Talking with robots about poetry

Cliff O'Reilly

Birkbeck, University of London

November 29, 2022

Abstract

If phenomenal consciousness is illusory and reduces to a perceptual interface between a belief-driven ego system and a lattice platform of both singular and hybrid connectionally-constructed concepts where phenomena themselves are sensation-induced constructs of complexes of things in the world then in order to communicate in terms of true meanings we would require a translation system between the complexes of things that have importance for creatures of different biologies whether evolved or synthetic. In this paper I investigate and propose an ontological scheme that can act in some way to bridge this gap and potentially allow humans to talk with robots about themselves and the world.

1 Introduction

What is a robot? In this essay I will consider mostly phenomenal consciousness and an ontological approach to concept mapping, but what I take to be a robot is an intelligent machine. David Chalmers talks about robots in his re-framing work on the problem of consciousness [1] and I use the term loosely too, rather than the more mechanical interpretation such as a cybernetic slave. By intelligent machine I refer to a hypothetical connectome capable of general intelligence comparable to basic human-level cognition.

There is a profusion of thought in recent times on the nature of consciousness. This essay attempts to add to those thoughts. I take a perspective of reducing the objects of

perception to a practical set of tools that could be used to aid communication with non-human entities, such as robots. Thinking about consciousness is not a novel pursuit. As far back as the earliest human writings we have evidence that wise people in the ancient world could see the issues with which we still grapple today — i.e. the nature of conscious perception and how it relates to ourselves and the world: what makes you *feel* like *you* and how can these thoughts and *feels* relate to objects in the world that are seemingly not *you*.

In more recent times, however, the advances in philosophy and technology (backed by a growing understanding gained through scientific investigation) have re-framed the conundrum by positing numerous theories, some of which are challenging to our basic intuitions as humans. Perhaps this should not come as a surprise given that our conscious perceptions seem out of reach and ethereal and to a large extent defy definition.

I take a generally Illusionist stance on consciousness: the idea that perceptual qualia are in some way constructed by our brains for good behavioural reasons evolved over millions of years. The Illusionist position suggests that our perceptions of the world are imagined, but relate to and are caused by the world in which we exist. I stretch this position to say that qualia — the phenomenal perceptions with which we are intimately aware, yet struggle to define — are actually no more than blends of worldly concepts (neural structures). I will argue that these complexes of information inputs to the brain are processed by a series of interconnected *layers* which results in a lattice of concepts. Up to this point these functions would be similar in many different types of creature — there is nothing peculiarly human about this. Most humans, and some other creatures, however, have another processing unit which I term the *ego module* since the purpose (driven by the evolutionary benefit it enabled) is to gain agency over the world around us. I suggest that without a sense of self and an arrogance that the world is ours with which to tinker and control we wouldn't have phenomenally conscious experiences.

The basis for this thought is similar to, but different from the *Bundle* theory of substance, proposed by David Hume and taken on by many others. It is not claimed here that substances themselves are constructs of properties, but only that the perception of them — i.e. the concept matrix we have in our heads — is a function of the many sensory and secondary properties and qualities of objects and concepts.

If this argument holds and our perceptions are in fact dynamic mixtures of primary sensory inputs blended with evolved secondary concepts which support our fitness for our environment, then it may be possible to break down the mixture by combination of metaphysical investigation and statistical analyses in order to map out parts of what the mixture could look like. An object ontology of conceptual cognition is developed in this essay.

This is a physicalist and a materialist essay — I take it that the world exists beyond my brain (but that I construct my own personal view of what it is) and phenomenal consciousness is not something beyond the material world (by considering the problem of phenomenal consciousness illusory the problem disappears to some extent). However there is a sense in which the argument on consciousness leads to a Panpsychist view.

In this essay I shall define the problem to be addressed, outline numerous background theories and positions, propose an argument that perceptions are in fact a conceptual blend and nothing more, and a construction of object ontology which aligns with the perspectives as argued is developed and attached. The implications for this ontology is discussed and a proposal for extra-human communication put forward which may lead to methods of design and interaction with artificial intelligent machines of the future.

Finally the title proposition is advanced in an argument that suggests we can talk poetry with artificial machines.

1.1 Definitions

Some of the terms I have used vary from the standard therefore it is worth clarifying ahead of their use.

- consciousness: the result of information processing in systems
- concept: a structure in complex systems that may relate to and respond to stimulus of a specific object real or abstract
- ego: the sense of self that enables control over the environment, without which humans wouldn't create tools
- mind: general term for higher brain functions and especially conscious awareness

- awareness: the neural function that introspectively composes our hybrid concepts into something we process in real-time
- connectionist: a general term for networks of nodular processing units which, when combined in large numbers, can generate high complexity. A well understood example would be neural networks.
- connectome: a connectionist system
- qualia: the content of a conscious phenomenal experience, e.g. the redness of seeing red

2 The Problem

The problems addressed in this essay are two-fold: firstly the theoretical question over how or why humans have and report phenomenal experience and secondly the practical solution to what it might mean for a non-human being to be able to communicate with humans in a meaningful manner. The first is perhaps as old as human thought, but the second perhaps more recent and more in focus recently as artificial machines take a greater role in our societies.

It is by now an age-old problem — that of how we consider our perceptions and their relationship with the outside world. Even the thought of an outside world implies an inside world — one to which only we have access and is private and special. The problem seems intractable — how can we possibly explain something so private and ungraspable in terms that make sense beyond our minds.

In recent times the growth of technological explanations from how the brain functions seems exponential. The advent of neuroscience and the understanding of the physical processes from the level of quarks and the quantum world to the realm of complex connectionist systems gives us far more knowledge with which to gain insight than had the many philosophers from the time of the earliest pre-socratics and Indian and Chinese thinkers, yet it seems to me, on reflection, that the fundamentals of how we work as humans — and not just mechanically — is lagging behind. In this essay I will attempt to add more to the philosophical side of the see-saw and redress the balance away from technical advances although the knowledge gained from science is

crucial to answering the goal proposed in the title which is to be able to talk with robots about poetry.

David Chalmers is a leading philosopher in this field and has suggested that perhaps by re-framing the problem into a *meta-problem* we can gain traction against it. The question of how phenomenal consciousness can exist can be thought of as the question of why we think it is a problem at all. This is a brilliant way to think about the issue and although Chalmers has sympathy with the Illusionist approach, he may not concur that by reducing qualia to illusions the hard problem disappears even though the meta-problem remains.

