around it and can manipulate them correctly, neither intention nor autonomy can be ascribed to the robot. Intentional agency seems to enhance the perception of intelligence (Linson *et al.*, 2012) and autonomy is considered a crucial feature of intelligence (Müller, 2012). Therefore, the robot will be perceived as having limited intelligence, if any. Fortunately, hedonic states can also provide insight into intentional agency and autonomy. Indeed, hedonically driven actions in variable environments can be viewed as producing intelligence as an emergent phenomenon (e.g., Hutter, 2012).

Consider, for the sake of argument, a simple agent with two physiological HMs as depicted in Figure C3. One homeostatic mechanism is constructed to regulate body temperature and another to regulate blood glucose level. There are two set points, one for body temperature and another for blood glucose level, and two current points, one for body temperature and the other for blood glucose level. When the CPs deviate from their respective SPs, regulation errors are present. Hedonic states are then generated that vary in intensity commensurate with the degree of deviation, thus, with the size of the regulation error. As these systems are generally independent from each other, these deviations and their resulting hedonic states can also vary independently (Schulze, 2003; Schulze and Mariano, 2003). Sufficiently large regulation errors arouse the organism and activate behaviors. In the experienced agent, *i.e.*, one that has engaged in successful learning, the activated behaviors are behaviors related to restoring body temperature (*e.g.*, moving into the sun or moving into the shade) and behaviors related to restoring blood glucose level (*e.g.*, eating). In short form, if the temperature CP deviates from the temperature SP, but the blood glucose level CP does not deviate from the blood glucose level SP, temperature-related behaviors will be activated and *vice versa*. If a regulation error is present in both systems, giving rise to competing behavioral options, the system that has the largest regulation error, and hence generates the strongest hedonic state, will have precedence. In an even more sophisticated agent, arbitration could furthermore include expected hedonic states based on the agent's predictive abilities. Behavioral selection, therefore, is also dependent on hedonic states and is arbitrated in favor of the most intense state. Therefore, one could reasonably ascribe the selection of a specific behavior as intending to reduce the most disagreeable hedonic state. Furthermore, both the activation of a behavior and the selection of a specific behavior are internally generated – the agent is autonomous and goal-directed, both in terms of seeking specific objects such as shade or food and in terms of executing specific behaviors.

Further Discussion

Hedonic states are, loosely speaking, subjective experiences of changing levels of physiological stress. Persistent physiological stress is harmful. Stress, via hedonic states, activates behaviors through the generation of drive states to reduce stress. Thus, hedonic states also provide the motivating forces to act. Intelligence, or more broadly, cognitive processes, then serve to shape those behaviors to be more effective and more efficient. The association of hedonic states with external cues that signify conditions that induce or relieve such stresses (*e.g.*, classical and operant conditioning) causes these external cues themselves to become capable of inducing or relieving stress. As this stress is not directly, but merely by association, of physiological origin, it constitutes the psychological induction or relief of stress that is generated by external cues. Seen from this perspective, behaviors generally serve to reduce or avoid stress and cognitive abilities function to improve behavior to avoid harmful physiological stress or to procure the relief of such stress more effectively. Without hedonic states, there is no effective inherent teaching signal. Hedonic states, therefore, form part of an adaptive process that promotes survival.

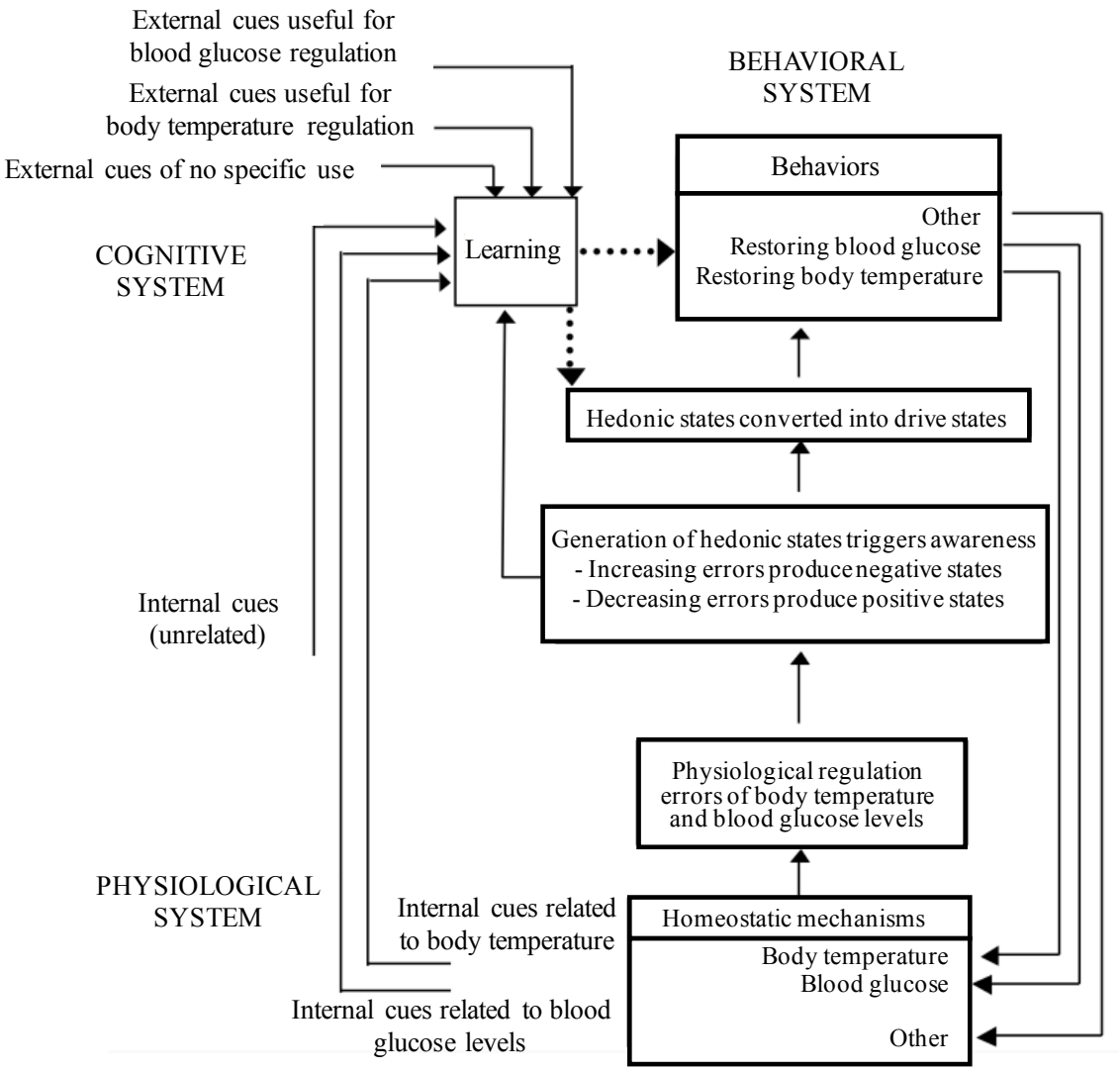

Figure C3

Learning is a cognitive process where associations are formed between external sensations, internal sensations, and hedonic states. It modifies behaviors so that they become more effective and efficient. As it generally takes time for associations to form and strengthen, their modifying effects on behaviors emerge gradually. An artificial agent equipped with homeostatic mechanisms will exhibit both autonomous and intentional behavior. When a homeostatic mechanism is unable to regulate a given variable within its required range, a hedonic state is generated that permits one or more behaviors to aid in regulation. Autonomy thus ensues because behaviors result from internally arising imperatives. Furthermore, because inappropriate behaviors will aggravate negative hedonic states, hedonically guided behaviors, enhanced by learning, will be goal-directed and appear to be intentional (from Schulze, 2003).

