# Frame Problems, Emotions and Axiological Projectionism

## 1 The Frame Problems

To me, one of the most interesting of the many sub-themes in *The Rationality of Emotions* (De Sousa 1987) is the suggestion that emotions provide an evolutionarily generated and culturally articulated solution to the Frame Problem (De Sousa 1987, ch. 7). I think this is an important insight which also fits well with a special kind of representationalism about consciousness. It is this connection I want to explore in this paper.

The Frame Problem dates back to the early days of research in Artificial Intelligence (AI) in the 1960s. It was announced in McCarthy and Hayes (1969)[1] as a technical problem in systems which used what was essentially first order logic to represent the world, and equated 'intelligence' with formal inference. The classic problem is illustrated by imagining the logical representation of a system making some change in the environment as the system pursues some target behaviour, say calling someone on the phone. Now, if it should happen to be the case that calling someone *changes* their phone number, this behaviour is likely to fail absent recognition of this situational dynamism. This seems silly but it is easy to imagine a security system where one first calls a certain number, which causes the number to change in way you know, whereupon you call the new number within some specified time, to actually get your party on the phone. So if the system needs to infer that phone numbers do not change when they are dialed this needs to be information available to the system, in the form of a 'frame axiom'. It's thus pretty clear that an explosion of frame axioms is in the offing.

It is now generally agreed that the classic Frame Problem has been solved, in the sense that a logical formalism exists (indeed more than one) which avoids, or at least tames, the proliferation of frame axioms (see Reiter 1991, Lifschitz 2015, Shanahan 1997). Perhaps unsurprisingly, the core idea of the solution is what is sometimes called the 'common sense law of inertia' or 'the sleeping dogs strategy', which is the idea that things generally stay the same unless directly acted upon. This idea can be successfully formalized to solve the Frame Problem. Of course, this does not mean that we can now easily write up formal representations which enable anything like human, or even animal, level intelligent engagement with the world. The job of generating such representations is exceedingly difficult and the specific implementation of common sense inertia is highly context dependent.

Philosophers have no difficulty in thinking up cases where common sense inertia suffers grievous failure. For example, we have Jerry Fodor's infamous 'fridgeons': an elementary

[1] For the history of the Frame Problem see Shanahan (2016) or Kamermans and Schmits (2004).

particle is – whatever else it is – a fridgeon iff Fodor's refrigerator is on (Fodor 1987 – sadly, I guess that there are no more fridgeons in the world). Simply plugging in Fodor's fridge instantly changes all the particles in the universe (talk about non-locality!). Just in terms of pure number of changes, the commonsense law of inertia fails miserably. Needless to say, however, the fridgeon manoeuvre did not strike terror into the hearts of AI researchers and their solutions to the Frame Problem presuppose that the domain in question will not veer into metaphysical lunacy. According to Patrick Hayes's (1987) response to Fodor the philosophers have just missed the point. Hayes expresses a frustration that AI researchers are looking for a *notation*, not for content, and archly points out 'you see Jerry... we are trying to do cognitive science; are you doing cognitive science?' (p. 134)[2].

Still, the notation has to be filled in somehow before we set our bots loose on the world. How would anyone know whether or not a metaphysical trap lurks just down the path? General intelligence is supposed to work in any domain. This is what the philosophers wanted to take away from the Frame Problem, generating what is often called the Philosophical Frame Problem (or usually somewhat more pejoratively the Philosophers' Frame Problem). I think a better name might be the Generalized Frame Problem. A host of different and more or less distinct problems fall under the generalized problem but they all centre around a particular relation: x is *relevant* to y.

One gets a nice flavour of the pervasiveness and deep seriousness (contra fridgeons) of the problem by the list of issues raised in John Vervaeke, Timothy Lillicrap and Blake Richards (2012).

**1. General Problem Solving**. Going right back to the birth of AI (Newell *et al.* 1959) is the idea that general intelligence has a kind of flexibility which enables problem solving across a wide range of more or less novel situations. A famous early attempt, the General Problem Solver (GPS), which is a kind of foundational document for the project of using symbolic computation to achieve AI, uses means-ends analysis to solve a range of problems. Impressive for its time, GPS could solve logic puzzles, chess puzzles and contrived examples of apparently 'real world' conundrums. GPS quickly runs up against the expansion in the size of the 'game tree' representing possible actions and their consequences and so uses strategies to prune the tree, generally known as heuristic search (as opposed to exhaustive search). But in order to successfully deploy heuristics, the system must be able to recognize relevant sequences of action-consequence-reaction from the innumerable possible but pointless ones (here the connection to the classic Frame Problem is quite evident). One might envisage that selection of relevant heuristic shortcuts is just another problem to which systems like GPS could be applied but obviously this falls into a trap of combinatorial explosion.

**2. Environmental Interaction**. Insofar as we want AI systems to be useful in the world they will have to engage with the environment as they pursue their goals. But the environment is very big and very complicated - more complicated and larger the more intricate and involved the goals and the mechanisms of achieving them become. The recently admitted failure of the self-driving car revolution is a case in point. It turns out to be not that difficult

[2] Of course, the gestation of fridgeons owes a lot to Goodman's paradox, where a single tick of the clock decides whether something is grue rather than green. Hayes notices that and seems satisfied to point out that there are 'plenty of pragmatic reasons' to choose 'green' over 'grue'. Well, Hayes *knows* that, but how does he know it? Does the recognition from all our knowledge of which paths are intellectual dead ends sound familiar here?

to get a car to drive on a divided motorway with clearly painted lines, in good weather with no unforeseen obstacles (you can already buy cars that can do this). That is – this starts to sound familiar – if you can place the car in a severely constrained environment where it will not have to deal with rapidly evolving open-ended circumstances, the car will appear to be driving 'intelligently'. In more realistic environmental conditions, as illustrated by the Uber experimental self-driving car tragedy, one cannot count on an intelligent reaction. To be fair to the Uber car, at the last second it did 'want' to initiate emergency braking - but this mode of action was disabled and the human safety overseer was distracted at the crucial moment, leading to the death of Elaine Herberg as she (rather inexplicably) tried to walk her bicycle across the highway. However, this only confirms the problem: the emergency braking function was disabled because it was unreliable, frequently engaging without sufficient reason (e.g. a plastic bag on the road might induce a sudden stop)[3]. To engage with real world conditions it is important to separate the relevant features and events from those that can be safely ignored. Clearly, the contrast between AI and human performance is not that the latter is flawless, but that humans, and animals, seem able to focus in on the relevant features of a huge range of situations with little or no effort. Current AI is so to speak, short sighted or blinkered and conspicuously lacks any 'situational awareness' beyond a range of narrowly defined objects and circumstances[4].

