Enaction, Convolution and Conceptualism: An AI-Based Exploration of Dharmakīrti's Perception and Conception

JUSTIN BRODY
Franklin and Marshall College
jbrody2@fandm.edu

Abstract: This paper attempts to give a preliminary account of concept formation that employs ideas from artificial intelligence to naturalize certain points of Dharmakīrti's philosophy. It argues for a form of conceptualism that employs enactive cognition and neural networks to ground a conceptualist understanding of concepts which are imputed on the basis of an agent's interaction with its environment.

Keywords: Dharmakīrti, artificial intelligence, enaction

DOI: https://dx.doi.org/10.15239/hijbs.03.02.01

Introduction

My goal in this paper is to give an account of perception and concept formation that is both grounded in contemporary artificial intelligence and compatible with Dharmakīrti's account of perception and conception. This account aligns with many projects aimed at 'naturalizing Dharmakīrti', notably in Arnold and Dreyfus, but arguably also present in Siderits and Ganeri.[1] I am largely in agreement with the claims of the aforementioned works; I see the main contribution of the present work as employing neural networks to defend a conceptualist reading of Dharmakīrti as espoused by Dreyfus in 'Apoha as a naturalized account of concept formation'. It is worth quoting Siderits at length here, who lays out the debate between nominalism and conceptualism with his usual clarity:[2]

> The conceptualist is understood as holding that what all cows have in common is just that they fall under the same concept, where concepts are thought of as mental contents of some sort or other. But if one holds that concepts are essentially linguistic in nature ... then conceptualism collapses into nominalism.
>
> Conceptualism might seem to represent a distinct position if one holds an abstraction theory of concept formation according to which one forms the cow concept by abstracting away the distinctive features of particular cows and conjoining only the similarities. Such a form of conceptualism would claim, *contra* the nominalist, that our use of 'cow' is governed by objective features of individual cows and not just by linguistic conventions. And since it does not explicitly employ the realist's cowness universal to explain our linguistic behaviour, it would appear to be distinct from realism as

---

[1] See Arnold, *Brains, Buddhas, and Believing*; Dreyfus, 'Apoha as a Naturalized Account of Concept Formation'; Siderits 'Apohavāda, Nominalism and Resemblance Theories'; Siderits, 'Śrughna by Dusk'; and Ganeri, 'Apoha, Feature-Placing, and Sensory Content'.

[2] Siderits, 'Apohavāda, Nominalism and Resemblance Theories', 153; emphasis mine.

> well. **What such a variety of conceptualism does require, however, is some account how two things *x* and *y* might resemble one another more than either resembles *z*.** And this turns out to require kind universals to avoid the difficulty raised by the fact that a given resemblance paradigm might always resemble other particulars in a variety of respects (the problem of paradigm overdetermination). Conceptualism is thus distinct from nominalism only when it takes the form of a resemblance theory, and this turns out to lead in the end to one or another type of realism.

I propose that neural networks, coupled with an enactive theory of concept formation, provide for just such an account of resemblance which does *not* require kind universals. This will be the fundamental feature of the proposed model of perception and concept formation. Additionally, I will argue that the account I give has the following features:

1. It will give a causal account of perception.
2. It will ground Dharmakīrti's observation of the relative richness of perception in information theory.
3. It will be compatible with (and thus suggest) a nominalist ontology about concepts.
4. It will be compatible with an external ontology of the ultimate that is grounded in infinitesimal bearers of properties.
5. The inability of concepts to capture the nature of reality will be expressed in terms of information theory and computability theory.
6. The theory will be grounded in physicalist notions of causality. While not committed to a physicalist ontology, it is compatible with one and hence can appeal to contemporary inclinations to reject both dualism and idealism.
7. The concepts supported by the theory will be based on *affordances* (the potential uses an object affords a viewer). This accords with Dharmakīrti's explicit writing and many suggestions in the Buddhist literature.
8. They will ground a causal explanation of intentionality.
9. The grounding of a philosophical theory in a computational

> model makes the model both concrete and explicit. This can allow for a framework which clarifies some philosophical questions.

The project here is similar in spirit to that of J. Ganeri in 'Apoha, feature-placing, and sensory content', which sketched a 'bottom-up' theory of concept formation and highlighted various consonances with Dharmakīrti's ideas. Perhaps the most salient differences here are that the current paper is explicitly grounded in mathematical constructs borrowed from artificial intelligence and, rather than demonstrating consonance with Dharmakīrti's notion of exclusion (*apoha*), I argue that the model affords a strictly positive account of concept formation that is thus more intuitive while maintaining the essential (soteriologically salient) aspects of Dharmakīrti's theory.

