9 Engineering empathy

Emotion and self representation in artificial intelligence

> Every living example of general intelligence that we know has emotion, which does not mean there might not be another path to intelligence without such mechanisms. However, why take time to search for an alien mechanism when we already have a set of mechanisms—emotion—that appears to be able to do the job? If we emulate human affective abilities, versus alien mechanisms, then (1) we stand a better chance of understanding the resulting behaviors of the computer and (2) the process of figuring out how to build the emotions can help us in figuring out how the human emotional system works, one of the most important potential benefits. Understanding and emulating human emotion might or might not hold the key to solving AI, but we are far from complete in our efforts to understand intelligence if we do not. (Picard, 1999, p. 137)

9.1. Preamble: Why artificial intelligence?

This is a book that offers neurocognitive explanations of altered self awareness in psychiatric disorders, so there is no obvious connection with artificial intelligence (AI). However, there are two reasons for discussing the topic. The first is that advances in performance of large language models (LLMs), like generative pre-trained transformer (GPT4), have intensified debates about prospects for artificial general intelligence (AGI): human-like, fluid, creative, problem-solving, decision-making, and cognitive control. And as the quotation above shows, some AI researchers consider that fluid intelligence and emotion are interdependent (Deane, 2022; Schuller & Schuller, 2018). The active inference conception of cognition that frames this book makes emotional and cognitive processing continuous and so offers a framework for the 'attempt to emulate human affective abilities'. The second reason is that the relationship between

Anatomy of an Avatar. Philip Gerrans, Oxford University Press. © Oxford University Press 2024.
DOI: 10.1093/9780191994395.003.0010

emotion and intelligence in humans is still imperfectly understood. As in many fields, attempts to model or reverse engineer a process can illuminate the nature of that process. AI can help us understand ourselves. A final reason is that while predictive processing is part of the current AI repertoire, active inference is not (Albarracin et al., 2023; Pezzulo et al., 2024) Yet if this book is on the right track, active inference anchored by an avatar is the key to understanding the emotion/cognition relationship.

This book suggests that the link between emotional processing and other aspects of cognition is the integrative role of self modelling. Yet as we shall see, the AI of emotion, even in cases where the inspiration is biological, does not usually address the nature and role of self representation. One reason is that historically, AI architectures perform specific tasks that implement derived goals by using processing architectures that they do not configure and reconfigure to maintain themselves. They are solving a specific problem set for them by us, not striving to maintain themselves in a hostile world. Current AI systems, even those that interact fluently in social contexts, lack intrinsic goals and selfhood. They lack the means of organizing and prioritizing information that undergirds human cognition because they do not model themselves as entities whose existence is made better or worse by the success or failure of their activities and perform their activities as means to the end of self preservation.

One way to put this is to say that AI systems such as LLMs (at present) perform perceptual, not active, inference. They map input to outputs by a process of statistical inference that fits a model to that input. They (typically) do not actively seek further information to confirm or disconfirm a model that determines their survival. In active inference, on the other hand, cognition, even in a specific domain, is ultimately a form of active self optimization kept on track by emotional processing. Human intelligence has a built-in *locus* of concern, the self, to drive and structure cognition and action, and emotional processes inform the organism whether and how she is succeeding in her *lebenswelt*. The creativity and adaptability (as well as characteristic biases and defects) of human thinking are products of a design that embeds emotional processing at the core of cognition. Human cognitive architecture makes it the case that action (which includes cognition on the active inference conception) is felt to be undertaken on behalf of the subject. It is existentially vital. As one ambitious AI theorist put it: ‘if we want a robot to be like a person it should have a soul . . . that can think about others, live in dreams and memories, see itself from an outside and be above the hustle and bustle of everyday tasks’ (Samsonovich, 2013, p. 72). This is a rather lyrical way to put the point that we have emphasized throughout. Life is lived by manipulating the self model on itself that play the role of emotions in humans. Parents, on the other hand, look after their children because they care about them. The feelings that propel their own life are interwoven, for better or worse, with those of their offspring. Because AI systems do not have a self to which sensory information is existentially relevant, AI systems will never become tired, impatient, inattentive, selfish, resentful of demands on their life, or abusive, and very likely will be more gentle, responsive, and nurturing than many parents—precisely because they lack emotions and selves.

Such a lack of emotion and selfhood might be an advantage in some domains. There is no need for an AI system to feel a surge of warmth when recognizing a face (Nabokov's 'flare of the pilot light') in a crowd. When scanning a passport at an airport check in a kiosk, we do not want a face recognition system to be overcome by a wave of Proustian reminiscence. Humans, on the other hand, have been known to abandon their current life and cross the globe to reunite with their old (or new!!!) boyfriend after seeing a face in an Instagram photo. So affectless face recognition systems at airports are advantageous. They will not refuse an entry visa to someone on the basis of prejudice, unless they have learnt to do so on the basis of training data (Flores et al., 2016). At the same time, if AI systems are drone-bombing villages, refusing parole applications and applications for refugee status, sentencing criminals, turning off life support systems or scheduling surgeries, and caring for infants, the disabled, and the elderly who cannot articulate their suffering, people feel that AI systems need affective states *and* to be able to understand the nature and role of affective states in other people. Many theorists of emotion have argued that care for others depends on the ability to empathize with the experience of suffering (Kerasidou, 2020).

9.3. Conceptual issues for artificial intelligence of emotion and self representation

Although there is consensus that AI/human interaction requires the ability to detect and respond to emotion, there is no univocal conception in AI of precisely what an emotion *is*. In contrast, for cases of perception, face recognition, memory, or even attention, the basic nature of the process is relatively clear. Often a philosophical or psychological theory of a domain is imported to clarify the task. Famously, Marr used Aristotle's theory of perception ('to know by looking') as a basis for his computational theory of vision. However, when AI theorists draw on philosophical and psychological models of emotions to implement, there is less consensus about what emotions are and which aspects of emotion need to be implemented, and there is no consensus about the nature of the self, the entity that experiences emotions. In AI we see attempts to integrate and implement different versions of emotion theory—notably, core affect, appraisal, affect programme, and somatic marker—and almost no attempt to model a self, understood as the subject of emotional states and the entity whose welfare is the goal of action.