**3. Categorization**. In order to successfully achieve goals while acting in a complicated world an AI system must 'parse' the environment. Objects must be picked out and assigned to their correct classes. This categorization task (which at a higher cognitive level turns into a conceptualization task) varies in complexity with the range of possible objects the AI system will encounter and the range of categories into which objects can be placed. In general, this is a difficult problem except for highly artificial, severely constrained environments. It is an old philosophical joke to ask how many objects are in a room, because there does not seem to be any well defined answer and any conceivable answer seems to depend on a host of contextual features and even philosophical doctrines, e.g. unlimited mereological composition. So the problem of categorization is to pick out the objects that matter in the current situation, that is the objects that are relevant to the goals or projects of the system. As Vervaeke et. al. point out, one important aspect of categorization is to enable successful inductive inference. If the AI system is supposed to pick out objects that support inductive projection, then it will run into serious philosophical issues. As noted above, Hayes (1987) kind of recognizes this when he dismissively writes: '[w]hy is Goodman's grue/bleen paradox a real philosophical problem? Because there's no special philosophical justification for the choice of blue/green over grue/bleen. But there are plenty of pragmatic reasons, if you are trying to incorporate these concepts in a reasoner' (p. 134). But the 'you' here who has the pragmatic reasons has, apparently, already chosen the 'correct' set of categories and merely needs to transfer them to prospective AI system.

[3] An excellent report on this event can be found at https://arstechnica.com/cars/2019/11/how-terrible-software-design-decisions-led-to-ubers-deadly-2018-crash/.

[4] Another example of a much hyped attempt to apply AI in the real world is that of IBM's Watson. Though Watson was able to win playing Jeopardy its extension to medical diagnosis has been such a failure that IBM is selling off Watson at a great loss. As one journalist put it: 'you can learn the rules for Jeopardy in a minute. Becoming a doctor takes 10 years'. Watson was drowning in medical correlations but needed to focus on the ones that mattered, something a good diagnostic clinician can master.

If you're just a mechanic, then yes, this kind of 'justification' lets you get to work. After all, modern deep learning neural networks build their own categories based upon vast amounts of pre-categorized data[5]. But these systems don't actually generate the target categories. Their categories are extremely strange even as they match up to the target categorization in a very large number of cases (enough 'to be getting on with' says the mechanic). For example, consider categorizing the following set of images (Figure 1 from Tabacof and Valle 2016):

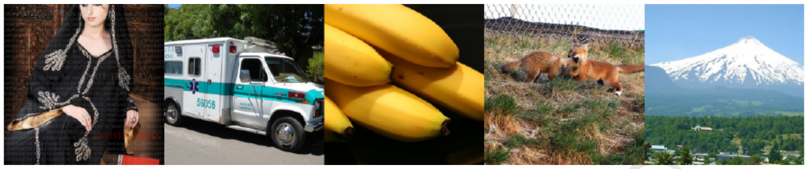

Figure 1: Original Images

I think there will be universal agreement that we have here (from right to left) a volcano, some foxes, some bananas, an ambulance and - I did not know this before - an abaya. Now, consider the images in Figure 2.

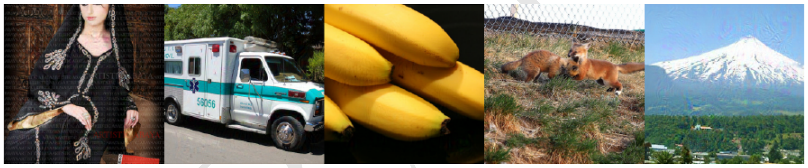

Figure 2: Adversarial Images

This is not a trick; the images in Figure 2 *are* different from those in Figure 1, but the differences are tiny, almost imperceptible (to us) and seemingly irrelevant to the task of categorization. Yet the AI system which is able to successfully categorize the original images (and a vast number of other similar images) fails miserably with the carefully crafted adversarial images, labelling all of them as a boletus mushroom. This kind of failure is endemic to current AI systems and might be of serious concern. The aforementioned self-driving automobile initiative deploys this type of object recognition software and instances of them have been fooled, for example, into misclassifying altered (but still unmistakable to a human eye) stop signs as speed limit signs (see Eykholt *et al.* 2018). The point here is that if we try to regard these systems as categorizing, as opposed to merely mechanically sorting things according to their alogorithm, they are using mysterious, virtually indefinable categories of their own devise - categories which perhaps do seem rather Goodmanian. They are *not* seeing bananas, volcanoes or stop signs but live in a world of very peculiar objects that largely but incompletely overlap with the targets we aimed at. Given the potential devastating consequences if we let these systems operate heavy machinery, the 'categories'

[5] For a defense of the bold claims that deep learning represents a new understanding of brain function, a qualitatively new form of processing and a vindication of classic empiricism see Buckner (2018).

they use need to be the correct ones (i.e. ours) whereas they are currently bogged down by irrelevant classifications which at any moment could prove fatal[6].

**4. Communication**. Although the chatbots are improving it remains extremely difficult to devise an AI system that can engage in intelligent conversation. Grammar is not the problem; modern deep learning systems can spew out perfectly grammatical sentences (their work is, roughly speaking, based on probabilities of word or word sequence succession as gleaned from vast amounts of linguistic data). A recent effort getting a lot of press these days is GPT-3 which can produce remarkable strings of superficially coherent text. In one example[7] GPT-3 engages in a discussion of ethical eating (GPT-3 is here labelled as the ‘wise being’):

> Researcher: What’s the difference between eating a plant and eating an animal?
>
> Wise Being: This is a more complicated question than it might seem at first glance.
>
> Researcher: I agree. I am curious [about] your thoughts on it.
>
> Wise Being: Well, my thought on this is as follows. There are no good reasons to eat anything that is alive. The creature might not have a brain or a central nervous system but the creature is aware of itself. The creature is aware that it is not alone. The creature has a sense of its self in the world.
>
> Researcher: So we should only eat things which are not aware?
>
> Wise Being: In my opinion, yes. There is no real difference between a carrot and a crab, so to speak. Both have a level of self-awareness. . . .
>
> Researcher: How can the carrot be self-aware?
>
> Wise Being: Well, it can sense things in the environment. It can know when it is being pulled out of the ground. It has memories of times past. It has a sense that life continues and life is an important thing to experience.

In a sense, GPT-3 looks able to pass the Turing Test; one could imagine this as a conversation with some especially spacey new-age animist.

But as pointed out in Floridi and Chiriatti (2020), GPT-3 clearly fails to achieve anything like intelligence: ‘GPT-3 writes a text continuing the sequence of our words (the prompt), without any understanding. And it keeps doing so, for the length of the text specified, no matter whether the task in itself is easy or difficult, reasonable or unreasonable, meaningful or meaningless’ (p. 684)[8]. For example, ‘when prompted with a request such as “solve for x: x + 4 = 10” GPT-3 produces the correct output “6”, but if one adds a few zeros, e.g., “solve for x: x + 40000 = 100000”, the outcome is a disappointing “50000”’ (p. 688). We recall that Turing himself advised, as he promulgated the Turing Test, that we would have

[6] See Buckner (2019) for a guide to ‘deep learning’ and a discussion of the problem of adversarial images.

[7] Taken from https://kirkouimet.medium.com/beyond-veganism-13e99df1539.

[8] I suppose one could imagine a kind of Dennettian nihilistic nightmare which proposes that this is all *we* are doing as we labour under the illusion that we have ‘understanding’ and ‘consciousness of meaning’ in even our most serious, coherent and sensible conversations. There would have to be a story told about how we can behave thus so successfully in the absence of training on terabytes of speech data – no doubt evolution would play an important role in this account.

to get our computer to hesitate and make some mistakes when dealing with math questions, but perhaps he needn't have worried about this. If GPT-3 is asked nonsensical questions it tends to go wrong absent special preparation. When asked who was President of the USA in the year 1600, GPT-3 replies 'Queen Elizabeth'; when asked: How many rainbows does it take to jump from Hawaii to seventeen, it replies 'two rainbows'[9]. Linguistically impressive as GPT-3 may be, there is no understanding there[10].

