Feeling is Believing

Why affective neuroscience rejects computational-representational attitudes and endorses a pluralistic embodied cognitive science

David Haslacher
3897559

M.Sc. Thesis
artificial intelligence

Utrecht University

First supervisor:
Jack van Honk

second supervisor:
John-Jules Meyer

45 ECTS

Abstract

A popular conception in the mind/brain sciences today is the metaphor of the brain as a computer. Philosophers usually interpret this notion as being a combination of a computational account of reasoning and a representational account of mind. According to this hypothesis, all relevant thought consists of executing some well-defined rules over representations which model (and have a direct correspondence to) the external world. This work revolves around challenging the computer metaphor of mind and advocating a pluralistic, biological approach to cognition which then guides the scientific use of computation in the mind/brain sciences and AI. Since attitudes (and intentionality in general) are staples of cognition, this case will be made from the perspective of affective neuroscience. Emotions allow us (and other animals) to evaluate whether entities in the environment are advantageous or not, which is the most basic form of reasoning possible. Psychology unsuprisingly confirms that rational cognition depends heavily on emotion. Affective neuroscience additionally makes a powerful case that emotions arise out of the seething, complex interactions taking place in the physical (biological) brain and body. Such phenomena extend all the way to the molecular level of description, and perhaps even lower. I will therefore examine neuroscientific theories of emotion to make two seperate but related cases. For one, cognitive phenomena require many levels of description, as many other phenomena in the natural sciences already receive (this is explanatory pluralism). Furthermore, cognition-related computation should implement naturalistic theories drawn from these levels rather than being seen as a general framework. Even if rules and representations can be empirically identified on any level, they would surely not explain every cognitive phenomenon - most likely, they would be a small subset of the bigger picture; tools in a much larger explanatory toolbox. I therefore propose that the mind/brain sciences and artificial intelligence should employ computation to simulate the physical non-symbolic basis of cognition, rather than seeing the brain as a computer.

Preface

Having come from a background in computer science, I understand and respect the awesome power that one can wield with rules and representations. However, I also respect science, and therefore feel that it isn't prudent to popularize the notion of the brain as an algorithmic system without the appropriate evidence. I believe that human ingenuity is a subset of nature's ingenuity, and that concepts of computation are a subset of the former. In order to truly understand human ingenuity, I think we must start with the natural sciences. This belief, combined with my interest in the relationship between emotions, cognition, the body, and their lower levels of description compelled me to write this thesis. Thank you, Prof. Meyer and Prof. van Honk, for providing valuable comments and thereby aiding me in expressing my opinions comprehensively. Lastly - thank you, Luiza, Mom, and Dad, for taking the time to care and read about my interests.

Contents

1 Introduction

Cognitive science is often defined as the interdisciplinary study of the mind, and therefore embraces many fields including psychology, artificial intelligence, philosophy, neuroscience, linguistics, and anthropology. [39] However, since the inception of the field, there has been a strong trend of reducing all manner of cognitive processes to symbols manipulated by some set of rules. This approach is known as cognitivism and relies heavily on the computer metaphor of the mind, where the focus is on discovering the algorithms underlying cognitive processes. The focus of this work, in contrast, is on suggesting a restoration of a diverse interdisciplinary epistemological framework for studying the mind/brain, where pluralistic explanations as featured in other natural sciences take precedence.

Since I will not be able to make my case from the point of view of all subdisciplines of cognitive science, I will focus on affective neuroscience. The reason for this is that emotions have long been regarded as secondary to cognition - a sort of thorn in the side of any rational human being - belonging to the domain of art and aesthetics rather than rationality. As will be described in section 3, research has revealed that this sentiment could not be more wrong. In fact, emotion is advantageous to all manner of thought, including the 'rational' sort. From the perspective of affective neuroscience, affective processes are quite literally the foundation of the mind/brain. They are the ultimate attitudes that humans and other mammals hold towards entities of any nature. Section 4 will describe emotions from the perspective of two prominent neuroscientists, Jaak Panksepp and Antonio Damasio, and section 4.3 will summarize some basic functional principles of affect from the neuroscientific perspective which will form the basis of my further argument. One major point of agreement here is that emotions are inextricably embodied.

The neuroscientific perspective will underscore a notion which has gained more attention in recent years; namely that the brain is not a computer but a complex system. Against this backdrop, section 5 will reject two cognitivistic concepts: propositional attitudes and the broader computational theory of mind. This is not supposed to be an indictment of computers becoming intelligent, however, as some philosophers like to argue. Rather, it is meant to caution against the use of algorithms as scientific models without empirical evidence of their existence, which would require a corresponding functional mapping to biological states. The brain has highly complex nonlinear interactions [29] which must be respected if one wishes to employ an algorithm to model cognition. In general, a computer will only be able to model a subset of the interactions taking place within the physical system that the brain is, and a reduced linearized model of cognitive phenomena (as traditionally employed in symbolic AI, and

especially machine learning) doesn't seem to move any field related to cognition towards the goal of understanding or reproducing sentience. I argue that computers and algorithms are meant to be seen as simulational tools, not as scientific models of the mind/brain. To that end, they should try to faithfully reproduce the underlying physical system as much as possible. Furthermore, the virtue of any model is always to be seen in relation to the phenomena to be described – therefore, even if some algorithm (be it linear or nonlinear) could be shown to operate at some level of description in the physical brain, it is highly unlikely that it would help explain phenomena which are affected by important dynamics at other levels.

Section 6 will address the issue that complex systems such as the brain feature many levels of organization, where phenomena at any level may be truly emergent - that is, not reducible to levels below. Therefore, I argue that the traditional tripartite computer-oriented levels of analysis of David Marr should be replaced by a pluralistic approach to the scientific study of the mind/brain which features an epistemological landscape with many levels of analysis vertically, as well as many domains/functions of cognition horizontally. In such a vision, bridges between theories are of course very helpful if they can be built, but reductionism is not the goal.

Finally, section 7 will call for embodied naturalistic (neuromorphic) architectures in artificial intelligence. In considering the requirements for constructing an emotional agent with the properties that the emotional human mind/brain has, I will argue that architectures drawn from just one level of organization are insufficient. Therefore, I reject purely connectionist models, and make the case for neuromorphic architectures which draw from multiple levels of organization. Generally speaking, I argue that artificial intelligence should build architectures which attempt to simulate as many theories within the pluralistic cognitive science landscape as are required for their application domain.

2 Two competing approaches to the mind

The study of the mind has undergone a number of paradigm shifts, which have been strongly influenced by the technical methodologies available to researchers at the time. The popular emergence of behaviorism in the early 20th century was largely due to an inability of proponents of logical positivism to specify experimental conditions and observations for directly measuring mental states and processes. [1] Thus, much research in psychology during this time relied on a methodology of associationism, whereby the key object of study was the relationship between stimulus and response. The mind/brain was essentially seen as a device for conditioning. By the 1950's, however, the advent of com-

puter science had fundamentally started to influence the study of the mind, and cognitivism was born. By the 1990's, a second revolution had begun, emphasizing embodied explanations of cognition over more abstract and amodal (independent of sensory data) ones.

2.1 Cognitivism

As a paradigm for studying the mind/brain, cognitivism denies the 'black box' analogy of behaviorism in a very important way; namely by specifying structures and processes operating in the mind to produce behavior. With its advent, the computer metaphor became the dominant analogy in the study of the mind, giving rise to the interdisciplinary cognitive science research field. Naturally, this paradigm shift was strongly influenced by likes of Alan Turing and his peers, pioneers in the field of computer science. The general assumption of a cognitivist approach to mind/brain is that there exists an underlying architecture of cognition that is more abstract than its physical implementation. Frequently, cognitivists will also claim multiple realizability, which is the notion that a single mental kind can be instantiated in multiple physical kinds. [2]

This claim is not a central tenet of the cognitivist's approach, however. Most important are explanations of cognition in terms of the rules and structures - kinds such as feature lists, semantic networks, and frames, as well as rules operating on them. [3] These explanations are often abstract and amodal, such that the brain is seen as a device for extracting and processing these structures independently of the specific stimuli they are derived from. Such an approach has worked well for explaining linguistic ability, for instance, with Noam Chomsky's generative grammars being hugely influential to this day. Whilst a language module may exist (although there is no direct neuroscientific evidence for it [11]), for the rest of this essay I will be attempting to describe the insufficiency of disembodied structures and processes independent of sensory data (i.e. amodal ones) in describing the mind/brain, specifically with respect to emotion.

2.2 Embodied Cognition

There are several shortcomings of a cognitivist approach. As is to be expected, one of the main criticisms is the neglect of the body - the mind is not a free-floating thinking machine, but a biological brain intimately connected to its 'slave systems' and the environment. In other words, thinking is intrinsically dependent on perception and action systems. This was the impetus for a large amount of empirical research on cognition in the last 10 years that explored this relationship. [6]

I will not review the evidence in detail here, as a plethora of studies can be found at [7]. It suffices to say that it appears that in higher cognition tasks, even off-line (in absence of perceptual input), the modal systems are activated in a simulation process. Concepts seem to be grounded in perception and action systems, such that associations between abstract and concrete properties are reactivated even in absence of the concrete stimulus - also known as multimodal simulation. This is central to both implicit and explicit memory processes. Thus, for instance, color associations are remembered even in absence of the color, and the appropriate grasping motion for delicate/tough objects is simulated at the simple mention of the object's name. Furthermore, experimental manipulation of bodily states affects performance on more abstract cognitive tasks - a good example of this is the implicit evaluation of stimuli/utterances dependent on pleasurable/displeasing stimulation.

Social cognition is also intrinsically gestural, such that action systems are activated by default in social tasks. Another mental phenomenon cognitive explanations tend to neglect is the role of emotion. As will be explained in detail in sections 3 and 4, emotions themselves seem to be inherently embodied. In Damasio's somatic marker hypothesis (section 4.2.1), the notion of multimodal simulation is applied in the form of an "as-if loop". In this case, simulation is not a case of an abstract property triggering an amalgamation of sensory experiences in a top-down manner, however. Rather, the "as-if" loop in the brain allows persons to bypass the physiological states necessary for emotions, allowing them to observe the would-be psychological effects of such sensations. This forms the basis of the ability to read others' feelings; mindreading and empathy. Multimodal simulation [3] , however, is usually explained as an abstracted perceptual symbol triggering various sensorimotor reactions - a top-down process bottoming out in various sensory modalities. Damasio's "as-if" loop, on the other hand, is a bottom-up process allowing the sensory modalities to be bypassed and higher psychological effects to be observed. In any case, there is strong evidence for a grounding of abstract notions in concrete sensory experiences.

These results imply that abstract, amodal processes alone are an insufficient [3] level of description for the mind. This has prompted a number of theoretical models in the mind/brain sciences, as well as artificial intelligence, which incorporate the body and sensory data into cognition. A notable example from robotics is Rodney Brooks' subsumption architecture, where sensory systems are coupled to higher cognitive layers in parallel to generate behavior. [6] A large part of this thesis will be dedicated to convincing the reader that emotions are an example of a quintessential cognitive phenomenon which requires description on a somatic (body) level. The next section will examine emotions

from a psychological perspective, to lead into their description by affective neuroscience.

3 Psychological theories of emotion

Our emotional experiences are complex and multifaceted. Emotions can be approached from a number of different vantage points, so it is not surprising that theories thereof are relatively varied and numerous. I will not so much consider the evolutionary or cultural contexts of emotions here, but will rather focus on the actual emotional processes as they are implemented in the brain. For the layman, the cognitive value of an emotion may not readily be apparent. It is a common misconception that emotions are a marker of irrationality, contrasted with 'cold' rational thought. From such a layman's perspective, emotions belong exclusively to the domain of phenomenology, their feeling being a byproduct of existing as a sentient mind. Granted; this viewpoint is very much culturally dependent - in some 'hot-blooded' societies, emotional evaluations are actually predominantly valued as utility estimators of various entities, both personally and in discourse.