The Architecture

Let us imagine two beings sitting in a room together. The first one (let's call him Milinda) turns to the second and comments, 'I think I see a cat'. Swivelling his head, the second, Menander, notes a cat and says, 'Yes, that's Fluffy. It's about her dinner-time so she's probably hungry–let me go feed her'. This exchange is unremarkable except that Menander is Menander 3000, Milinda's 3000th attempt at building an intelligent robot. While Menander's recognizing the cat as Fluffy, reasoning about her hunger and taking responsibility for helping her are simple matters for humans, implementing them in a general way is far beyond the state of the arts for today's robotics. We will give a sketch of some of the capacities Milinda might have needed to develop in Menander and how he might have implemented them.

One of the most crucial things Menander will need to develop is a grounded set of concepts. He will need to have proficiency with concepts like 'cat', 'Fluffy', and 'dinner-time'. This will mean understanding, for example, when a particular thing he encounters falls under the concept of {\it cat} as well as facility with general reasoning about cats: to conclude, for example, that because she is a cat, Fluffy is not likely to want fresh carrots for dinner.

In Joel Parthemore's article, 'The Unified Conceptual Space Theory', a list of defining qualities appropriate for an agent in a 4E framework[3] is given. Thus, per Parthemore, concepts should be systematic, productive, compositional, spontaneous and subject to revision. His explanations for these properties is slightly essentialist for the present context,[4] so we give a more deflationary account of each of these:

1. **Systematic:** the same concept can be applied across unboundedly many potential or actual contexts;
2. **Productive:** a finite number of concepts can be used to create an unbounded number of complex concepts and (at least in the case of linguistic agents) propositions;
3. **Compositional:** concepts fit together in reasonably reliable and predictable ways to form complex concepts, and at least certain concepts can be decomposed into component concepts;
4. **Spontaneous (in the Kantian sense):** concepts are part of the agent's mental stream and are not controlled by external conditions; and
5. **Subject to revision:** particular concepts are not fixed but rather adapt with the agent's experience.[5]

We add to this that concepts should be both *intentional* and *grounded*—that is they should be *about* something and causally connected to what they are about.

While a fully robust conceptuality is beyond the current capabilities of AI systems, there is enough in present technology for us to sketch a model of what a concept using system might look like. This starts with the binding problem: for a given concept like *cat*, how are particular objects of the world bound to it and seen as instances of it? Here contemporary AI can give compelling answers, based on the

---

[3] This framework will be discussed later in the paper.

[4] My thanks to Douglas Duckworth for pointing this out to me.

[5] Modified from Parthemore, 'The Unified Conceptual Space Theory'.

recent (and often stunning) success of neural network models which have been trained to recognize various images as belonging to particular categories and further divide those images into coherent objects according to their category.

Basic Perception

For simplicity, let us assume that our robot is equipped with a simple low-resolution camera. The camera divides its viewing window into a set of square pixels, say 100 pixels wide and 100 pixels long. As light falls on each square pixel, an estimate of the amount of red, green and blue components is made and transmitted to the neural network. This is the initial contact between the perceptual system and the world and already involves (perhaps infinite) information loss—the infinite possibilities of the world are collapsed into a finite set of pixel values (Figure 1).

The neural network itself proceeds in two stages. First, convolutional layers detect basic features of the sensory input—usually this takes the form of detecting various types of edges in different orientations. The output of the first layer is a kind of map of where such edges occur; for example, a detector that looked for horizontal edges would output a representation of where in the image such edges were found. The data from these edge detectors is then fed into the next layer, which looks for particular patterns of edges. For example, a second layer might look for corners formed by a horizontal edge meeting a vertical one. A third layer might look for patterns amongst those patterns and so on. The final convolutional layer will thus contain information about the distribution and configuration of various combinations of basic patterns within the image (Figure 2).

The final stage of the neural network will take this information and assign a numerical score to each of a number of possible categories. These scores are turned into relative probabilities, and a typical readout for the network might be that it believes it is looking at a cat with ninety-five percent probability, a rat with two percent probability and three percent probability for anything else. We further assume that the neural network operates with some kind of contemporary

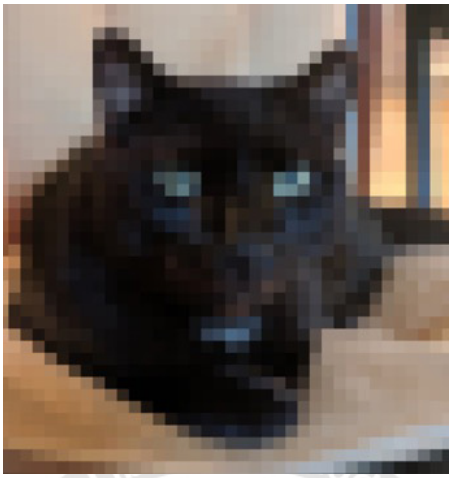

FIG. 1 A highly pixelated image of a cat. Image by Justin Brody.