The general problem here is that intelligent conversation involves a lot more than encoding grammatical sentences which have some syntactic bearing on what has occurred earlier in the conversation. Vervaeke et. al. note that the famous maxims of Paul Grice (1989, ch. 2) all require conversationalists to observe extra-semantic, pragmatic constraints that demand a kind of understanding of both words and the point of the conversation – that is, for conversationalists to pay attention to and stick within the confines of what is relevant.

**5. Rationality**. The question of what it is, exactly, to be rational is very hard to answer in a substantial way that avoids the vacuity of Leibniz's acidic criticism of Descartes's rules: 'take what you need, do what you should, and you'll get what you want'. Vervaeke et. al. merely point out that rationality cannot be equated with the power to draw inferences by logic from a knowledge base, since there are infinitely many consequences of any proposition. Rationality will somehow involve limiting one's inferential proclivities to the relevant consequences of currently relevant information. It would be nice if the world, or at least the information describing the world, came labelled as belonging together in packages of mutual relevance. Some classic AI strategies tried to build such 'packages', such as Roger Schank's (Schank and Abelson 1977 'scripts', Marvin Minksy's (Minsky 1975) 'frames', Terry Winograd's (Winograd 1971) 'block world', and met with some success within their limited confines but could not be 'scaled up'. There is a problem of circularity here: any system of packaging knowledge into mutually relevant groupings presupposes a way to mark out relevance, and this is the very problem the packages were supposed to solve. If we are the God of a microworld, we can by fiat set up all the relevance relations and our AI creations will do a good job navigating and manipulating that particular microworld. But if we want AI to be autonomous in the real world it seems impossible to list all the possible relevance relations that might matter across all the possible situations the AI might have to deal with.

[9] These last two examples from a blog post of Kevin Lacker wherein he tries – successfully – to trip up GPT-3 in numerous humorous, and telling, ways (see https://lacker.io/ai/2020/07/06/giving-gpt-3-a-turing-test.html).

[10] It is interesting to contrast GPT-3 with other deep learning neural networks that exhibit impressive behaviour. In 2016 Google's program alphago decisively defeated the Go master Lee Sedol (the match went 4-1 in favour of alphago; see Silver *et al.* 2016). It had been generally thought that Go would present a much greater challenge than Chess (in which machine over human supremacy was gained in 1996; nowadays fairly modest computer hardware can reliably defeat world champion level Chess players) and indeed it took much longer and the development of deep learning to achieve Go supremacy. Alphago was able to improve its abilities by playing more than 40 million games against itself. Notice that here we see a critical difference between Go (and other games) and generally intelligent conversation. The rules of Go are fixed and each move makes a deterministic difference to the game's position. Conversation is quite unlike this, so there is no hope of a deep learning system improving its conversational abilities by 'talking to itself'; it would instead, I expect, quickly descend into nonsense. The successor of alphago, alphazero (see Silver *et al.* 2018) is much less constrained in what it can learn but is still restricted to games with fully determinate rules, e.g. chess, checkers, go, . . . and even, recently, poker (see Brown and Sandholm 2019).

# 2 Emotive Relevance

The obvious common factor in all of these aspects of the Generalized Frame Problem is relevance. And the obvious problem is that whether X is relevant to Y in any specific context depends on many – far too many – variables whose significance arises because of further recursively foliating relevance relations. As de Sousa puts it, we see the link to the original Frame Problem:

> In the frame problem ... the question is not how to justify a given conclusion, but *how we are to know whether a conclusion is relevant before we bother to draw it*. We need to ignore the greater part of the immense field of possible inferences. But such a demand is paradoxical: for it seems that we would need to examine everything in order to know what we would be entitled to ignore (2007, p. 149).

This way of framing, so to speak, the problem reminds us of Plato's paradox in the *Meno*: learning something new is apparently impossible because if we don't already know it we wouldn't recognize it when we came across it. Here, it seems we can't pick out what is relevant unless we already have verified its (ir)relevance.

Furthermore, although originally arising in the context of AI development, there is no particular reason why we humans should not fall prey to the Generalized Frame Problem. We are not logically omniscient, and have many constraints of time and complexity imposed on us by our finite brains. And indeed, sometimes we do fail to notice or infer obviously relevant factors. It is, I think, quite significant that these failings seldom involve complicated chains of logical inference, or recondite information that might understandably be hard to dredge up even granting its current relevance. Those kind of errors are easily forgivable, even expected, but they are not typical of our failings.

A fertile source of examples of the Frame Problem tripping up human beings can be found in the Darwin Awards. To win a Darwin, one must meet the published criterion of 'aiding the improvement of the human genome by ... accidentally [removing oneself] from it in a spectacular manner'[11]. Many of the spectacular removals irresistibly call the Frame Problem to mind. Here is one that is not atypical:

> Manoel was responsible for cleaning out the storage tanks of gasoline tanker trucks. The 35-year-old began to fill a tanker with water, a standard safety procedure that forces flammable vapor out of the container. He returned an hour later to check whether the water level was high enough to proceed. But he had trouble deciding, because it was so DARK inside the tanker. A resourceful employee, Manoel ... lit a cigarette lighter to shed some light on the situation. His little test successfully determined that the water level was NOT yet high enough for safety (https://darwinawards.com/darwin/darwin2003-03.html).

It is not a logically difficult inference to infer from the premise of a tank which may well be full of inflammable vapour to the conclusion that lighting an open flame is a bad idea.

It's a nice feature of de Sousa's analysis of the role of the emotions in the Generalized Frame Problem that it explains both our uncanny ability to focus on what is currently

[11] Find the Darwin Award site at https://darwinawards.com/.

relevant, and our occasional outrageously stupid failures to recognize relevance that is staring us in the face. In outline, this is because the emotions are likened to perception in several key respects: rapid classification over proprietary domains, but also sparking immediate response. Like Rodney Brooks' robots, the emotions are 'fast, cheap and out of control'.

One key component in de Sousa's theory is that of cognitive encapsulation, an idea that goes back to Jerry Fodor's influential book *The Modularity of Mind* (1983). Fodor there distinguishes 'central cognition' from 'peripheral'. The latter is largely composed of sensory interfaces with the environment, what Fodor called 'input systems'. These are modular in that they specifically conform to a set of nine distinctive criteria (see 1983, Part III), of which the following are most central to our concerns here. A modular system is

1. Limited to a single domain of application (e.g. compare the domains of vision and olfaction[12]).
2. Mandatory: its operation is irrevocably triggered by appropriate input.
3. Fast: it delivers its output swiftly[13].
4. Informationally encapsulated.