As many studying the mind/brain long enough would probably agree with, acknowledging the very basic function emotions have in rational thought processes and shaping our intuitions is necessary for any field that models cognition. They function as a primary evaluator of situations, allowing the brain to make rapid decisions under time and resource constraints - it seems quite clear that they have an important differentially valenced (positive or negative) appraisal function. One might say that they are a kind of cognitive optimization by evolutionary processes. Additionally, emotions are usually distinguished from moods, which are unspecific (not object-directed) feelings persistent over time. Furthermore, in addition to the subjective feeling of an emotion, a bodily response is usually a component of the experience. The above points are relatively uncontroversial. However, theories of emotion begin to clash when one considers the specific causal process that underlies affective arousal. Roughly speaking, traditional accounts of emotion can be divided into top-down (cognitive) and bottom-up (non-cognitive) perspectives. The former perspective contends that an emotional response is the result of a cognitive evaluation of stimuli, whilst the latter places the affective response as an automatic process before any cognitive evaluation. [8] I will briefly describe these theories before describing a third (feedback) approach in the next section which I believe to be more accurate in the face of all the evidence.

3.1 Cognitive theories

Two motivations can be stated for positing that emotions are the result of a cognitive evaluation. [8] The first is that the response to the same stimulus can vary across individuals and points in time. Thus, memory (and other experience-dependent associations) seem to be an important factor in the affective response. Activities that may have been exciting during a person's teenage years often lose their affective appeal later in life. Secondly, abstract categories can be the basis of the triggering of a certain emotion, whereby the specific stimulus instances of that category can vary wildly - fear could be elicited by both alligators and bears on the basis of their being aggressive animals. Thus, higher cognitive capacities seem to determine the emotional response. Cognitive theories of emotion come in two flavors - the judgment theories and the appraisal theories. [8] The former are mainly philosophical, making heavy use of folk-psychological concepts (which are propositional attitudes, addressed section 5.1). In these theories, the physiological response is usually not a necessary component of an emotional experience. In contrast, appraisal theories come from scientific psychology. Whilst they usually incorporate more nuanced categories of mental states, they essentially also refer to folk psychology in the form of beliefs, desires, intentions, and the likes. The cognitive models are varied, and attempt to delineate causes of specific emotions by, for instance, types of circumstances and types of motives of the person. [8] Additionally, the physiological response is considered an essential component of an emotional response, as opposed to the case of the judgment theories.

3.2 Non-cognitive theories

On the other hand, bottom-up theories of emotion attempt to describe the mechanisms by which emotions are automatically elicited by some stimulus, without deliberation. Some of these posit that all emotions are non-cognitive, whilst others grant the existence of cognitive emotions. Paul Griffiths, for instance, is of the latter camp, and conceptualizes non-cognitive emotions as "affect programs". [8] Essentially, these contain both a memory mechanism by which low-level associations can be remembered to belong to a certain emotional response category, as well as a physiological response mechanisms. The exclusively non-cognitive accounts (such as that of Jenefer Robinson [8]) of emotion are similar, except that they exclude emotions being elicited by cognitive processes. That is not to say that cognitive processes are unable to influence emotional responses under these theories, however, as they may still modify the initial bottom-up emotional memory. However, this supposedly must occur before the stimulus is responded to, so cognitive appraisals may not influence emotional responses in

'real-time'. Naturally, an exclusively non- cognitive account of emotions begs the question why it must be exclusively non-cognitive if cognitive appraisals clearly are able to modify the process in some way. This seems like an unnecessary postulation in the model without any real explanatory power over a hybrid account. In summary, cognitive theories claim that emotions are produced by higher cognition in the moment, whilst non-cognitive theories claim that a local, low-level memory saves and automatically reproduces affective associations that may have been previously saved through cognitive influence. Both of these perspectives will be integrated in the neuroscientific account of emotion in section 4, especially in Antonio Damasio's somatic marker hypothesis.

3.3 Feelings-as-information

A psychological account of the relationship between emotion and cognition wouldn't be complete without a mention of Norbert Schwarz' feelings-as-information theory [9], being as influential as it is. It provides an illuminating account of the functional advantages of having 'rational' processes coupled to emotional processes, and also depicts some cases where this coupling may fail to produce the desired advantageous behaviour/cognition. Both these advantages and failures will be reflected in the next section on the neuroscience of emotions, especially in Antonio Damasio's somatic marker hypothesis, making this psychological theory a neurally realistic account of the rationality/emotion relationship.

As the title suggests, the main postulate of the feelings-as-information theory is that people use their feelings as a source of information. Feelings are varied, so the type of information communicated depends on the feeling - this is quite self-evident when one considers that people tend to avoid causes of negatively valenced feelings and approach causes of positively valenced feelings (known as classical conditioning). According to the theory, people attribute a feeling as being about whatever is in the focus of attention, irrespective of whether a causal relation between the entity and the feeling actually exists. Thus, affective mechanisms are rife with the potential for misattribution. That is not to say that persons completely ignore this potential - "when a feeling is attributed to an incidental source, its informational value is discounted; conversely, when it is experienced despite perceived opposing forces, its informational value is augmented." [9]

Furthermore, feelings can induce different problem-solving strategies depending on the perceived complication of the situation - thus, 'rigorous' analytic thought processes are only triggered if there is a perceived necessity for them. In summary, whilst embodied affective reactions are not flawlessly attributed to their causes, they do provide an essential source of information. This is true es-

pecially also for metacognition, such that feelings convey important information about how the self is progressing. Thus, the impairment of affective mechanisms can be crippling for a person's general and social problem-solving and learning abilities, as well as potentially causing issues of self-perception.

In a nutshell, the feelings-as-information theory is a higher version of Damasio's neurally based somatic marker hypothesis. In both cases the triggering of somatic (body) states results in a valenced feeling (and evaluation thereof), which influences the task currently in working memory. Both Damasio [16] and Schwarz [9] state that negatively valenced feelings tend to produce slow, deliberative, and finicky styles of problem-solving. Positively valenced feelings, on the other hand, produce emotionally driven intuitive (heuristic) decisions. Additionally, the Iowa gambling task (the chief source of empirical evidence for the somatic marker hypothesis) reveals that when strong emotional states are induced in subjects prior to the task, decisions are strongly influenced and the tendency is to decide less advantageously. [16] Thus, subjects implicitly 'apply' their emotional states to the task at hand because the psychological mechanism is deeply ingrained. It is an advantageous mechanism, because emotions and their causes usually co-occur. However, they do not always - in the case of affective disorders, for instance, there may be both an unusual emotional response as well as an undue persistence of such states over time. These feelings would interfere with unrelated tasks, and present a malfunction of the system.

4 Neuroscientific theories of emotion

In the following section, two neuroscientific accounts of emotions will be considered, in both of which a recurrent loop of interpretation and recall of subneocortical somatic (body) representations by higher-level cognitive systems produces the totality of embodied and cognitive emotional experience. Furthermore, the somatic representations interact with motivational circuits, a dynamic which allows for certain body states to be felt as positive or negative, and certain behaviors to be elicited in response. These areas essentially implement the basic drives a mammal wants fulfill. Generally speaking, Antonio Damasio and Jaak Panksepp are in agreement about a number of principles which underlie our sense of self, our emotional lives, and the larger architecture of the mind/brain. These principles will form the basis of my larger argument that models of cognition (both human and artificial) must be based on emotion that is based on self- representation in an intimately embodied fashion. Furthermore, there are a few differences in the two theories which I will also address for posterity's sake.

First, however, the next two subsections separately describe the two neuroscientists' theories of consciousness and emotion. Although the focus of this

work is on the importance of embodied emotional processes in the mind, affective processes cannot be separated from the issue of consciousness, as the two phenomena are far from being categorically distinct - they share much of the same biological substrate and mechanisms. In the following section, somatic states will refer to the physiological body changes and their representation in the brain. There are (technically speaking) somatic representations, which represent the exterior surface of the body, and visceral representations, which represent the internal organs. The former has a higher resolution in the brain, as detail on the outer surface is more important for navigating in complex environments. For this work, however, I will refer to both types of body states and their brain representations as somatic. The feeling of an emotion will denote the consciously experienced changes in somatic states, whilst I use the term emotion as a summary of the whole process. Note that this terminology does not correspond with all literature – there is no universally used standard.

4.1 Jaak Panksepp

Being the main pioneer of a "cross-species affective neuroscience," [11] Panksepp has extensively investigated the neural basis of core emotional processes. The focus of this definition of affective neuroscience is the sub-neocortical level of the brain, where a set of ancient brain structures lie which all mammals share. Evolutionarily speaking, our brains diverged from those of other mammals only after these basic structures were established. It is in these brain structures that the basic instinctual behaviors are dictated; those which make human behavior similar enough to that of an ape, a dog, or even rats. Animal models therefore are a valuable empirical source of experimental data in such a paradigm. This should be taken with a grain of salt, however, as the next section will detail. Whilst the rough organization of these subcortical emotional areas do seem homologous across most mammals, it is questionable to what level of detail the analogy holds up. In humans, for instance, the wiring is much more dense [58]. Panksepp's cross-species affective neuroscience places strong emphasis on these homologous brain structures, although the question of the extent to which animal data is applicable to humans is questionable. The significantly more interconnected wiring within the subcortex, as well as between the sub- and neorcortex in humans should make the dynamics of our lower brain networks quite different from that of other mammals. Additionally, there is considerable variability across individuals in both size and connectivity of brain structures in humans, largely due to the fact that (even at the subcortical level) there is much experience-dependent plasticity. [58] These criticisms will be discussed in the next section, however.

According to Panksepp, there seems to be sufficient evidence for seven distinct primary- process emotional circuits: SEEKING, FEAR, RAGE, LUST, CARE, PANIC, and PLAY. [11] This is naturally an oversimplification, however, as these neural circuits are composed of various neuropsychological components, such that it would always be possible to create a more nuanced neural model – which is what the critics of this 'basic emotion' approach emphasize. According to the basic-emotion approach, on the other hand, all emotional experiences are composite of these basic components. The main point is that these subcortical circuits, independently of how complex they truly are, control behaviors necessary for immediate survival of all mammals. It should be stated that one of these tasks is the homeostatic regulation of the body - physiologically, all life requires very specific conditions in order to maintain the processes conducive to its preservation. In this view, the brain is just an organ to ensure successful homeostasis. The basic emotional systems are thus connected to the action systems of the brain. These basic emotional circuits also interact with a kind of homunculus – I do not intend to imply a miniaturized cognitive agent pulling all the strings, however. Rather, this homunculus refers to the complete moment-by-moment mapping of somatic states and their changes. It is affected by the primary-process circuits and vice-versa, an interaction which allows the relevance or valence of somatic changes to be determined. According to Panksepp [11], these self-representational areas, located in the deep midline areas of the brain, are where the feeling of an emotion may arise.

Affective neuroscience thus paints a picture of primary-process behavioral circuits which are intrinsically connected to representations of body states and action areas. These first-order representations of physiological changes, according to Panksepp, give rise to consciously experienced emotions in both humans and animals. There exists a significant body of empirical research corroborating both the notion that these subcortical brain structures cause the feeling (rather than mere behavior) of an emotion, and that they are to a certain extent functionally equivalent in humans and other mammals. [11] Regarding the ability of other mammals to have valenced experiences, research has shown that animals seem to have emotional experiences in response to drugs which cause similar attraction/aversion behaviors in humans. [11] Since these animal models lack the neocortical structures which humans have, the emotional experiences must arise from subcortical structures. Regarding the functional equivalence of these subcortical structures, imaging studies have revealed a passive correspondence, and experimental stimulation of these areas have revealed similar reactions in both human and animal models. [11] However, these results should be taken with a grain of salt as they are far from conclusive. In fact, as discussed in the next section, it may even be the case that for every study finding similarities,

there are multiple studies finding differences. An idea that affective neuroscience manages to solidly support, however, is that many animals have the ability to consciously experience emotional feelings in a fashion that is very similar to us - an idea which many humans gladly (but stubbornly and falsely, in the face of the evidence) reject. An overview of how this basic subcortical consciousness interacts with higher brain function can be found in figure 1.