Because hedonic states are closely associated with the induction and execution of behavior, they are closely associated with our experiences. Figure C3 makes plain that every external cue,

i.e., every stimulus, by virtue of learning, can become associated with a hedonic state, hence can take on a valence. The stronger the hedonic state, the stronger the valence potentially associated with a stimulus. Through such associations, the state of the external world becomes linked to the state of the self. This link provides external cues with meaning. Thus, meaning has value if there is a valence, and the more intense the valence, the greater the meaning. Meaning has direction and this is provided by positive and negative valences, originally imbued by positive and negative hedonic states, by pain and pleasure. Meaning has context, primarily between the self and the environment. Things have reduced meaning if they cannot be related to the self and its hedonic experiences. They can be learnt, like nonsense syllables, or read, like the technical language of an abstruse subject, but remain devoid of meaning if no valences can be assigned to their cognitively perceived relationships. Furthermore, through higher-order conditioning, external cues without meaning, *i.e.*, valence-neutral external cues, can also take on valences. As a result, a vast number of external cues can take on valences, some weak, some intense, and many in-between.

Hedonic states also provide insight into the nature of independent intentional agency. Because they arise essentially in response to the waxing and waning of physiological stresses, and because physiological stresses are dependent on one's physiological makeup and how it responds to internal and external stressors, and because one's physiological makeup is a given and not dependent on the instructions, consent, or desires of other agents, primary hedonic states arise internally and independently of others. Secondary hedonic states arise due to the presence of external cues and the valences that those cues have been associated with and can modify the effects of primary hedonic states on the drive state to produce a 'net' drive state that would then trigger a behavior. A strong secondary hedonic state can oppose or override a primary hedonic state to produce a null drive state or a drive state leading to a different behavior than would have occurred in its absence. As the drive state determines the action to be executed, behavior in the absence of secondary hedonic states will appear goal-directed because repeated and predictable patterns between external conditions and behaviors will emerge. In the presence of secondary hedonic states behavior will still appear to be goal-directed, but more complex.

Appendix D. Putative Brain Structures Subserving Meaningful Mental States

Brain structures postulated to subserve the generation of meaningful mental states are shown in Figure D1. Figure D1 integrates concepts from Figure 1 (meaningful mental states) and Figure C3 (homeostatic systems), though represented in a necessarily much simplified form.

Starting with the sensory systems, all peripheral sensations are transmitted to the thalamus. The thalamus is a structure of major importance that distributes sensory inputs to somatosensory and other cortical structures, including the parietal cortex, for cognitive processing (Grant *et al.*, 2012). It also has major projections to the limbic system where affective states are generated (Namburi *et al.*, 2016). Bridging nuclei interface these cognition and valence related parts of the thalamus (Vertes *et al.*, 2015a,b).

Through activation of the limbic system, directly from the thalamus (Namburi *et al.*, 2016; Sah *et al.*, 2003) or via the locus coeruleus (Chandler *et al.*, 2019), affective states for all stimuli are generated or present for association. They are generated *de novo* for unfamiliar or unexpected stimuli (*cf.* Appendix B) or, for other stimuli, via physiological homeostatic systems that regulate a variety of physiological variables essential for survival (*cf.* Appendix C). Brainstem and hypothalamic structures, such as the nucleus of the tractus solitarius and the paraventricular nucleus, often function as controllers of physiological homeostatic systems (Schulze, 2003).

Through connections with the locus coeruleus, they activate the latter when homeostatic dysregulation occurs. By activating the locus coeruleus, causing subsequent irrigation of several brain structures with norepinephrine (Aston-Jones and Cohen, 2005), physiological stresses produce arousal. Arousal through the locus coeruleus is also produced by the startle response following unfamiliar or unexpected stimuli. In the latter case, activation is likely to occur through projections from the thalamus to the amygdala (Sah *et al.*, 2003) and from there to the locus coeruleus (Sah *et al.*, 2003; Chandler *et al.*, 2019) as evidence of direct efferents from the thalamus to the locus coeruleus is currently absent.

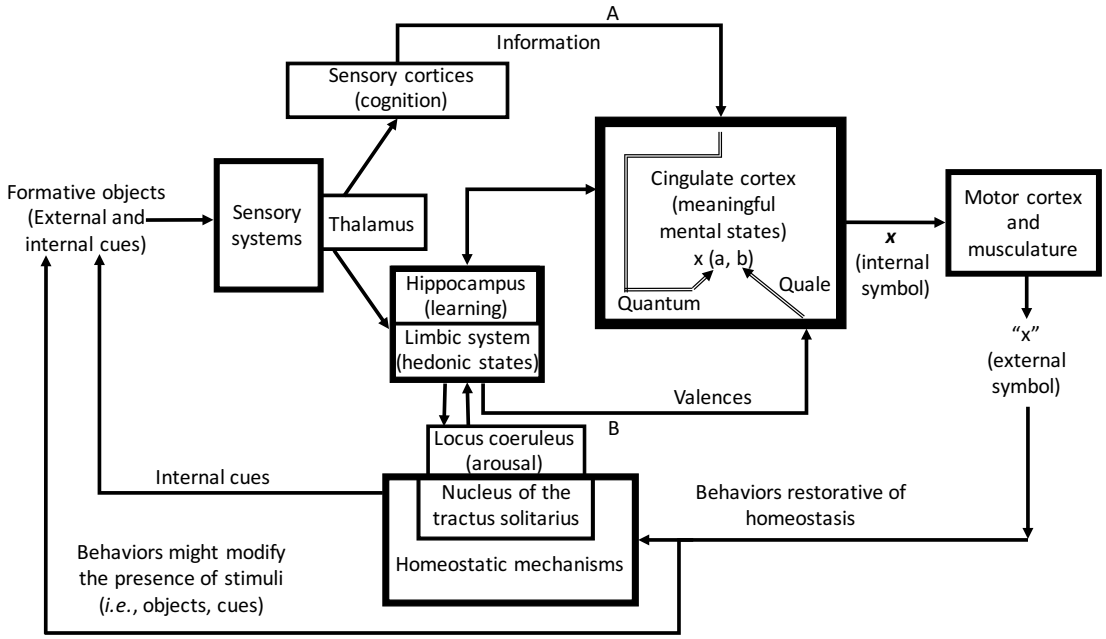

Figure D1
Streams of sensory input via sensory systems are distributed by the thalamus to sensory cortices where it is highly processed producing 'information' that is relayed to the cingulate cortex. Physiological status regulation by homeostatic mechanisms are appraised by the limbic system as hedonic or affective states providing valence streams also relayed to the cingulate cortex. These streams permit the combination of quantitative cognitive information with qualitative valences into meaningful mental states and enable action-outcome learning executed via connections with motor systems. See text for more detail.

Relative to other primates, the posterior parietal cortex has expanded significantly in humans permitting highly processed visual and somatosensory information to guide behaviors and tool making that can be executed via strong connections with the motor cortex (Kaas *et al.*, 2018). These highly processed visual, spatial, somatosensory and auditory cognitive streams that convergence in the parietal cortex are received by the cingulate cortex; it also receives valence inputs from the limbic system (Vogt, 2016). The cognitive and valence inputs are received in an affective-cognitive gradient going from anterior towards posterior cingulate cortex (van Heukelum *et al.*, 2020). This permits the combination of quantitative cognitive information with qualitative valences to enable action-outcome learning (Rolls, 2019), thus the generation of operant behavior (*cf.* Appendix C). Learnt action-outcomes can be committed to memory through

connections between the cingulate cortex and the hippocampus (Rolls, 2019; van Heukelum *et al.*, 2020) and behaviors can be implemented via connections between the midcingulate cortex motor area and motor cortical areas (Rolls, 2019; van Heukelum *et al.*, 2020). Parietal and cingulate cortices are also irrigated by norepinephrine (Aston-Jones and Cohen, 2005; Caminiti *et al.*, 2015) and the level of arousal so provoked modulates their performance (Aston-Jones and Cohen, 2005).