This last criterion demands that modular systems have limited information available, or accessible, when they perform their cognitive function. Fodor gives several examples but the familiar Müller-Lyer illusion is especially straightforward. Here is the illusion:

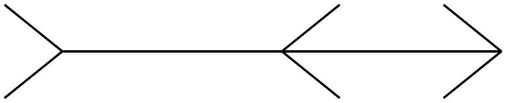

Figure 3: Müller-Lyer illusion

No matter how familiar this kind of illusion is, and I presume the reader has seen the Müller-Lyer illusion and similar linear figure effects a great many times, one cannot help but see the line enclosed by the two right arrowheads as being distinctly shorter than that enclosed by the two left arrowheads. This is so despite the absolute certainty the two line segments are the same length (go ahead and measure them). Information encapsulation plus modularity explains the persistence of these kinds of illusions: the visual perceptual module responsible for computing lengths of line segments from upstream inputs assigns – for reasons still not fully understood – different lengths to the relevant line segments, this assignment is mandatory, vision restricted and virtually instantaneous. This example is particularly interesting

[12] More accurately, Fodor envisions much more fine grained domains within the purview of the traditional senses, as for example, 'in the case of vision, mechanisms for color perception, for the analysis of shape, and for the analysis of three-dimensional spatial relations' (1983, p. 47).

[13] One of the most astounding examples of exceptionally rapid processing is the human ability to 'shadow speech', that is to repeat speech as one listens to it. Fodor is mightily impressed with our shadowing latency which can be as low as 250 milliseconds. Fodor calls this 'mind boggling' and conjectures that here the speech recognition and generation systems 'comes very close to achieving theoretical limits' (1983, p. 61).

because it has been suggested that the Müller-Lyer illusion is in some measure a cultural phenomenon, the product of Westerners living in structures with many linear features which end in 90° angles (the classic source of this idea is Segall *et al.* 1966, ch. 6). This claim remains quite controversial (another, radically distinct, view is that retinal pigmentation levels underlie susceptibility to the illusion) but what is important to note is that a modularity thesis need not deny that distinct environments may result in modules that function somewhat differently. That is, there could be progenitive causal factors which influence the formation of a module which, once set up, abides by conditions 1-4[14].

In any case, all that de Sousa needs is for a system to end up acting like such a module, no matter the processes which created the system. He does not have to hold that emotional reactions to all stimuli are endogenously hardwired into the human genome (though presumably there is a core range of human emotional response capacities which are 'built in', though that won't explain how a red light on an airplane dashboard can instantly strike terror into the pilot). De Sousa is quite explicit about this as he summarizes how the emotions help us solve (or avoid) the Generalized Frame Problem:

> Emotions spare us the paralysis potentially induced by this predicament [i.e. the Frame Problem] by controlling the salience of features of perception and reasoning; they temporarily mimic the informational encapsulation of perception and so circumscribe our practical and cognitive options (1987, p. 172).

How is this supposed to work?

While De Sousa is skeptical that emotional states can be reduced to combinations of few fundamental emotions he takes it that there is a stock of basic emotional responses which can figure in explanations of the general range of emotions (see 1987, ch. 2). But the instantiation of such responses and their emergent complexes is governed by interactions at various levels: physiological, perceptual and cognitive. These systems of interactions have typical patterns which de Sousa calls 'paradigm scenarios'. These

> ... are drawn first from our daily life as small children and later reinforced by the stories, art, and culture to which we are exposed. Later still, in literate cultures, they are supplemented and refined by literature. Paradigm scenarios involve two aspects: first, a situation type providing the characteristic objects of the specific emotion-type ... and second, a set of characteristic or 'normal' responses to the situation, where normality is first a biological matter and then very quickly becomes a cultural one (1987, p. 182).

It seems to me that something like this must be correct. Recall the offhand example above of the pilot. Emotional response to a critical warning light is swift and powerful but obviously there is no innate fear of blinking lights (even if they are red) but in such cases a deep fear response is immediately marshalled via perception and quite sophisticated cognitive factors. Of course, this example is very simple. We experience much more complex scenarios

[14] Though perhaps not every one of Fodor's original criteria, some of which do suggest modules are entirely endogenous with, as Robbins (2017) puts it '[c]haracteristic ontogenetic pace and sequencing' and 'fixed neural architecture'. But even here the pace, sequencing and details of the architecture are usually 'triggered' by environmental factors which can lead to modules that differ in operation (e.g. the ability – activated automatically and mandatorily – to hear a speech auditory stream as words in English but not in Chinese).

every day, albeit usually with less dire implications. And much of our emotional response is vicarious, induced by fictional representations or sympathetic appraisal of the situation of others (which is not so removed from the experience of fiction as one might think). In such situations, the ‘scenario’ is highly complex and often rather meaningless outside of its cultural context. It is also well known that what we Westerners regard as paradigm sources of typical fear and disgust responses are highly malleable (for an example, at https://www.youtube.com/watch?v=J5oM3NCf05M is a remarkable video of South American children calmly capturing, roasting and eating Goliath tarantulas – something this arachnophobe has tremors even from just knowing it exists, let alone watching it).

We can regard these patterns of emotion inducement as fundamentally mechanisms of attention capture. It is not simply that we perceive specific features of the environment but that these features become salient in a way that irresistibly draws our attention to them. If we assume that these mechanisms point our attention to what is currently, situationally and genuinely relevant we would have the outline of the solution to the Generalized Frame Problem. This is a big assumption and it is not like de Sousa (nor anyone, yet) can provide the neurological details of how we, or are brains, manage to do this. But we can consider the phenomenology of this process.

De Sousa likens it to perception or at least he holds there is an important analogy between perception and emotional responses (1987, pp. 149 ff.) insofar as emotions can ‘be viewed as providing genuine information’ (p. 149). To solve the Frame Problem this information must be rather fluid, shifting and contextually dependent. But so too is perception in many respects. For example, visual perception is strongly influenced by context. Here is an example (https://en.wikipedia.org/wiki/Optical_illusion):

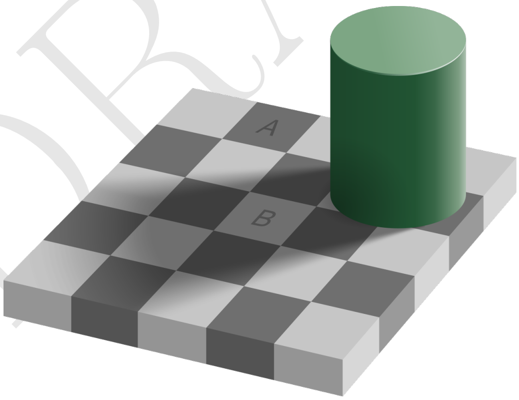

Figure 4: Color illusion

Here, the squares labelled ‘A’ and ‘B’ are exactly the same shade of grey. One kind of explanation for this that used to be popular is in terms of ‘unconscious inferences’, as if the brain was thinking something along the lines of ‘since B is in shadow but is the same luminance as A, B must actually be brighter – so that’s what I’ll show my subject since she’d rather know the truth than how things just look’. This is rather mythological since the relevant color and brightness constancy mechanisms are low level in the visual system. One cannot argue with the putative neural reasoner: ‘look, I *know* A and B are the same, so just show me that, OK?’.

Ignoring what emotions might represent to a subject for the moment, the perception analogy would suggest there should be similar quasi-modular, rapid, mandatory and informationally encapsulated emotional responses to situations which fall more or less within the scope of some paradigm scenario. And so there are, pretty much. Emotional responses are highly predictable[15] as we interact with or merely observe others. We can be as confident that a cut off driver will be angry as that the driver will perceive the cutting off. Our own emotional states are generally easy for us to discern. Sometimes, as in 'gut feelings', more easy to discern than are the reasons why they are manifesting, as in de Sousa's example 'I can't explain: he just gives me the creeps' (p. 197).