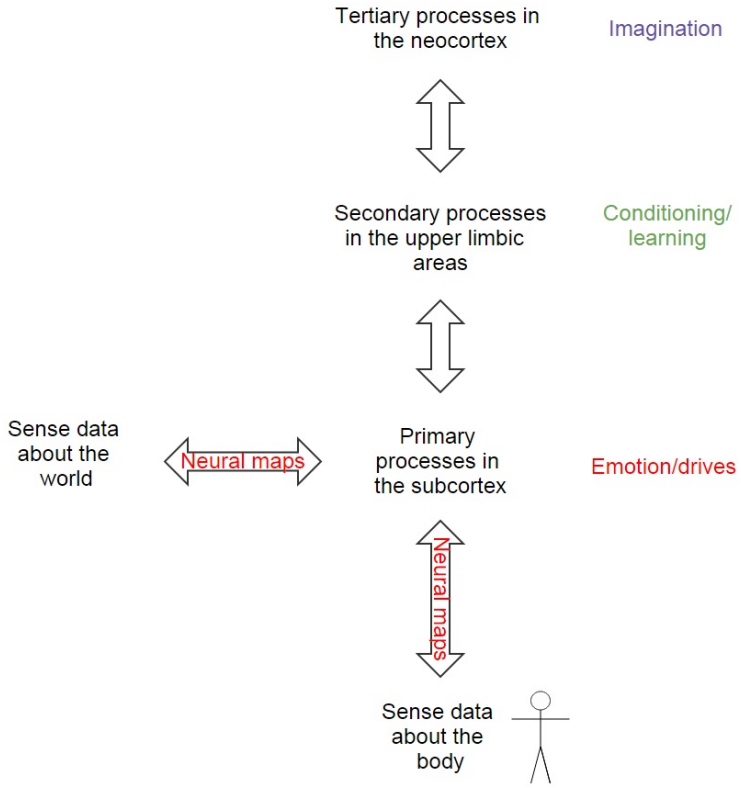

Figure 1: Three tiers of the mind according to affective neuroscience. Adapted from [10]

The division of the mind into primary, secondary, and tertiary processes has a history in the cognitive sciences, and follows a more or less hierarchical organization of the brain. These two points are relatively uncontroversial, depending on how specifically they are formulated, and are also the basic model of the mind that affective neuroscience works with. However, it is largely the tertiary processes which have been studied in the cognitive sciences. A more detailed outline of the respective functions of these layers and their location in the brain can be found in figure 2.

Tertiary Affects and Neocortical Awareness Functions

a) Thoughts and planning

b) Emotional thoughts and regulation

c) Free will - revising one's behavior

Secondary-Process Affective Memories and Learning

a) Classical conditioning

b) Instrumental and operant conditioning

c) Behavioral and emotional habits

Primary-Process, Basic-Primordial Affective States

a) Valenced feelings resulting from the external world

b) Valenced feelings resulting from the body (for homeostasis)

c) Coupling of emotion and action

Figure 2: Functions of the primary, secondary, and tertiary processes. Adapted from [10]

Developmentally speaking, the tertiary processes are a blank slate at birth from the perspective of affective neuroscience, and are shaped through experience by primary and secondary processes. Only after a long series of rewards, punishments, and other learning effects arising from subcortical emotional processes do the neocortical functions become crystallized. Notice that in figure 1 there is a recurrent relationship between each layer, indicating that after neocortical functions have been shaped, both bottom-up and top-down effects are at work. This means that the mind is never free of its more basic, animalistic urges, although we are distinct from other animals in our ability to exert a significant amount of top-down control over them. It is by this top-down control that we are able to trigger embodied emotional reactions to abstract concepts and categories. Our will is therefore a lot more free of basic instinctual emotions than that of, say, a dog. In a dog, the basic drives would have a larger amount of control over emotional reactions. That is not to say that the more ancient structures are superfluous to our cognition - in fact, secondary and tertiary processes without primary processes would be a both soulless and useless. The primary processes also continue to provide the very foundation of our consciousness and intelligence throughout a lifetime. After all, (without endorsing the basic-emotions approach) how would we socialize without PLAY, and how would we be ambitious without SEEKING? Furthermore, as mentioned in section 4.2.1, emotional reactions (possibly initiated in a top-down manner) are a large influence as to which cognitive problem-solving style is applied. For instance, recall of a sad memory may trigger tears and a depressive mindset in which cognition is slow and focused.

4.1.1 The controversy of discrete (basic) emotions and the limits of functional homology

As mentioned above, it is by far not clear whether it is realistic to construe emotions as discreet dispositional circuits (interacting with self-representational areas and thereby the body) that combine to form complex expressions. In other words, it is debatable whether emotions can be considered natural kinds. Philosophically speaking, natural kind categories can be defined according to either the common observable properties of their members (analogy), or their common causal substrate (homology). Furthermore, concerning what can potentially be observed, one must distinguish between external emotional behaviors and more complex emotional (phenomenological) feelings and their cognitive effects. Valenced feelings result from the complex interplay of basic needs/action areas and self-representational areas, and should definitely be considered an important observation of the phenomenon that we call emotion. Of course, interactions with higher brain areas also influence these feelings. Therefore, from the existence of apparently separable action patterns one cannot conclude that the concurrent phenomenological experiences and relevant wider brain effects are similarly describable in discreet terms. It can definitely be argued, additionally, that both should be considered relevant observations under the definition by analogy of a natural kind.

Furthermore, under the definition by homology of a natural kind, all instances of a basic emotion should share the same biological substrate – both across humans, and when comparing humans and their animal models. However, it is easily arguable that the neural machinery involved in generating the totality of an emotional experience is much too dependent on an individual's experience, generating highly individualistic interactions that also depend intimately on the person's specific biochemistry (most relevant here is the abundance of various neurotransmitters, which varies from person to person). This machinery is also much more densely wired in humans than it is in the animal models that provide much of the empirical evidence presented in favor of the basic emotions approach, so the cross-species homology has undeniable limits. Therefore, it may not be possible to assign a common neural basis to discreet categories of emotions. Jaak Panksepp and Lisa Barrett have provided an extensive dialogue through various publications highlighting the question of whether emotions can be construed as natural kinds. In my view, the machinery generating an emotional experience and its wider effects is much too complex to be described in discrete terms – regardless of whether one defines natural kinds by analogy or homology.

The definition of emotions as natural kinds on the basis of homology is one

of the central arguments of Jaak Panksepp's position. However, the homology between cases of emotional expression in humans and animals (as well as within the human population) breaks down at the relevant level of neural detail. To reiterate; Panksepp presents causal evidence in favor of basic emotions in animal models because such causal evidence would be unobtainable in humans – stimulation of subcortical areas would be problematic. Therefore, his causal evidence relies on the basic system he is manipulating and its similarity to other systems which he draws conclusions about. Plasticity creates differences between humans, and a significant difference in the density of subcortical wiring between humans and animals makes comparison and categorization of emotional experiences difficult. [58] The basic-emotions model Panksepp endorses relies on a subcortical neural substrate which is sufficiently homologous between all human individuals, as well as between humans and the animal models from which the evidence is derived. However, although we uncontroversially share a similar rough architecture of the mind, it is by far not clear whether our subcortical emotional systems are sufficiently similar to warrant such a conclusion. We have evolved different brains from the animal models presented as evidence, and our brains are each highly unique due to modification by experience (plasticity).

To conclude, there are two major issues facing the basic-emotions approach. The first is the reliability of the causal evidence presented in its favor, due to the subcortical homology which probably does not hold at the required level of detail. The second is the question of how accurate such a discrete description is, assuming that the evidence holds up. Considering the complex interplay of emotional components in the subcortex about to be described in the next section, as well as the uniqueness of our individual brains, it seems unlikely that a discrete description of emotional circuits is possible.

4.2 Antonio Damasio

The approach to consciousness and emotion that Antonio Damasio offers is in many ways very similar to that of Jaak Panksepp, although his approach is from a cognitive neuroscience perspective, which stresses the importance of neocortical areas. Damasio begins his account of consciousness [14] by stressing the evolutionary/biological purpose of the brain, like Panksepp: it is primarily a tool for survival of the body, which means that its primary task is the maintenance of homeostasis (biological equilibrium) in a complex environment. Fulfillment of basic emotional needs is primarily for this purpose. Furthermore, his account is heavily reliant on the notion that the brain is at the most basic level a machine for topographical representation - making maps. This includes maps of the body, the environment, and the self in relation to the environment, as well

as maps of changes in all of the aforementioned. Such a process never ceases, and both perceptual/physiological changes as well as top-down recall (imagination) influence the generation of these mental images. In this way, cognition is incessantly embodied in these mental maps, which are the building blocks for emotional feelings. An emotional feeling results when the brain makes maps of these mental maps interacting with dispositional circuits, which are the equivalent of Panksepp's basic drives. For a review of the evidence for topographic mapping in the sensory areas of the brain, as well as an interesting discussion in its possible functions in higher cognition, see [15]

The key difference between Panksepp's account of the mind and Damasio's is that the latter's requires a 'cortical readout' [12] of the representations of physiological changes in order for them to become consciously experienced emotional feelings. In his terminology, emotions are the basic physiological changes that are experienced (I refer to these simply as somatic states, however). The feeling of an emotion, as with Panksepp, is the first-order representation of these physiological changes interacting with dispositional circuits, which lends the somatic representations their valence. However, Damasio posits the need for second-order representations located in the neocortex for conscious experience: "[it is] only thereafter that an organism that is responding beautifully to its environment begins to discover that it is responding beautifully to its environment" [13]. Damasio's overall model presents a protoself, a core self, and an autobiographical self, the respective functions of which are summarized in figure 3.

Autobiographical self
Memories generate a core self that is embedded in memory
and linked in a large-scale coherent pattern
Core self
Maps of the body are brought into relation with maps of the environment
Protoself
Primordial (raw) feelings are generated from body maps

Figure 3: Damasio's stages/levels of the self. Adapted from [14]

The stage of conscious awareness occurs in the core self, which does not yet require the use of memory, for instance. The higher cognitive functions only come into play in the autobiographical self, which places the core self into the context of language, life experiences, and other functions which Panksepp would label tertiary processes. It goes without saying that there are reciprocal interactions between each of these layers.

4.2.1 The somatic marker hypothesis

Against the backdrop of his theory of consciousness, Damasio proposes a theory of the influence of emotion on decision making which is essentially a more comprehensive neuroscientific version of Norbert Schwartz' feelings-as-information theory [9]. Many of the postulates of the feelings-as-information theory can be explained in terms of the somatic marker hypothesis (SMH). The most key notion of the SMH is that emotions provide a valuable source of information for making fast and advantageous decisions in complex environments. Previously, theories of human decision making, especially in economics, assumed that people were rational calculators of some utility function. Emotions were ignored in the decision process, partly because of a lack of a useful framework in which to specify them, as well as their relationship with higher cognition. However, it has become abundantly clear that one can not simply model away the very substrate of human cognitive abilities and only focus on secondary or tertiary processes.

According to the SMH, every object and event of the environment or imagination has the ability to induce a set of physiological reactions in both body and brain. Damasio calls the automatic bottom-up (stimulus-driven) causes "primary inducers" and the top-down (imagination- driven) causes "secondary inducers" [16]. The primary inducers should be thought of as activating those evolutionarily shaped primary-process emotions which Panksepp stresses - These reactions, perceived in a positively or negatively valenced manner, guide a person's decision about a situation currently in their attention. In the body, reactions can take the form of, for instance, muscle contractions and heart rate changes. In the mind/brain, reactions can be fourfold [16]:

(a) Neurotransmitters are released, which can affect working memory (e.g. dopamine, serotonin, noreadrenaline, etc.)

(b) Somatic maps of the body ("as-if" states) in the somatosensory cortex are modified

(c) Transmission of signals from body to somatosensory cortex is modified

(d) Motor system is biased or inhibited for certain actions

Emotional reactions are therefore more than a matter of mere neural signaling in the mind and physiological reactions in the body proper, but also a matter of a modulation of how information is processed neurally by neurotransmitters. Overall, the different possible causal interactions in the theory are displayed in figure 4.