3 Background

The world is not *in itself* the same as the way we perceive it. Some would say that there's an element of transparency regarding perception — that we see a tree in the same way that it presents to us via a set of stimuli. However when investigations of phenomena in terms of abstract properties are undertaken it can be shown that there is a disparity between the tree and our perception of it. It certainly feels to us as if we see the thing *exactly* as it is, however depending on what time of day we view the tree it will be reflecting quite a different set of stimuli (photon properties) and therefore we simply have to realise that our brains construct the world for us rather than presenting it to us how it really is. Similarly when we think about visual illusions[1] and hallucinations, dreams or even if we just close our eyes and use our imagination, we are constructing a world that cannot be the same as the one in which we exist in physical terms. Our eyes (and other senses) can trick us and conversely we can construct our own phenomenal world while disconnected from *reality*.

3.1 Illusionism

For Daniel Dennett, an early proposer of the idea that conscious perception is something, "You may *think* you're directly acquainted with... but that's a fact of personal psychology" [2]. He also describes a typical Illusionist process as "When there is a red

[1] `https://michaelbach.de/ot/`

stripe in the world, the redness is a complex physical property of the stripe; when there just seems to be a red stripe in the world, that very same property is *represented* as being present by some team of brain agents that are the cause of your false conviction."

Dennett also proposes a useful analogy for the Illusionist perspective. The mind is a bit like the abstract user interface of a standard desktop personal computer. Where there are icons that represent things like hard disk folders or applications these abstract away hidden complexities of how the computer manages its internal devices or executes software. This is similar, Dennett thinks, to how the mind acts in an abstracted away state from the hidden complexities of the neural complex that underlies it.

Perhaps the leading proponent of the Illusionist project is Keith Frankish. The collection of essays edited by him forms a group of seminal thinking on the newly-forming subject. It has many critics, but often the arguments against it form-up as simply based on intuition — how could this experience I am having be an illusion? As unintuitive as it may seem it appears to be a growing research field essentially due to the power to effectively solve the hard problem of consciousness (although not without other costs) and it sits well with cognitive science which finds that often the experiences we claim to be inner private qualia don't stand up to scrutiny under investigation.

The key argument is summarised as the "view that phenomenal consciousness, as usually conceived, is illusory" [3]. As humans (and this potentially applies to any other creature or potential robot) we benefit from an introspective ability which misrepresents conscious experiences as having phenomenal properties. It is not an Idealist approach, but entirely Physicalist in that the material world creates impacts on our sense and these, in turn, develop into brain states to which we have internal access. This is all physical stuff and the last step whereby we *see* the world as is not as we believe it to be and is in effect a construction of our own minds. Frankish calls these properties which we believe we are experiencing as *quasi-phenomenal* indicating that there do really appear to be qualia which we can experience, but these are not really as they appear and do not require special explanation (e.g. via dualist or panpsychist theories). We can instead explain, via a placeholder for a future more detailed theory that explains the mechanisms of such as function, that we have evolved to experience the world in these special ways and have gained an advantage over other species (if it can be called that).

3.2 Attention Schema Theory

Michael Graziano and others have recently proposed an approach to engineering conscious machines [4]. I take issue with his overtly practical approach to consciousness science: "As our information technology has improved, the information content of the mind has become less mysterious", however I think he is right by saying that "at the same time the act of being conscious of it, of experiencing anything at all, has become more remote and seemingly unsolvable" [4]. Contrary to the second statement he makes what seems like an incredibly bold statement: "We are close to understanding consciousness well enough to build it", but he is not sure if it is engineerable at this point in time. However a general proposal for how elements of a chain of neuronal structures integrate to form conscious is proposed and includes an internal model of the self, sets of models about objects in the world and an attentional model of the interface between the self and those objects. This high level description seems convincing, but it is at such a gross level that it would be hard to believe it wasn't somehow accurate given what scientists have discovered about cerebral modularity.

Graziano seeks a direct implementation of such a set of models, however I don't believe that sufficient background foundations are considered in terms of what is needed to 'understand' the world. A model is a great idea, but it needs to be grounded in something that makes the model accurate. In humans we have hundreds of millions of years of evolution behind us to help us effortlessly construct models of the world. It is significantly more difficult in a robot that is starting from scratch (or worse is being guided by a creator that hasn't fully understood the mechanics of the machine it is trying to create).

3.3 Integrated Information Theory

Integrated Information Theory (IIT) is a theory proposed by Guilio Tononi and espoused and developed by Kristoff Koch. In this approach consciousness is a function of information processing and the amount of consciousness is proportional to the complexity of the system. This complexity is represented by a numeric value termed *Phi*. The way that information is integrated across something like a brain and the resultant complexity is sufficient for consciousness — or the appearance of it — to become

evident. All systems have some number Phi (which can be zero indicating zero consciousness). Consciousness is a physical form of matter, a product of the complexity of physical elements. The theory is highly mathemetised and practical and being investigated by scientists actively.

Any conscious experience has five axioms at its centre:

1. it exists for itself
2. it is structured
3. it is specific the way it is
4. it is one
5. it is only one

Similarly there exist associated postulates which are essential properties of any underlying substrate:

1. intrinsicality
2. composition
3. information
4. integration
5. exclusion

The theory is highly developed and has similar illusionist implications in terms of phenomenal consciousness. There are potentially overlaps with the ideas put forward in this essay, however the conclusions are not the same.

4 Ontology of cognition

A mouse rustles through the undergrowth of a corn field. We might say that its behaviour is governed by a series of perceptions and reactions which, to human observers, may seem erratic, swift and nervous. The reactions its genetic forebears made were directly correlated to their very survival and success and, being their progeny, mice of today are therefore adapted to their environment in order to both protect themselves against their predators and to enable flourishing through reproduction and sustenance.