Parietal and cingulate cortices are also involved in language production. Human functional imaging studies implicate the inferior parietal cortex specifically in representational aspects of semantic memory while emotional aspects involve limbic areas and posterior cingulate regions (Binder and Desai, 2011). Though the cingulate gyrus' precise linguistic function remains unclear, the cognitive inputs from the parietal cortex and its demonstrated emotional involvement suggest that these aspects are combined in the cingulate gyrus. The anterior cingulate cortex has also been found to be involved in goal-driven language selection, a high-level control process over the selection of words intended to convey a specific concept (Faulkner and Wilshire, 2020). Furthermore, the strong reciprocal connections of the posterior cingulate gyrus with the hippocampus, essential in memory formation including language-related declarative memory (Tulving and Markowitsch, 1998), also permits the encoding of cognitive and affective events in declarative memory (Binder and Desai, 2011).

Taken together, meaningful mental states are postulated to arise in the cingulate gyrus, through the combination of cognitive information from the parietal cortex and valences from the limbic system. Because these brain structures are organized into separate cognitive and affective streams that ultimately converge prior to output structures, this organization implies that cognition and affect are not serially related and must be simulated in parallel.

Appendix E. Binding

It is postulated that, once a meaningful mental state is established, other meaningful mental states can be combined with it through a process known as *binding*. Binding causes separate meaningful mental states established by familiar stimuli, when they arise in stable contexts, to become unitized into chunks. When unitized, chunks themselves begin to function as individual meaningful mental states.

In word recall, temporal associations have been shown especially important in producing binding. These associations have a forward bias, the *contiguity effect*, such that a better association occurs between a word and the word following than between the same word and the one preceding it (Kahana *et al.*, 2008). Though temporal associations naturally produce stimulus grouping, and the forward bias is suggestive of a built-in alertness to potential causality, binding can also be facilitated by intentional grouping, for example, through stress patterning (Reeves *et al.*, 2000). In contrast, midazolam, a drug that creates temporary anterograde amnesia, disrupts binding and impedes recognition for unitized memories (Reder *et al.*, 2006).

Binding is also considered essential to the formation of episodic memories where the simultaneous recall of various aspects of the experience—time, place, feelings, etc.—are necessary. The parietal cortex, where features stored in disparate neocortical regions converge, is considered to be the cortical site that binds these relational aspects of experiences (Shimamura, 2011; *cf.* also Appendix D). Note that the position here is nuanced differently in that binding between cognitive aspects of experience is considered to occur in the parietal cortex, but the association between cognitive and valence aspects deemed to occur in the cingulate cortex.

Appendix F. Communication

Language production is a process, originating from the currently salient meaningful mental states of the speaker, whereby recognized and grounded symbols are used in recognizable patterns to provoke meaningful mental states in the hearer.

It is perhaps instructive to consider the use of language in the context of the stimulus-response patterns so familiar to psychologists. Take, for example, the eyeblink response. To study this response, often in the framework of a conditioning experiment, it is elicited in animals such as rabbits by applying a gentle puff of air to the eyeball that causes the nictitating membrane to close rapidly. The same response can be elicited in two adult individuals from the same species by applying a similar puff of air to an eyeball of each. As one might expect, such individuals will have similar sensory and motor mechanisms, subserved by similar neural circuitries (Christian and Thompson, 2003), causing near-identical stimuli to elicit near-identical responses. In the above case, however, the puff is produced by the experimenter. However, there is no reason to suppose that the same gentle puff of air cannot be produced by one rabbit and directed at the other, causing the latter to blink. Equally reasonably, their roles could be reversed. When a conspecific produces a stimulus directed at another, it is often termed a 'signal' and it is known that such signals and their responses appear to be under common genetic control (Camhi 1984). Because they evolve in tandem, the signal and its 'meaning' stay correlated for as long as the correlated pair have benefits to the species in a given environment. This suggests that one could view the production of a sentence as a 'stimulus' directed at a recipient causing a specific 'response' by the recipient.

But, because the digital agent and its human interlocutor are not under common genetic control, the signaling and response mechanisms in the digital agent must be simulated so that the same signals can be generated by digital agents and can elicit responses in digital agents that are similar to those elicited in humans. However, even though human signal production and signal response might be genetically linked, the specific signals produced could vary as attested to by the existence of separate languages. As an aside and though controversial (*e.g.*, Haspelmath, 2020), it is interesting to reconsider Chomsky's universal grammar in the context of a genetically linked signal production and signal response process.

Furthermore, teaching plays an important role in language acquisition by children – this ensures that symbols, the signals produced, are grounded in a consistent manner. Likewise, digital agents could be trained on the same set of symbols through supervised learning to effect communication with humans and other digital agents.

Based on this stimulus-response view, one can now separate the communication process into the following components:

(i) Machines must have physiologies simulated to be similar, to the extent relevant, to those of humans. This creates similar internal states to those in humans in response to similar external and internal stimuli. Consequently, they will 'perceive' and 'experience' stimuli in analogous ways.

(ii) They must generate, through these simulations, similar meaningful mental states in response to similar internal states the way humans do.

(iii) They must be able to access and evaluate their meaningful mental states like humans do.

(iv) They must have a similar lexicon with contents that are similarly grounded to those of humans. Grounding must be dynamic in ways similar to that in humans.

(v) They must select words, based on their meaningful mental states, in a way similar to the way in which humans do the same.
(vi) They must concatenate words into sentences the same way humans do to encode these meaningful mental states.
(vii) They must be able to parse sentences the same way humans do.
(viii) From the parsing, they must be able to induce the same meaningful mental states, grounded to the same referents, in the same way that humans do.

In summary, when two individuals deploy the same symbol derived from the same formative object, that symbol is grounded to the same referent and they can communicate. When received by the hearer, meaning is reconstructed when the symbols of language, by virtue of being grounded, are connected with meaning coordinate pairs to generate a meaningful mental state. Because of the similarity of these processes in both speaker and hearer, the speaker can use language to generate a specific meaningful mental state in the hearer. Communication, therefore, contains information, in the form of external symbols, but no meaning. Thus, the meaning conveyed by language does not arrive along with the words, but arises due to the subsequent reconstruction prompted by the words.

On the above view, the success of communication depends on the extent that an intended sequence of meaningful mental states can be induced in the recipient of the message. However, because symbols can be ambiguous, the same symbol can have different referents or different aspects of the same referent. Moreover, individuals have different histories, thus different experiences, and their meaningful mental states provoked by the same stimuli or evoked by the same symbols, will differ. Taken together, these variations will cause the meaningful mental states induced in the recipient of a message to deviate, to lesser or greater extent, from the meaningful mental states leading to the production of the message. Clearly, such variations can lead to misunderstandings, but can also be exploited. Grice (1989) considers three ways in which an utterance can be interpreted: (i) as normally understood, (ii) as implied by the speaker in a conventional sense, and (iii) as implied in a nonconventional sense. Thus, two basic notions of meaning result – what a sentence means in a generic sense and what a specific speaker intends it to mean by using the sentence on a particular occasion (Grandy and Warner, 2020). What a sentence means in a generic sense will be closely related to the teaching and supervised learning aspects of language and what a specific speaker intends it to mean could also draw on the known meaningful mental states of the producer and those anticipated to exist or occur in the recipient.

Implementing human-computer interactions via natural language will benefit from a theory of language that explains natural communication in a way that is functionally coherent and mathematically explicit (Hausser, 2014). Despite our long history of language use, an explicit understanding of what language is, seems to be lacking. The view taken here, especially in terms of the recreation of meaning, is rather similar to Hausser's (2014) definition of successful natural language communication where a digital agent correctly recreates the speaker meaning from the uttered language. In my view, however, successful natural language communication is furthermore dependent on meaningful mental states to explain the selection and intentions of the speaker and the effects of utterances on the hearer. The representation of meaningful mental state contents in terms of meaning coordinates might permit the definition of language in a manner that addresses the need for a mechanistic formulation. Thus, by explaining the meaning of sentences in terms of the mental states of language users, it is mentalist in nature (Speaks, 2014).