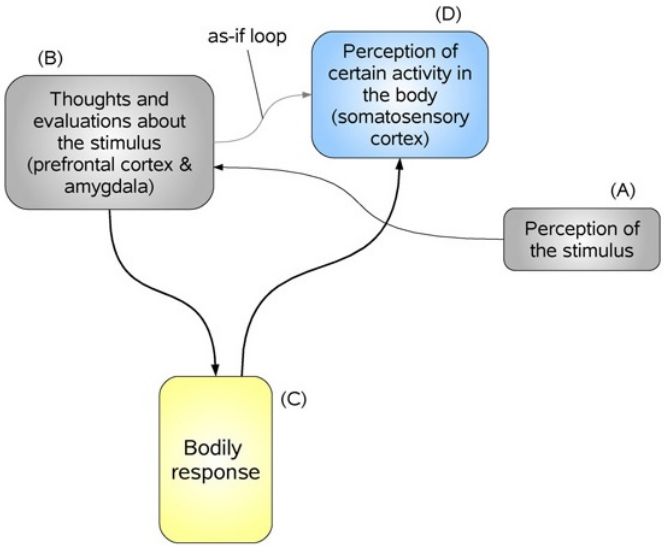

Figure 4: Functional areas of the somatic marker hypothesis. [8]

One of the most significant brain areas in this interaction is the ventromedial prefrontal cortex, labeled 'B' in figure 4. It is in this area that abstract conceptual associations with concrete emotional states are saved. Based on this architecture, there are roughly speaking two possible chains of events: a "body loop" and an "as-if loop". [16]. The former denotes regular perception of a stimulus (A), thoughts and evaluations of the stimulus (B), the bodily response (C), and finally the perceived feeling of said activity (D). The latter is a form of offline cognition in which there is no stimulus and the body is bypassed, but where imagination causes thoughts and evaluations of a concept (B), which has the ability to cause an 'as-if' feeling in the somatosensory cortex (D). Of course, these two pathways are usually not separable, as an actual stimulus will also involve the 'as- if' part of the loop. In general, however, the body loop will cause significantly stronger reactions than the 'as-if' loop. Intuitively, this is to be expected: the thought of a crocodile advancing on one's position shouldn't feel as scary as an actual crocodile advancing on one's position.

At this point in the story, however, the question remains as to how these emotional mechanisms bias higher cognition beyond creating a phenomenological experience and influencing immediate biologically ingrained reactions. The answer lies in the major neurotransmitter systems (e.g. dopamine, serotonin, noreadrenaline, and acetylocholine), which are connected directly from the brain stem to brain systems underlying decision making and working memory. At these sites of higher cognition, the neurotransmitters which are released on the basis of emotion-based stimulation modulate the synaptic activity of neurons

affecting higher cognition. It is by this mechanism that the contents of working memory are influenced, leading to "objects" and "response options" being either endorsed or rejected - in other words, neurotransmitters "help bias the options for plans and action" [16]. At this point it should be stated that the evidence for the SMH is not unequivocal. Most of the supporting evidence comes from the Iowa Gambling Task and a handful of other experimental paradigms - in other words, the hypothesis has not been extensively tested. Furthermore, the apparent confirmations of the theory have alternative explanations in more simple reward and punishment mechanisms via the prefrontal cortex and amygdala. For a critical discussion of the empirical evidence, see [18].

4.3 Summary

In this section I will reiterate some of the chief points of agreement between Damasio, Panksepp, and the general neuroscientific community on the issue of the function of emotions and their relation to higher cognition. Furthermore, my own conclusions will reinforce what I consider to be lessons that cognitive science can learn from affective neuroscience. These principles will form the basis of my argument that abstract cognitivistic models are insufficient to model brains/minds and general intelligence, and that the computer metaphor of the mind should be replaced by more suitable naturalistic models that are able to integrate the implications of embodiment and emotions. Furthermore, these principles should all be valuable in designing artificial agents, even though it may be far from straightforward as to how one could integrate them into an implementable cognitive architecture. First, however, one major point of disagreement between Damasio and Panksepp should be highlighted, which is mostly relevant to the issue of consciousness: the issue of cortical readout of emotions. Essentially, researchers in cognitive neuroscience, such as Damasio, usually subscribe to the notion that in order for body representations to be consciously experienced they must be transferred to second-order representations in the neocortex. This would imply that many mammals do not in fact experience emotions subjectively, but only exhibit emotional behaviours [12]. Panksepp and other affective neuroscience researchers disagree with this notion, and urge the use of animal models for basic affective experience in humans. For the purposes of my argument against cognitivism and the computer metaphor of mind, however, this difference is moot. Neuroscience as a whole has some very strong points of agreement on the embodied function of emotions, which I would like to highlight in the following.

4.3.1 The hierarchical brain

The following may seem like a platitudinous statement to many, but it is worthy of highlighting in any discussion of cognition. It is generally accepted that the brain is organized in a more or less hierarchical manner, with basic emotional and perceptual mechanisms operating on a low level and higher cognitive functions (located in the neocortex) operating on a higher level. In the previous sections, this notion has been reflected in the primary, secondary, and tertiary processes of Panksepp, as well as the protoself, core self, and autobiographical self of Damasio. This implies that cognition is intrinsically layered, and that there are causal forces operating between each of these layers. The more simple layers accomplish more basic tasks such as directing motion, perceptual processing, and providing an interface to the body. This is where the bulk of emotional processes take place. As one moves upwards in the hierarchy, the functions become more complex because of an increase in neural resources which can learn a larger variety of behaviors than their subordinate levels - the evolutionary impetus for providing mammals with the neocortex. Thus, our current brains are the result of control systems iteratively packed onto each other, each layer providing an additional abstraction level for the parameters of the control system beneath it. [19]

This has several advantages, including a resistance to damage or change (robustness) as well as the ability to dynamically rewire parameters in an inter-level fashion (adaptability). [19] In this fashion, humans have been able to develop a large amount of self-control, namely for the purposes of advantageous social life. Culture, art, and science are all the result of increasingly abstract layers of thought grounded in, ultimately, in the most basic somatic, motivational, perceptual, and motor systems. Of course hierarchy is just one organizational principle in the brain, and a notion of hierarchy in neural networks is not nearly sufficient to describe the emergence of the complex behaviors that we have. It is, however, a necessary principle, and an oft overlooked one.

The symbols, schemas, plans, and rules of cognitivism abstract away from the innate hierarchical nature of cognition, and thereby forgo the robustness and adaptability advantages such an architecture provides. Furthermore, cognitivism attempts to model tertiary processes directly, and in abstract fashion, without differentiating between more basic and more abstract thought. A consequence of this is a stark rigidity of possible behaviors/concepts, as well as the symbol grounding problem [20]. In a hierarchical model of cognition, however, symbol grounding is natural - the perceptual and motor faculties at the bottom provide it, as Barsalou [3] describes in his perceptual symbol system model. Within such a framework, I contend that the emotional component of

meaning comes naturally – an abstract symbol can elicit somatic states, which, along with sensory inputs, are also the basis for embodied conceptual representation in a perceptual symbol system. These basic somatic states interact with low-level circuits which dictate the cognizer's most urgent needs, providing a bedrock of meaning upon which further cognitive functions are built.

4.3.2 No modularity

In cognitive science a concept based on the computer metaphor of mind that one encounters relatively frequently is Jerry Fodor's notion of modules and transducers. According to Fodor [21], modules are localized functional areas of the brain which possess nine key characteristics, amongst which are domain specificity, informational encapsulation, and impenetrability by other cognitive processes. Fodor [21,22] posits that only low-level processes (primary processes such as perception, action, and emotion systems) are modularized, and that secondary and tertiary processes are implemented by a general purpose processing system. Transducers are low-level conversion units processing sensory data into a ready-made input format to the modules, in this vision. According to affective neuroscience, however, it seems that both notions are unwarranted. Fodor did say that the notion of a module may be useful to cognitive science insofar as the nine properties are fulfilled to "any significant extent" [21] - however, the above mentioned three properties seem to be an inappropriate description of basic emotional mechanisms. I focus on these three because they are key in the notion of a module, and because emotional systems seem to not fit this description.

The perspective of Jaak Panksepp [10] would give a certain amount of credence only to the first claim - that emotional processes are in some sense domain specific. He states that SEEKING, FEAR, RAGE, and other primary emotional circuits are discrete and do fulfill quite 'specific' functions conducive to the survival of the organism. This is what Panksepp means when he says they are "modularized" [10]. However, it is questionable (as outlined in section 4.1.1.) whether a discrete description of basic emotional circuits is even possible at any realistic level of detail. The notion is highly debated in the field therefore, and the discreteness of basic emotions is not accepted by many neuroscientists. Furthermore, whether one adheres to Panksepp's notion of basic emotions or not, they do not fulfill all of Fodor's requirements to be called modules; they would not be encapsulated if they existed, for instance. What is quite sure is that basic emotions are neither informationally encapsulated nor impenetrable by other cognitive processes - they strongly interact with (and are modulated by) somatic representations and higher cognitive processes. Furthermore, mod-

ules are meant to be physiologically distinct [21], whereas basic affective circuits are largely overlapping. [10] Basic affective mechanisms cannot be transducers either, because their function is not just low-level input and output, but also a global homeostatic one (and more).

4.3.3 Simulation

One of the key reasons why primary process affective mechanisms cannot be modules, at least not in a Fodorian sense, is because of their strong integration with higher cognitive processes. As is emphasized in Damasio's somatic marker hypothesis, offline cognition can produce simulations of affective states via the 'as-if' loop. Even if the somatic marker hypothesis were to turn out to be incorrect in its current formulation, it is clear that conceptual recall involves some kind of reenactment of affective states. Simulation is a proposed neuroscientific theory which operates at many levels of cognition - initially, it was started with the discovery of "mirror neurons". These neurons are activated in the process of observing external behavior of a third party, causing an inferred emotional reaction in oneself [23]. Supposedly, this mechanism underlies the ability to apply folk psychology, i.e. to infer the thoughts of others. This ability is also known as the theory of mind (TOM). It has been suggested [23] that this process operates at a conceptual level; that observed behavior triggers a flow of imagination which tertiary level processes deem relevant to the trigger, allowing for an inference about the autobiographical self of the observed individual. Naturally, as with any proper scientific investigation, the function of mirror neurons is debated [23].

However, the simulation notion has been highly influential, and not just for explaining TOM. It has been applied to explain the very nature of concepts grounded in modality, which seem to have an undeniable primary process affective component. Barsalou's perceptual symbol system [3] is a quintessential formulation of this notion. In his view, supported by a large amount of empirical research, the activation of an abstract 'symbol' is passed in a top-down manner through the various levels of control in the mind to ultimately reactivate an amalgamation of parts of previous sensory experiences. In this way, concepts are very much fluid, as Douglas Hofstader passionately argues [24]. Simulation can thus be viewed as the top-down embodied reenactment of some abstract conceptual symbol, undeniably involving affective associations. This stands in contrast to the notion of an amodal symbol system which is so often advocated by cognitivists within the framework of a computer metaphor of mind, for which there exists little to no empirical evidence at all [25]. It should be noted that a main competitor to the simulational account of TOM is the so-called theory-

theory. This perspective essentially grounds TOM in the analytical ability of people to discern various folk-psychological states in others, and to connect them in a coherent theory of their mental processes. Whilst this stands in contrast to the psychological reenactment that simulation theory posits, the two do not necessarily have to be mutually exclusive in humans. For artificial agents, however, the latter has naturally been employed. Architectures based on the belief-desire-intention model of cognition are quite widespread, and as such are much more amenable to the theory-theory. For an overview of formalizations of TOM in artificial intelligence, as well as an approach based on propositional dynamic logic, see [52].

It is doubtful, however, whether rules and propositions are sufficient for describing human cognition, especially emotions. The simulational account of TOM is backed by a large amount of empirical evidence, such as for mirror neurons. The perceptual symbol system that simulation theory goes hand in hand with also provides an epistemologically coherent and naturalized account of the mind. Furthermore, the rules and representations of a theory-theory starkly reduce the complex interactions that take place in the mind/brain: as is the focus of this work, this applies especially to emotional states. I do not assert that theory-theory should be abandoned – rather, I contend that simulation theory provides a more realistic account of TOM (as well as other mental phenomena).

4.3.4 Dynamic layered topographical representation

In addition to evidence for embodied simulation, neuroscience has uncovered the use of topographic maps (ordered projections) for representing the body (egocentric) and the environment (allocentric) [26]. These maps are constantly updating, such that the mind perceives the body to be a continuous (over time) entity in a continuous environment, the latter of which is segregated into meaningful entities determined by perceptual filters which are themselves maps upon maps. Both Damasio and Panksepp agree that emotion is in part the result of somatic maps (of the whole body) interacting with dispositional circuits which dictate basic drives. Furthermore, maps are layered and interconnected, such that various properties of the changing environment can be represented, such as spatial and temporal relationships between objects and self. [14] Interconnected sensory maps, representing an integration of topographically organized information received about the body and the environment, can be seen as the basic unit of representation in Barsalou's perceptual symbol system. The complex nonlinear interactions [29] therein present a further problem for processing by rules on structures which often features in cognitivistic models of the mind. Damasio [14] justifiably argues that mapping is a basic principle upon which

cognition is built, and it certainly is a valuable guiding principle for embodied models of cognition, both natural and artificial.