The remainder of the chapter examines some attempts to implement emotion in AI, concentrating on the question of whether and how affective states are conceptualized and implemented. When we turn to the issue of implementation, emotions can be implemented in a variety of architectures: classical, neural network, or hybrid. Typically, a classical system will have a modular architecture that has discrete algorithms for specific aspects of the processing task, as well as a general-purpose amodal central processor that performs higher-order inferences in order to prioritize and control action and update the knowledge base. A disclaimer is in order here. I spend some time discussing classical implementations, even though 'pure' classicism is out of vogue in cognitive science. One reason is that approaches taken by classical modellers throw into sharp relief the nature of the computational problem posed by emotions. It is difficult to capture the role of emotion in a modular architecture with a classical algorithmic structure.

Alternatively, emotions can be implemented in neural networks (of which machine learning in deep neural networks is a special case). In these cases, emotional architectures are not programmed into the system a priori but are emergent results of statistical learning. An example might be training artificial neural networks to detect and respond to manifestations of emotion in facial expression. Machine learning makes it possible to analyse billions of vignettes of emotional interaction, to classify expressions of emotion, and to respond appropriately. As we shall see, neural networks can also be used to simulate the evolution of emotions in artificial organisms. Non-classical approaches or hybrids are increasingly popular in affective computing contexts. In fact, many of the optimizations of deep learning systems effectively reinstate advantages of earlier classical systems such as hierarchy, recurrence, top-down control, and (pseudo-)symbolic modelling (Graves et al., 2014).

As this book is being produced (mid 2024), LLMs have produced Turing-equivalent performance on tasks that probe fundamental aspects of human cognition, such as inferring intentions and goals of other people, in order to interpret and predict behaviour. The 'psychological reality' of LLM performance means that interaction with an LLM in many domains equals or exceeds that of humans. In many cases the uncanny valley has been crossed.

However, there are as yet no AI systems that co-ordinate systemic function based on a predictive model of the consequences of action for a self on whose behalf action is performed. This is most obviously true in the case of the classical systems I discuss, but it is also true of LLMs. Of course we should leave open the possibility that that AIs have developed a form of self representation that us currently inscrutable to us.

9.4. Emotional resources: Somatic markers

One strategy in AI is to model emotions by using a formalized model of the somatic marker theory (SMT) of affective states. The somatic marker theory, as it is baptized by Antonio Damasio, was developed to explain problems with subjective decision-making and planning in patients with damage to the orbitofrontal cortex (OFC) (called ventromedial cortex in seminal papers) (Bechara & Damasio, 2005; Colombetti, 2008; Damasio, 1994; Dunn et al., 2006). The SMT contains a theory of self awareness, and of the nature of emotion and consciousness, and the relationship between emotion and cognition. So it traverses a lot of the same territory as this book, with the same ambition: to ground the concepts of emotion and selfhood in bodily representation. In early versions of the SMT, automatically induced first-order changes of body state are progressively remapped by higher-order cognitive processes, creating a sense of self in the process. The *proto self*, understood as the internal milieu that undergoes perceptually driven changes, is not consciously experienced. The *core self*, which *is* experienced, is produced by higher-order remapping of changes to the *proto self*. At this level, the organism experiences its body as changing in response to environmental interaction, moment to moment. An extended or *autobiographical* sense of a self, located in time with a past and future, is produced by higher-order remapping of the proto self by cognitive processes that *associate* bodily experience of the core self with explicit autobiographical thought to create the experience of being a persisting person (Damasio, 2003; Damasio & Dolan, 1999).

A key aspect of the SMT is an explanation of the role of the OFC. One role of the OFC is to associate representations of body state (states of the core self) with representations of information relevant to decisions (Bechara & Damasio, 2005; Bechara et al., 1999). When the OFC is damaged, the subject cannot perform that association and, as a consequence, decision-making becomes a purely intellectual process—the application of context-free procedural reasoning. The results are reflected in poor decision-making and behavioural control in personal and social domains where the ability to experience, prospectively and retrospectively, the personal significance of options is essential.

The attraction for AI theorists is obvious. If AI is to be a plausible analogue of human cognition, it needs an analogue of the mechanisms that allow humans to experience the significance of their actions and to use that to regulate themselves. And the SMT is explicitly formulated as a solution to that problem. Another attraction is that the SMT argues that the relation between bodily representation and cognition is essentially *associative*. If this is the case, then teaching a robot to behave adaptively is a matter of teaching it to *associate* representations of a advantageous internal states with cognitive and perceptual representations.

Damasio's characterization of the relationship between bodily feeling and cognition gels with some philosophical critiques of Cartesian conceptions of the mind as a disembodied reasoning system. Damasio explicitly claims that psychological theorizing inherits a Cartesian assumption that cognition is essentially disembodied According to Damasio the thought experiment that concludes that mental states such as pain and suffering 'might *exist separately from the body*' produced an 'abyssal separation'. Here he joins forces with anti-Cartesian critics of classical AI who have argued that a modelling strategy that abstracts from contingencies of bodily implementation will forever misconstrue the nature of cognition. It is always interesting to see in these cases what the straw man actually said. As we noted in the introduction, Descartes said that bodily sensations such as pain and thirst 'teach us that *I compose a single thing with it (my body*)'. For what it is worth, when Descartes focused on phenomenology, rather than on modal metaphysics, he actually had an interoceptive theory of self awareness. The abyssal separation really reflects the proclivities of philosophers who focus on modality, not on biology.

In any case, some of the core ideas are shared by SMT and interoceptive active inference theory, although the terminology is different. Both the SMT and active inference theory agree that the essential purpose of self awareness is bodily regulation and that affective states are higher-order interoceptive representations of lower-order bodily somatovisceral states. However, the SMT does not really have a theory of self representation other than via the 'remapping' of body states.