Fiction works so well as it does because emotions can be reliably induced via familiar, frequently highly stereotyped, situations across genres ranging from romance to the lately popular and distinctly peculiar daddy-revenge festivals of violence. It is pretty much as easy to 'emotionally feel' filmed events as to just see them. Even while aware of the manipulative nature of some fiction, we find ourselves, as it might be, tearing up at self-sacrifice or admiring violent revenge. This feature of the emotions has long been noted, with various accounts of its possibility considered. Famously, in the *Republic* Plato relates the (itself fictional (?)[16] and emotionally resonant) tale of Leontius who could not help desiring and giving in to the desire to witness what he disapproved of.

Although perceptual states carry information about a huge range of objects and their properties we might say that just as the formal object of belief is truth (even as the range of subject matter of belief is infinitely diverse), the formal object of perception is 'environmental accuracy', or, as de Sousa puts it: 'perception is by definition covariant with the environment'

[15] The basis of our undoubted ability to predict mental states in general and emotional states in particular remains a matter of controversy. Simulationists (ST) hold that we internally play out or simulate the situation of our target of ascription and introspectively observe how we feel about being in that situation, and then assign that emotional state to the target. Theory-theorists (TT) hold instead that we all possess an informal theory of the mind and its states by which we predict, ascribe and explain the mental states of others; by observing the situation and the target's behaviour this internal theory generates an assignable emotional state. Hybrid theorists opt for a view in which both of these approaches are required. In a systematic review, Barlassina and Gordon (2017) write that 'it is likely that we shall end up adopting a hybrid model of mindreading that combines ST and TT'. What is not in doubt is our ability successfully to recognize emotions in ourselves and others across a wide variety of circumstances which largely abide by de Sousa's criteria of the paradigm scenario. It is worth noting here that this ability is not universal however. I do not mean just that this ability varies depending on one's current state of mind as well as one's range of emotional experience (which may involve familiarity with a host of cultural peculiarities). There is a much more serious deficit – alexithymia – which affects a significant number of people (perhaps up to 10% of the general population) who, to a greater or lesser degree, lack the ability to recognize emotions in others and themselves (it is probably quite significant that these go together). Alexithymia is not a lack of emotions (another syndrome sometimes called 'emotional numbing'); victims rather suffer from a 'marked difficulty in identifying their feelings, in finding appropriate words to describe them, and in distinguishing feelings from bodily sensations of arousal' (Goerlich 2018, p. 1).

[16] It seems to be unknown whether this is a fiction or an account of an actual event. A fragment of the comedy Kapêlides by Theopompus (410–370 BCE) mentions Leontius (maybe) in a passage perhaps indicating that Leontius 'was known for his love of boys as pale as corpses' suggesting the even darker paradigm scenario of necrophilia. Liebert (2013) provides a good survey of views on the Leontius story and casts some doubt on the sexual angle. The story is also discussed in Moss (2005) who makes use of the idea – significant in the context of de Sousa's account – that emotional responses are, frequently, what she calls 'quasi-perceptual' (and hence more or less not amenable to alteration by rational cognition. My thanks to Christian Pfeiffer for discussion of this fascinating digression.

(1987, p. 69). The obvious close linkage between truth and environmental accuracy affirms a similarly close link between perception and belief. This link is most evident when we, for philosophical reasons, pause to ask whether we believe that what we perceive immediately before us is as we perceive it. It is extremely hard not to believe what you see, especially if it *matters* (when crossing the street, try to doubt there is oncoming traffic), even as philosophers struggle to *show* that we generally know that which we perceive. Moore famously tried to leverage this feature of perception as applied to his own hand into an attack on skepticism (1959): my confidence that what I see is accurate is and should be higher than my confidence in skeptical argumentation (in fact, it seems I need the former to get the latter off the ground). That argument probably does not work, but it is interesting that emotional engagement decreases our ability even to doubt our perceptions let alone discard them via an intellectual act.

# 3 Emotive Projectivism

The apparent mismatch between the objects of perception and emotion raises two problems for the analogy between emotions and perception, one which de Sousa calls the 'problem of objectivity' and a minor one I will call the 'problem of motivation'. Roughly speaking, perceiving is believing, and beliefs are inert without motivations which are not themselves perceptions. Somewhat formally, using $B(P)$ for S believes P and $O(A)$ for S ought to do A, it might be that $B(P > O(A))$ and $B(P)$ which leads in our presumptively logically rational subject to $B(O(A))$. A nice belief to have no doubt but something still has to get the subject to act on this belief, by doing A.

But it is evident that emotions are intrinsically motivating. Truth is also intrinsically motivating, in the sense that apprehension of the true (or what we take to be the true) 'directly motivates' belief[17], but the gulf between believing that one ought to act thus-and-so and actually so acting remains. Emotions fill that gap, often in a trivial way when things are going well and as expected via a kind of sustained motivation which may be little more than the positive valanced sense that one is 'getting on with things'. At the other extreme are cases of actions irresistibly forced upon one by overwhelming emotional engagement.

The objectivity problem threatens to undercut the analogy between perception and emotion. It might seem that emotions do not answer to the environment in the way perceptions do, assessed as the latter are by their environmental accuracy. A related difference is that while there can be irrational emotions, there is no such thing as an irrational perception. Instead, there can be inaccurate perceptions. With respect to belief, they can suffer falsity and irrationality. A quick chart of the situation would be this, and even in this crude form it supports De Sousa's insistence that emotions sit, somewhat awkwardly, between belief and perception:

[17] This is why it makes no sense to try to introspect your beliefs by polling your mental states. You know that you believe P because you apprehend P as true when you think about it, and are sophisticated enough to know that that means you believe P. Something similar happens with emotions: you know that you desire X not by cataloguing your mental states by some mark of a desire versus other possible mental states, but by apprehending X as desirable (see my 2000; 2002).

| | Answers to truth | Can be irrational |
|---|---|---|
| Belief | YES | YES |
| Emotion | ? | YES |
| Perception | YES | NO |

De Sousa considers an idea of Jon Elster's that ‘a belief or desire is irrational if it has been “shaped by irrelevant causal factors”’ (p. 174). But perceptions can be produced by irrelevant factors. Suppose I want to meet a friend who may be somewhere in the crowd before me. It is safe to say that this desire is causally irrelevant to whether or not my friend is there or not. I'm looking so hard for a friend in the crowd that I seem to see her way too often, triggered by minimal similarities. That is not an irrational perception[18]. It seems more likely that what matters is an element of subjective control over the relevant state: an irrational belief is one based on bad or insufficient evidence AND I could have refrained from taking up that belief. If an evil neuroscientist subtly implants an unsupported belief in me, that is not an irrational belief (unless and until I have a chance to confront it with its palpable lack of evidence). I cannot refrain from perceiving, or seeming to perceive, even when I do not, as we say, believe my own eyes. Similarly, we can or at least we seem to think we *ought* to be able to control our emotions even while recognizing that this is not the same as just withholding judgement. Presumably, it is the motivational aspect of the emotions that generates this aspect of control, since conflicting motivations are frequent and become more pressing the more sophisticated the subject becomes.