4.3.5 Reward and punishment grounded in homeostasis

One of the central functions of mapping the body is the maintenance of the complex equilibrium of physiological processes known as homeostasis. Biologically speaking, maintaining this process is one of the central goals of cognition, and should certainly feature in a holistic set of theories which attempt to explain natural cognition in all its facets. The question one could pose, however, is whether it is really necessary to model homeostatic interactions for general intelligence, and furthermore, for which specific cognitive phenomena this aspect should play a role. The notion that brains and other intelligent systems need to be embodied can be easily defended on the grounds that cognition is essentially bilateral interaction with an environment. However, where does this leave the notion of homeostasis? Arguably, part of the answer can be found in the reward and punishment mechanisms which emotions implement. From a top-down perspective of cognition, there needs to be some standard or set of ideals by which higher processes operate, for which they can be rewarded for following, or punished for disobeying. This is essentially what homeostasis is to brains. Without having a notion of purpose, one cannot have a notion of intelligence. I argue that for brains and behaviour, the modeling of homeostasis (to some extent) is key, because cognition and homeostasis are biochemically linked, quite intimately so. Watt [27] displays the general relationship as such:

Homeostasis
⇕
Emotion
⇕
Cognition

This shows itself not only in emotionally valenced behaviours, but also in various other cognitive phenomena, such as mood disorders. [27] In general, the physiological state of the body can induce quite varied styles of cognitive processing. One could ask, additionally, in what cases embodied artificially intelligent systems need to model their body at such a level of explanation. Perhaps for artificial general intelligence, many other types of non-human bodies could suffice, depending on what environment the system finds itself in. Such systems could even depict human-like intelligence by replacing the heomeostasis component of reward/punishment with a more general notion of predictive equilibrium, as will be discussed in section 7. Whether such ventures would be

successful, however, is of course highly speculative. What is certain is that if science wants to understand human behavior, or even attempt to teach machines to do so, it should not ignore the homeostasis which underlies our emotional experiences.

4.3.6 Neurochemicals

Back in the brain itself, neurochemicals have been shown to influence a wide variety of cognitive and brain phenomena, from neural plasticity to emotional feelings and moods. [28] They are clearly integral to explaining human cognition, but are completely neglected by information processing approaches traditionally employed in cognitive (neuro)sciences. Similarly, an equivalent component may be necessary for neuromorphic architectures to dynamically modulate their behavior. Insofar as modeling the emotional brain is concerned, however, it is clear that neurotransmitters are invaluable. For instance, the key mechanism by which emotional processes bias decision making is found in the major neurotransmitter systems, the cell bodies of which are located in the brain stem (which receives somatic signals from the body) and whose axons converge all over the telecephalon, which has been implicated in decision making and other higher cognitive tasks. Additionally, there are connections between the prefrontal cortex and amygdala and these cell bodies, allowing for neurotransmitter release during offline cognition. [16] This is part of the mechanism by which Damasio's 'as-if' loop operates, and, more generally, one mechanism by which imagination can implicitly modulate our feelings and behaviour. For these reasons and more, it is imperative to model human cognition, especially emotional cognition, below the neural level. Perhaps similar mechanisms will prove indispensable to neuromorphic artificial models of cognition as well.

4.3.7 Nonlinear dynamics

The interactions between the brain and environment that are stressed by an embodied perspective have, amongst other intra-brain phenomena, drawn researchers to modeling techniques originally from physics. It has become an increasingly popular realization in the field of brain and mind research that transitions between 'mental states' are impossible to describe by some number of rules operating on the current state, because the brain is in many ways a complex system which is open, contains feedback, is time-sensitive, and where the relationship between cause and effect is not proportional. [29] This is especially true for the interactions that determine emotional behavior, as theorized in the somatic marker hypothesis. The feedback interaction between body, prefrontal cortex, and somatosensory cortex is highly unpredictable and constantly modu-

lated by stimuli from the environment and higher brain areas. This is true even for a pure neurocomputational perspective, where lower levels of modeling (such as the neurochemical one) are not yet regarded. Panksepp states that "the basic emotional systems may act as "strange attractors" within widespread neural networks that exert a certain type of 'neurogravitational force' on many ongoing activities of the brain." [30] It is becoming increasingly clear that the system from which all of the rather impressive human capabilities emerge from is intrinsically complex, such that cognitive science should not continue attempting to compute cognition in a linear fashion. The linear systems approach, however, is a staple of the cognitivistic perspective of the mind.

4.3.8 Plasticity

Finally, a notion not stressed very much in the writings of either Damasio or Panksepp, but very important in any discussion of the mind is that of the brain's plasticity. Many models of cognition implicitly assume a static neural substrate, which is simply not the case. The brain is highly adaptive, in part, because of its ability to modify its own architecture as a response to experiences. This is reflected in the paradigm of neuroconstructivism, which asserts that "[the] architecture of the brain and the statistics of the environment [are] not fixed. Rather, brain-connectivity is subject to a broad spectrum of input-, experience-, and activity-dependent processes which shape and structure its patterning and strengths...These changes, in turn, result in altered interactions with the environment, exerting causal influences on what is experienced and sensed in the future." [50] Studies, correspondingly, have demonstrated a "profound impact of environmental events in shaping the neural circuitry of emotion." [51]

5 Is the embodied emotional brain a computer?

On the basis of the principles outlined in the previous section, I will now discuss to what extent emotion can be viewed as computation.

5.1 Propositional attitudes

A central theme of many cognitivistic approaches to mind is the notion of a propositional attitude. Although propositional attitudes have been criticized by many researchers from many areas of cognitive science, I will give my own brief indictment thereof on the basis of the notion that embodied emotions as discussed above provide a substantial component of what we call 'attitudes'. First, however, a definition is needed. According to Paul Churchland [31], a

propositional attitude denotes the relation that a mind holds towards an abstract proposition expressing some state of affairs. This could be a belief that 'x' holds, or a desire to bring about 'y'. Thus, propositional attitudes can be expressed as binary predicate statements such as 'a believes x' or 'b desires y', and can consequently be embedded into a logic expressing general law-like relations between the attitudes, because one can quantify over both term positions. This system seems neat and homologous to statements often formed in folk psychology, but is unfortunately scientifically untenable. A well-known version of propositional attitudes can be found in the belief-desire-intention model of human practical reasoning developed by Michael Bratman. [53] His philosophical theory attempts to explain human practical reasoning in terms of epistemic constructs such as beliefs, desires, and intentions – which are intended to be states of mind. Influential especially in artificial intelligence, the BDI model of human reasoning has generated the BDI agent architecture. In such software, the current 'state' of the agent can be summarized in a list of predicate statements denoting its current epistemic attitudes.

One can attack propositional attitudes on many grounds, not the least of which being the inexistence of such a thing as an abstract proposition - there exist utterances, which are merely strings of symbols until they are interpreted. Furthermore, their interpretation (and relations between these interpretations) are utterly resistant to formalization, because they are the result of the complex (affective) interactions between brain and environment. Attitudes should therefore not be described in logic. Churchland [31], for these reasons, advocates an elimination of all folk psychological description of attitudes in favor of a neurocomputational approach. Whilst I share the general sentiment of stressing the neural level, I do not believe folk psychology to be totally false, nor do I see the neurocomputational level of description as the definitively cognitive level. In general, it seems that ascertaining which level of explanation is definitely cognitive is a lost cause – for every level of granularity, there will be exceptionally detailed cognitive phenomena which require a lower level of description, as well as more coarse phenomena that do not require such depth. I have stressed that emotional processes are at the core of how brains interpret higher symbols and decision options in a valenced manner, i.e. they are the basic determinant of our attitudes and the foundation of cognition.

Considering the overall phenomenon 'emotion' then, for instance, one could ask what the definitively emotional level of explanation is. On the one hand perhaps the beliefs, desires, and intentions of folk psychology can be validated to some extent as attractors in the complex emotional processes as described on a neurocomputational level. On the other hand, it may be that the neurocomputational level is insufficient on the grounds of subneural phenomena, such as

the influence of neurochemicals - which seem indispensable for emotional regulation. Making these assumptions, one could say that all levels are valuable because they are not necessarily completely reducible to eachother and each contributes a subset of explanations of the total set of cognitive phenomena. Even assuming the potential for complete reduction, it would make sense to preserve all levels because each may provide the most efficient means of description of certain phenomena. Besides, it is generally speaking far from clear which levels of description are relevant for describing emotional attitudes, especially also because they are intrinsically in a feedback loop with our physiology and the environment.

5.2 Computational theory of mind

The question of if and how cognition can be described computationally has been central to discussions of the mind and brain since the beginning of the computer science field. If the complex nature of cognitive processes leaves natural language and mathematical logic out of the question as modeling tools, then how complex would a computational architecture of the mind/brain have to be? I assert that the drive in cognitive science research to produce a single underlying architecture is misguided, and fuelled by the metaphor of the brain being a computer instead of a complex open physical system. The computational position has spawned the levels of analysis of David Marr [33], which essentially reduce brain and behavior to three explanatory levels:

Computation

⇕

Algorithm

⇕

Implementation

This framework has been highly influential in cognitive science, and has become the default reference for how one can explain cognition. It is, however, intrinsically married to the computational theory of mind (CTM). If the framework is applied to the mind as a whole, the algorithmic level implies that there exists a set of representations (mental states), also known as a language of thought, or 'mentalese' [34], over which are defined a number of production rules to produce new representations. This is referred to as the CTM, which essentially equates the brain with a Turing machine, or a weaker form of symbolic computation. [35] The question of whether the brain implements any algorithm, however, is an empirical one: in order for it to do so, there must exist a mapping of physical states to representations, and the brain must follow the sequences of states over time that the production rules dictate. This will be described in slightly more detail in the next section. It seems highly unlikely, in the face of the evidence, that it can ever be shown for the complex nonlinear system that the brain is [29], however. Another reason to reject Marr's levels as giving an account of general cognitive architecture is that an account of the algorithmic level requires an account of the computational level - that is, one must know what is being computed. However, one hallmark of our cognitive abilities is to adapt to arbitrary environments and tasks.

Thus, I contend that there is no obvious computational or algorithmic level for emotions, not to mention other domains of cognition. Until there is, it seems premature to adopt the computational levels of description and apply them generally across all cognitive domains. Perhaps for certain subsets of cognitive abilities that are more straightforward computationally, such as those of the visual system described in Marr's original work [33], it is sensible to adopt such levels of description. However, as a scientific framework to explain the brain in its entirety, these levels of description are clearly lacking. The emotional processes described in this work, for instance, do not admit to such a description. Furthermore, they are relevant both for the 'rational' processes of cognitive science and the 'irrational' processes described, for instance, by psychopathological research. For both fields, it would be a disservice to adopt Marr's explanatory framework in favor of many more levels of explanation able to capture all the levels of organization in which phenomena appear which one wishes to explain. The brain is, after all, not a computer but a complex system

shaped by the laws of nature instead of design.

5.3 Computational explanation versus computational simulation

The above argument is entirely compatible with the 'weak AI' thesis, which states that an algorithm can, at most, simulate any arbitrary cognitive process, but that it can never be a cognitive process in the way that brains implement them. I assert that the embodiment of emotional processes supports this thesis. The interaction between environment, body, and brain that determines emotions and other cognitive phenomena is highly complex, such that it is questionable whether any one algorithm will be able to explain the properties of the brain in their entirety. This is why the crucial difference between computational simulation and computational explanation should be stressed. I conceive of the former as the practical use of a computer to implement some theory of cognition, and of the latter as the scientific theorizing of the brain as a particular computational symbol system. If computational explanation were possible for the totality of cognitive phenomena, the 'strong AI' thesis would be true.