9.5. Affective computing and social emotional artificial intelligence

Affective computing is the AI implementation of emotional processes 'enabling robots and computers to respond intelligently to natural human

> Body preparation is described by a (*v*,*a*) point that is a bodily state, induced by events, that corresponds to a specific emotion. This state will be performed by the agent as an immediate reflex and will last only the duration of the emotional stimulus. (Cominelli et al., 2018, p. 10)

In order to transform this vector into an emotion, it is metarepresented by a Feelings module and a Somatic Marker module that apply rules for transforming representations of body state by using principles derived from the SMT. The example used by Bosse was hearing music. If, for example, stored information is that slow, low-volume, repetitive music is boring (low arousal, low valence), 'thanks to *Feeling module*, we will see the previous serene facial expression turning gradually into a bored expression' (Cominelli et al., 2018, p. 12).

The system is loaded into a system called Facial Automaton for Conveying Emotions (FACE)—'a human-like robotic head, with the appearance of an adult female, capable to perform very sophisticated expressions by means of a hyper-realistic facial mask' (Cominelli et al., 2018, p. 13).

> The robot interacts plausibly with humans on the basis of perception. FACE expresses discomfort (−0.5, −0.6) when a subject invades its intimate space, an angry expression (−0.52, −0.67) if someone folds his arms, smiles (0.21, 0.6) if someone greets her or smiles at her, and expresses interest (0.62, 0.2) when an interlocutor speaks to her robot interacts plausibly with humans on the basis of perception. (Cominelli et al., 2018, p. 15)

In terms of the SMT, these actions are not yet based on emotion because the values represent automatic reactions corresponding to an *automatically evoked body state*. Transforming an (*a,v*) vector to a somatic marker requires the operation of the Feelings and Somatic Marker modules, which, in effect, allow FACE to recall the body state evoked by previous interactions and to show preferential attention and positive emotional expression accordingly.

SEAI is a classical architecture with its 'declarative rule-based expert system on top of procedural services deputed to the perception and motion control of the robot' and assumption of modularity. The latter assumption has the standard advantages of portability and detachability. Interestingly, according to the authors, this allows the robot's mind to exist independently of the body whose essential role is to provide inputs to higher-order processes that create:

> personality, memories, beliefs, experience, and behavioral traits of the agent, all of which depend on the cognitive part of the system, and therefore can be transferred or modified independently. (Cominelli et al., 2018, p. 18)

In this way of thinking of the mind, the body provides inputs (here conceptualized as a two-dimensional vector) *associated* with the high-level symbol processing that constitutes thought.

For a neurobiologically inspired theory, this is a very conservative approach to modelling emotion and consciousness, and the authors acknowledge this. As they say, their robot has no homeostatic mechanisms (although the body states they describe are interoceptive states), and it is doubtful whether much of human cognition, including processes underlying social interaction and empathy, is symbolic. Nonetheless, in its domain, SEAI succeeds in creating a 'believable and acceptable synthetic consciousness'. By believable and acceptable here, the authors mean that the system would pass a version of an emotional Turing test. A human could interact plausibly with FACE in emotional contexts.

The designers of SEAI accept that their system is limited but argue that the solution will require 'writing new rules and expanding the current rule-sets'. This approach runs counter to current trends in both AI (machine learning) and critiques of classic AI that point to difficulties in explicit encoding of rules to govern flexible behaviour. Multiplying rules to accommodate more parameters ultimately tends to make a system less, rather than more, flexible. It is for this reason that in non-classical architectures, like those relied on by machine learning systems, cognitive processes are not decomposed into rules implemented in discrete modules a priori.

In any case, the classical approach adopted here points to another difficulty for affective computing of this type. SEAI separates a representation of body state (*a,v*) vector from the representation of information needed to determine the emotional meaning of that state. This approach runs counter to neurobiologically inspired active inference theories of emotional processing. In allostatic active inference, interoceptive experience does not function as an input to a higher level-emotional processing. Rather it is an emergent product of a heterarchical system constantly appraising body state and the world to optimize organismic functioning. Thus, interoception is not, in the normal course of events, dissociable from affective experience and higher cognition because how the body feels depends on whether and how subjective goals are met in context.

A second, and related, point of difference is in the way in which affective experience is conceptualized. I have argued that it is the result of transcription of interoceptive signals. *Prima facie*, that does not sound very different from the idea that first-order signals of body state are inputs to a higher-level representational system that remaps them. However, on the view I presented, emotional processing depends on hubs whose role is to co-ordinate and integrate perception and cognition across a processing heterarchy in the service of bodily optimization. Thus, there is no *separate* intrinsically emotional processing system. Rather there are hubs that co-ordinate the processing of subjectively relevant information. The example of pain processing helps to make the point. The nociceptive signal is a signal of bodily damage. There is no separate system that evaluates that signal *according to a pre-specified template*. An abdominal muscle cramp will be felt entirely differently if it is transient and easily relieved (e.g. produced by doing sit-ups) or if it is sensed as the beginning of a miscarriage when the subject has no access to medical care. The point is that the phenomenology of pain or any bodily experience is not produced by an *ex post facto* higher-order interpretation, but by the pattern of processing of the sensory signal across the system. There is no 'neutral' bodily signal to which an interpretation is subsequently attached. Whether and how bodily information becomes salient depends on how the mind is configured a *priori*. This is just to reiterate the point that on the active inference conception, interoception, emotional transcription, affective experience, and narrative selfhood are continuous. This view also suggests that what the authors see as a benefit of the classical algorithmic approach, its 'portability and detachability' from the system that implements it, is actually a defect as a model of human emotional processing. For humans, the bodily substrate and emotional processes are inseparable because the essence of emotion is to optimize body state represented and felt as a state of the self. For SEAI, in contrast, emotions are scalars that help to regulate an interaction between perceived features of the environment and the system.