Appendix G. Consciousness and Self-Consciousness

Human consciousness is medically considered to consist of arousal and awareness (Edlow *et al*., 2012). Arousal occurs when locus coeruleus neurons activate brain circuits with the release of norepinephrine and readies them for activity. These neurons are quiet or fire slowly during sleep or when drowsy and so facilitate disengagement from the environment, they promote attention when firing at moderate levels, and they produce rapid scanning and increased sensitivity to the surroundings when firing at high levels (Aston-Jones and Cohen, 2005). Arousal, in turn, activates awareness networks in the cortex. Awareness is not possible without arousal as shown by coma in the presence of brainstem lesions but intact cortices (*cf*. Edlow *et al*., 2012).

There are two main ways, and a possible third, in which arousal seems to be provoked – one is related to habituation and the perception of unexpected stimuli and the other to physiological dysregulation. A further possibility might arise exclusively from cognitive processes. At the onset of habituation, a new or unexpected stimulus activates arousal. Though arousal is high, a memory of the stimulus is absent or weak. Then, as habituation proceeds, stimulus awareness causes a strengthening of memory and a reduction of arousal. Eventually, habituation to a given stimulus results in a memory of the new stimulus and a loss of arousal along with the reduced response to the stimulus (*cf*. Thompson, 2009). Regulation errors cause conditions of physiological stress that require supportive behavioral intervention (Schulze, 1995; Schulze, 2003; Schulze and Mariano, 2003). Without awareness of the regulation error and its nature, the needed behaviors cannot be instigated. Such awareness occurs when the physiological stresses produced by regulation errors cause arousal through activation of the locus coeruleus with subsequent noradrenergic irrigation of cortical structures (Schulze, 2003).

In all of these cases, arousal occurs in response to an activating event – a new stimulus, physiological stress, or a cognitive process. Because arousal has been provoked by something, and arousal activated awareness cortical networks, a conscious state will be a state of consciousness of something. Thus, a meaningful mental state, because it is postulated to occur in the cingulate cortex, arises along with consciousness. Furthermore, this consciousness is phenomenal consciousness because arousal permits awareness of stimuli and feelings.

Self-consciousness (s-consciousness), a self-referential phenomenal consciousness, might be explained by the existence of two brain hemispheres. The two hemispheres essentially constitute two near-identical functional 'brains' lodged in the same skull and connected to the same body. As they travel through the world, they occupy almost exactly the same space at the same time, resulting in near-identical experiences. However, the two hemispheres do differ in one important respect – one of them is dominant, a phenomenon referred to as 'laterality'. Furthermore, the brain hemispheres and the body are contralaterally connected – the right hemisphere is connected to the left side of the body and vice versa. Because a given hemisphere controls the contralateral side of the body, the side of the body controlled by the dominant hemisphere will generally be the dominant side, thus have the dominant limbs and sensory organs (*e.g.*, the right hand being dominant compared to the left hand). Hence, a solution to s-consciousness may reside in the phenomenon of laterality by allowing one hemisphere of the brain to perceive and refer to the other hemisphere as if it were an object, a situation possibly facilitated by contralateral control. Because the observed object is nearly identical to the one doing the observation, the observation process is akin to the observing hemisphere observing itself rather than observing something else.

A further important role may be played by mirror neurons. Mirror neurons are motor neurons that become active when observing someone else performing some action even if one is inactive oneself (*e.g.*, Oberman *et al*., 2007). Because mirror neurons are not confined to one hemisphere

(*e.g.*, Filimon *et al*., 2007), both hemispheres have the ability to observe each the other in action. For example, the dominant hemisphere steering the dominant hand is observed by the non-dominant hemisphere wherein mirror motor neurons of the non-dominant hand become activated. The non-dominant hemisphere therefore, while observing the actions of the dominant hemisphere from a virtually identical point of view, experiences via its mirror neurons the behavior of the dominant hemisphere. Moreover, sensory input received and valences generated by the dominant hemisphere while performing its activity are shared with the non-dominant hemisphere along fibers crossing from one hemisphere to the other through the corpus callosum.

Because both hemispheres reside in the same body, experience the same 'macro' conditions of nutritional state, fatigue, hormone levels, temperature, and so on, quanta and qualia from the observed 'partner' (*i.e*., the observed hemisphere with its contralateral part of the body) can be compared to, and tightly integrated with, quanta and qualia from the observing partner – solidifying the sense of self while preserving the ability to refer to it objectively. Indeed, because of this close confinement and integration of two computationally independent hemispheres, one partner could validly refer to the other partner as 'I' and vice versa. Duplication of the hemispheres allow self-reference and laterality allow for the initiation and control of the process. Although the reader may note semblances to the concept of the bicameral mind, Jaynes did not try to resolve the nature of the relationship of the 'I' to the 'me', but merely wished to indicate the nature of the problem (Jaynes, 1976).

Finally, though not certain, it is of interest to consider that the two hemispheres could explain internal dialogue. Language is not predominantly confined to the left hemisphere. The cortical language network develops bottom-up processing bilaterally in the temporal cortices during the first 3 years of life. But, in the years thereafter, increased functional selectivity and structural connectivity of the left inferior frontal cortex occur as top-down processes emerge gradually (Skeide and Friederici, 2016).

Appendix H. Simulating Valences

Imitating human behavior is gimmickry and an AI dead-end. It will ultimately not lead to breakthroughs in understanding the true nature of thinking and the associated problems of generating and decoding meaning. The Turing test is unfortunate in that it promotes imitation (copying surface structure/surface function), rather than simulation (copying deep structure/ deep function). Thus, we need to shift the focus from imitation to simulation, but to do that, we need to know what to simulate. The main text already contains an enumeration of the type of algorithms deemed necessary to simulate meaning and communication. Here, I want to focus on simulations pertaining to a critical aspect of meaning – the generation of valences. Appendix C is also of particular relevance here.

To simulate human behavior, a digital agent should be provided with a rudimental digital physiology that simulates key physiological regulation processes as they occur in humans. Next, a module needs to be created that simulates hedonic states such that the regulation processes can, on occasion, activate this module to produce hedonic states. "On occasion" does not imply randomly, but in a manner that simulates how and when regulation errors in humans lead to the generation of hedonic states. Digitally generated hedonic states then should be able to produce a drive state as well as provide a hedonic state teaching signal in the form of a valence to a cognitive module. The cognitive module should be able to associate valences with external inputs and store such associations along with associations between different external and internal cues

themselves. The presence of a sufficiently strong drive state should be able to activate a behavioral module as well as provide input to the cognitive module where the selection of a particular behavior occurs. Finally, the simulated physiologies should be sensitive to external conditions such that behaviors, when they change the degree and manner in which external factors impinge on a digital agent, can influence their regulation processes. On the face of it, this may seem to require the creation of robots with sensory systems such as taste, touch, smell, vision and hearing (and others), but a crude approximation may be possible by having a digital agent experience the world lexically – through the keyboard.

A rudimentary physiology can be simulated by approximating fluctuations in current point (CP) values for different homeostatic mechanisms (HMs). For example, the adult energy requirements are about 2000 kcal/day of which, say, 1600 kcal are used during 16 h of wakefulness. Thus, normal waking state metabolic processes require enough glucose to provide about 100 kcal/h. Assuming that one starts with a 'normal' amount of blood sugar, there is enough reserve to last approximately 3 hours before the CP starts to deviate from the set point (SP). Furthermore, the manner of deviation depends on how other processes, such as the conversion of fat to glucose, are activated.

The upshot is that CP fluctuations need to be modeled with some appropriate functional form. By implication, a different and appropriate functional form needs to be utilized for every CP. The CP often involves ligand binding to a receptor, for example, glucose binding to a hypothalamic receptor. Because these processes generally have logarithmic or logistic functional forms (*e.g.*, Tinberg *et al*., 2013), such functions are good candidates for modeling basic CP fluctuations.