I caution against the use of computational explanation for any domain (not to mention the entirety) of cognition, however, without appropriate empirical evidence. Furthermore, I contend that this evidence must come in the form of a demonstration of a strict notion of implementation of some algorithm by the physical brain. I say 'strict' because weak interpretations of what it means for a physical system to implement some computation usually results in relatively useless notions such as that every rock implements every finite state automaton (FSA) [54]. This conjecture is utterly devoid of explanatory value, because it necessarily employs a notion of implementation that is too easily satisfied – by just about any physical system. Specifically, it was raised by Hilary Putnam [55] as a criticism of computation as an epistemological foundation for the study of the mind. However, a rock only implements every FSA if one is too flexible with how physical states are mapped to abstract computational/representational states. In Putnam's construction, a rock implements every FSA because for any period of time, it will transition through a number of unique states which can be interpreted as causally mirroring any FSA's states. This argument relies on the noncyclical behavior of the physical system, which Putnam contends is axiomatic. However, the similarity between the physical system and the FSA is only superficial in such an interpretation, because it does not take into account counterfactual states. Clouds happening to take the shape of a smiling face at the same time as one's own happy thoughts usually don't lead one to think the sky shares any similar emotional or cognitive pro-

cesses - but one might reconsider mocking New Age culture if they smiled if and only if one thinks something happy.

I therefore adopt the stricter notion of implementation advanced by David Chalmers [54], which necessitates counterfactual satisfaction. Informally, he states that "a physical system implements a given computation when there exists a grouping of physical states of the system into state-types and a one-to-one mapping from formal states of the computation to physical state-types, such that formal states related by an abstract state-transition relation are mapped onto physical state-types related by a corresponding causal state-transition relation." [56] The physical state-types, together with their causal state-transition relation, must thereby follow the computation for every configuration, not just for one possible eventuality. It must be (theoretically) impossible for the physical system to disobey the computational laws it supposedly implements. An illustration of correct versus insufficient implementation under this definition can be found in figure 5, where the colors indicate a mapping. In both the implementation and the non-implementation, it is possible to reproduce a run of the abstract computation. However, the non-implementation does not support counterfactual conditions. In the non-implementation, the physical system is able to disobey the abstract computation by jumping from A' to C with event e2. Thus, applying e1 would not have the consequent state expected by the abstract computation anymore. This event (e2) is allowable in the correct implementation, because it only leads to states A or A', which are equivalent to state A in the abstract computation, and thereby have all the same consequent states with respect to e1.

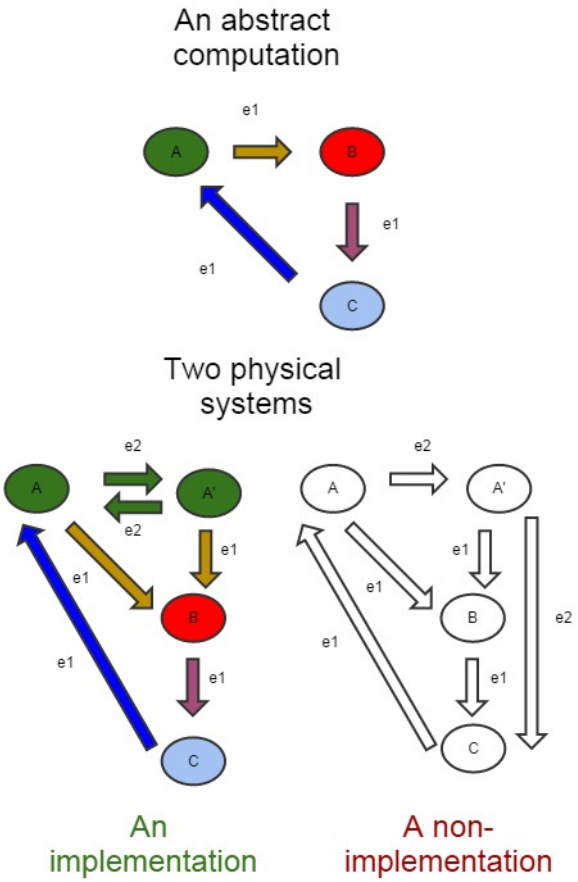

Figure 5: Valid and invalid implementation

This definition of implementation can be made technical and concrete by employing a specific model of computation, such as the combinatorial state automata that Chalmers further introduces [54]. However, the notion of states and transitions is sufficient for this analysis. Before representational states and rules internal to the brain and explaining all of cognition are empirically identified, it is wrong to say that the brain is a computer. Only an identification of intrinsic rules and representations lend a CTM explanatory power, which is what we want of a scientific theory. [36] Until this happens, brain science should not become an applied computer science.

That does not mean, however, that computers aren't indispensable in advancing our scientific knowledge of cognitive processes - computer simulations have been extremely helpful in a wide variety of natural sciences. The question of whether a certain cognitive phenomenon one wishes to describe admits a computational description by simulation is one of degree, however. Computer

simulations are approximate, and their success depends on the degree of accuracy with which one wishes to describe the phenomenon in question. Since embodied processes central to emotional processes and higher cognition operate at a very detailed level of description, lower than that of neural networks, it seems that current connectionist computational frameworks are insufficient to describe affective processes appropriately. Additionally, it is questionable whether any one simulation will be able to describe affective phenomena for a wide range of applications - simulations depend very much on what property one is simulating. As for the CTM, unless an empirically based functional decomposition of affective (or other) processes is discovered, the thesis cannot hold as a scientific explanation. That is, the statement that 'the mind is a computer' should be replaced by 'the mind is a complex physical system' until a mapping of physical states to representations, along with production rules, are identified. Considering the complex feedback interactions one finds in all manner of brain processes, for instance in the somatic marker hypothesis, this seems unlikely.

5.4 Summary

Cognitive sciences have increasingly come to realise that without the basic emotional building blocks, human and artificial models of cognition will never be able to represent the drives that underlie all complex human behaviours and cognitive capacities. With emotions being a complex system of embodied interactions, the notion of propositional attitudes is clearly outdated. Furthermore, there seems to be no clear set of symbols and syntax operating to produce these intimately physiological phenomena, such that the computer metaphor seems equally outdated in describing the brain, which, in reality, is a complex biological system. For this reason, the next section is dedicated to addressing the foundational issues that arise from the switch in epistemological perspective from computer to complex system, especially how this changes which levels of description science should employ to describe cognition. Emotional processes will be situated within these new explanatory levels, and it will be discussed how artificial intelligence should draw architectural principles from this framework in an application-dependent (or general) manner.

6 Explanatory pluralism in cognitive science

As Karl Popper once remarked, "we are not students of some subject matter, but students of problems. And problems may cut right across the borders of any subject matter or discipline." [38] In many ways, this has been the credo of cognitive science, which has often been simply defined as the "interdisciplinary

study of the mind, embracing psychology, artificial intelligence, philosophy, neuroscience, linguistics, and anthropology." [39] However, since its inception in the 1950's, the field of cognitive science has had a strong tradition of attempting to reduce cognition to some form of (often symbolic) computation. This has undoubtedly been fueled by the strong computational backgrounds and interests of some of the pioneers in the field - people such as Marvin Minsky, Allen Newell, and Herbert Simon. Each had made highly significant contributions (essentially creating the fields) to cognitive psychology, cognitive science, and artificial intelligence. Whilst having both interdisciplinary training and mindsets, their approaches all had in common a goal of computational modeling of cognitive processes.

As such, there was always an implicit belief in the reduction of psychological processes to an algorithm. This was made explicit, for instance, in Newell's call for "unified theories of cognition." [40] According to Newell, the scientific study of psychological processes could only profit from a unified framework which could explain as much as possible of available experimental data - and when it would be significantly falsified, it could be recreated. His Soar architecture [40] is an example implementation of such a framework, and has been used by researchers in psychology, cognitive science, and artificial intelligence. The inability of such classical (symbolic) models to explain cognition then spawned connectionism, a paradigm which models cognition in abstracted neural networks. [41] This paradigm is also highly reductionist, however, and makes a large number of assumptions both about the functional properties of neural networks and their sufficiency in producing cognition. The affective processes described in this work, which are indispensable for minds, make use of subneural and embodied phenomena not describable by connectionism. The new field of computational neuroscience comes closest to providing modeling at levels as low as the molecular - however, within such a paradigm, it becomes questionable if high-level phenomena can be explained, or how the environment is to be integrated. Essentially, computational neuroscience is also a form of reductionism.

As stated earlier, I believe the notion of a general computational framework of cognition to be impossible. In explaining a specific cognitive phenomenon, it is of course possible and desirable to use computational simulation to make valid predictions. However, such simulations will always be restricted to a subset of cognitive phenomena, and will only be as good as the physical theory underlying them. Setting aside the issue of computation, the general lack of any one distinctly cognitive level of explanation has led a number of researchers to propose an epistemological framework for the cognitive sciences which has been at the heart of the study of complex systems in other natural sciences for a long time: explanatory pluralism. [42,43] Most essentially, higher (by measurement

scale) levels of explanation are constituted by more complex entities, which are in turn composed of entities from lower levels of explanation. Each level features its own set of mechanistic rules, which may translate in a straightforward manner such that reduction is possible. However, complex systems often feature emergent phenomena - those which cannot be described as the sum of their parts. In the brain, such phenomena seem plentiful, such that a multi-level explanatory framework is indispensable. An overview of some of the levels relevant for cognitive science can be found in figure 6. Notice that of the levels that are displayed (which are by far not exhaustive), the greatest common level of description between brain and body is the molecular one. Thereby, assuming one works only with these levels, brain-body interactions require a molecular level of description.

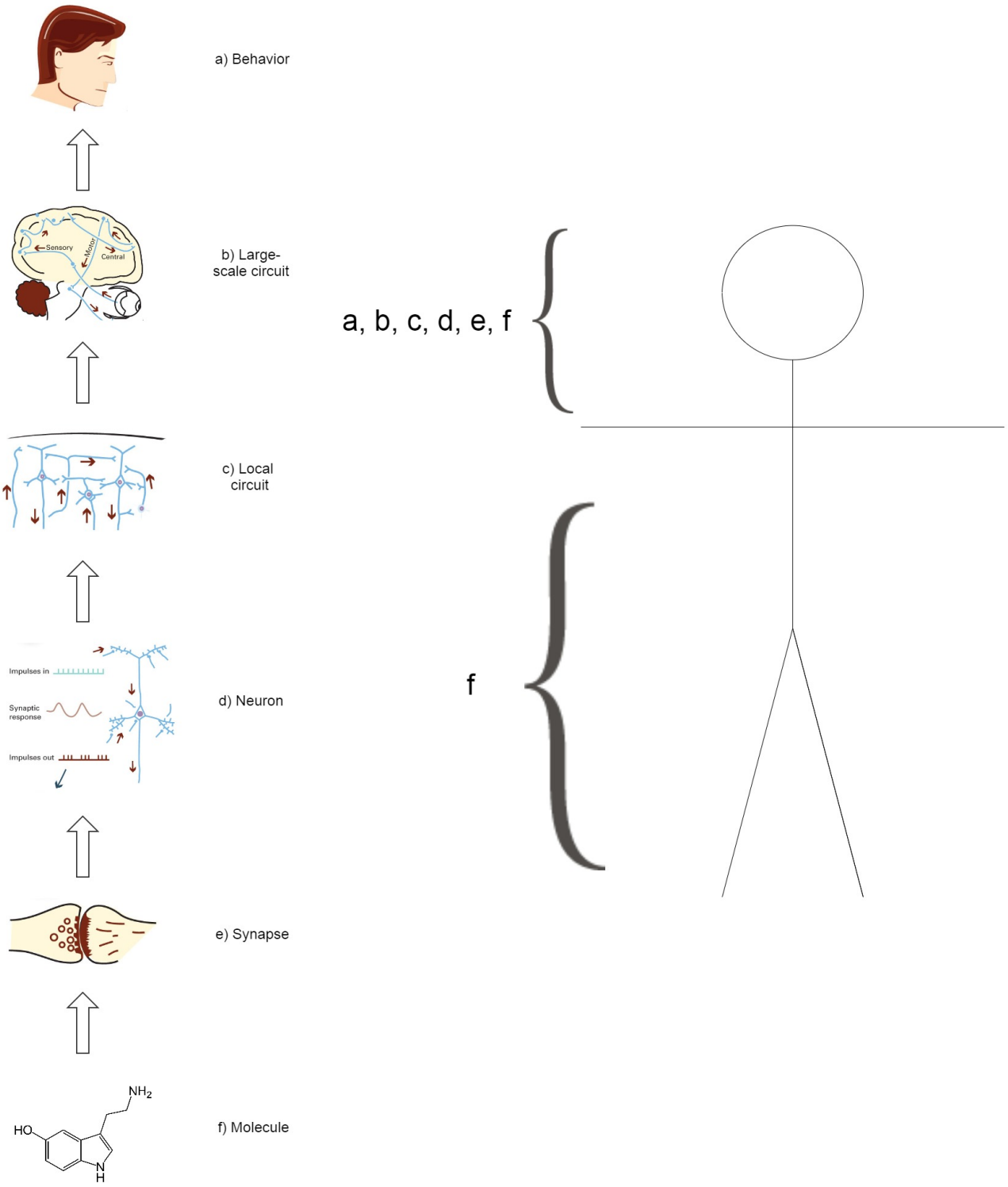

Figure 6: Possible levels of explanation in a pluralistic cognitive science. [45]

The emotional processes described in this work can be explained at the behavioral, systems/circuit, and molecular level. In order to describe how the somatic marker hypothesis functions, for instance, one must explain the behavioral modification of somatic marker, but also how they are actually implemented on a neural and sub-neural level. These latter two require reference to neurotransmitter systems, basic affective and somatic circuits, as well as the effects of the neurotransmitters themselves at the synaptic and molecular level.