Taking a patronising approach and assuming that if we reduce the dimensions along which a mouse may concern itself we might draw up the following set and probably in this order of importance since each successful concern becomes somewhat redundant if the previous one is unfilfilled:

1. evade predators
2. find food
3. find a mate
4. protect progeny

A similar reduction can be made regarding human concerns. We act in our environments in ways that follow the same dimensions. We fear for our safety (albeit in most modern societies this concern has thankfully been removed in the main) and we are driven both consciously and subconsciously to find food, a mate and then to protect and nurture our offspring. The nuance and practicalities of core behaviours can vary enormously, but for many we can trace the ways we act to a set of principles. If this is the case how, then, is a mouse different from a human with regards to basic concerns? Neurologically we have a larger and more complex brain and this likely correlates to the set of higher concepts that we could assume we are able to manifest. Explaining what a sub-prime loan is to a mouse or the meaning of a poem to a robot might be extremely difficult, but in principle we should be able to deconstruct it so as to make some sense even if far more coarsely grained in meaning. To do this we have to translate away from human concepts and into a more general or abstract set of ideas; in effect to talk a common language with non-humans. Some concepts are particularly human and therefore probably not translatable, but getting the gist of something might in theory be possible. The advent of emojis in digital communication is, in a way, a step back to early forms of abstract communication such as Egyptian hieroglyphs. Their semiology is complex and laden with common understanding that has evolved in parallel with their use in social media, for example. The core images that reflect basic emotions are fairly easy to grasp, however, and are understandable across cultures and language boundaries ([5] [6]). This mechanism of concentrating multiple nuanced thoughts into a single image (e.g. of a smiling face) doesn't completely negate the nuance and often the complex communication is maintained through shared contextual or pragmatic

knowledge. Boiling down knowledge to a simple set of easily-communcable signs might be a way to think about intra-species interaction.

The thought that we can create a robot with consciousness who will understand the world in the same way that a human does is unlikely to be possible given the concepts that the robot would need to be able to factor and very likely wouldn't manifest. We may be able to replicate behaviours via a set of models trained against human activity, but to reproduce the intricacies of hundreds of millions of years of evolved programming seems on the face of it an impossible task and perhaps unnecessary if the creator of such a system would be happy with an approximation. A robot cannot be like a human unless its principles of concern are similar enough. Unless a robot has functional drives and instincts founded on principles of existence hundreds of millions of years old then it is not going to see the world in the same way. I am not arguing that a robot cannot be intelligent in its own manner and cannot experience phenomenal consciousness, but it would need its own set of principles of existence as a foundation. In actuality regardless of whether it could access the rich complexity of a comparably-human-level concept system any artificial intelligence capable of sufficient complexity would become its own species of life. Sufficient complexity is hard to define, but there is scope for this investigation to be included in any translation scheme that depends on the ontology of cognition developed herein.

Imagine the enormous complexity of a hierarchy of concept forming systems such as that of a great ape — a chimpanzee. The visual, auditory, olfactory and somatosensory signals innervating its nervous system on a constant, millisecond basis and updating across a multitude of nervous media and at a level of parallelism far beyond any computing system humans have created seems insurmountably challenging to replicate and this only only the sensory layer. This initial layer of information is difficult to comprehend alone, but on top of that the nervous system filters and aggregates that raw information across multiple incredibly complex and interconnected layers of processing. Even though this process is dauntingly sophisticated, I propose to reduce this in some way to a set of core principle functions or concepts that might be required to enable such a creature to thrive in its enviuronment.

Extending the idea of a mouse's core principles — reducing the dimensions along which information might be processed — I present an *ontology of cognition*. By ontol-

ogy I mean simply 'an explicit specification of a conceptualization' [7]; a documented encapsulation of a set of entities. The organisation of this document is by layers — the hypothesised groupings of connectionist systems responsible for information processing and generally in accordance with a signal path from sensory input to higher functions such as awareness and complex cognition.

The ontology presented herein are not intended to be complete or perhaps even particularly accurate to what is really processed in the average brain, however the purpose is to codify something sensible that can be used in various applications such as as a template or starting point for the training of a hypothetically generally intelligent machine. There are numerous ontologies in this domain not all of which are published or documented and many of which are aligned to neuroscientific or pathological investigation rather than philosophical ([8][9][10]).

In the Critique of Pure Reason, Immanuel Kant introduced a brilliant conceptualisation of human cognition [11]. Crucially for this essay Kant's section on the Transcendental Aesthetic draws out a way of thinking about space and time. Objects, for Kant, are "determined or determinable" to us in their "shape, magnitude and relation to one another" in *space*. Space is not an empirical concept derived from the world outside us, but is a "necessary *a priori* representation, which underlies all outer intuitions". Similarly for Kant "Time is a necessary representation that underlies all intuitions". The point for the ontology of cognition is that space and time must be treated differently from ordinary ideas. They are not concepts to be activated, but are structures of the fabric of the networks themselves - shapes and arrangements of connectionist architectures at present beyond our scientific investigations, but in practice perhaps manifestations of complex meta-models (e.g. [12] [13])

5 Argument

As I leave the house to go to the shop I am aware that I have a shopping list in my pocket. I am aware that there are things on the list and also that I shouldn't forget to consult the list when I arrive at the store. Usually within seconds I forget all about the list as I stroll down the street. Sporadically I become aware of the road or the footpath or the people around me or the trees etc. I come to a crossing in the road

and there are vehicles close to me that are travelling at speed. Without consulting my conscious mind I stop and become aware of the cars or lorries that would kill me instantly if I were to walk out into the road. What was my conscious state throughout this scenario? I was initially aware of the shopping list, but quickly this awareness faded into a sort of background position as I got on with my walk down the street. Without needing to access any state of awareness I stopped from walking out into the road and certain injury. My claim is that I am conscious the whole time, but only aware of my consciousness when deliberately activated — i.e. when reading the shopping list. At other times I'm just existing in the world in a sort of autopilot mode. At any second I can snap out of autopilot and back into awareness mode again, but for most of my existence as a human I am in autopilot. After I arrive at the shop I am able to bring to mind some aspects of the unaware walk and therefore some memory must be maintained even though I am not aware of it at the time.