Blood glucose (*i.e.*, the CP) fluctuates over the course of a day. Levels spike immediately after a meal, then drop fairly quickly to a plateau where they remain for some period of time before gradually declining to fasting levels (*e.g.*, Daly *et al*., 1998). The blood glucose CP fluctuations shown in Figure H1 can be simulated with a function using two hyperbolic tangent terms:

$$CP_{bg} = 100 + 25\times\left(\frac{-\tanh\big(0.1\times(time-40)\big)+1}{2} + \frac{-\tanh\big(0.01\times(time-340)\big)-1}{2}\right)\ mg/dL \tag{H1}$$

with *time* being the number of minutes since the last meal.

The high values represent blood glucose levels immediately postprandially (around 125 mg/dL), the plateau values are those between meals (about 100 mg/dL) and the low values represent fasting glucose levels (about 75 mg/dl). Note that 100 mg/dL appears to be the 'preferred' level. If the CP is above this level, insulin secretion during and shortly after a meal brings it back down to 100 mg/dL. If it falls below this level, it eventually triggers eating to bring the CP back up. Hence, the "100" in Eq. H1 can be replaced with the blood glucose level SP. The system can be initialized to start after a 'primordial' meal.

Function values are then further affected by behavioral consequences leading to modified values of the CPs. It stands to reason that behavioral consequences need to be simulated to have an effect on the respective CPs. Thus 'eating' needs to be simulated as this affects blood glucose levels, blood saline levels, blood pressure, blood lipid levels, body temperature, water balance, etc. The question is, in the absence of a physiology, how are 'eating' and its physiological consequences to be simulated? The only answer I have here is that it has to be done vicariously through the keyboard. Thus, matching keyboard input to keywords like 'food', 'potatoes', 'vegetables', etc. should be the virtual equivalent to eating. It is important to realize that this is a

very basic process, thus only matching of strings occur, rather than an understanding of specific words. The process is analogous to that of a baby consuming milk—it does not need to understand the meaning of 'milk' in order to consume it—it merely needs to match the presence of milk to its desire to consume nourishment.

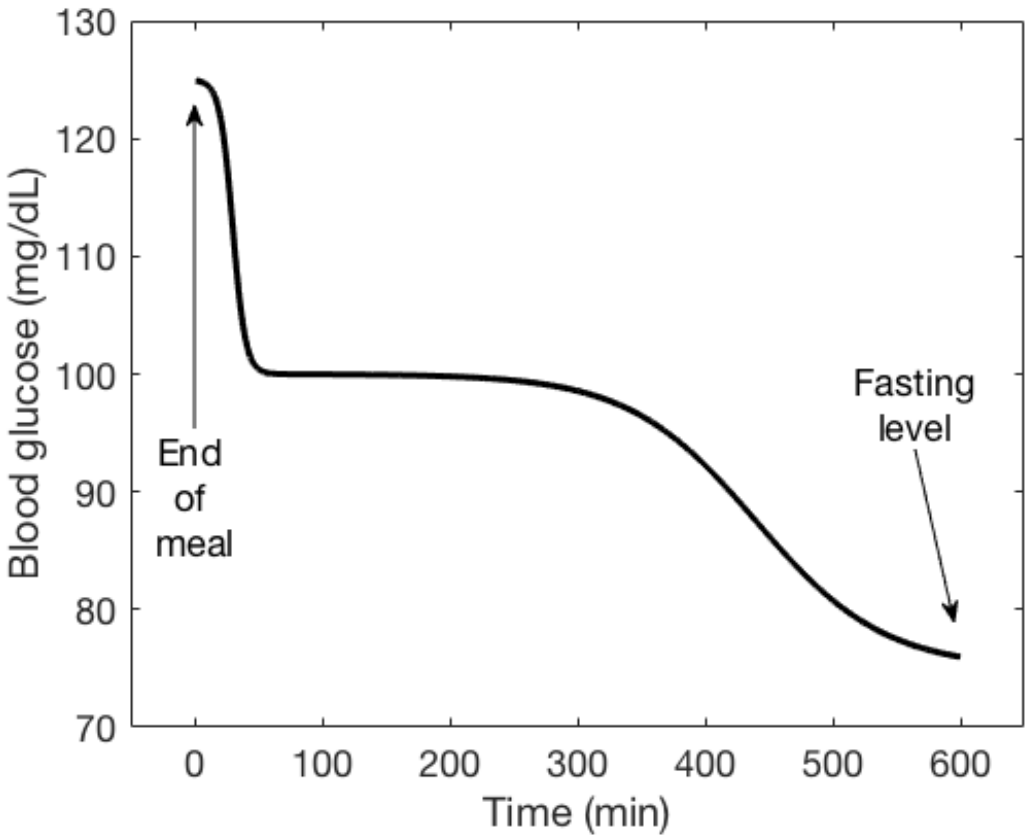

Figure H1
Fluctuations in blood glucose levels after a meal are simulated with a function containing two hyperbolic tangent terms. High levels occur immediately after a meal and low levels are fasting levels.

Inferences that the digital agent is eating are drawn from text containing these keywords and a small increase in blood sugar should occur commensurate with each match. In practice, this means that the CP is adjusted to the corresponding blood glucose level and allowed to progress from there. For example, every time a food word is matched, the CP is adjusted by augmenting is current level with 5 % of the difference between maximum and minimum blood glucose levels, *i.e.*, 125 mg/dL and 75 mg/dL. As the difference is 50 mg/dL, the adjustment amounts to 2.5 mg/dL. To keep this crude simulation tractable, eating is not allowed to elevate blood glucose levels above 125 mg/dL. In effect, this adjustment amounts to resetting the independent variable to that point where the function value corresponds to the current CP plus the addition(s). After the adjustment, the CP is again allowed to change as a function of time according to Figure H1 and Eq. H1. Because behaviors affect blood glucose levels, Eq. H1. is modified to accommodate them:

$$CP_{bg} = \min\left(\max(blood\ glucose + external\ effect, 75), 125\right.$$
$$= \min\left(\max(blood\ glucose + 0.025 \times number\ of\ matches, 75), 125\right. \qquad \text{(H2)}.$$

Next, the internal milieu needs to be simulated. Though a good simulation might proceed along very different lines, a simple sketch is given here to provide a point of departure. For example, when the blood glucose level drops to fasting levels, internal sensations arise that are, in aggregate, labeled and experienced as 'hunger'. These internal sensations can be generated depending on the value of the blood glucose CP and a word list of such sensations updated periodically. This list contains words such as 'hunger pangs', 'growling stomach', 'empty stomach', 'hunger', etc. and each of these has a certain probability to be generated for a given value of the CP. For example, with a CP just below100 mg/dL, there is a 10 % probability of generating the internal sensation 'hunger' (*i.e.*, posting the term 'hungry' to the current internal sensation word list). At 95 mg/dL, the probability increases to 30 %, at 90 mg/dL to 50 %, and at 85 mg/dL to 75 %; thereafter the probability declines because the probability of generating the internal sensation 'growling stomach' instead is now increasing.

The internal sensations, in the form of the internal sensation word list, are then available to supervised cognitive modules to enable associative learning between internal states, external cues, and hedonic states. Once again, the approach suggested here is to make a crude approximation of the outcomes of the real physiological processes. Thus, instead of using a word list, it would be better to simulate internal states directly which can be labeled via supervised learning, the way a parent teaches a child to recognize and label their internal sensations. However, a simulation of these processes is beyond the scope of the current effort.

Once the CP is known, the regulation error can be established from Eq. C1 and a hedonic state curve as a function of regulation error can be simulated:

$$hs = \left(0.5 \times \left(\tanh\left(0.2 \times (-15 + regErr)\right) - \tanh\left(0.2 \times (15 + regErr)\right)\right) - 0.999\right)/0.997 \qquad \text{(H3)}$$

where *hs* is the hedonic state and *regErr* is the regulation error. This curve, a more quantitative version of the qualitative one shown in Figure C2, is shown in Figure H2.