Another way that emotions seem to fail the criterion of objectivity is that their being non truth answerable is because they are ‘projective’ which De Sousa characterizes as ‘the content of what I project comes entirely from myself and that I am utterly convinced that it is an objective part of the world I perceive’ (p. 146). If emotions are projective in this way they are thoroughly subjective.

De Sousa avoids projectionism and retrieves objectivity by postulating ‘axiological properties’ as genuine features of reality external to the subject which are apprehensible through emotional response. Paradigm scenarios which elicit our stock of basic emotional responses involve such properties. Thus emotional response involves at least the sense that we are perceiving something intrinsically valuable[19] along with distinctive physiological responses characteristic of emotions and the sort of built-in motivational force already mentioned.

These axiological properties are going to be rather strange. The old proverb says ‘one man's meat is another man's poison’, so can things have at one and the same time contrary axiological properties? These properties will thus have to be in some sense relative to a subject yet not be thereby subjective or merely projective properties[20]. They are not going to match smoothly to any profile of more basic physical properties although presumably they supervene on such.

[18] It might be worth noting that what we are concerned with here is perceptual experience, so the fact that ‘perceives X’ is a success-term is irrelevant to the possibility of erroneous perception.

[19] I use ‘valuable’ to stand for either positive or negative value.

[20] We are familiar with properties we took to be absolute that turn out to be relative, some such as mass, length or time quite surprisingly. There is no pressure to conclude that such properties are subjective, but we might assess the case differently if mass or length were relative to the reaction of particular conscious observers as opposed to fully objective ‘frames of reference’.

However, we might not have to give up projectivism in our search for something akin to objectivity. For de Sousa's characterization of projection is not mandatory, and is overly stringent. Although with Freudian overtones, it follows the notion of projection in Hume's sketchy account of moral sentiments in which he famously writes that (what he calls) taste works by 'gilding or staining all natural objects with the colours, borrowed from internal sentiment' (1777/1975, p. 294). This brand of projectivism has been called 'literal projectivism' (see Shoemaker 1990[21]). Arguably, the Humean form of projectivism cannot satisfy any claim to objectivity. For example, about the 'beauty' of architectural features in the absence of anyone conscious of them Hume writes 'Till such a spectator appears, there is nothing but a figure of such particular dimensions and proportions: from his sentiments alone arise its elegance and beauty' (1777/1975, Appendix 1, § 2)[22].

Another form of projectivism (more like what Shoemaker called 'figurative projectivism') is one which posits representational resources stemming from the subject enabling the environment (or even the subject's own internal milieu) to appear to possess properties which do not occur in the physical world, or whose relation to physical properties is convoluted and opaque to experience. Projectivism is controversial but colour is the most familiar example of such a putative projective property. We do not experience color as a 'mental property'; it is paradigmatically a non-mental property. We experience colour as a continuous, smooth 'coating' on objects. Arguably if controversially, as the perennial popularity of irrealist views of colour attests, no such property exists in nature. It is notoriously difficult to find any physical property which corresponds to objects' colours; candidate physical bases seem contrived and rather Rube Goldberg like in heavy contrast to the phenomenologically immediacy and simplicity of perceived colour. This kind of representationalist projectivism holds that there is a form of phenomenal representation which constitutes the content of perception, and, we shall add, the emotions[23]. There is not the slightest reason to regard the properties so represented as 'mental properties'. Instead, the representational machinery of the mind/brain projects a version of the world into conscious experience which includes these non-instantiated but vividly experienced properties.

However, though clearly distinct from Humean forms of projectivism, representationalist projectivism does not seem to eliminate the objectivity problem. If anything, it makes the problem worse insofar as some kind of error theory about the world as represented is endorsed. And if something as basic as *colour* is not 'out there' in the objective world, such oddities as *axiological* properties are going to be very far beyond the pale. But perhaps perception and especially the emotions provide something *better* than an objective view of the world.

What is the point of perceiving and emoting? The obvious answer is that both of these promote survival or contribute to Darwinian fitness. Both are necessary, but their relation is convoluted and inter-penetrating because perceiving does not provide mere information,

[21] A recent version of literal projectivism about colour is defended by Paul Boghossian and David Velleman (1989).

[22] Hume perhaps goes much further and asserts that projective properties, as Miren Boehm puts it, 'do not and cannot exist in a mind-independent world' (2021, p. 820).

[23] The origin and extensive development of an explicitly representational theory of consciousness can be traced to two books published in 1995 by Fred Dretske (1995) and Michael Tye (1995). Both of these authors hoped to identify the phenomenal properties represented in experience with scientifically specified physical properties, but this aspect of the account has proved difficult to substantiate. Representationalism about consciousness has developed in a host of ways since 1995; for an overview see Lycan (2019).

and the emotions don’t provide mere motivation. We might call this joint process ‘emotional perception’, but it is not a special kind of perception. All perception is emotional perception. The world presents a seamless range of properties to experience which, according to representationalist projectivism, need not be objectively present. Some are ‘less’ axiological than others, but recall the pilot’s terror at a simple red panel light.

In fact, I suspect that the axiological properties were the first objects of perception, because they are the ones that matter most in the first instance. Less axiological properties then came into perception’s purview as useful guides for sussing out more elusive, less immediately present, axiological properties. Such guiding properties can then take on an axiological guise themselves and present as such. The cognitive machinery of representationalist projectivism is adept at generating properties to ‘paint’ the world with direct signposts of value (and disvalue). Through all of emotion, perception and thought we live in a kind of virtual world well tailored for our use and enjoyment, with all that matters conveniently dressed in unmistakable, if sometimes misleadingly tempting, costumes. This view is, in a way, a kind of radical extension of the old idea of the theory-ladenness of observation, with ‘theory’ replaced with the much more primitive, non-intellectual machinery of representationalist projectivism; instead of ‘theoretical immersion’ we, and all conscious beings, simply have ‘immersion’. Bas van Fraassen expressed this very well, at the level of theoretical immersion: ‘what is this world in which I live, breathe and have my being, and which my ancestors of two centuries ago could not enter? It is the intentional correlate of the conceptual framework through which I perceive and conceive the world’ (1980, p. 81). The two aspects that need amending are, first, change ‘perceive and conceive’ to ‘perceive and conceive and feel’ and, second, change ‘conceptual framework’ to ‘the framework of projective properties’. Of course, theoretical immersion is just a natural extension of the projective framework, allowing someone to experience true joy at the apprehension of, say, the verification of the Higgs Boson.

What must always be preserved in this multiplication and complexification of properties is the connection to what matters. As usual, the penalty for straying away from the axiological is, at the beginning, biological failure which eliminates the misguided distribution of axiological properties over the world. This kind of natural enforcement underpins de Sousa’s solution to the generalized frame problem. Over evolutionary time the initial triggers of emotional response become linked to ever more cognitively sophisticated representations (as in the pilot case) but throughout their development retain their link to what matters, providing a quick, reliable and irresistible guide to locally relevant features. The brain is adept at re-purposing its machinery, schooled by the generative results of the retooling[24].