Many aspects of the mind are argued to be irreducible to the brain, also known as the position of property dualism, so perhaps one could posit an additional level of explanation on top of the behavioral. [44] Applicable to each level of description, theories are also stratified horizontally into different domains, and for each domain and its sub-domains there may be many competing or complementary theories in existence. These domains could be vision, memory, perception, embodiment, language, emotion, social cognition, and so forth - and one could come up with many more nuanced subdivisions. Explanatory pluralism holds that this heterogeneity is to be celebrated, as it allows for plenty cross-fertilization between levels, domains, and theories. Thus, in explaining a certain cognitive phenomenon, the researcher should analyze theories and methodologies of those horizontal domains which are relevant for the phenomenon. Furthermore, how high or low of a level the theories from these domains are drawn from depends upon the granularity required for describing the observations of the phenomenon in question. Lastly, it is important to emphasize that there is no specific computational level in such a framework. This goes hand in hand with my earlier argument in section 5.3 that computation should be considered a tool for simulation until it can be shown to be a tool for explanation. Any given set of theories may be converted into an algorithm, even in combination - but as of yet, there is no convincing reason to suppose an algorithm should feature as a theory.

7 Naturalistic artificial intelligence in three territories of research

For posterity's sake, it should be stated that this work is in agreement with the so-called 'weak AI' thesis'. [46] This position states that a computer program can at most simulate cognitive processes, but not emulate them, such that a computer can never be fully intelligent in the way a human brain can. I contend that this is true on the grounds that a computer simulation will never be able to reproduce the complex physical system that the brain is, at least not in its entirety. It may become tantalizingly close with time, however, and be able

to accomplish many human tasks at a much greater speed. In order to reproduce general cognition, however, one would need an account of all (relevant) domains of human cognition. Then, one would need to know which levels of description these domains span in order to design a system which reproduces them faithfully. Then, a coherent algorithm simulating all of the appropriate dynamics at these levels would be necessary. The possibility of this feat seems staggering and questionable, and would depend on understanding the brain in its entirety. Designing domain-specific AI systems, however, is an easier task. Just as the cognitive scientist may draw from some subsection of the landscape of levels and domains in explaining cognition, so may the artificial intelligence researcher in designing architectures for cognitive agents. Although the following statement may be rejected by many working in the artificial intelligence community, I believe that intelligent agents will necessarily need to be based on neuromorphic computing in order for the field to progress. In considering the intricate machinery producing human emotions, there seems to be no classical/symbolic architecture which could be able to reproduce a wide range of affects so flexibly. Additionally, judging by the functional importance of neurotransmitters in generating emotional behavior and cognition in humans, as discussed in section 4, even current connectionist models seem to be insufficient.

7.1 Strong AI, applied AI, and cognitive simulation

It can be said that there are three major strands of artificial intelligence research [57]. The first, applied AI, has been most successful in producing industry-viable research because it aims to create smart, (usually) domain-specific systems by any means available. The second, strong AI, aims to produce machines with the full cognitive capacity of humans, i.e. ones that should ideally be indistinguishable from their natural counterpart. The third, cognitive simulation, is most relevant for brain research because its goal is to test empirical theories of cognition by implementing them. This work is relevant namely for strong AI and cognitive simulation, although I believe applied AI could profit equally from the neuromorphic and naturalistic principles stated herein.

This work is namely about understanding and reproducing distinctly human cognitive phenomena, with emphasis on emotions, as their modeling has remained elusive. Human cognition is surely not the most efficient solution to all problems of applied AI, so it would be senseless to employ these principles as a general rule in that particular subdiscipline of AI. For instance, a naturalistic cognitive architecture for a well-defined game would probably be counterproductive, as much more efficient means for calculating the optimal strategy are usually available in these cases. However, where the goals of ap-

plied AI overlap with those of strong AI (i.e. where the designed system is specifically supposed to reproduce some subset of distinctly human behavior) it becomes advantageous to implement cognition as the brain does it. An example of such an application would be an emotional agent whose purpose is realistic social interaction. Ideally, such an agent would have a sophisticated virtual or physical human-like body, as this is necessary for understanding the embodied emotions of its conversational partners. The most realistic software model of said human's emotional state would then consist of a neuromorphic architecture most closely resembling the target's actual brain. Assuming a number of clues about the target's emotional state have been gathered, the best guess as to the further cognitive effects thereof would be to feed them into a realistic model of the target's brain as stimuli. This is exactly what happens when mirror neurons are activated in simulation when biological brains meet socially. Since the human brain will probably remain the system which understands human emotions most closely and efficiently for a long time, if not forever, the theorized agent should simulate it. Naturally, the computational cost of any whole-brain-simulation venture is currently untenable, but I believe that this is merely a practical problem and a question of time before it is solved.

In strong AI and cognitive simulation, I believe software architectures should be created according to their biological counterpart as closely as possible. Since the explicit goal of these research fields is to reproduce and understand human cognition, respectively, implemented models should span as many biological levels of explanation as possible, with more weight in detail given to those levels of explanation which are most relevant for the domain(s) of cognition being modeled. To some extent, this distribution of resources will be determined by the state of the art in the mind/brain sciences – the more we know about which biological systems are most relevant for any given cognitive phenomenon, the more informed the designer's choice will be. Where the correspondence between biological system and cognitive phenomenon is not well detailed, the designer will have to make 'creative choices'. However, such choices only further the purpose of cognitive simulation as they implement an experiment which will allow for refining of the scientific models. I would also like to address the question of whether truly strong AI is even possible, which is a legitimate objection one might raise against the program of creating an artificial machine with indistinguishably human cognitive abilities.

I contend that the answer hereto, as with many things in life, is a matter of degrees rather than being categorical. The future behavior of two humans in identical environments is only as similar as the state of their brains are, and to a lesser extent of influence, their bodies. In the same way, the difference in behavior between a real and an artificial brain can only be as large as the

difference between the mechanisms that are implemented in each. One could of course still say at this point that there are physical mechanisms that are impossible to model on a classical computer, such as quantum phenomena. This, however, is an open empirical question – and there exists no widely accepted evidence suggesting a reason some brain biology relevant for cognition can't be modeled in principle. The suggestion that an implemented virtual brain wouldn't phenomenally feel (as in have qualia) is equally moot – and would only pose a problem if the simulation or 'strong' agent was intended to reproduce the phenomenality of experience. It will inevitably be impossible to reproduce all phenomena associated with cognition in a computer – for instance, cracking one open with a hammer will probably never produce the same wet, gooey feeling that it would in the case of a real brain, unless computer engineers became very creative. There isn't, however, any reason to doubt that the gap between computer simulations and brains can be closed – and the way ahead is to respect the biological system as closely as possible.

7.2 Emotional agents

For now, I would like to consider the hypothetical case of designing a neuromorphic emotional agent architecture. I will assume the agent to be designed is embodied (physically or virtually), as any sufficiently intelligent system needs to be to flexibly learn and interact with its environment. For the sake of identifying plausible architectural structures, I will refer to Panksepp's delineation of the brain, as explained in section 4.1. In order to believably implement a continuous range of emotions, I believe such an agent would inevitably require a constant mapping of its 'physiological states,' such that it would require the analog of a nervous system. Something akin to the major neurotransmitter systems in humans, which reach directly from the lowest levels of the brain to higher cognitive areas, could provide the vehicle by which cognition can be modified rapidly as opposed to slowly. This would endow the neuromorphic architecture with the basics for intuitive reactions, as opposed to relying only on learning mechanisms which operate over a longer period of time. As Daniel Kahnemann illuminates in his widely popular "Thinking Fast and Slow" [60], human ingenuity benefits from these two fundamentally different means of dealing with problems – I believe artificial cognizers could too. Furthermore, 'primary-process' control systems to dictate basic needs and would be required. These would interact with secondary- process learning layers, which would link perceptual symbols with the agent's positively or negatively valenced past experience to keep track of which concepts have been advantageous to it. These basic layers would interact with higher layers of control, which I will not discuss

here. Thus, when such an emotional agent would encounter emotional situations or stimuli, it would respond in a manner like Barsalou's perceptual symbol system: initially, the percept would cause re-activation of associated past perceptual experiences. These would have a learned relevance to its basic primary-process needs, which would cause associated 'physiological' reactions. Due to the neuromorphic nature of such a hypothetical architecture, a complex variety of mixtures of primary process and physiological activations could be observed, lending the agent a believable landscape of observable emotions. This somatic response would trigger 'neurotransmitter' release, such that higher cognitive areas would instantly be affected, akin to the way in which Damasio's somatic markers supposedly influence cognitive problem-solving styles in humans. Such an agent would thus exhibit both believable observable emotional responses, as well as a modulation of rational processes located higher up in its control hierarchy.

A possible question one could pose at this point is how the basic needs of such an agent could possibly be sculpted/trained in such a way that the interaction with higher control systems would result in intelligently adaptive behavior. Furthermore, how could one sculpt these emotional behaviors to support the task one wishes to design the agent for? It has been suggested that the human brains are essentially prediction machines, "bundles of cells that support perception and action by constantly attempting to match incoming sensory inputs with top-down expectations or predictions." [47] Furthermore, the framework of neuroconstructivism stresses the plasticity inherent in all human learning - in other words, the human neural architecture is only fixed to a limited extent, and "cognitive development can thus be understood as a trajectory originating from the constraints on the underlying neural structures." [48] Thus, our hypothetical neuromorphic architecture need not be fixed - to the contrary, a substrate with the appropriate dynamics would mold its emotional responses to be able to better predict future experiences on the basis of past experiences. It has even been suggested, using physically plausible modeling based on the free-energy principle, that the positive or negative valence of emotional reactions in humans directly reflects prediction error (on a neural level). The free-energy principle is essentially the notion that living systems strive to reduce disorder (free energy) in order to maintain themselves – from this perspective it would make sense for disorder to feel painful and order to feel pleasant, at a neural level. A lower prediction error would go hand in hand with less free energy, and should thus feel more pleasant. The result of this modeling was that "when sensations increasingly violate the agent's expectations, valence is negative and increases the learning rate. Conversely, when sensations increasingly fulfill the agent's expectations, valence is positive and decreases the learning rate." [49]

In general, the relationship between the brain as a prediction machine, emotions, and situated agents is highly interesting and warrants attention from all cognitive sciences.

8 Conclusion

The aim of work has been to examine the nature of emotions as seen through the lens of neuroscience, and to draw conclusions with respect to the computation-oriented epistemological framework underlying much of the work done in cognitive science. It is clear that emotions are highly relevant to the interdisciplinary study of the mind, as they form the most basic attitude we can hold; they dictate the most primal behaviours and are indispensable for the higher cognitive functions. From the point of view of affective neuroscience, the emotional brain is shared homologously by all mammals and forms the very foundation upon which the rest of the brain is built. This is true for explaining human behaviour, as a large variety of both pathological and 'rational' cognition requires explanation in terms of how we feel about stimuli and concepts. Furthermore, a general modeling of intelligence can also be said to depend intimately on emotions. For one, learning intelligently about the environment would not be possible without basic drives such as SEEKING and PLAY. Also, it seems quite clear that emotional markers act as a kind of heuristic in helping the brain make decisions under complex circumstances - even if Damasio's somatic marker hypothesis turns out to be empirically unfounded. Essentially, emotional valence is the currency of the brain, and a signal of what is to be valued negatively or positively.