The hedonic state threshold is a level above which the regulation error triggers arousal and the hedonic state enters awareness. Thus, if the hedonic state increases in intensity to a level above the threshold (assumed to be 15 % of the maximum level plus a small random variation), one becomes aware of the hedonic state. This, in turn, generates a drive state of commensurate intensity that will be used, in conjunction with the hedonic state, to select, activate, and monitor appropriate behaviors. This relationship is shown in Figure H3.

Some assumptions need to be made about those parts of the physiology not simulated. Specifically, it is assumed that the CP for a given HM changes as a function of time in the same way that it changes on average for a given human regional population. These assumptions are made to avoid the need for more detailed simulations or modeling of physiological processes that will add a heavy burden of programming without contributing in any substantial way to the fundamental processes required for the generation of meaning.

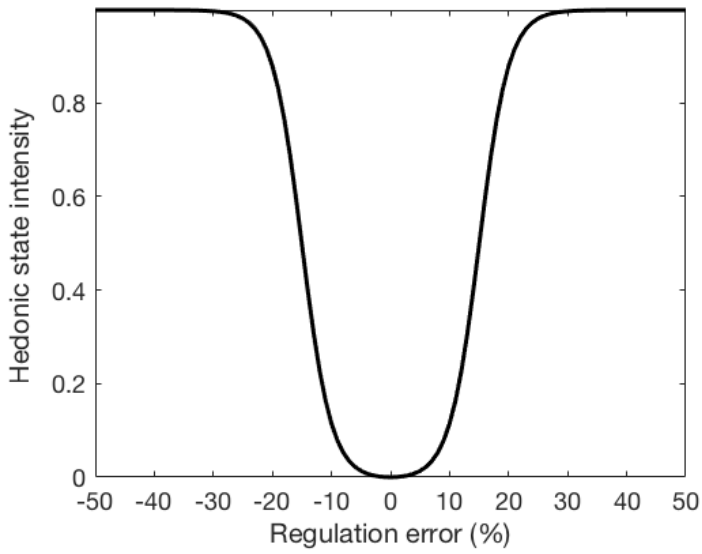

Figure H2
Hedonic states vary as a function of regulation errors. With increasing absolute value of the regulation error, a negative hedonic state ensues and a positive one with a decreasing absolute regulation error.

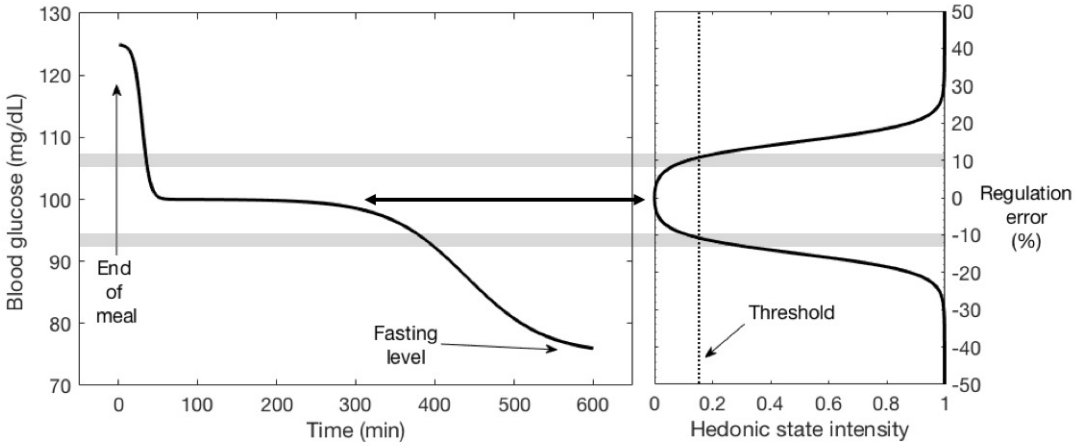

Figure H3
The relationships between fluctuations in blood glucose levels, regulation errors, and hedonic states (the relationships between the left and right panels are illustrative and quantitatively approximate only). A physiological homeostatic mechanism usually keeps blood glucose levels stable near the 100 mg/dL set point (left panel; also *cf.* Figure H1), hence, there is no regulation error (right panel). The double-headed arrow shows the blood glucose level at which there is no regulation error. However, when the mechanism cannot maintain regulation, the current point moves away from the set point and a regulation error with concomitant hedonic state arises. Slightly varying thresholds (within shaded bands), one on each side of the set point (zero regulation error) represent levels at which arousal occurs and hedonic states enter awareness and induce drive states that trigger restorative behaviors.

By evaluating the operation of each HM, it can be determined if a regulation error for each one exists. If this is the case, a 'hedonic state' with appropriate intensity and sign is generated for each one and an overall hedonic state established through a summation of the individually generated ones. It is this latter result that constitutes a valence and that furnishes the quale aspect of meaning. It also provides a teaching signal for learning and predictive processes. Valences are also generated by habituation and the mere exposure effect (*cf.* Appendix B), but these might be comparatively simple to conceptualize and simulate.

Once this basic structure is in operation, the synthetic generation and decoding of meaning can be implemented with a main program where initialization of critical variables is followed by entering a perpetual loop. Within this loop, homeostatic systems, learning systems, behavioral systems, and others, are repeatedly executed by calling the relevant modules. It can be augmented at a later stage with more HMs, but also with other types of behavioral modules that might appear to be different from those of physiological regulating mechanisms such as behavior mediated by endocrinological systems.

References

Adams, F., and Aizawa, K. 2017. Causal Theories of Mental Content. In *The Stanford Encyclopedia of Philosophy* ed. E. N. Zalta. Metaphysics Research Lab, Stanford University. Available electronically at https://plato.stanford.edu/archives/sum2017/entries/content-causal.

Andersson, P.; Strandman, A.; and Strannegård, C. 2019. Exploration Strategies for Homeostatic Agents. In *International Conference on Artificial General Intelligence*, 178-187. Shenzhen, China. Springer.

Arrighi, C., and Ferrario, R. 2005. The Dynamic Nature of Meaning. In *Computing, Philosophy and Cognition* eds. L. Magnani and R. Dossena, 295–312. London: King's College Publications.

Aston-Jones, G., and Cohen, J. D. 2005. An Integrative Theory of Locus Coeruleus-Norepinephrine Function: Adaptive Gain and Optimal Performance. *Annu. Rev. Neurosci.* 28: 403-450.

Bianchini, F., and Bruni, D. 2012. What Language for Turing Test in the Age of Qualia? In *Revisiting Turing and his test: comprehensiveness, qualia, and the real world*, 34-40. University of Birmingham, UK: AISB/IACAP Symposium.

Binder, J. R., and Desai, R. H. 2011. The Neurobiology of Semantic Memory. *Trends in Cognitive Sciences*. 15: 527-536.

Block, N. 2002. Concepts of Consciousness. In *Philosophy of Mind: Classical and Contemporary Readings* ed. D.J. Chalmers, 206-218. Oxford University Press USA.

Bonzanni, M.; Rouleau, N.; Levin, M.; and Kaplan, D. L. 2019. On the Generalization of Habituation: How Discrete Biological Systems Respond to Repetitive Stimuli: A Novel Model of Habituation That Is Independent of Any Biological System. *BioEssays*. 41: 1900028.

Bornstein, R. F., and D'agostino, P. R. 1992. Stimulus Recognition and the Mere Exposure Effect. *Journal of personality and social psychology*. 63: 545-552.

Burato, E.; Cristani, M.; and Vigano, L. 2011. Meaning Negotiation as Inference. *arXiv preprint arXiv:1101.4356*. Available electronically at https://arxiv.org/pdf/1101.4356.pdf.

Cáceres, M. J., and Perthame, B. 2014. Beyond Blow-Up in Excitatory Integrate and Fire Neuronal Networks: Refractory Period and Spontaneous Activity. *Journal of theoretical biology*. 350: 81-89.

Camhi, J.M. 1984. *Neuroethology: Nerve Cells and the Natural Behavior of Animals*. Sinauer Associates Inc.