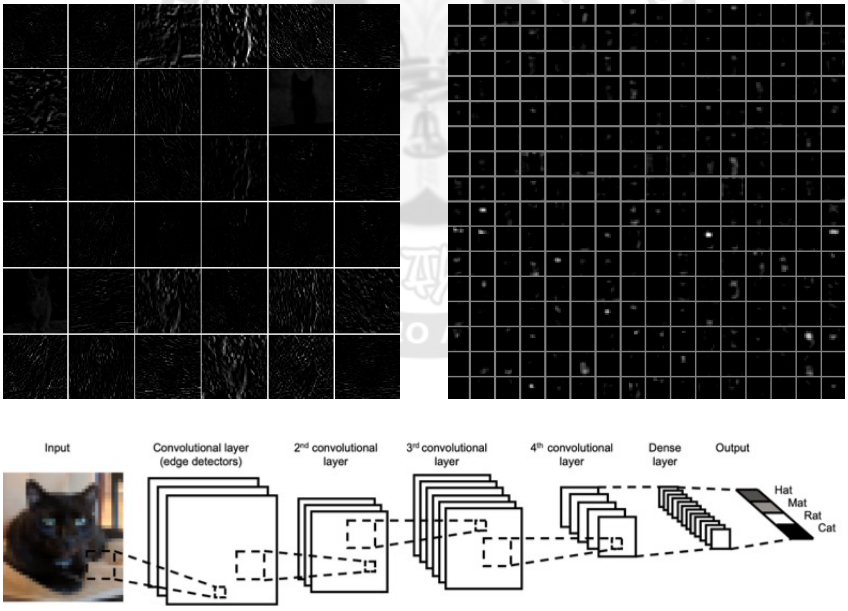

FIG. 2 A simple neural network. Above: the first and final layers of a network looking at a cat. From: http://cs231n.github.io/understanding-cnn/. Below: The network processes the image and outputs probabilities representing the degree to which it believes each given class is represented in the image.

segmentation algorithm;[6] this will allow the network to detect not just what objects are in its visual field, but where they are as well.

We take such a network as a grounding for a concept of 'cat'. We might then feed the output of such a network into a symbolic processing system. These are classical AI systems which work with symbolic information; in this case, we may want to generate symbolic information representing something like 'cat in visual field at position (10, 25)'. Such a system could then perform symbolic reasoning; knowing that cats need to be fed and believing itself responsible for this particular cat, the robot could conclude that it should get up and feed the cat.

While the concepts in such a system (call it Menander 1500) would seem to meet the criteria laid out by Parthemore; I would argue that they are not fully grounded and intentional. Arguably, any intentionality such systems display is borrowed from the humans who create them.[7] If Menander 1500 is able to correctly pick out images of cats, it is not because such a concept is inherently meaningful; it is mostly because he was shown a massive quantity of images and was able to find effective boundaries in 'image space' between images which it was told were of cats and images which were not.

To make the leap to full-fledged intentionality, Milinda would need to make these concepts more meaningful and less based in simple algebraic contingencies. Some of this is attained at the higher levels of Menander's processing system—part of the meaning of the symbol *cat* is determined by the role the symbol plays in Menander's knowledge base.[8] But a much richer and more deeply grounded intentionality comes from employing principles of enactive science—it will be precisely these capacities which distinguish Menander 3000 from Menander 1500.

---

[6] See, for example, Redmon, et al., 'You Only Look Once'.

[7] This point is made forcefully by John Searle. See Searle, 'Minds, Brains, and Programs'.

[8] We can think of this as a *functional* level of the symbol's meanings.

Enactive Perception and Symbolic Reasoning

Enactive cognitive science is a part of the so-called 4E paradigm—cognition is embodied, embedded, enactive and extended. The ideas grow out of original work by Franciso Varela and Humberto Maturana,[9] with heavy influence from phenomenology (via Merleau-Ponty) and J. J. Gibson's ecological psychology. The theory has a number of critical components, but here we'll focus on the ideas that: 1) meaning-making is something that an agent does on the basis of a division of phenomena between self and environment[10] and 2) that perception is a form of action.[11] These twin insights guide our attempt to extend Menander 1500 into an agent with its own intentionality.[12]

A basic (and much-simplified) picture that emerges from the tradition is that an agent learns[13] some kind of boundary that distinguishes it from its environment. Based on that, as it explores the world it lives in it learns the effects of certain actions on the world (*affordances*). These affordances guide the agent's process of segmenting its perceptual experience into objects—essentially, a thing is reified as a whole that can be acted on in a particular way (for example, a chair is a kind of thing you can sit in). Perceiving such an object is not a passive process of receiving and processing information (as Menander 1500 basically does) but rather of actively probing the object, perhaps simply by looking at it from many different angles (e.g. through rapid eye saccades). It is crucial to note that the kinds of affordances detected will depend on the kind of body and

---

[9] See Varela, et al., *The Embodied Mind*; Maturana and Varela, *The Tree of Knowledge*.