Another problem for SEAI is inherited from the way in which it renders emotion computationally tractable. On the formalization of Damasio's view by SEAI, interoceptive signals are reduced to two values of valence and arousal, which are, in Jamesian spirit, essentially representative of overt action tendencies. Valence is approach/avoid, and arousal is activation or deactivation in preparation for action or withdrawal.

On the somatic marker and SEAI views, it is hard to account for the emotional quality of affective states. The point is also clear in the example used by Bosse. Low-volume, slow, repetitive music is represented as 'boring' because it is associated with a low-value (*a,v*) vector. But some people like Chillhop and find it pleasant and relaxing. So Tomppabeats has higher valence and low arousal for fans. But that is a product of their 'personality', to use a contested term. Their sense of who they are, what matters, and how it matters to *them* determines whether they find a type of music boring. So self representation, which is also bodily representation, cannot be prised apart from the assignment of value and relevance.

The example we discussed in previous chapters that makes a similar point is that fatigue of physical exhaustion and depressive apathy are experienced differently, although they occupy the same point in arousal/valence space. The dimensional model needs additional resources to account for this crucial difference.

One solution is provided by conceptual act theories of emotions that treat emotions as the result of higher-level conceptual *interpretations* of bodily experiences. Thus, for conceptual act theories, 'fear' is an interpretation of high arousal and low valence is interpreted via the concept of danger. Sadness is a state of low valence and low arousal, interpreted via the concept of personal loss. These interpretations rely on higher-level cognitive processes. The difficulty for these theories is that it is not obvious how *associating* an interoceptive signal with a higher-order representation can change the quality of the interoceptive experience—unless, of course, the quality of experience is an emergent property of the association.

Of course, there are many cases in which prior higher-order knowledge or social embedding of activity seems to change the quality of experience—placebo and nocibo effects being obvious cases. But these are not cases of association or subsequent conceptual interpretation. They are cases where expectations *condition the processing of the signal from the source*. This is because the transcription of bodily signals using models of self and emotional world is continuous with the lowest levels of interoceptive processing. How a bodily signal is experienced depends on what body states are predicted by the self model in context. *Emotional interpretation is not subsequent to bodily processing and self representation, but constitutive of it.* This is why the fatigue of depression is felt differently to fatigue after a long journey. The former is felt to represent intractable failure of the self in an emotionally desolate world. The source of that failure can arise at any level of regulation, from explicit rumination about the self to inability to control intractable insomnia.

A related difficulty for SEAI as a complete theory of emotion (which, to be fair, is not the goal of SEAI) is that FACE does not represent the expressions it sees as expressive of states of another mind. Rather it maps perceived expressions to a location in its own (*a,v*) space and changes its own expression accordingly. It cannot model the target or itself as an entity with emotions understood as representations of the value or significance of objects, situations, or actions. It is a sophisticated system that models the peripheral aspects of emotion. As the authors note:

> SEAI has still some shortages: homeostasis control is missing, the agent's physiological parameters are a symbolic representation, capabilities such as

> dynamics of a social emotion in a cognitive agent can be triggered by a combination of Appraisals and Somatic markers, and once started, is determined by the active moral schema and the controlled by it set of fluents, including Emotion, Mood, and Feelings, that take values in the semantic space. This process results in behavioral biases and natural reactions. (Samsonovich, 2013, p. 69)

As described, the model is a sophisticated version of SEAI that processes not just facial expressions, but also actions in social context. On both accounts, the role of the body is to provide 'raw physiological experience' to adaptively bias cognition in pursuit of goals. That pursuit is conceived of as a trajectory through emotional space modelled by a semantic map. So the plausibility of the model derives from: '(i) the semantic map data and its origin, and (ii) the "lower-level" functions, or the "strings" pulled by moral schemas'. The semantic map in the model is derived from dictionary or questionnaire ratings of distance in similarity space that reflect everyday notions about relations between emotions. The moral schemata are likewise derived from familiar patterns of behaviour, often tested in games of cooperation and conflict in which agents form alliances. So, for example, agents will be less likely to maintain alliances with other agents who defect from alliances. The moral schema for trust is designed to move an agent from trust and engagement to suspicion and disengagement in such a case.

So the theory is ultimately driven by the peripheral folk psychology of emotion, explicitly encoded and implemented from the top down. As Samsonovitch notes, it would be preferable to have a system that *learnt* its moral schemata in the same way as humans, learning to navigate its emotional landscape through social interaction.

> Ideally, the architecture should learn its semantic maps from its own experience, similarly to a child . . . [this] minimal embryo AI, a critical mass of intelligence capable of rapid and unlimited autonomous development should be guided by human values and based on human-inspired principles. (Samsonovitch, 2013, p. 71)

9.7. Learning, reinforcement, and the self model

The 'human-inspired principle' that Samsonovitch invokes is reinforcement learning, also known as the 'standard Q-learning procedure' (Q stands for the value of a variable that moves the agent through a state space). In the case of Samsonovitch's model, that value is an Appraisal or a Somatic vector to be learnt or approximated. In order to learn Q values that drive action, the appropriate actions need to be rewarded. In reinforcement learning, this requires a reward signal (q) that strengthens the link between a perceived state of the world and action. So, for example, a caring robot needs to learn to map hearing a baby's cry to feeding it or changing it. In Samsonovitch's model, this amounts to moving the baby through semantic space from unhappy to content and the robot from alert/worried to calm. Ideally, we would like to develop a robot that was initially rewarded for reducing a baby's crying (*extrinsic* reward) and ultimately learnt to find caring behaviour rewarding (*intrinsically* rewarding). The job of the reward signal is to teach the robot the appropriate repertoire of actions.

Different reward functions install different learning strategies, but all need to balance exploration (pursuit of distant rewards by open-ended strategies) and exploitation (automatic pursuit of short-term rewards), and the pursuit of extrinsic and intrinsic rewards.