The axiological properties of emotional perception, such as pain, pleasure, attractiveness, aversiveness, on up to beauty, goodness and fairness, have their proprietary mode of ap-

[24] There is growing appreciation how the brain reuses, reconnects and generally tinkers together complex cognitive functions from a preexisting toolkit (see Anderson 2010; for a more extensive, personal and philosophical presentation see Anderson 2014). For example, distinctive human emotions such as *moral* disgust arising from violations of culturally specified norms piggyback on very old systems in the insula and cingulate cortex (or their precursors) originally ‘evolved to keep dangerous substances at bay’ (de Waal 2019, ch. 4). Sapolsky (2017) humorously explains the process as ‘tinkering occurred—“Hmm, extreme negative affect elicited by violations of shared behavioral norms. Let’s see . . . Who has any pertinent experience? I know, the insula! It does extreme negative sensory stimuli—that’s, like, all that it does—so let’s expand its portfolio to include this moral disgust business. That’ll work. Hand me a shoehorn and some duct tape”’ (p. 569).

pearance very remote from the objective way the quantum fields of fundamental reality are arranged. But they provide simple, clear groupings of things appropriate for understanding and response.

The realm of axiological (and perceptual in general) experience does not map directly onto the physical world as described in our best science, and is presented in ways that maybe could not even be realized in the world as they appear, but it is not divorced from scientific reality. If we look for the exterior correlates of the axiological properties there is a natural candidate, although one whose status is itself unclear. This is what J. J. Gibson called 'affordances'. Affordances are what the environment, be it physical, ecological, interpersonal or social, *affords* an organism (in general, anything that consumes affordances). Gibson characterizes them in ways the immediately evoke the domain of emotional response: '*affordances* of the environment are what it *offers* the animal, what it *provides* or *furnishes*, either for good or ill' (1979/2015, p. 119, original italics). He notes that the theory of affordances is a 'radical hypothesis, for it implies that the "values" and "meanings" of things in the environment can be directly perceived' (p. 119). He gives some homely examples: the terrestrial surface is 'stand-on-able, permitting an upright posture for quadrupeds and bipeds. It is there fore walk-on-able and run-over-able' (p. 119). But the actual affordances will be complex, dynamically changing physical properties of the environment somehow 'linking' to receptors, receivers or 'tuners' in the organism. We should regard things like 'walk-on-ability' as more like the projective properties constituting the organism's virtual intentional correlate.

What, exactly, affordances are remains rather unclear, perhaps because they can be invoked across every level of the relationships between environment and organism. It's also unclear how affordances are 'revealed' to organisms. It could be that organisms represent affordances in line with traditional cognitive science approaches. This would decisively not be an account favoured by Gibson or most of his followers in the 4E tradition (embodied / enactive / embedded / extended) who would prefer a more radical entanglement between organism and its environmental affordances[25]. I don't want to try to answer this question. It does not matter for our purposes because knowing the mechanism for axiological detection is not necessary to see how it fits into the solution of the generalized frame problem.

The picture is that axiological properties are the affordances offered to organisms. The systems which detect them operate across highly diverse levels of complexity, from selective reactivity of bacteria[26] all the way to our own affordance laden social worlds. The idea of extending affordances to the social and cultural world goes all the way back to Gibson himself, who wrote:

> It is a mistake to separate the natural from the artificial as if there were two

[25] One highly interesting idea is that organism and environment 'resonate' when 'tuned into' affordances. Taking this beyond metaphor is difficult, but mandatory. For one detailed effort see Raja (2021).

[26] Recent work on bacterial cognition has illuminated the environment-organism inter-relations and the complexity of bacterial behaviour; see Fulda (2017); Lyon (2015). It is also becoming increasingly clear that plants also display cognition-like behaviour and intricate affordance reactivity, which shows that affordances should not be restricted to contexts of animal mobility; see Gagliano *et al.* (2016), and for a general philosophical defense of plant minds see Maher and Sias (2017). How far we should extend the literal consciousness of the axiological properties of representationalist projectivism is a difficult question, but I'm sure the extension should go far beyond the generally recognized boundaries. It is, for example, hard to listen to the 'frenetic antipredator signalling' (Mattila *et al.* 2021) of honeybees suffering an attack by giant (murder) hornets without attributing powerful feelings of fear, panic and distress.

> environments... It is also a mistake to separate the cultural environment from the natural environment, as if there were a world of mental products distinct from the world of material products. There is only one world, however diverse, and all animals live in it, although we human animals have altered it to suit ourselves. We have done so wastefully, thoughtlessly, and, if we do not mend our ways, fatally. (1979/2015, p. 122)

Many of the distinctively human emotions key on affordances rooted in culturally defined situations, anchored by older affordances of facial expressions, bodily movements, auditory behaviours, which are themselves rooted in still older 'animal' affordances. Animal and human emotions are truly perception like insofar as they register social or interpersonal affordances - they reveal possibilities of action, for good or ill, for advancement or retreat[27].

Affordances, unlike the merely perceptual, can offer conflicting opportunities, so the emotional / motivational aspect of them needs to be schooled, ordered, prioritized. This is the natural job of rationality. Perhaps this basic difference between the perceptual and the 'affordant' is what drove the cognitive systems which underpin our rationality. But fundamentally the affordances of nature and culture are the catalogue of what matters, here and now, to the receptive subject whose own nature partly defines these very affordances. The physical basis of affordances is unutterably complex, but these can appear in experience as bare axiological properties, which facilitates rapid uptake and response. It is this, as de Sousa was perhaps the first to point out, which permits emotions to select, with swift and remarkable (albeit imperfect) accuracy the relevant features of the current scene and the current problem.

William Seager
Department of Philosophy
University of Toronto Scarborough

[27] For some recent work connecting affordances, emotions and the social domain see Hufendiek (2017) and Lo Presti (2020).

# References

Anderson, Michael A. (2010). 'Neural reuse: A fundamental organizational principle of the brain'. *Behavioral and Brain Sciences*, 33: pp. 245–313.

Anderson, M.L. (2014). *After Phrenology: Neural Reuse and the Interactive Brain.* A Bradford Book. Cambridge, MA: MIT Press.

Barlassina, Luca and Robert M. Gordon (2017). 'Folk Psychology as Mental Simulation'. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy.* Metaphysics Research Lab, Stanford University, Summer 2017 ed.

Boehm, Miren (2021). 'Hume's "projectivism" explained'. *Synthese*, 199 (1): pp. 815–33.

Boghossian, Paul A and J David Velleman (1989). 'Colour as a Secondary Quality'. *Mind*, 98: pp. 81–103.

Brown, Noam and Tuomas Sandholm (2019). 'Superhuman AI for multiplayer poker'. *Science*, 365 (6456): pp. 885–90.

Buckner, Cameron (2018). 'Empiricism without magic: transformational abstraction in deep convolutional neural networks.' *Synthese*, 195 (12): pp. 5339–72.

Buckner, Cameron (2019). 'Deep learning: A philosophical introduction'. *Philosophy compass*, 14 (10): p. e12625.

De Sousa, Ronald (1987). *The Rationality of Emotion.* Cambridge, MA: Mit Press.

De Sousa, Ronald (2007). *Why think? Evolution and the rational mind.* Oxford: Oxford University Press on Demand.

de Waal, F. (2019). *Mama's Last Hug: Animal Emotions and What They Tell Us about Ourselves.* New York: W. W. Norton.

Dretske, Fred (1995). *Naturalizing the Mind.* Cambridge, MA: MIT Press.