[10] Thompson, *Mind in Life*.

[11] Noë, *Action in Perception*.

[12] It is worth noting that while something like Menander 1500 could presumably be built in a few years using existing technology today; the extension to Menander 3000 delves into uncharted territory.

[13] The term 'learn' here is standard in AI and simply refers to a computer program algorithmically computing statistical information. It does not imply any kind of agency or consciousness.

interests the agent has—a snake will gain no sitting affordance from a chair.

A robotic agent based on these ideas might start by developing a self-model. In the first place, this will require the agent to develop senses of *ownership* and *agency*[14]—fundamental knowledge of which of its percepts correspond to its body and its actions. Based on these, an agent can develop a sense of how its actions affect the world and learn affordances. A full self-model will also include more cognitive elements—the agent should be able to observe its own cognition, draw conclusions about the structure of that cognition, and fundamentally draw a distinction between the vehicle and content of mental events.

In the end, an enactive perceptual theory might replace the three-dimensional convolutions in Menander 1500 with four-dimensional convolutions that are action guided. Thus, rather than learning static features of a static scene, the agent will learn dynamic features that correspond to expectations from executing certain actions. For example, rather than detecting a single vertical edge as we did before we might now have a thin three dimensional cylinder that changes according to specific robot movements. Based on learning how these visual features change in accordance to simple movements, the robot might learn how things change in accord with more complex movements and this composition corresponds to affordance learning.

As these basic objects are conceptualized in a symbolic theory, we introduce something like a (deflationary) self-awareness into this as well. First-order logic has long been recognized as having the power to make meta-assertions; that is statements in this formal system can be about other statements.[15] This is an important feature of cognitive systems, as it gives agents the capacity to work with *propositional attitudes* and express, for example, the statement that: *I believe that*

---

[14] Gallagher, *How the Body Shapes the Mind*.

[15] This requires a number of technical details, and in fact is only a feature of *some* theories within first-order logic. The basic ideas (which were developed and exploited by Gödel in his incompleteness theorems) are well-explained in Nagel and Newman, *Godel's Proof*.

'the cat food is in the kitchen'. The capacity to thus quote one's own beliefs also gives agents the crucial capacity to change them.[16] As such, they would seem to be essential ingredients of any intelligent system and would need to be included, along with self-modelling, in Menander 3000's architecture.

It is important to note that the contents of these perceptions are neither arbitrary nor based on external essences; rather, they are based on the needs, wants and goals of the organism. In John Dunne's 'Key Features of Dharmakīrti's Apoha Theory', he writes:

> we use concepts not simply out of some pernicious habit, but rather with a specific purpose or goal in mind. ... Dharmakīrti is obliged to show how our words and concepts yield useful information that enables us to act effectively in the world, even without the presence of any real universal.[17]

Our model of perception begins to provide such an account—the fundamental way in which the world is divided into objects is based precisely on the agent's need to act effectively in the world. These in turn ground the words and concepts which the agent explicitly employs in its activities.

Analysis External Ontology

We begin our philosophical analysis of this architecture by noting that it is largely neutral toward the ontological status of external entities. Following Siderits,[18] let us assume a *Sautrāntika*[19] ontology and take a position of 'mereological reductionism'—thus reducing

---

[16] See Cox and Raja, eds., *Metareasoning*.

[17] Dunne, 'Key Features of Dharmakīrti's Apoha Theory', 90.

[18] See Siderits, 'Apohavāda, Nominalism and Resemblance Theories'.

[19] My use of the term '*Sautrāntika* ontology' derives from Tibetan exegesis, and for our purposes mainly refers to the non-idealist reading of the ontology ascribed to Dharmakīrti.

any ultimately existing and causally efficacious external entities to infinitesimal properties. This immediately gives a mathematical formulation of the inability of any conceptual apparatus to capture the true nature of the external world, since the amount of information in any continuous region of space will necessarily be infinite. We sketch the argument here from a couple of different perspectives.

Let us start by breaking a region of space into one metre cubes. According to mereological reductionism, in any whole-part relation the whole cannot exist independently of the parts. Since the one-metre cube of space represented is a whole made up of eight half-metre cubes it cannot be real and any statement about the one-metre cube is at best a summary of the states of the eight sub-cubes. Anything about each of these eight sub-cubes is at best a summary of eight sub-sub-cubes, so that any statement about the original cube is really a summary of sixty-four sub-cubes.

Now, let us assume that each sub-cube can be in at least two states. Then a full description of the state of all sub-cubes would take at least sixty-four bits if each state has equal probability—this quantity is called the *entropy* of the system. As the process continues indefinitely, the number of bits of information required goes to infinity. As any brain or machine can presumably only work with a finite number of basic concepts, these cannot capture the entire situation.[20] Our final ontological commitment is thus that reality is at most composed of infinitesimal particles and a conceptual structure that supervenes on this reality.