The epistemological implications of emotion complement the emerging framework of embodied cognitive science. Emotion is an intrinsically embodied process, and requires a level of modeling that goes beyond abstract rules and representations as theorized by the computational theory of mind. Evolved out of a need for maintaining biological homeostasis, the basic (good or bad) feeling of what happens is a cornerstone of cognition. In fact, the level of modeling necessary for emotional behaviors goes all the way down to the neurochemical - neurotransmitter systems are an important mechanism by which emotional circuits influence higher cognitive areas. Because embodiment and emotions require an emergence of phenomena on different layers of complexity, this work has argued for explanatory pluralism in cognitive science. Under such an epistemological framework, there can be no unified theory of cognition. Rather, a diverse patchwork of theories across different vertical levels of measurement and different horizontal domains of cognition serve to explain the properties of the brain. The embodied perspective fits nicely within such a naturalistic

framework, where abstract associations (concepts) are ultimately grounded in low-level perceptual and emotional experience.

This view of the brain as a complex natural system is contrasted with the computational theory of mind, which posits the existence of an all-explanatory algorithm of thought. This work has expressed pessimism towards such a project on the grounds that the required rules and representations would need to be identified in the physical substrate. I believe we can say that affective neuroscience has strongly suggested that attitudes are too complex and span too many levels of description to be effectively reducible to some algorithm. For instance, it seems that many emotion-related interactions taking place within the body and brain are necessarily molecular. In general, homeostasis, emotions, and attitudes are very closely related. Rather than seeing it as a scientific model of the mind, I have suggested that computation be viewed as a tool for simulating any collection of theories from the pluralistic framework. This notion has been discussed with respect to artificial intelligence, where more naturalistic models of cognition have already been gaining in popularity in the last decades, in contrast with its 'good old fashioned AI' beginnings. I do not deny the possibility that algorithmic states and transitions may be mappable onto certain levels of description, in the sense of David Chalmers [54]. I simply contend that this must be shown, and express skepticism that such discoveries could explain a wide range of cognitive phenomena. I therefore believe that a pluralistic framework will be necessary for the systematic progression of the mind/brain sciences and AI.

Even today, the insistence on describing cognition in terms of abstract processes and structures remains popular in cognitive science and AI. Marvin Minsky's "The Emotion Machine" [61] makes a great point in emphasizing that emotions are just one of the many domains of thinking, and that they are to be included in any representation of a concept. Furthermore, they are a necessary step in any thought process. This is an important message, and Minsky's book contains quite a few of these. However, in my view, by describing emotions discretely and discussing rule-based emotional reactions in the abstract he discredits himself by ignoring the biological basis from which these dynamics supposedly arise. I have argued that specifically this is not the case – emotions and their resulting attitudes are very much a biological phenomenon to be described on many levels. One can not reliably say that they are also a computational phenomenon, unless one identifies such a computation on any level.

9 Sources

[1] Graham, George, "Behaviorism",The Stanford Encyclopedia of Philosophy (Fall 2010 Edition), Edward N. Zalta (ed.), URL = http://plato.stanford.edu/archives/fall2010/entries/behaviorism/.

[2] Bickle, John, "Multiple Realizability",The Stanford Encyclopedia of Philosophy (Spring 2013 Edition), Edward N. Zalta (ed.),
URL = http://plato.stanford.edu/archives/spr2013/entries/multiple-realizability/.

[3]Barsalou, Lawrence W. "Perceptions of perceptual symbols." Behavioral and brain sciences 22.04 (1999): 637-660.

[4]Thagard, Paul, "Cognitive Science", The Stanford Encyclopedia of Philosophy (Fall 2012 Edition), Edward N. Zalta (ed.), URL = http://plato.stanford.edu/archives/fall2012/entries/cognitive-science/.

[5] Thagard, P., 2005. Mind: Introduction to Cognitive Science, second edition, Cambridge, MA: MIT Press.

[6] Wilson, Robert A. and Foglia, Lucia, "Embodied Cognition", The Stanford Encyclopedia of Philosophy (Fall 2011 Edition), Edward N. Zalta (ed.), URL = http://plato.stanford.edu/archives/fall2011/entries/embodied-cognition/.

[7] Barsalou, Lawrence W. "Grounded cognition." Annu. Rev. Psychol. 59 (2008): 617-645.

[8] "Internet Encyclopedia of Philosophy." Emotion, Theories of. N.p., n.d. Web. 26 June 2014. http://www.iep.utm.edu/emotion/.

[9] Schwarz, Norbert. Feelings as information: informational and motivational functions of affective states. Guilford Press, 1990.

[10] Asma, Stephen, et al. "Philosophical Implications of Affective Neuroscience." (2012).

[11] Panksepp, Jaak. "Affective consciousness: Core emotional feelings in animals and humans." Consciousness and cognition 14.1 (2005): 30-80.

[12] Lewis, Michael, Jeannette M. Haviland-Jones, and Lisa Feldman Barrett, eds.Handbook of emotions. Guilford Press, 2010.

[13] Damasio, Antonio R. The feeling of what happens: Body, emotion and the making of consciousness. Random House, 2000.

[14] Damasio, Antonio. Self comes to mind: Constructing the conscious brain. Random House LLC, 2012.

[15] Thivierge, Jean-Philippe, and Gary F. Marcus. "The topographic brain: from neural connectivity to cognition." Trends in neurosciences 30.6 (2007):

251-259.

[16] Damasio, Antonio R., B. J. Everitt, and D. Bishop. "The somatic marker hypothesis and the possible functions of the prefrontal cortex [and discussion]. "Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 351.1346 (1996): 1413-1420.

[17]Damasio, Antonio. Descartes' error: Emotion, reason and the human brain. Random House, 2008.

[18] Rolls, E.T. (1999). The brain and emotion. Oxford: Oxford University Press

[19]Baev, Konstantin V. Biological neural networks: Hierarchical concept of brain function. Birkhäuser, 1998.

[20] Harnad, Stevan. "Symbol-grounding Problem." Encyclopedia of cognitive science (2003).

[21]Fodor, Jerry A. The modularity of mind: An essay on faculty psychology. MIT press, 1983.

[22] Fodor, Jerry A. The mind doesn't work that way: The scope and limits of computational psychology. MIT press, 2001.

[23] Gallese, Vittorio, and Alvin Goldman. "Mirror neurons and the simulation theory of mind- reading." Trends in cognitive sciences 2.12 (1998): 493-501.

[24] Hofstadter, Douglas R. Fluid concepts and creative analogies: Computer models of the fundamental mechanisms of thought. Basic books, 2008.

[25] Barsalou, Lawrence W., et al. "Grounding conceptual knowledge in modality-specific systems." Trends in cognitive sciences 7.2 (2003): 84-91.

[26] Petersen, Rasmus S., and Mathew E. Diamond. "Topographic maps in the brain." eLS (2002).

[27] Beauregard, Mario, ed. Consciousness, Emotional Self-regulation, and the Brain. John Benjamins Publishing, 2004.

[28] Webster, Roy, ed. Neurotransmitters, drugs and brain function. John Wiley and Sons, 2001

[29] Faure, Philippe, and Henri Korn. "Is there chaos in the brain? I. Concepts of nonlinear dynamics and methods of investigation." Comptes Rendus de l'Académie des Sciences-Series III- Sciences de la Vie 324.9 (2001): 773-793.

[30] Panksepp, Jaak. Affective neuroscience: The foundations of human and animal emotions. Oxford university press, 1998.

[31] Churchland, Paul M. "Eliminative materialism and the propositional attitudes."The Journal of Philosophy (1981): 67-90.

[32]Chalmers, David J. "A computational foundation for the study of cognition." (1993).

[33] Marr, D. "Vision, 1982." Vision: A Computational Investigation into the Human Representation and Processing of Visual Information (1982).

[34] Aydede, Murat, "The Language of Thought Hypothesis", The Stanford Encyclopedia of Philosophy (Fall 2010 Edition), Edward N. Zalta (ed.), URL = http://plato.stanford.edu/archives/fall2010/entries/language-thought/.

[35] Horst, Steven, "The Computational Theory of Mind", The Stanford Encyclopedia of Philosophy (Spring 2011 Edition), Edward N. Zalta (ed.), URL = http://plato.stanford.edu/archives/spr2011/entries/computational-mind/.

[36] Piccinini, Gualtiero. "Computational modelling vs. Computational explanation: Is everything a Turing Machine, and does it matter to the philosophy of mind? 1."Australasian Journal of Philosophy 85.1 (2007): 93-115.

[37] Dale, Rick, Eric Dietrich, and Anthony Chemero. "Explanatory pluralism in cognitive science." Cognitive science 33.5 (2009): 739-742.

[38] Popper, Karl. Conjectures and refutations: The growth of scientific knowledge. routledge, 2014.

[39] Thagard, Paul. "Being interdisciplinary: Trading zones in cognitive science."Interdisciplinary collaboration: An emerging cognitive science (2005): 317-339.

[40] Newell, Allen. Unified theories of cognition. Harvard University Press, 1994.

[41] Garson, James, "Connectionism", The Stanford Encyclopedia of Philosophy (Winter 2012 Edition), Edward N. Zalta (ed.), URL = http://plato.stanford.edu/archives/win2012/entries/connectionism/.

[42] Dale, Rick. "The possibility of a pluralist cognitive science." Journal of Experimental and Theoretical Artificial Intelligence 20.3 (2008): 155-179.

[43] de Jong, Huib Looren. "Levels of explanation in biological psychology. "Philosophical Psychology 15.4 (2002): 441-462.

[44] Robinson, Howard, "Dualism", The Stanford Encyclopedia of Philosophy (Winter 2012 Edition), Edward N. Zalta (ed.), URL = http://plato.stanford.edu/archives/win2012/entries/dualism/.

[45] "." . N.p., n.d. Web. 26 June 2014. http://www.cambridge.org/features/bermudez/images/slides/Chapter4_(Integration_challenge).ppt .

[46] "Internet Encyclopedia of Philosophy." Artificial Intelligence. N.p., n.d. Web. 26 June 2014. http://www.iep.utm.edu/art-inte/.

[47] Clark, Andy. "Whatever next? Predictive brains, situated agents, and the future of cognitive science." Behavioral and Brain Sciences 36.03 (2013): 181-204.

[48]Mareschal, Denis, et al. "Neuroconstructivism-I: How the brain constructs cognition." (2007).

[49] Joffily, Mateus, and Giorgio Coricelli. "Emotional valence and the free-energy principle." PLoS computational biology 9.6 (2013): e1003094.

[50] Sporns O. What neuro-robotic models can teach us about neural and cognitive development. In: Mareschal D, Sirois S, Westermann G, Johnson M.H, editors. Neuroconstructivism: perspectives and prospects. Vol. 2. Oxford University Press; Oxford, UK: 2007

[51] Davidson, Richard J., Daren C. Jackson, and Ned H. Kalin. "Emotion, plasticity, context, and regulation: perspectives from affective neuroscience. "Psychological bulletin 126.6 (2000): 890.

[52] Sindlar, M. P. "In the eye of the beholder: explaining behavior through mental state attribution."SIKS dissertation series2011 (2011).

[53] Bratman, M. E. (1999) [1987].Intention, Plans, and Practical Reason.CSLI Publications.ISBN1-57586-192-5.

[54] Chalmers, David J. "Does a rock implement every finite-state automaton?."Synthese108.3 (1996): 309-333.

[55] Putnam, Hilary, and Hilary Putman.Representation and reality. Vol. 454. Cambridge, MA: MIT press, 1988.

[56] Chalmers, David. "The varieties of computation: A reply."The Journal of Cognitive Science13 (2012): 211-248.

[57] Copeland, J. (n.d.). What is Artificial Intelligence? Retrieved August 5, 2014.

[58] Barrett, Lisa Feldman, et al. "Of mice and men: Natural kinds of emotions in the mammalian brain? A response to Panksepp and Izard."Perspectives on Psychological Science2.3 (2007): 297-312.

[59] Panksepp J. Neurologizing the psychology of affects: How appraisal-based constructivism and basic emotion theory can coexist.Perspectives in Psychological Science.2007;2:281–296.

[60] Kahneman, Daniel.Thinking, fast and slow. Macmillan, 2011.

[61] Minsky, Marvin. "The emotion machine."New York: Pantheon(2006).