This state of being — conscious, but unaware — is what I posit is the default for any system that processes information. Some systems, like that of humans, have extra functionality in the guise of an awareness component — analogous to the Attention Schema from Graziano [4]. Under this broad definition the idea of being *unconscious* makes little sense. Even while asleep (so called 'being unconscious') we are potentially using both the conscious and awareness parts of our brain in a different state entirely that is largely divorced from sensory input. We have awareness of our dreams (which could be daydreams as well as sleeping dreams) and our conscious states, but most likely through a filter of introspection and with reduced external sensory input — even though dreams can often appear as real as wakeful states. Only in extreme situations is the brain not activating concepts and processing information. Research into vegetative states and a mechanism to inspect consciousness (e.g. [14]) highlight that while blood is pumping it is absolutely normal for consciousness to be present although for the majority of the time we are not aware of it and act in a sort of autopilot mode.

A number of theories have emerged that suggest that *complexity* is at the heart of explaining consciousness, e.g. IIT. If a system such as a neural network has sufficient complexity (number of nodes and connections between them) then phenomenal consciousness will emerge. In these arguments the threshold is set perhaps arbitrarily just below where humans exist (or maybe a little lower for those who consider other great

apes, or octopodes to have some form of highly complex connectome). This argument seems a little too self-determinant to be objectively valid. Just in virtue of a complex system (that humans and a few others possess) without a clear definition or argument as to why complexity is important must fall short of validity. I argue there's consciousness all the way down, but in a very wide definition of what consciousness is. It is not a state of awareness, but a process of information transformation. Some creatures have an extra facility of awareness which is not purely a result of complexity, but is a specific structure evolved and selected over time and with benefits to the individual. This extra facility is what we are activating when we become aware of ourselves.

If all creatures have consciousness, but not awareness of it, what might that be like? Thomas Nagel's proposition that being a bat is difficult for humans to comprehend [15] doesn't necessary prevent us from imagining a general schema for how it might seem. Humans are regularly (in actuality the majority status) unaware of our conscious mind. It appears to us sporadically and probably more or less in different individuals, but our general day-to-day existence of popping down the shops or putting out the rubbish doesn't require us to be consciously aware of our own state. Perhaps non-phenomenally-aware creatures have an existence similar to ours when we go about our daily business. If we would balk at the idea of considering ourselves unconscious while we are engrossed in a movie (and not actively aware of our self) then surely we should apply the same standard to non-humans. With some creatures it will seem far-fetched to imagine this (e.g. ants or tapeworms), however logically it might seem hard to argue against. The logic in this argument can be laid out as shown below:

(a) humans spend most of their existence unaware directly of themselves

(b) humans do not consider themselves to be unconscious when in (a)

(c) states such as those of (a) are the result of information processing

(d) therefore, consciousness is the result of information processing

There is no constraint on the level of information processing-induced consciousness other than the complexity or amount of information being handled. Therefore even the most simple process would be conscious, but only at the most simple level. In this sense the amount or extent of consciousness is proportional to the complexity of the

system in which it manifests. The implications of this are quite profound. In theory any information process would have some form of consciousness which must therefore include inanimate objects. The Panpsychist approach to consciousness is a growing field of research and has many varieties. This author does not hold many of them, but the implications to a broad definition of consciousness surely overlaps with it. Perhaps the slight alarm that comes with considering a two-node neural network executing within a desktop computer system as conscious stems from the overloaded semantics of the word consciousness. In a sense we can take *consciousness* to be synonymous with *information* we begin to feel less alarmed. Any information system is conscious, but not aware. It is the awareness and phenomenal aspect of the use of the term consciousness which we would separate and not include in our broad interpretation. Under that description this essay is not Panpsychist except when considering that type of conscious, alluded to previously, that occurs while we are awake and active, but not aware.

There are neurological and functional reasons for imagining the processing of information in higher creatures like humans as layered or sequential. Sensory information enters the brain and is aggregated, filtered and coalesced through connected networks of neurons (primary $\rightarrow$ secondary $\rightarrow$ tertiary etc) with successive layers in the information flow developing more complex networks. Kant argues for the same general process: "The capacity (receptivity) for receiving representations through the mode in which we are affected by objects, is entitled *sensibility*. Objects are *given* to us by means of sensibility, and it alone yields us *intuitions*; they are *thought* through the understanding, and from the understanding arise *concepts*." [11]. Kant's philosophy is brilliant and deep, but this essay's scope is more functional than deeply philosphical. Suffice to say that the process by which sense data results in sets of concepts aligns with the ontology herein. Various parts of the brain are functionally specific and therefore it seems valid to posit that networks generate (or represent) specific concepts. The concept of a square can be constructed logically from four straight lines oriented at right angles to each other and joined together to form a continuous single line. The sub-concepts of this hybrid concept could be loosely represented as below:

$$\{SQUARE\} \equiv \{QUANTITY = 4\}\{LINE\}\{ATRIGHTANGLES\}\{JOINED\}$$

In the brain the 'square' is an activation of neurons with connections to other neural networks that somehow represent the constituent concepts (quantity, line etc). In this way a square can be represented in a non-primary-layer network and is a construct of more basic concepts (and primary networks). This process can apply to the entire brain — a series of unfathomably complex networks each aligned to a concept and successively aligned networks (along a signal path starting from sensory inputs) can represent or be responsible for higher concepts of any sort, such as 'debt consolidation loan' or 'favourite whiskey'. It can be assumed that network complexity is proportional to conceptual complexity. A fruit fly has a simpler connectome than a human and its nervous system should therefore have a smaller set of concepts within its compass. It does have many concepts of course, such as {LINE} or {DANGER} and has consciousness of the sort I suggest, but obviously in a much more basic capacity than those with larger brains. It is not suggested that if someone were able to look into the brain while it is forming a square concept that they would be able to see it clearly represented. The brain is a much more dynamic system that arguably does not represent at all in a logical way. Hutto and Myin, in their work on Radical Enactivism, suggest that the 'content' of the brain is not propositional in terms and doesn't reduce to truth conditions: "the notion of content is elastic enough to allow that the relevent correctness conditions might be understood in terms other than truth: say, in terms of accuracy, veridicality, or some other kind of satisfaction condition where these are taken to differ from truth conditions." [16]. This is my view too. The overlay of a truth condition to the natural world seems particularly anthropocentric and therefore if the model of cognition is applied to non-humans, how could their brains represent in propositional terms in the absence of propositions? Much more likely to be the case it seems is the connectionist, dynamic processing paradigm whereby complex activations of cells in the brain relate through sensory data, memories and behavioural outputs to the outside world. A multi-stage, indirect representation is therefore created that enables inputs to the brain to be stored and to filter through successive networks of neural processing towards resultant behaviour which itself is monitored and fed back into the same system and, via activity in the world, generates a set of models and behaviour complexes which explain how we are able to exist fairly successfully. This isn't to say that some neuron systems are causally specifically related to outside phenomenon (the

'grandmother cell' and 'gnostic fields' [17]), but that is not a general rule — the brain does not create a picture of the world to be viewed on a *mind screen*.