Perception

We begin by noting that, in our model, perception is an inherently *causal* process. That is, infinitesimal particulars (the real) interact with Menander's sensors, and a set of a causal chain of (physical)

---

[20] The assumption of equal probability in the above argument is not fully justified, and it seems likely that there are ways of assigning probabilities to the substates that would not result in infinite total entropy. This warrants further investigation.

events which end with some categorical neurons being 'on' (in the sense of having a high voltage running through the associated memory). This is fundamentally in accord with Dharmakīrti's thinking, which sees perception as causally active and conception as causally inert. In our scheme, the contents of symbolic registers will be inert even though the registers themselves will be causally active. As pointed out by Dan Arnold,[21] this dichotomy seems necessary to preserve the kind of mental causation that needs to be accounted for. Specifically, we need to give an explanation for how my thoughts about a cake in my kitchen can spur me into action if concepts are causally inert. The most obvious answer is that the moment of mind (for Dharmakīrti) or the symbolic reasoning system (for this model) is causally active, even if the universals which these systems nominally insatiate are not.

What exactly is it, then, that is perceived? It cannot be the full cascade of pixels and features processed by the network,[22] it can only be the end results of those process that are made available to awareness. Let us, provisionally, take as the result of perception the high-level categorizations that correspond to the end neurons; along with place information which is easily acquired with contemporary segmentation algorithms.[23]

Physicalism

While not committed to a physicalist ontology, the account of conceptuality given here is compatible with one. In chapter 1 of *Brains, Buddhas, and Believing*, Arnold reads Dharmakīrti as committed to dualism in his Sautrāntika manifestation and to idealism when he takes a Yogācārin position. These are both anathema to many AI

---

[21] See Arnold, *Brains, Buddhas, and Believing*.

[22] Although, it is said that *arhat*s can perceive the *dharma*s that constitute reality.

[23] For an account of how features with place information can ground binding and full concept formation, See Ganeri, 'Apoha, Feature-Placing, and Sensory Content'.

researchers while physicalism is seemingly anathema to many Buddhists. Thus there would seem to be a real ontological tension here, and one might justifiably ask whether a physicalist reading of Dharmakīrti is a coherent project. However, in this same section, Arnold argues that Dharmakīrti's *epistemology* is precisely what is invariant whether taking his realist or idealist ontology; thus, one might argue that Dharmakīrti's primary concerns are epistemic rather than ontological. As such, one can expect to find something fundamental in Dharmakīrti's work that can be preserved across ontological readings and this paper could be read as the beginnings of trying to give a reading of Dharmakīrti which is compatible with physicalism. That said, we do not want to develop a model which is completely ontologically neutral since we wish to maintain a commitment to mereological reductionism in particular and some form of emptiness in general.

Another reason Arnold gives for Dharmakīrti's commitment to dualism is that *karma* requires a subjective grounding for experience. While this is reasonable, a commitment to subjectivity need not imply a commitment to dualism. Indeed, my own research (in collaboration with Don Perlis of the University of Maryland) centers on the creation of computational models of subjectivity. There do not seem to be any obvious reasons why such a model could not accurately reflect classical notions of *karma*. Indeed, *karma* is primarily understood by Arnold as deriving from intentions (or *cetana*). Yet the functioning of karma can just as easily be explained as operating based on emergent *cetana* as ultimately existing *cetana*, at the possible cost of a demotion of karmic force from the ultimate to the conventional (this is hardly obvious though—if causality itself is taken as conventional, all causality can be taken as between fictional entities, perhaps in line with the Madhyamika's use of 'mere' causality).

Also, I note that on my reading of Arnold, rejection of physicalism rests on the necessity of a previous moment of conscious experience as a cause for any given moment of consciousness. However, this seems to be a somewhat dubious assumption. For example, modern psychology admits many *unconscious* causes of cognition (as, apparently, does Waldron's reading of Yogācāra).[24]

Ethical Robots

There are some deep ethical questions around the creation of artificial intelligence which are brought up by many other papers in this panel. I will consider some of the issues raised by Douglas Duckworth, Charles Goodman, and Joshua Stoll.[25]

Underlying many of these discussions is whether or not an AI can ever be truly conscious—the so-called strong AI debate. Duckworth and Goodman seem to admit this possibility for different reasons, while Stoll ultimately dissolves the question in a *Prasangika* analysis. I am perhaps most closely aligned with Stoll on this question but with simpler arguments. I do not know whether an artificial mind can ever be conscious, but think that one of Dennet's great contributions[26] was to show that a great number of questions that seem to intersect with consciousness do not rely on the actual presence of qualitative states. Questions of intelligence, ethics, self-awareness and subjectivity can all be separated from the question of whether or not a machine is conscious. Dennet demonstrates this as part of his program of qualitative eliminativism. But you do not need to be an eliminativist to take this potential separation from his arguments—he has (effectively in my mind) shown that these phenomena can be studied in non-qualitative systems. The questions of whether qualia are (in the case of humans), or essentially must be, present in such systems can be bracketed.