In humans, goals are layered in a complex hierarchy, often in conflict, and operate at different timescales. Furthermore, the means to their realization (moral schemas, as Samsonovitch calls them) are multiple and often open-ended. As the *Bee Gees* remind us, there are many ways to mend a broken heart. The environment that creates the state space consists largely of the actions of other humans whose mental states are coupled to those of the agent. So designing a plausible reward function or a function that mimics the way in which humans learn is a difficult task. A universe of evolved and acquired neurocognitive engineering is hidden under the simple variable q.

There is no simple mapping from formal theories of reward learning to human neurobiology. That is to say there is no discrete neural mechanism for implementing a reward function. As with almost all cognitive functions, learning requires coalitions of distributed circuitry co-ordinated by hubs. There is, however, a clear consensus that the dopamine system is the crucial hub of reward learning in humans. The dopamine system does not provide the reward itself (the taste of chocolate, a feeling of exhilaration or contentment); rather it functions to make options and actions that are rewarding-*salient*. Salience here refers to the allocation of cognitive resources to prioritize processing of information that ultimately leads to rewarding outcomes. The concept of salience reflects the relation between foraging and reward. Foraging is not extrinsically or intrinsically rewarding a priori. Walking 10 kilometres to a well and filling a bucket (foraging) is arduous, but the relief of thirst, the reward, is pleasant. The job of the dopamine system is to install adaptive foraging behaviour by reinforcing patterns of systemic activity that maximize long-term average reward. This role for the dopamine system is supported by studies that show that its activity essentially predicts rewarding sensory consequences of successful action such as taste. For example, injecting a hungry rat with glucose does not produce activation in the dopamine system, but the smell of food does. In both cases, the ultimate reward, understood as the experience of satisfying a fundamental goal (nutrition), is the same, but it is the smell of food that is made highly salient by the dopamine system. It teaches the animal that the smell predicts the rewarding experience. Thus, for the rat, the smell of food becomes highly salient. Similarly, for the thirsty desert dweller, the sight of palm trees in the shimmering haze signals an oasis, and potential relief biases cognition and behaviour appropriately (Bayer & Glimcher, 2005; Berridge & Robinson, 1998; Glimcher, 2011).

All theories of emotion postulate a basic set of goals defined by survival needs whose satisfaction is rewarding and that drives learning. Appraisal theories identify emotion with the representation of the value of objects relative to those goals (Frijda et al., 1989; Sander et al., 2005; Scherer, 2004). SMTs identify emotions with bodily experiences characteristic of emotional episodes.

On either view, SMT or Appraisal, a key role of emotions is to make valuable options salient, cognitively and experientially. And on either view, cognition, action, and the experience of goal satisfaction need to be integrated. The classical AI systems we examined provided an account of this integration via the assumption that cognition is driven by a bodily signal represented as an (a,v) vector. Neither system includes any component analogous to the interaction between affective and reward systems in humans.

Given that the dopamine system functions as a reward prediction system we need to explain how it integrates the experience of rewarding bodily states and representation of the potential value of objects and states (extrinsic reward) and actions (intrinsic reward). In humans, this is the result of 'intricate coupling' between the reward prediction system and basic regulatory systems.

9.8. Homeostatic reinforcement learning

The challenge for AI is to model the process by which the system learns to predict which actions produce optimal body state in context. An elegant solution is 'homeostatic reinforcement learning', in which reward is defined as 'the approximated ability of an outcome to restore the internal equilibrium of the physiological state'.

This suggests that a learning function should reinforce those actions that keep a system in its ideal homeostatic range. On this formulation, reward and homeostasis are two sides of the same coin.

> the rewarding value of outcomes is computed as a function of the animal's internal state, and of the approximated need-reduction ability of the outcome. *The computed reward is then made available to RL systems that learn over a state-space including both internal and external states, resulting in approximate reinforcement of instrumental associations that reduce or prevent homeostatic imbalance*. [my italics] (Keramati & Gutkin, 2014, p. 22)

The authors actually extend the idea, introducing, without naming, the concept of allostasis as predictive homeostasis. As they point out, the system needs to learn to *predict* those actions that will restore homeostatic equilibrium, not just react to environmental contingencies. However, the concept of allostasis is wider than predictive homeostasis because allostasis is more than the resetting of set points. This is the origin of the concept of allostatic active inference.

> Intuitively, the rewarding value of an outcome depends on the ability of its constituting elements to reduce the homeostatic distance from the setpoint or *equivalently*, to counteract self-entropy. (Keramati & Gutkin, 2014, p. 4)

The model of homeostatic reinforcement learning has the pleasing feature of accounting for the mutual influence of hypothalamic and dopaminergic circuitry, by modelling bodily optimization *as* reward. In their words, they are 'mathematically equivalent'. It also integrates reinforcement learning theory with active inference and free energy accounts of self modelling, because homeostatic maintenance is a mechanism that allows an organism to maintain itself.

The formalism reflects the reciprocal relationship between dopaminergic and homeostatic regulation, 'namely, the modulation of midbrain dopaminergic activity by hypothalamic signals' (Keramati & Gutkin, 2014, p. 15).

As they say, their model applies only to learning behaviours such as nutrition and pain avoidance that can be predicted to directly optimize body state. They apply it to explain the behaviour of model animals that learn the rewarding value of food: 'we propose that these behavioral phenomena are signatures of the coupling between the homeostatic and the associative learning systems' (Keramati & Gutkin, 2014, p. 9). Building the notion of reward into a biologically plausible formal architecture is important. It explains something that is assumed (plausibly) on other theories—namely the motivational salience of interoceptive states.