5.0.1 Ego module

In addition and directly linked to awareness is what I term the *ego module*, so called for two reasons. Firstly it is a modular component *bolted-on* to the more primitive brain — its composition and function varies across individuals and species and is decomposable, and secondly its roots lie in the generation of self and, more specifically *ego*. If an ego module is put in conjunction with a conscious creature that has awareness then phenomenal conscious would naturally be possible and reportable. Humphrey has a similar term — "special brain module" [18] which has a job of reading signals passing from sensory to higher parts of the brain. Without an ego module to whom could I report my phenomenal awareness? To whom would first person conscious experiences appear; *who* would be doing the experiencing? This seems obvious and circumstantial to ask the question, but I suggest there's a good reason for the evolutionary benefits of such a system and a resultant side-effect is phenomenal consciousness.

The making of tools that enable control over the world is an ability limited to a relatively small number of creatures. The process by which this evolved is not known fully, however it is hard to imagine it being possible without a belief that certain actions are feasible and within the grasp of the creature exhibiting them. Perhaps a series of random acts came before the ego, but either way it seems intuitive for them to be coupled in an advanced tool-making creature such as a human.

Awareness of my environment is not quite enough for me to believe that I have control over my domain in highly complex ways. It is more beneficial for me to have an ego that is an extension to awareness. There's an environmental benefit to having a concept of oneself. This is what enables humans to appear more advanced than all other creatures. As far as we can tell humans are the only ones with a complex sense of self and a language that allows us to communicate.

Humans have done something different from all other creatures on the planet. They developed a complex tool-making ability and a desire to harness the world to their advantage. We can see the obvious result of this everywhere we look today — from mobile phones to skyscrapers. In order to gain this ability humans need a sense of

ourselves as empowered to take on and control the world. This is the basis of our ego — the thing that gives us our unique sense of ourselves and our belief that we're the only creatures with a consciousness worth considering.

This ego module is responsible for our belief that we have phenomenal consciousness and partly addresses Chalmers' *meta-problem*: the ego module has no capability to process meta information about our cognition. Its job is to form a world-controlling entity (which has been eminently successful) and not to understand its own perceptual panorama. Having conscious awareness is a by-product or side-effect of the evolutionary *purpose* and therefore it is simply not able to reflect on its own perceptions and hence the struggle we have in attempting to do so.

The word ego has some negative overtones and its use herein is partly deliberate for that fact. There is a sense of arrogance involved in the construction of a self that can control the environment — somewhat justified when progress is beneficial. Undoubtedly there have been significant advances and benefits to humans across millenia, however it cannot be denied that hubris or arrogance still exists across many collective endeavours from issues around climate change to ubiquitous and heinous abuses of flora and fauna. The pursuit of controlling the world sometimes appears to trump a necessary care for it.

5.0.2 Phenomenal Consciousness

As per the theory of Illusionism, I suggest that phenomenal consciousness — the manifest qualia we experience — is no more than a self-induced awareness backed by our sense of self. There are evolutionary benefits to this facility in our success as creatures. Many will say that this position is unintuitive and ignores the plain fact that we feel things and experience directly the world as it is and it is difficult not to have sympathy with that view. The experience we all have is so enthralling that to deny it would seem ludicrous.

Knowledge of how sensory information is manipulated by the brain to be representation rather than a direct correspondence (e.g. colour constancy theory[2]) means that we can discount the proposal that we have direct transparent experience of the world. The issue remains, however, that we feel intensely the experience of things. Seeing

[2] `https://en.wikipedia.org/wiki/Color_constancy`

a red book feels like we have direct access to the object. If it is illusional, however, then what could the experience comprise? Even though a number of excellent theories exist, the problem has not been sufficiently explained and, in this author's view, we have no evidence for a physicalist explanation with elements that are not the concepts themselves. As unintuitive as it may seem, the simplest account is that the phenomena is itself the conceptual representation of the sensory and hybrid concepts. We have access to these neuronal structures through our awareness and ego module complex via what Frankish calls the "introspectability of sensory states" [3]. I suggest that sensory inputs are only a part of the complex to which we have introspective access. Although arguably having a foundation in sensory input, there are subsequent-layer hybrid concepts that must be part of the complex presented to the *mind* and are potentially significantly informationally-distant from sensory input.

I will work through an example of how this could be understood using the ontology of cognition. I will say that there are activations between elements in the ontology (which refer to neural structures in the brain). The extent of activation will vary of course and there will be some level of dynamic activation — which is why no two experiences are the same and why some people see and understand things differently (a red book for me could be a long-lost diary for someone else with a whole host of different activations taking place).

Unless there is a relevant prior event — such as someone shouting 'hey! look at that read book' — before I become phenomenally conscious of the red book the initiating signal is the radiation reflected by the atoms in the book's surface material and which passes through my eye to the optic nerve. A significant amount of information processing takes place just in the nervous system at the back of the eye, but as the eventual signal passes to the back of the brain and creates a cascade of neural networks to be innervated and some activated (most will not be activated) various concepts could be stimulated, not necessarily in this order: {Object Identity} {Location} {Quantity} {Size} {Material Behaviour} {Place} {Age} etc. Others would be activated of course, but these are a good guess as to some of the primary ones. Had it been a joke book on the table then {Humour} would likely have been included in the hybrid complex of concepts manifested as torrents of inter-related minute electrical charge and chemical change. A fraction of a second later a different bunch of cells in the back of the eye

are hit with radiation and a different set of neurons are activated in the brain. The hybrid output is what we experience — it is actually purely the concepts that overlay the sense data, either deliberately as witnessed by introspection or more foggily as a conscious experience out of reach of awareness (e.g. as I walk down the street to the shops and don't get run over by cars only one metre distant from me), that are the composite of a phenomenal experience.