From a Buddhist-studies point of view, this is perhaps most poignantly highlighted by Stoll's evidence suggesting that Buddhas, in fact, do *not* have qualitative experiences. That said, there are still very crucial ethical questions at play which *do* intersect fundamentally with the possibility for machine consciousness.[27] Thus contemporary scholars are beginning to ask what kinds of rights conscious robots

---

[24] Waldron, *The Buddhist Unconscious*.

[25] Duckworth, 'A Buddhist Contribution to Artificial Intelligence?'; Goodman, 'Machine Learning, Plant Learning'; Stoll, 'Buddha in the Chinese Room'.

[26] Dennett, *Consciousness Explained*.

[27] Goodman, 'Machine Learning, Plant Learning'.

need to be afforded and whether it is acceptable to create a new species of enslaved beings.[28]

Perhaps most saliently, as the potential for super-intelligent machines grows closer, the need to endow these machines with some kind of ethics becomes clearer. As Goodman notes, there are many who see this as an existential question—if we do not get this right, our human species may not survive. I would propose that, rather than relying purely on a set of codified rules for ethics (these seem to always miss corner cases), we can push ethical behaviors into a more instinctive kind of framework. In Menander 3000, the symbolic processing level is in dynamic interplay with the perceptual/behavioral level. The latter level is much more informationally rich and captures phenomena that would be practically inexpressible at the symbolic level. Thus one might find some hope in training ethical behavior at this more intuitive level by training it in the same way we train it as a perceptual system.

Enactive Representation?

The picture outlined above raises the question of whether or not the perceptual system is representational. The answer seems to depend on how one defines a representation. The highest levels of a convolutional network certainly form a relatively static coding of the robot's environment. However, if we conceive of a representation as a straightforward and static copy of the external world, then clearly these are not representations. In the first place, if we employ mereological reductionism, then there are no cats in the world to reflect. What exists in the world are at most infinitesimal particles which interact causally with the robot's sensory system. Ultimately what the robot learns to detect are not static essences but sensorimotor contingencies; and agents with different bodies and minds will learn different sensorimotor contingencies. This parallels the idea in Buddhism that what humans see as water, *deva*s see as nectar and *preta*s see as pus. Presumably the difference here is not in the environment itself, but

---

[28] Gunkel, *Robot Rights*.

in the way in which the agent might interact with that environment and what is afforded—bliss for the *deva*s and thirst-quenching for the humans. Further, different classes of beings will presumably be members of different linguistic communities, which can perhaps be read as providing a particular set of sensorimotor constraints to its members.

So, this is not at all a way of picking out real, mind-independent properties of the world. To the degree that there is any ontological grounding for final concepts, they rest in sensorimotor contingencies, which are clearly a product of both the agent and the world. Mathematically, a representation is a relationship that preserves some kind of structure (i.e. a homomorphism). What is the relationship between Menander 3000's representation of *cat* and the real world? Let us provisionally assume that the world is composed of infinitesimal dharmas which have a deterministic effect on the sensors. In the first place, the cumulative effects of every dharma are incomputable, so the function is ultimately intractable (i.e. indescribable). So we may, perhaps, speak of a supervenient representation, with the understanding that the same reality will be superveniently represented by different agents in perhaps fundamentally different (but constrained) ways.

As a practical matter, it seems difficult to dispense with representations entirely. We note that Wilson and Foglia argue that such a move brings back many of the same problems of behaviorism[29]. From a Buddhist perspective, I would argue further that since our concepts and representations are a fundamental part of the human condition, some account of how they arise and function conventionally is needed—this seems difficult on a purely non-representational account.

It is worth noting that the model here can support some of the features of 4E cognition noted by Duckworth in the present issue.[30] In particular, because the features learned by the system are based on multiple senses and adaptable, it is easy to imagine that the perception of a person would change fundamentally when a Menander is holding a warm cup—the features and affordances present would

---

[29] Wilson and Foglia, 'Embodied Cognition'.

[30] Duckworth, 'A Buddhist Contribution to Artificial Intelligence?'.

change. Similarly, Menander can adapt his body schema to allow for prosthetics (or more prosaically, tool use).