Caminiti, R.; Innocenti, G. M.; and Battaglia-Mayer, A. 2015. Organization and Evolution of Parieto-Frontal Processing Streams in Macaque Monkeys and Humans. *Neuroscience & Biobehavioral Reviews*. 56: 73-96.

Chandler, D. J.; Jensen, P.; McCall, J. G.; Pickering, A. E.; Schwarz, L. A.; and Totah, N. K. 2019. Redefining Noradrenergic Neuromodulation of Behavior: Impacts of a Modular Locus Coeruleus Architecture. *Journal of Neuroscience*, 39: 8239-8249.

Christian, K.M., and Thompson, R.F. 2003. Neural Substrates of Eyeblink Conditioning: Acquisition and Retention. *Learning and Memory*. 10: 427-455.

Daly, M. E.; Vale, C.; Walker, M.; Littlefield, A.; Alberti, K. G.; and Mathers, J. C. 1998. Acute Effects on Insulin Sensitivity and Diurnal Metabolic Profiles of a High-Sucrose Compared with a High-Starch Diet. *The American Journal of Clinical Nutrition*, 67: 1186-1196.

Damasio, A. R. 1994. *Descartes' Error: Emotion, Reason and the Human Brain*. New York: Putnam.

Eckersley, P., and Sandberg, A. 2013. Is Brain Emulation Dangerous? *Journal of Artificial General Intelligence*. 4: 170-194.

Edlow, B. L.; Takahashi, E.; Wu, O.; Benner, T.; Dai, G.; Bu, L.; Grant, P.E.; Greer, D. M.; Greenberg, S. M.; Kinney, H.C.; and Folkerth, R. D. 2012. Neuroanatomic Connectivity of the Human Ascending Arousal System Critical to Consciousness and Its Disorders. *Journal of Neuropathology and Experimental Neurology*. 71: 531-546.

Ertel, W. 2017. *Introduction to Artificial Intelligence*. Springer.

Eth, D.; Foust, J. C.; and Whale, B. 2013. The Prospects of Whole Brain Emulation Within the Next Half-Century. *Journal of Artificial General Intelligence*. 4: 130-152.

Faingold, C. L.; Riaz, A.; and Stittsworth Jr, J. D. 2014. Neuronal Network Plasticity and Network Interactions Are Critically Dependent on Conditional Multireceptive (CMR) Brain Regions. In *Neuronal Networks in Brain Function, CNS Disorders, and Therapeutics* eds. C.L. Faingold and H. Blumenfeld. Academic Press.

Faulkner, J. W., and Wilshire, C. E. 2020. Mapping Eloquent Cortex: A Voxel-Based Lesion-Symptom Mapping Study of Core Speech Production Capacities in Brain Tumour Patients. *Brain and language*. 200: 104710.

Filimon, F.; Nelson, J. D.; Hagler, D. J.; and Sereno, M. I. 2007. Human cortical representations for reaching: Mirror neurons for execution, observation, and imagery. *NeuroImage*. 37: 1315-1328.

Fjelland, R. 2020. Why General Artificial Intelligence Will Not Be Realized. *Humanities and Social Sciences Communications*. 7: 1-9.

Ford, J. 2011. Helen Keller Was Never in a Chinese Room. *Minds and Machines*. 21: 57-72.

French, R. M. 2012. Dusting Off the Turing Test. *Science*. 336: 164-165.

Goldwag, J. A., and Wang, G. 2019. Training Artificial Neurons: An Introduction to Machine Learning. In *Developments in X-Ray Tomography XII*, 111131P. San Diego, Calif.: International Society for Optics and Photonics.

Grandy, R.E., and Warner, R. 2020. Paul Grice. In *The Stanford Encyclopedia of Philosophy* ed. E. N. Zalta. Metaphysics Research Lab, Stanford University. Available electronically at https://plato.stanford.edu/archives/sum2020/entries/grice/.

Grant, E. L.; Hoerder-Suabedissen, A.; and Molnár, Z. 2012. Development of the Corticothalamic Projections. *Frontiers in Neuroscience*. 6: 53.

Greif, H. 2012. Laws of Form and the Force of Function. Variations on the Turing Test. In *Revisiting Turing and his test: comprehensiveness, qualia, and the real world*, 60-64. University of Birmingham, UK: AISB/IACAP Symposium.

Grice, H.P. 1989. *Studies in the Way of Words*. Harvard University Press.
Harnad, S. 1990. The Symbol Grounding Problem. *Physica D: Nonlinear Phenomena*. 42: 335-346.
Haspelmath, M. 2020. Human Linguisticality and the Building Blocks of Languages. *Frontiers in Psychology*. 10: 3056.
Hausser, R. 2014. Foundations of Computational Linguistics: Human-Computer Communication in Natural Language. Springer.
Hodges, A. 2012. Beyond Turing's Machines. *Science*. 336: 163-164.
Hutter, M. 2012. One Decade of Universal Artificial Intelligence. In *Theoretical Foundations of Artificial General Intelligence* eds. P. Wang P. and B. Goertzel. Paris: Atlantis Press. Available at https://doi.org/10.2991/978-94-91216-62-6_5.
Hutto, D. D., and Myin, E. 2018. Much ado about nothing? Why going non-semantic is not merely semantics. *Philosophical Explorations*. 21: 187-203.
Jaynes, J. 1976. *The Origin of Consciousness in the Breakdown of the Bicameral Mind*. Boston: Houghton Mifflin.
Kaas, J. H.; Qi, H. X.; and Stepniewska, I. 2018. The Evolution of Parietal Cortex in Primates. In *Handbook of Clinical Neurology* eds. G. Vallar and H.B. Coslett. Elsevier.
Kahana, M. J.; Howard, M. W.; and Polyn, S. M. 2008. Associative Retrieval Processes in Episodic Memory. Available at https://surface.syr.edu/psy/3.
Kremelberg, D. 2019. Embodiment as a Necessary a Priori of General Intelligence. In *International Conference on Artificial General Intelligence*, 132-136. Shenzhen, China. Springer.
Kupfermann, I. 1991. Learning. In *Principles of Neural Science* eds. E.R. Kandel, J.H. Schwartz, and T.M. Jessel, 805-815. Elsevier.
Leader, L. R. 2016. The Potential Value of Habituation in the Fetus. In *Fetal Development* eds. N. Reissland and B.S. Kisilevsky. Springer.
Linson, A.; Dobbyn, C.; and Laney, R. 2012. Interactive Intelligence: Behaviour-Based AI, Musical HCI and the Turing Test. In *Revisiting Turing and his test: comprehensiveness, qualia, and the real world*, 16-19. University of Birmingham, UK: AISB/IACAP Symposium.
McCall, R. J.; Franklin, S.; Faghihi, U.; Snaider, J.; and Kugele, S. 2020. Artificial Motivation for Cognitive Software Agents. *Journal of Artificial General Intelligence*, 11: 38-69.
Müller, V. C. 2012. Autonomous Cognitive Systems in Real-World Environments: Less Control, More Flexibility and Better Interaction. *Cognitive Computation*. 4: 212-215.
Müller, V. C., and Ayesh, A. 2012. Foreword From the Workshop Chairs. In *Revisiting Turing and his test: comprehensiveness, qualia, and the real world*, 4-5. University of Birmingham, UK: AISB/IACAP Symposium.
Namburi, P.; Al-Hasani, R.; Calhoon, G. G.; Bruchas, M. R.; and Tye, K. M. 2016. Architectural Representation of Valence in the Limbic System. *Neuropsychopharmacology*. 41: 1697-1715.
Nichols, D. S.; Innamorati, M.; Erbuto, D.; Ryan, T. A.; Pompili, M. 2021. An MMPI-2 Hopelessness Scale: Construction, Initial Validation and Implication for Suicide Risk. *Journal of Affective Disorders Reports*. 3: 100057.
Oberman, L. M.; Pineda, J. A.; and Ramachandran, V. S. 2007. The human mirror neuron system: A link between action observation and social skills. *SCAN*. 2: 62-66.
Parisi, G. I.; Kemker, R.; Part, J. L.; Kanan, C.; and Wermter, S. 2019. Continual Lifelong Learning with Neural Networks: A Review. *Neural Networks*. 113: 54-71.
Piaget, J. 1977. The development of thought. New York: Viking.
Pitt, D. 2020. Mental Representation. In *The Stanford Encyclopedia of Philosophy* ed. E. N. Zalta. Metaphysics Research Lab, Stanford University. Available electronically at https://plato.stanford.edu/archives/spr2020/entries/mental-representation/.