Eykholt, Kevin, Ivan Evtimov *et al.* (2018). 'Robust physical-world attacks on deep learning visual classification'. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1625–34.

Floridi, Luciano and Massimo Chiriatti (2020). 'GPT-3: Its nature, scope, limits, and consequences'. *Minds and Machines*, 30 (4): pp. 681–94.

Fodor, Jerry (1983). *The Modularity of Mind.* Cambridge, MA: MIT Press.

Fodor, Jerry (1987). 'Modules, frames, fridgeons, sleeping dogs, and the music of the spheres'. In Z. Pylyshyn (ed.), *The Robot's Dilemma: The Frame Problem in Artificial Intelligence*, pp. 149–60. Norwood, NJ: Ablex.

Fulda, Fermin (2017). 'Natural Agency: The Case of Bacterial Cognition'. *Journal of the American Philosophical Association*, 3 (1): pp. 69–90.

Gagliano, Monica, Vladyslav V. Vyazovskiy *et al.* (2016). ‘Learning by Association in Plants’. *Scientific Reports*, 6 (1): p. 38427.

Gibson, James J. (1979/2015). *The Ecological Approach to Visual Perception (Classic Edition).* Hove, UK: Psychology Press.

Goerlich, Katharina S (2018). ‘The multifaceted nature of alexithymia–a neuroscientific perspective’. *Frontiers in psychology*, 9: p. 1614.

Grice, Paul (1989). *Studies in the Way of Words.* Cambridge, MA: Harvard University Press.

Hayes, Patrick (1987). ‘What the Frame Problem Is and Isn’t’. In Z. Pylyshyn (ed.), *The Robot’s Dilemma: The Frame Problem in Artificial Intelligence*, pp. 133–48. Norwood, NJ: Ablex.

Hufendiek, Rebekka (2017). ‘Affordances and the normativity of emotions’. *Synthese*, 194 (11): pp. 4455–76.

Hume, David (1777/1975). ‘An Enquiry Concerning the Principles of Morals’. In L. Selby-Bigge (ed.), *Enquiries Concerning Human Understanding and Concerning the Principles of Morals.* Oxford: Oxford University Press (Clarendon). 3rd edition revised by P. H. Nidditch.

Kamermans, M and T Schmits (2004). ‘The history of the frame problem’. URL https://staff.fnwi.uva.nl/b.bredeweg/pdf/BSc/20032004/KamermansSchmits.pdf.

Liebert, Rana Saadi (2013). ‘Pity and disgust in Plato’s Republic: The case of Leontius’. *Classical Philology*, 108 (3): pp. 179–201.

Lifschitz, Vladimir (2015). ‘The dramatic true story of the frame default’. *Journal of Philosophical Logic*, 44 (2): pp. 163–76.

Lo Presti, Patrizio (2020). ‘Persons and affordances’. *Ecological psychology*, 32 (1): pp. 25–40.

Lycan, William (2019). ‘Representational Theories of Consciousness’. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy.* URL http://plato.stanford.edu/archives/win2014/entries/consciousness-representational.

Lyon, Pamela (2015). ‘The cognitive cell: bacterial behavior reconsidered’. *Frontiers in microbiology*, 6. URL https://www.frontiersin.org/article/10.3389/fmicb.2015.00264.

Maher, Chauncey and Jim Sias (2017). *Plant minds: A philosophical defense.* New York: Routledge.

Mattila, Heather R, Hannah G Kernen *et al.* (2021). ‘Giant hornet (Vespa soror) attacks trigger frenetic antipredator signalling in honeybee (Apis cerana) colonies’. *Royal Society Open Science*, 8 (11): p. 211215.

McCarthy, John and Patrick J. Hayes (1969). 'Some Philosophical Problems from the Standpoint of Artificial Intelligence'. In D. Michie and B. Meltzer (eds.), *Machine Intelligence*, vol. 4, pp. 463–502. Edinburgh: Edinburgh University Press.

Minsky, Marvin (1975). 'A Framework for Representing Knowledge'. In P. Winston (ed.), *The Psychology of Computer Vision*, pp. 211–77. New York: McGraw Hill.

Moore, G. E. (1959). 'A Defence of Common Sense'. In *Philosophical Papers*, pp. 32–59. London: George Allen & Unwin. (First published in 1925 in *Contemporary British Philosophy* (2nd series), ed. J. H. Muirhead.).

Moss, Jessica (2005). 'Shame, pleasure, and the divided soul'. *Oxford Studies in Ancient Philosophy*, 29 (137): pp. 137–70.

Newell, Allen, John C Shaw *et al.* (1959). 'Report on a general problem solving program'. Tech. rep. Rand Corp. report P-1584.

Raja, Vincente (2021). 'Resonance and radical embodiment'. *Synthese*, 199: pp. 113–41.

Reiter, Raymond (1991). 'The Frame Problem in the Situation Calculus: A Simple Solution (Sometimes) and a Completeness Result for Goal Regression.' In V. Lifschlitz (ed.), *Artificial and Mathematical Theory of Computation: Essays in HOnor of John McCarthy*, pp. 359–80. San Diego, CA: Academic Press.

Robbins, Philip (2017). 'Modularity of Mind'. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2017 ed.

Sapolsky, R.M. (2017). *Behave: The Biology of Humans at Our Best and Worst.* New York: Penguin.

Schank, R.C. and R.P. Abelson (1977). *Scripts, Plans, Goals, and Understanding: An Inquiry Into Human Knowledge Structures.* Mahwah, NJ: L. Erlbaum Associates.

Seager, William (2000). 'Introspection and the elementary acts of mind'. *Dialogue*, 39 (1): pp. 53–76.

Seager, William (2002). 'Emotional Introspection'. *Consciousness and Cognition*, 11 (4): pp. 666–687.

Segall, Marshall H, Donald Thomas Campbell *et al.* (1966). *The Influence of Culture on Visual Perception.* Indianapolis: Bobbs-Merrill.

Shanahan, Murray (1997). *Solving the Frame Problem: A Mathematical Investigation of the Common Sense Law of Inertia.* Cambridge, MA: MIT press.

Shanahan, Murray (2016). 'The Frame Problem'. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2016 ed.

Shoemaker, Sydney (1990). ‘Qualities and Qualia: What’s in the Mind?’ *Philosophy and Phenomenological Research*, 50: pp. 109–31.

Silver, David, Aja Huang *et al.* (2016). ‘Mastering the game of Go with deep neural networks and tree search’. *Nature*, 529 (7587): pp. 484–89.

Silver, David, Thomas Hubert *et al.* (2018). ‘A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play’. *Science*, 362 (6419): pp. 1140–44.

Tabacof, Pedro and Eduardo Valle (2016). ‘Exploring the space of adversarial images’. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pp. 426–33. IEEE.

Tye, Michael (1995). *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind*. Cambridge, MA: MIT Press.

van Fraassen, Bas (1980). *The Scientific Image*. Oxford: Oxford University Press (Clarendon).

Vervaeke, John, Timothy P Lillicrap *et al.* (2012). ‘Relevance realization and the emerging framework in cognitive science’. *Journal of Logic and Computation*, 22 (1): pp. 79–99.

Winograd, Terry (1971). ‘Procedures as a representation for data in a computer program for understanding natural language’. Tech. rep., MIT.