Conception

Strictly speaking, the outline of concept formation sketched above has attempted to be neutral as to the ontological status of universals—at least the gross universals like *cat-ness*. In this section, I will argue that it offers a coherent account of concept formation that does not rely on real universals. One might conclude that the latter serve no explanatory purpose and that this account thus supports a nominalist ontology. I will discuss this in relation to many of the issues raised by Georges Dreyfus[31] and give an account of how a neural network might instantiate a version of Dharmakīrti's *apoha* theory, but ultimately argue that a more positive account is more natural while fulfilling (at least some of) the basic needs that seem to have led Dharmakīrti to *apoha* in the first place.

Bridging the Perception—Conception Gap

Perhaps the most vexed questions that arise from any theory which separates percepts and concepts arise from questioning how the two are related. If they are really distinct categories, how can they be connected? If universals are causally inert, how can one account for mental causation? This problem is summarized well by Dan Arnold, who writes:

> How, in other words, do we get *from* (nonconceptual, causally describable) perception *to* the kind of semantically contentful thought for which that is supposedly foundational.[32]

This seems to have been a puzzle for Kant, who, according to Arnold, seems to have thought a full solution impossible. I claim that the

---

[31] Dreyfus, 'Apoha as a Naturalized Account of Concept Formation'.

[32] Arnold, *Brains, Buddhas, and Believing*, 121.

model provided by Menander offers at least a partial solution to these puzzles, which I will detail below. The basic idea is that conception operates in two ways: the actual mechanisms of our symbolic reasoning system are purely causal, but their semantic contents are not. We take a fairly deflationary view of intentionality—a symbol's referring to things in the world is just its capacity to be activated by external environments (this will be discussed more fully in the next section). Therefore, in a real sense there is no semantically contentful thought to get to; only constraints on the causal operation of the symbolic system. That said, it is also important for Dharmakīrti that perception is non-conceptual, which Arnold reads as meaning that perception 'by itself does not (indeed cannot) have the kind of content that makes perceptions as giving *reasons* for acting one way or another'.[33] This level of non-conceptuality could be defended for my view of percepts—as sub-symbolic entities they cannot participate directly in the kind of compositional symbolic reasoning that occurs at a higher level. Indeed, as we argued before, there is no tractable description of the operations of the convolution network.

What do we mean by a concept? We gave Parthemore's description above; a simpler description is given by Dharmakīrti, who, according to Dunne, says that 'a concept is a cognition with a phenomenal appearance that is capable of being joined with a linguistic expression'.[34] This would seem to count any symbolic reasoning system as a concept—and indeed symbolic AI owes much to Fodor's notion of a 'language of thought' which grounds all cognition and makes natural language possible.

Finally, we note that a crucial function of concepts is to construe percepts as identical. There is something of a tension here. On a traditional account (as instantiated by Menander 1500), unifying objects temporally would indeed have to occur in the symbolic reasoning system. Specifically, the perceptual system can detect instances of 'cat' at a particular location at distinct times but it would be up to the conceptual system to reify these into the same cat. With the enac-

---

[33] Arnold, *Brains, Buddhas, and Believing*, 120.

[34] Dunne, 'Key Features of Dharmakīrti's Apoha Theory', 87.

tive perception embodied by Menander 3000, the picture is slightly more complicated. Because this system has spatio-temporal perception, the high-level features picked out by the perceptual system would be reified both spatially and over a short duration of time; it would still be up to the conceptual system to reify the percept over longer periods. In short, the classical perceptual system would reify objects in space but not time, while the enactive perceptual system will unite objects spatio-temporally.

Conceptualism

Our main reference for this discussion will be Georges Dreyfus' essay on *apoha* as a form of conceptualism.[35] *Apoha* refers to Dharmakīrti's theory of concept formation—its fundamental tenet is that concepts such as *cow* are determined by the exclusion of everything else (e.g. *non-cow*). In his essay, Dreyfus addresses the question of whether the *apoha* theory, which he brands as a form of conceptualism, can give a coherent account of universal terms without reifying universals. The main critique leveled at so-called resemblance theories is that they replace real universals with real similarities, and thus subtly maintain the realist ontologies they are seeking to refute. Dreyfus defends *apoha* on the ground that it allows for concepts to be understood not via real similarities but only mentally determined ones. We will show that the neural model outlined above offers the same possibility.

We first note that a strictly conceptualist account of concept formation is afforded by a neural network model.[36] When presented with an image of (what is conventionally called) a cat, the early layers of the network will detect types of edges present in the image, higher layers will note the ways in which those edges are configured, and final decision layers will determine probabilities for the image belonging to various classes. If the image is of a cat and the network is well-trained, then there is nothing in this account that corre-

---

[35] See Dreyfus, 'Apoha as a Naturalized Account of Concept Formation'.