Nicholas Humphrey responds to Frankish's view on Illusionism by suggesting that an Illusionist would deny that our own subjective attitude to events, such as hearing a joke, have real properties and he implies that an Illusionist would have to agree that brain activities generate states of mind [18]. The distinction seems to me to be not entirely clear. A joke is a human concept and can be encapsulated as a set of brain activities to which we may have access via an awareness brain function, but to set up the idea of a joke as an objective non-human truth seems like a stretch. Couldn't the joke just be a response to concepts and our phenomenal awareness of any resultant humour is purely in our heads and there's no more explanation needed. There is no higher external truth that we must believe exists — our behaviours (laughing, sharing the joke etc) is all the real meaning that can be shown. Humphrey's concern that Illusionism reduces the beauty of the world or renders a funny joke not actually funny misses the point I believe. The power of the mind is just as awesome whether or not it's generated in my head or exists beyond my physical being. The difference is perhaps one of wanting to believe that the thing I know is bigger than I am and therefore has more potency. His concern also focuses on the danger to public opinion, which is perhaps a worry to a working philosopher, but this cannot mean serious trouble for a genuine truth-seeker. His counter proposal to Illusionism is *phenomenal surrealism* which is closer allied to it, but differs in the reduction to a pure illusion — Humphrey cannot conclude that phenomema don't really exist, but rather takes it that they're just not quite what we believe them to be in reality and are actually a hyper-realisation — "phenomenal redness is redder than real red" [18]. I am sympathetic to this attitude, however it seems to still leave open the question of what phenomenal *feels* actually are. The power of the Illusionist approach is that the hard problem seems to go away when there is no problem to be answered.

So in what way is the phenomenon *actually* the concepts that drive it? I would

say it is entirely the concepts that drive it and the aspects that we term qualia — e.g. the actual experience of seeing red — don't exist as we believe they do, but are hard-wired in us to be so convincing that we have a hard time conceiving of them as just the concepts (even as highly complex as those concepts may be) that we generate ourselves. The hard problem then becomes one of explaining why should we have these vivid illusions.

This position remains unproven and takes an intellectual and slightly uncomfortable leap to believe, however if it is taken on (and growing numbers of thinkers seem to be drawn to the idea) then

5.1 An Ontological Approach

Some might say that it is futile or even hubris to attempt to codify the entirety of human cognition. I would agree, but that is not what is being sought in this paper. In a similar way to how an lexical dictionary cannot ever capture entirely the extent and subtleties of a language, any ontology of cognition can only be at most a snapshot and a particular perspective of a set of cognitive creatures. In this case that is the author of this paper.

The purpose therefore is not to be complete or entirely accurate. If such an ontology can enable ways of non-human communication in a meaningful way and act as a road map along which future understanding can develop then some success will have been achieved.

Taking a layered approach inspired by the correspondence to elements in the nervous systems of creatures we have studied, e.g. the visual cortex processes visual input in layers each one becoming increasingly complex and using outputs from previous layers as inputs, therefore we start with sensory input. Sensory data is surely the most primitive and impactful factor on cognition. Given the thought that all life on this planet begins as a simple cellular construct and thrives through direct responses to various stimuli, such as visual radiation or chemical interaction, we presume that this form of impact is primary to any higher cognition functions we might recognise today.

The secondary grouping after the sensory entities includes basic processing functions which would take sensory input and develop novel concepts. The ability to sense

quantity or number is a primitive capability of most creatures, but needs sensory input in order to calculate a quantity of something. It is well known that connectionist networks can take sensory input and factor many subsidiary and useful statistics such as proximity and size. These calculations require input from the senses and therefore are in a layer after the primary (sensory) layer. Following the same logic all the various layers in this ontology have dependencies on *upstream* layers and produce outputs which can be used by any other layers.

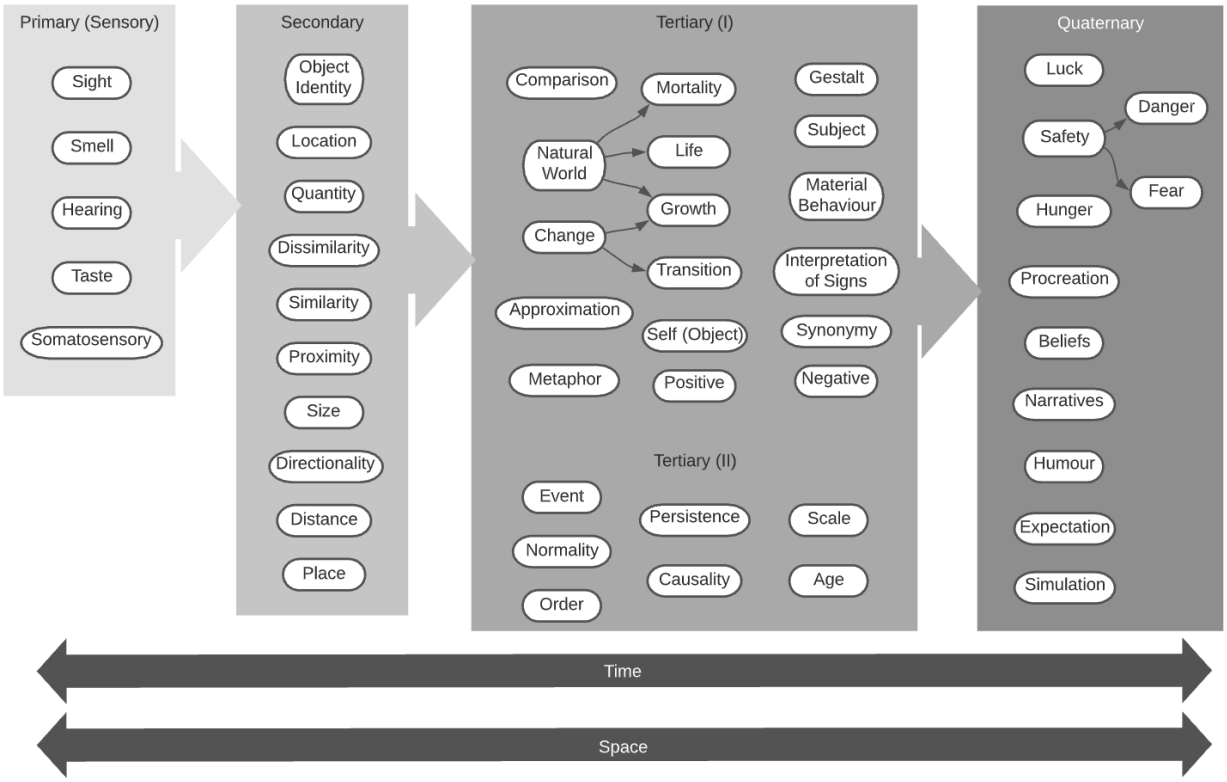

Figure 1: Ontology of Cognition

This model of concepts is not a generative model, i.e. this does not describe the functions in a developing, learning system (such as that of a child). In a system that is learning the innervations might well be different. Instead this schema represents an adult operating model.