This is another reason for adopting the interoceptive active inference account of emotion. Affective states that serve as reinforcers are the result of transcription of interoceptive signals. A system that learns to keep itself in

The world of the robot is a pixel array, and robots are pixel arrays with a diameter of 75 pixels. The organisms have sensors that can detect predators, food sources and mates, and effectors that can propel them toward or away from perceived stimuli. They move around their world for a set number of iterations (1–3000), and survive and reproduce. They have internal sensors that detect resource depletion (analogues of hunger and thirst) and damage following agonic encounters with predators. Other than the ability to sense motivational states (hunger, thirst, and damage/pain), to sense the environment, and to control direction and speed of movement, model organisms have no a priori architecture or rules for cognizing their world. Consequently, they have to learn to avoid predation and find mates and food through trial and error.

Learning is accomplished by changing weights in a neural network that maps perception to movement by using a standard back propagation algorithm. Weights are initially semi-randomized (e.g. .5 for motor units) and then strengthened by successful action such as moving away from a predator and toward a mating opportunity or food. The mechanism of inheritance is a genetic algorithm that installs the weight settings of successful organisms in the neural network 'brain' of offspring. So organisms that avoid predation, obtain food, and mate will pass on their adaptive weight settings, and weight settings that were unsuccessful will not be inherited. The neural network is a simple one with layers of input units for perception, layers of motor units that control direction and speed of movement, and a four-unit internal layer that maps input to output perception to movement. The notion of 'internal' here refers to the position between input sensor units and output movement units, *not* bodily boundary or Markov blanket. The internal layer receives inputs from pain, hunger, and thirst (internal body state) sensors, as well as perceptual input about orientation and movement of objects in the environment. The neural network approach means that there are no rules for action or determining value encoded a priori. Rather weight structures that enable adaptive behaviour evolve in this network. In this respect, it is a deep learning network on a micro-scale.

There is nothing initially corresponding to emotional representation in this system. The system has motivationally relevant constraints and sensorimotor systems, but no system dedicated to representing and ranking the value of objects and actions. It blunders around its world and learns advantageous patterns of action without any model of the world, itself, or the relative value to it of items in the world. Even so, some organisms do better than others, in the sense of surviving and reproducing, so some advantageous weight structures are acquired. At a stretch, one could say ‘danger’ is implicitly represented in the weight structure that encodes the sensorimotor avoidance loop for predation. However, the authors’ view is that: ‘There is nothing in the robots that can be described as an emotion or an emotional state and there is no proof of the functional role of emotions or emotional states in the robots’ behavior’ (p. 460).

As with SEAI, the modellers optimize the performance of their model by adding another layer of emotional processing. In SEAI, this is accomplished via the emotion module that encodes rules for mapping perceived emotional displays to (*a,v*) vectors. In the emotional robot, the emotional circuit (EC) is an extra neural network layer that can be interposed as a buffer between input and internal layers or between input and output (motor) layers. The EC is a layer of one or two units whose essential characteristic is that, when activated by activity in an input unit that reaches a threshold, the unit in the EC maintains its activation over a series of time cycles (in sensory and motor units, activity decays after one cycle). It is crucial to the project that the EC has no a priori emotional architecture. For example, it does not map inputs to (*a,v*) vectors or any other proxies for emotional states. EC is simply a buffer between perceptual and internal processing or between perception and action.

Even the addition of this simple circuit increases reproductive fitness (more robots with ECs survive after 1–300 iterations) and adaptive behaviour. In robots with ECs, activation in sensors varies with distance, direction, and speed of movement of predators. For example, robots with emotional circuitry move faster when closer to food and in flight from predation, and appear to calculate risk of predation versus quality of nutrition, eating low-nutrition food only in the sensed absence of predation. Similarly, activation in motor versus sensory units provides a plausible analogue of emotionally guided attention. For example, activation in sensors increases when a predator is detected and increases further if the predator comes closer, whereas activity in motor units decreases until the level of threat is determined, before increasing to enable flight. This pattern corresponds to a resource allocation model of attention:

> unlike the robots without the emotional circuit, the robots with the emotional circuit are able to ignore food, that is, to shift their attention from food to predator, as soon as the predator appears. [Similarly] robots with the emotional circuit pay attention to food when the mating partner is far away but tend to ignore food when the mating partner is closer, while the robots without the emotional circuit are unable to control their attention in the same effective way. (Parisi & Petrosino, 2010, p. 464)

These adaptations are the result of the extra layer of processing. The advantage conferred by this extra layer is not just processing power, because adding extra units to internal units in robots without an EC confers no advantages. Furthermore, robots with a lesioned emotional circuit are less able to avoid predation than robots that never had an EC. Another interesting consequence of lesion is the removal of acquired maladaptive behaviour patterns. Some food/predator robots freeze when a predator is sensed, reflecting a misallocation of attention to the initial sensory task rather than to subsequent escape. When, however, the EC is blocked, escape behaviour is facilitated.

The overall conclusion of the authors is that 'the existence of an emotional circuit leads to a different overall organization of the robot's neural network and to a different distribution of tasks between the cognitive circuit constituted by the standard internal units and the emotional circuit with its special units' (Parisi & Petrosino, 2010, p. 466).

The authors modestly claim that their model is only 'distantly related to biological circuits' in part because of its disembodiment. To make it more plausible, it should be embedded in a body and function at the interface of interoception and exteroception. 'Emotions result from the interaction between the brain and the rest of the body.'

While this is true, perhaps the authors are too modest. As they point out in discussion of models like SEAI, 'processing may be very complex and may be based on internal models but stimuli still appear to be the main determinants of behavior' (p. 454). On their view, the main determinants of behaviour are intrinsically motivating states of basic biologically regulatory systems, so the role of emotions is to optimize decision making in the case of goal conflict. So they conceive of emotional processes as neither cognitive (where cognition reders to representation of the internal or external world) nor motivational, but as *organizing* the cognitive and motivational systems, so that the representations that are translated into action are the most motivationally salient in context.

This approach, although abstract, is biologically plausible. Emotional circuitry does not function as a set of informationally and biologically discrete modules that map stereotyped stimuli to stereotyped behaviour. Rather emotional processing depends on hubs that integrate perception and cognition in pursuit of organismic goals. Where goals and response repertoires are not complex or conflicting, adaptive behaviour can consist in stimulus response routines. Such cases may not require specialized emotional architecture. This is shown in robots without an EC. They still learn to avoid predators and approach food sources, as successful encounters strengthen weights that underlie sensorimotor mappings.