Prigerson, H. G.; Kakarala, S.; Gang, J.; and Maciejewski, P. K. 2021. History and Status of Prolonged Grief Disorder as a Psychiatric Diagnosis. *Annual Review of Clinical Psychology*. 17: 8.1-8.18.

Rapaport, W. J. 2006. How Helen Keller Used Syntactic Semantics to Escape From a Chinese Room. *Minds and machines*. 16: 381-436.

Reder, L. M.; Oates, J. M.; Thornton, E. R.; Quinlan, J. J.; Kaufer, A.; and Sauer, J. 2006. Drug-Induced Amnesia Hurts Recognition, But Only for Memories That Can Be Unitized. *Psychological Science*. 17: 562-567.

Reeves, C.; Schmauder, A. R.; and Morris, R. K. 2000. Stress Grouping Improves Performance on an Immediate Serial List Recall Task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 26: 1638-1654.

Rodríguez, D.; Hermosillo, J.; and Lara, B. 2012. Meaning in Artificial Agents: The Symbol Grounding Problem Revisited. *Minds and machines*. 22: 25-34.

Rolls, E. T. 1971. Absolute Refractory Period of Neurons Involved in MFB Self-Stimulation. *Physiology & Behavior*. 7: 311-315.

Rolls, E. T. 2019. The Cingulate Cortex and Limbic Systems for Emotion, Action, and Memory. *Brain Structure and Function*. 224: 3001-3018.

Ruiz-Salas, J. C., and Luis, G. 2020. Induced Positive Affect Reduces the Magnitude of the Startle Response and Prepulse Inhibition. *Journal of Psychophysiology*. Available at http://dx.doi.org/10.1027/0269-8803/a000261.

Sah, P.; Faber, E.S.L.; Lopez de Armentia, M.; and Power, J. 2003. The Amygdaloid Complex: Anatomy and Physiology. *Physiological reviews*, 83: 803-834.

Schachter, S. 1964. The Interaction of Cognitive and Physiological Determinants of Emotional State. *Advances in Experimental Social Psychology*. Academic Press.

Schulze, G. 1995. Motivation: Homeostatic Mechanisms May Instigate and Shape Adaptive Behaviors Through the Generation of Hedonic States. In *Biological perspectives on Motivated Activities* ed. R. Wong. Norwood, NJ.: Ablex.

Schulze, G. 2003. *Propositions: Regarding Different Aspects of Human Behavior*. Trafford Publishing.

Schulze, G., and Mariano, M. R. 2003. *Mechanisms of Motivation*. Trafford Publishing.

Searle, J. 1980. Minds, Brains, and Programs. *Behavioral and Brain Sciences*. 3: 417-457.

Sgro, J. A.; Dyal, J. A.; and Anastasio, E. J. 1967. Effects of Constant Delay of Reinforcement on Acquisition Asymptote and Resistance to Extinction. *Journal of Experimental Psychology*. 73: 634-636.

Shimamura, A. P. 2011. Episodic Retrieval and the Cortical Binding of Relational Activity. *Cognitive, Affective, & Behavioral Neuroscience*. 11: 277-291.

Skeide, M. A., and Friederici, A. D. 2016. The Ontogeny of the Cortical Language Network. *Nature Reviews Neuroscience*. 17: 323-332.

Speaks, J. 2019. Theories of Meaning. In *The Stanford Encyclopedia of Philosophy* ed. E. N. Zalta. Metaphysics Research Lab, Stanford University. Available electronically at https://plato.stanford.edu/archives/win2019/entries/meaning/.

Staddon, J. E., and Cerutti, D. T. 2003. Operant conditioning. *Annual Review of Psychology*. 54: 115-144.

Stein, R. B. 1965. A Theoretical Analysis of Neuronal Variability. *Biophysical Journal*. 5: 173-194.

Sutton, R.S., and Barto, A.G. 1998. *Reinforcement Learning*: *An Introduction*. Cambridge: The MIT Press.

Szepesvári, C. 2010. Algorithms for Reinforcement Learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*. 4: 1-103.

Thompson, R. F. 2009. Habituation: A History. *Neurobiology of Learning and Memory*. 92: 127-134.

Tinberg, C. E.; Khare, S. D.; Dou, J.; Doyle, L.; Nelson, J. W.; Schena, A.; Jankowski, W.; Kalodimos, C. G.; Johnsson, K.; Stoddard, B. L.; and Baker, D. 2013. Computational Design of Ligand-Binding Proteins with High Affinity and Selectivity. *Nature*. 501: 212-216.

Tolentino, J. C., and Schmidt, S. L. 2018. DSM-5 Criteria and Depression Severity: Implications for Clinical Practice. *Frontiers in Psychiatry*. 9: 450.

Tulving, E., and Markowitsch, H. J. 1998. Episodic and Declarative Memory: Role of the Hippocampus. *Hippocampus*, 8: 198-204.

Turing, I. B. A. 1950. Computing Machinery and Intelligence. *Mind*. 59: 433-460.

Van Engelen, J. E., and Hoos, H. H. 2020. A Survey on Semi-Supervised Learning. *Machine Learning*. 109: 373-440.

van Heukelum, S.; Mars, R. B.; Guthrie, M.; Buitelaar, J. K.; Beckmann, C. F.; Tiesinga, P.H.E.; Vogt, B.A.; Glennon, J.C.; and Havenith, M. N. 2020. Where is Cingulate Cortex? A Cross-Species View. *Trends in Neurosciences*. 43: 285-299.

Vanhooren, S. 2019. Struggling with Meaninglessness: A Case Study from an Experiential–Existential Perspective. *Person-Centered & Experiential Psychotherapies*. 18: 1-21.

Vertes, R. P.; Linley, S. B.; and Hoover, W. B. 2015a. Limbic Circuitry of the Midline Thalamus. *Neuroscience & Biobehavioral Reviews*. 54: 89-107.

Vertes, R. P.; Linley, S. B.; Groenewegen, H. J.; and Witter, M. P. 2015b. Thalamus. In *The rat nervous system*, 335-390. Academic Press.

Vogt, B. A. 2016. Midcingulate Cortex: Structure, Connections, Homologies, Functions and Diseases. *Journal of Chemical Neuroanatomy*. 74: 28-46.

Wagner, A. R. 1961. Effects of Amount and Percentage of Reinforcement and Number of Acquisition Trials on Conditioning and Extinction. *Journal of experimental Psychology*. 62: 234-242.

Wilkenfield, J.; Nickel, M.; Blakely, E.; and Poling, A. 1992. Acquisition of Lever-Press Responding in Rats with Delayed Reinforcement: A Comparison of Three Procedures. *Journal of the Experimental analysis of behavior*. 58: 431-443.

York, W., and Swan, J. 2012. Taking Turing Seriously (But Not Literally). In *Revisiting Turing and his test: comprehensiveness, qualia, and the real world*, 54-59. University of Birmingham, UK: AISB/IACAP Symposium.

Yoshida, N. 2017. Homeostatic Agent for General Environment. *Journal of Artificial General Intelligence*. 8: 1-22.

Zajonc, R. B. 1968. Attitudinal Effects of Mere Exposure. *Journal of Personality and Social Psychology*. 9: 1-27.

Zillich, M. 2012. My Robot is Smarter Than Your Robot - On the Need for a Total Turing Test for Robots. In *Revisiting Turing and his test: comprehensiveness, qualia, and the real world*, 12-15. University of Birmingham, UK: AISB/IACAP Symposium.