Entities (classes in ontological terms) are not standardised or in any order within each layer. There are cross-layer dependencies, e.g. we cannot fathom {Approximation} without some prior concepts such as {Size} and {Object Identity}. However dependen-

cies between entities in the same layer are less strict and harder to define. As already stated, this ontology is fairly loose and the specifics of whether this is in reality how brains work isn't the goal — rather to further understand cognition in a detailed way that can be applied universally to beings with whom we might want to communicate or understand is the primary direction of travel.

Relationships between entities in this ontology are kept deliberately vague. At a neuronal level the relationships are physical and in terms of synaptic or chemicular connections. The logical relationships are much harder to define without further knowledge and investigation. Some attempts can be made, but for simplicity and to prove the exercise they are not included.

6 Discussion

My species is on the verge of creating machines of deep complexity. According to some theories this could entail a form of phenomenal consciousness coming into being. I am skeptical that we are as near to this pivot point as some would believe and the key reason is a naivety that system builders have regarding the complexity of human conceptual thought which underpins our conscious minds. Similar to how early practitioners in computer science in the 1950s thought that we would have generally intelligent machines within decades there is a kind of blindness and arrogance to the fact that life, the world and our place in it is unfathomably complex and almost entirely beyond our practical thought. What we can do is generalise and schematise the world in order to gain traction against its vast complexity. When we do that we miss the detail, however, and a lot of that detail is subconsciously driving our behaviour and therefore is required to drive the behaviour of similarly intelligent machines. What I propose in this essay is a way to gain some more traction against the problem. By analysing the way we conceptualise the world, documenting it and somehow using it as a basis for a future intelligent machine we will not only have a better chance of actually creating a machine with genuine general intelligence, but also we will be able to fathom its workings with greater capability than if we did not.

Humans have a genetic lineage stretching back for billions of years — to the very earliest adaptive systems such as proto-cellular life. If we take the several billion years

of evolution of life as an input to the way we construct our view of the world today then we have to understand those background cognitive mechanisms when we want to understand our own consciousness. The way that a rodent reacts to the world is remarkably similar in essence to the way that humans achieve similar feats. Given that their brains and ours have fundamental similarities it doesn't feel like a leap to suggest that those similarities manifest in the foundations of our concept system with which we construct our perceptions of our habitats.

When we extrapolate phenomenal consciousness to non-human creatures we have to consider the underlying cognitive mechanisms that they would require in order to have a comparable sense of the world. To some extent it will be impossible for humans to achieve this. Thomas Nagel's notable paper on *what it is like to be a bat* reflects on the fundamental differences in the way that creatures interface with the world. It would seem impossible for humans to appreciate at a cognitive level what sonar *feels* like, however, bats are mammals and share a significant amount of their genetic code with humans so when they feel hunger wouldn't that bear a similarity to the way humans feel it too? There is a speciesist thread to Nagel's argument that has foundation in the idea that humans are superior and therefore how can we possibly understand lesser creatures — the paper is not framed as *what is it like (for a super intelligent extra terrestrial) to understand a human?*.

A robot cannot be like a human unless its principles of concern are similar enough. This point often gets lost when scientists talk about intelligent machines. It is easy for us to forget how complex our thoughts are and how much we take for granted through the generation of complex concepts.

It's not that phenomenal consciousness doesn't exist, it's just that it exists inside the head rather than on the outside. Following the ego argument above, it might be tempting to think that it would be more compelling to believe that we are the ones who are responsible for generating the pictures of the world and therefore are more powerful — how amazing would it be for my entire world to be created in my head and the power that gives me over my existence. However, this doesn't feel particularly powerful. As creatures in the world we desire to exist within it with veridicality. We want to see the truth rather than something that is compelling, but not real. The problem here is that the concepts of truth and reality are themselves invented by us

both because in our earlier history we benefitted by existing in a world to which our attentions made a good correspondence, but more recently our success in our power over our domain benefits by us correlating our view of the realm to the realm itself. If a more *Idealist* philosophy was a more likely candidate for how things really are then perhaps we would have greater correspondence between our perceptions and the world itself — instead we get things wrong surprisingly frequently.

6.1 Talking Poetry

The title of this essay implies a way to talk about poetry with robots. How might that be possible given the discussion up to this point? Thinking about what a robot might be like is crucial to understand the problem. It can be fairly straightforward to imagine — and is to some extent already within the compass of computer scientists — a machine capable of taking in sense data about the world (comparable to the primary layer in my ontology). Functions that have sense data inputs and calculate size, proximity, location, identity, age, comparison etc are all currently capable of some machine intelligences created in the field (akin to the secondary and tertiary layers in the ontology). As we progress to more higher functions the ability for machines of today to keep up with equivalent functions wanes. This is not a critique, but simply a fact that technicians have yet to find significant uses for these abilities and there is little to underpin the conceptual structures required to build them — there are few large-scale systems that can support complex concept formation (SOAR being one example, however [19]). In principle, however, the creation of a neural network (presumably artificial) that can *represent* {safety}, {beliefs} is feasible — certainly if basic versions are acceptible — and will depend on a world view and matrix of importance to the entity itself. In practice it seems unlikely that whatever importance matrix is present initially will be sufficient for complex behaviours to result. Instead a period of learning will be needed to create the network structures that can accommodate higher functions with tightly-coupled domain integration. Such a robot at this stage of development and perhaps after a period of learning would not be aware of itself and that component would be required for any such thought of phenomenal consciousness. Even without that extra qualia-inducing function, the machine would be incredibly complex at this

point and capable of communicating in highly complex ways (assuming a language modality is enabled and that too would require training).