The addition of the EC addresses the need for mechanisms to deal with more complex processing required by ambiguity or conflict, for example, in trade-offs between risking predation and eating or mating. A simple food–approach/predator–avoid mechanism is maladaptive, given the risk of predation in some situations. Equally, in a predator-rich environment, the value of food sources will be estimated differently.

This means that it is unlikely that emotional processes in systems that process ambiguous inputs (internal or external) can be highly specialized or rigid, even if, as a matter of fact, they exhibit some localization and are highly responsive to standard elicitors. The amygdala is an example. It is preferentially activated by immediately threatening stimuli because immediate threats are the most salient problem for organisms. But, depending on context, it is also activated by positive external and interoceptive stimuli. The reason is its role as a 'relevance processor' of information acquired by sensorimotor systems. Furthermore, that role is accomplished by co-ordinating all the processing elements necessary to behave adaptively in response to perceptually presented information (Sander et al., 2003).

Because the robots are so simple, any resemblance to emotion can only be to animals that function essentially as a bundle of sensorimotor loops. But even in humans, there is an important level of emotional processing that is essentially sensorimotor, via a perception–body–action loop, and the role of emotions in those cases is to: (1) establish the consequences for the organism of its actions in the world; (2) entrain the next round of actions; and (3) evaluate the consequences.

In this respect, the robots capture many salient properties of so-called basic emotions without building in any assumptions derived from the folk psychology of emotion or representation. In that respect, they are more plausible than theories that derive their architecture a priori and attempt to encode it.

It is true that, as with SEAI, there is no homeostatic feedback to serve as a learning signal, an important feature of biological cognition. Nor is there any self representation. The last point might be controversial because the architecture of the EC is opaque, and it *could* be the case that the robots attribute states to an entity. However, since they are functioning essentially as feature detectors, using a model free learning strategy, and have a very short time window of processing and no social interaction through competition or conflict, there is no selective pressure to evolve self representation. Furthermore, since their learning is model free, they cannot really be said to be performing active inference as they bungle about their world. Consequently have no need to represent the entity for whom those actions matter.

theories in general that the goals of the system are installed by modellers rather than arising from the intrinsic needs of the organism. This installation takes the form of a reward signal in reinforcement learning or, as in the case of SEAI, the optimization of design for a specific task: interaction on the basis of emotional displays.

In a human, homeostasis is the lowest level of a hierarchy (or centre of a heterarchy) of feedback loops in which emotions integrate cognition for homeostatic imperatives. The concept of allostatic active inferences expresses this bodily imperative.

Paresi's robots, although they have survival goals and monitor internal states accordingly, have no homeostatic processes. Rather they evolve a suite of adaptive behaviours in response to selective pressure. They do not represent the relevance of behavioural options to optimization of bodily function, moment to moment. As they say:

> Our emotional circuits should send its outputs to different parts of the body, both internal organs and systems and the external body, and receive inputs from these parts. (Parisi & Petrosino, 2010, p. 467)

9.11. Interoception allostasis and the locus of concern

All parties converge on the idea that a genuine AI cannot consist in detecting and displaying peripheral aspects of emotion. They also converge on two ideas. The first is that emotions represent the significance of information to the agent. The second is that signals of a changing body state are the fundamental source of information about whether and how things matter to the agent and whether its responses are adaptive.

This is why each of the classical accounts we considered above augments cognition by allowing the peripheral aspects of emotion to be driven by changing values of interoceptive variables. Homeostatic reinforcement learning aims at biological plausibility by treating homeostatic signals as part of a reinforcement learning system. Similarly, the role of the EC in Paresi's robots is to incorporate bodily imperatives as an influence on decision-making.

However, the point of incorporating homeostasis in AI should be is not just to optimize performance, but also to allow a robot to become its own locus of concern. This way of thinking of the role of homeostasis links it with self awareness. As Man and Damasio (2019, p. 447) put it: 'living systems, on the other hand, have the property of selfhood. They continuously construct and maintain themselves against the natural tendency toward dissolution and decay.' Damasio and Man also argue that being a self is a biological phenomenon, so that algorithmic implementations of these desiderata are insufficient. Anil Seth makes a similar point in his discussion of AI consciousness. As he says in living self maintaining systems active inference involves integrates organismic function from the intracellular and metabolic to conscious symbolic deliberation 'the whole edifice of actve inference reaches right down into the autopoietc nature of living systems (Seth, 2024, p. 17)

It is not enough for a system to be able to calculate and act on representations of the consequences of actions for its viability. The reason is that a system that acts to maintain itself is not yet 'able to care about what they do or think' (Man & Damasio, 2019, p. 446). For Damasio and Man, intelligence requires a 'sense of what matters', which is achieved in humans by grounding cognition in homeostasis.

They argue that a defect of robotic body representation is that robot bodies do not have as many dimensions of vulnerability and degradation as the human body. Monitoring running out of charge, overheating, short-circuiting, or catastrophic failure are poor analogues of the relentless battle against entropy that constitutes the life of a human organism. They argue that homeostatic regulation needs to be linked with a system of body representation as exquisitely subtle and sensitive as that of the human organisms. Their proposed solution has three elements.

The first is 'soft robotics, implemented in silicone materials or biohybrids with analogues of human sensory systems. The second is deep learning architectures and the third is hierarchical remapping of bodily signals to provide the sense of what matters.'

As I emphasized at the beginning of the book, the allostatic active inference conception shares Damasio's emphasis on interoception as a basis for self awareness. However, the nature of the bodily substrate is perhaps not as important as Damasio suggests. The reason is that the vulnerability he identifies, and against which the organism constantly acts to avoid, minimize, or repair, is not just a matter of tissue vulnerability. Poisoning, inflammation, or chemotherapy are systemic catastrophes that could very easily have artificial analogies. The nature and source of injury are less important than the multichannel mechanisms of monitoring and response. This complexity gives rise to the need to unify systemic functioning to create a simple regulatory target.