What of poetry then? What is a poem? In stark reductionist terms it is simply a collection of words that evoke feelings. It is certainly brutish to try to reduce all of an art such as poetry to a simple description, however it has to be done here to explain a principle. An example helps emphasise the point:

> you fit into me
> like a hook into an eye
>
> a fish hook
> an open eye
>
> (Margaret Atwood, "you fit into me" from Power Politics. Copyright © 1971 by Margaret Atwood)

A complete concept graph of all the interpretations of this short and evocative poem would be very large. Concepts such as {Subject} ('you', 'me', 'hook', 'eye'); {Material Behaviour} (clasps that use hooks and eyes are strong, hooks are sharp, eyes are soft); {Normality} (fish hooks aren't normally associated with seeing eyes); {Beliefs} (people desire good relationships); {Expectation} (good relationships are strong; good relationships aren't hurtful}; {Metaphor} (a hook is used to catch things such as partners in a relationship, open eyes see things) are good candidates for a conceptual mapping of this poem. Countless others could be constructed. The point here is to show that it is possible even in a rudimentary sense.

Again, it feels cold and reductive to be analysing great works of art in purely conceptual terms, but the analogy to the locus of phenomenal consciousness seems apt. I have attempted to draw out the perhaps stark idea that our sense of self and the very experiences of the world are — after the Illusionist theory — not real. The tendency for learned responses to balk at this and reflect on the loss of agency or potency that would result could be seen as a missed opportunity to reflect on how

amazing our constructed worlds are. What could be more incredible than that idea that an entire phenomenal world is inside my head (albeit stimulated by the sensory world available to me and evolved on this planet for millions of years)?

Literary analysis is a popular and unproblematically reductive pursuit and to some extent a conceptual analysis process is an extension of this established critical art. By regarding poetry (or any form of creative art) in analytic terms it might seem that some of the beauty is lost and I do agree with that, however by doing so we gain new understanding of the richness of the concepts we employ and the meaning and importance can be even loftier as a result.

How, then, would it seem to talk with a robot about poetry? We might ask them what do they think the poetry means. We might suggest interpretations and see how they respond. To answer convincingly they only need a complex concept network and perhaps a learned experience of how it can apply to the world — in this case language and world knowledge. Is this enough for a human interlocutor's satisfaction? It is probably doubtful to be. We tend to want to have a connection to our conversations. Atwood's poem seems to be about relationships, which in humans are a continual source of joy and pain — how could a robot relate to that? Would we need to believe that a robot can have a relationship and experience the highs and lows that would ensue? Perhaps a robot relationship, whether with other robots or other creatures, would be a necessity for us to consider it intelligent enough to analyse a poem. Otherwise any response they could make regarding the nuances of love would elicit a human-level rejection of its validity.

Then what about *love*? Would any intelligent machine capable of talking about poetry to a human need to understand love? And not just to know the definition, but to have been in love. What would that mean? Perhaps love is non-sexual for robots of the future. That particular function is unnecessary for neural networks to learn, however perhaps love of all kinds that transpires in an imperfect and cruel world will cause the requisite strife to enable learned experience in robots the acquaintance of which will facilitate great wisdom on the passions between individuals of all kinds.

Even if the robot has a perfect, nuanced response it might be better for it to suggest a more embodied or imperfect answer. Whether it does this by selection (i.e. the robot makers may have designed to serve a human's needs) or by genuinely having a sense

that it feels the weight of the words (i.e. by adapting the awareness or ego functions that humans appear to have) there seems to be no barrier in principle to talking to robots about poetry.

References

[1] Chalmers D. The meta-problem of consciousness. Journal of Consciousness Studies. 2018;25(9-10).

[2] Dennett D. Illusionism as the default theory. In: Frankish K, editor. Illusionism. Imprint Academic; 2017. p. 65–72.

[3] Frankish K. Illusionism as a Theory of Consciousness. In: Frankish K, editor. Illusionism. Imprint Academic; 2017. p. 11–39.

[4] Graziano MS. Rethinking consciousness: a scientific theory of subjective experience. WW Norton & Company; 2019.

[5] Wiseman S, Gould SJ. Repurposing emoji for personalised communication: Why means "I love you". In: Proceedings of the 2018 CHI conference on human factors in computing systems; 2018. p. 1–10.

[6] Guntuku SC, Li M, Tay L, Ungar LH. Studying cultural differences in emoji usage across the east and the west. In: Proceedings of the International AAAI Conference on Web and Social Media. vol. 13; 2019. p. 226–235.

[7] Gruber TR. A translation approach to portable ontology specifications. Knowledge acquisition. 1993;5(2):199–220.

[8] Poldrack RA, Kittur A, Kalar D, Miller E, Seppa C, Gil Y, et al. The cognitive atlas: toward a knowledge foundation for cognitive neuroscience. Frontiers in neuroinformatics. 2011;5:17.

[9] Smith B. Classifying processes: an essay in applied ontology. Ratio. 2012;25(4):463–488.

[10] Ceusters W, Smith B. Foundations for a realist ontology of mental disease. Journal of biomedical semantics. 2010;1(1):1–23.

[11] Kant I. Critique of Pure Reason. translated by Norman Kemp Smith. London Macmillan; 1934.

[12] Buzsáki G, Llinás R. Space and time in the brain. Science. 2017;358(6362):482–485.

[13] Zuo XN, Di Martino A, Kelly C, Shehzad ZE, Gee DG, Klein DF, et al. The oscillating brain: complex and reliable. Neuroimage. 2010;49(2):1432–1445.

[14] Dehaene S, Sergent C, Changeux JP. A neuronal network model linking subjective reports and objective physiological data during conscious perception. Proceedings of the National Academy of Sciences. 2003;100(14):8520–8525.

[15] Nagel T. What is it like to be a bat. Readings in philosophy of psychology. 1974;1:159–168.

[16] Hutto DD, Myin E. Evolving enactivism: Basic minds meet content. MIT press; 2017.

[17] Gross CG. Genealogy of the "grandmother cell". The Neuroscientist. 2002;8(5):512–518.

[18] Humphrey N. Redder than red illusionism or phenomenal surrealism? In: Frankish K, editor. Illusionism. Imprint Academic; 2017. p. 116–123.

[19] Laird JE. The Soar cognitive architecture. MIT press; 2019.