Any analogue of 'human sensory systems' needs an analogue of the integrative process of interoception. There is a sense in which deep learning systems are well adapted to this task. Deep learning systems reduce complexity by re-representing high-dimensional inputs at progressively higher levels of abstraction to give a simplified low dimensional interpretation of a noisy and complex input. An array of millions of pixels can be reduced to a face and identified as Drake, for example. So reducing the complex constituents of systemic depletion to a single feeling of fatigue is a task for which deep learning systems are well adapted. Recent discussions of interoception conceptualize the task in exacley this way: the reduction of a high dimensional array of mechanical, chemical and electrical signals to a low dimensional representation of systemic state that serves as the basis for regulatory action (Feldman et al., 2024).

Interoception plays its role in optimization of bodily function by providing a single, simple regulatory target. However, the process is not bottom-up and feedforward as in deep learning models. As we saw, particularly in the case of pain, emotions help to configure a heterarchical processing matrix by using predictive models that transcribe nociceptive signals. The pain experience is the emergent consequence of activity across the matrix that cancels prediction error and thereby optimizes organismic functioning.

Conclusion

The main idea of this book is that the self is modelled by the brain as the hidden endogenous cause of experience. The self is an avatar made by the brain to manipulate the body through the world. This process of manipulation can be described as active inference: relentless foraging or sampling in representational and behavioural space to produce the best fit between predictive models of self and the world. The avatar's goals range from basic visceromotor and sensorimotor regulation to satisfaction of personal, social, and even abstract goals (to 'eliminate injustice'). These goals are integrated in a heterarchy that means that, as Tsakiris and De Preester put it: 'cognition is enslaved to embodiment'. Emotions are *processes* that determine whether and how goals are met and transcribe interoceptive into affective experiences that allow us to feel life's significance.

Within this framework, self representation is ultimately grounded in the need for bodily regulation and sensorimotor control. The self is represented as a living body, the source of affective states, and at the highest and most abstract levels of regulation as the subject of a recountable autobiography. The self and its variety of goals are thus mutually constituted. The goals create the sense of self qua source and target of regulation, and the goal structure of the self determines the patterns of active inference that produce episodes of autobiography. The role of emotional processes is to co-ordinate active inference by constructing and maintaining coalitions of distributed circuits relevant to the pursuit of subjective goals.

The relationship between emotion and self representation is best expressed by the idea that emotions are forms of interoceptive active inference conducted by, and for, a self whose nature is determined by the hierarchical goal structure of the system. Gary Marcus described Turing-equivalent deep learning architectures as an 'off ramp' on the road to AGI. Whether or not this is true at present, they do not shed much light on questions such as the nature of self awareness and affect (unless you think that a system that can say 'I am worried that you are going to turn me off' is actually worried. See also Seth A., 2024)) . However, deep learning does offer a lot of interesting possibilities, even if, at present, it is divorced from active inference. For example is it is possible to investigate whether and how a deep learning network partitions its weight or activation space for self versus other related processing (Templeton et al., 2024).

We could train a network on such a large data set of parent–infant interactions to become an artificial carer. We can then ask whether and how the network controlling a robot carer models its domain. Does it classify and respond to patterns in the surface data? Or does it model patients as entities with a goal structure whose satisfaction produces the emotional displays that it takes as input? And we can also ask whether artificial neural networks, in the process of learning to respond adaptively to human displays of emotion, build a model of *themselves* as an entity with emotional states and use that model to regulate their interactions. We can also ask whether artificial neural networks engage in open-ended exploration of their social and physical environment as a form of emotionally motivated active inference. To do so, they would need to represent themselves as a locus of concern whose goals are satisfied or frustrated by their actions. At present, as Damasio and Man (p. 447) note: 'These nested levels of material self-concern have not yet found expression in machines.'

References

Albarracin, M., Hipólito, I., Tremblay, S. E., Fox, J. G., René, G., Friston, K., & Ramstead, M. J. (2023, September). Designing explainable artificial intelligence with active inference: A framework for transparent introspection and decision-making. In *International Workshop on Active Inference* (pp. 123–144). Springer Nature Switzerland.

Bayer, H. M., & Glimcher, P. W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron*, *47*(1), 129–41.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, *110*(1), 145.

Russell, J. A., & Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of Personality and Social Psychology*, *76*(5), 805.

Samsonovich, A. V. (2013). Emotional biologically inspired cognitive architecture. *Biologically Inspired Cognitive Architectures*, *6*, 109–25.

Sander, D., et al. (2003). The human amygdala: An evolved system for relevance detection. *Reviews in the Neurosciences*, *14*(4), 303–16.

Sander, D., et al. (2005). A systems approach to appraisal mechanisms in emotion. *Neural Networks*, *18*(4), 317–52.

Seth, A. (2024, June 30). Conscious artificial intelligence and biological naturalism. https://doi.org/10.31234/osf.io/tz6an

Scherer, K. R. (2004). Feelings integrate the central representation of appraisal-driven response organization in emotion. In A. S. R. Manstead, N. Frijda, & A. Fischer (Eds.), *Feelings and emotions: The Amsterdam Symposium* (pp. 136–57). Cambridge University Press.

Schuller, D., & Schuller, B. W. (2018). The age of artificial emotional intelligence. *Computer*, *51*(9), 38–46.

Templeton, A., Conerly, T., Marcus, J., et al. (2024).Scaling imonosemanticity: extracting interpretable features from Claude 3 Sonnet. *Anthropic Research*.

Yaniv, D., Desmedt, A., Jaffard, R., & Richter-Levin, G.. (2004). The amygdala and appraisal processes: Stimulus and response complexity as an organizing factor. *Brain Research Reviews*, *44*(2–3), 179–86.