[36] For simplicity, we will discuss this in terms of Menander 1500's simpler perceptual model. The same comments apply to Menander 3000.

sponds to trying to detect a universal 'catness'; rather, the network is trained to detect statistical regularities. This is quite explicit—the network is presented with millions of images and is repeatedly told whether or not it has the right answer for what kind of object it is viewing. By adjusting the network weights throughout this process, it learns a set of features and discriminations amongst those features that will maximize its chances of making the correct determination. This can fail in a spectacular way if a new image is presented which does not correspond statistically with the cat images the network has been exposed to.[37] Often, however, these networks work amazingly well.

The main idea of *apoha* theory is that concepts are recognized not directly by identifying real properties but rather by excluding other categories. There is a sense in which this idea might be made to fit with the network model just given. Ultimately, a designation of 'cat' is made because the cat neuron outputs a number that is significantly higher than the output from the other neurons in the final layer. Thus, one way of configuring this network would be to have the detection of features that lead the network to output 'cat' suppress other category-level neurons rather than exciting the 'cat' neuron. Mathematically and functionally, such a scheme would behave equivalently. Fundamentally, however, it seems unimportant to try to align this explicitly with *apoha*. The purpose of the latter, it would seem, is to lay a foundation for knowledge that does not require real properties. I argue that the neural model outlined here also achieves this.

One of the critiques of conceptualism that Dreyfus addresses in 'Apoha as a naturalized account of concept formation' is that it replaces real properties with real similarities. That is, rather than having a real property of 'catness' that makes Fluffy a cat, we are left in a situation in which the real similarity between all particular cats

---

[37] Experimenters have noticed that changing a few pixels in an image of one animal can lead the network to make dramatically different judgements about the category the image belongs to. These changes are imperceptible to humans and would not cause a person to change their judgement.

makes Fluffy suitable to be called a 'cat'. Dreyfus goes on to defend Dharmakīrti's *apoha* from this line of attack on the grounds that the similarities which unite particulars in that system are all mentally constructed. I will argue that a similar defense is at hand under the Menander model, and that indeed there are no real similarities in this model. For example, the notion of similarity which implicitly grounds Menander's use of the term *cat* may differ subtly from another robot's use of the term, even while both may have started with identical programming and both use the term in accordance with convention.

If we examine the specific conditions that will cause Menander to attribute the term 'cat' to two different particulars, we see that this will correspond to the network attributing the label cat to its sensory input with a sufficiently high probability. This, in turn, corresponds to the output of one neuron (the 'cat neuron') outputting a number that is sufficiently higher than the outputs of all the other potential categories. Importantly, the combination of factors that lead to a particular value for that neuron need not be the same. So Menander may focus on the presence of whiskers while his brother (with identical hardware and initial software) may focus on the shape of the feet combined with the presence of a tail. In general, a vast number of different configurations can lead to exactly the same classifications—these will correspond to different robots agreeing on exactly the same classification while disagreeing robustly on the necessary and sufficient conditions which determine that classification.[38] An even greater number will agree on the conventionally recognized cats (which is what they are trained on) while disagreeing at so-called edge cases; images which are not clearly of cats or not.[39] These configurations define the appropriate notion of similarity—we can reasonably say

---

[38] To even call them such is somewhat misleading; because the conditions do not correspond to things like the discrete presence or absence of human-recognized features; rather they correspond to vast arrays of network value which often defy human interpretation.

[39] In Fodor, 'Against Darwinism', Fodor gave an argument which might be paraphrased as noting because two agents might refer to the same particulars with their concepts does not make them the same (intentional) concepts.

that two images are similar if two robots which agree on the conventional classifications of images agree on the class for the two images. Thus, two robots may have massively different concepts of what a 'cat' is, yet both use the term proficiently (in accordance with convention) in most cases. No essential universal or similarity unites these concepts, they are mostly the same in that both agents happened to learn a set of largely overlapping sensorimotor contingencies.

It seems, however, that there is still a sense in which real properties are maintained. One of the interesting facets of this story is that it makes it very clear that there is still some kind of statistical structure that remains. We owe some kind of account of this. On what basis are there statistical regularities (or sensorimotor contingencies) to pick up? There are mathematical results like the Central Limit Theorem and the Law of Large Numbers, which tell us that repeated applications of randomness can begin to look like order. While intriguing, even these results are based on some kind of constancy—if phenomena themselves are random, their randomness must be regular for these results to apply. Why should the world behave in such a way? In short, although we may have done away with the kind of universals the *Nyaya* ascribed to phenomena, the fundamental question about whether real structure is necessary for cognition seems to persist, albeit in a significantly subtler way.

A final ontological question raised by Charles Goodman[41] is whether or not the presentation here involves a reification of set theory. I maintain that it does not, and that even though the arguments here are couched in the language of formal set theory, they are neutral to its ontology. This is similar in spirit to Kurt Gödel's use of set theory in proving his famous Incompleteness Theorems—in particular he uses 'naive set theory' as a kind of meta-language to prove results about formal set theory as an object theory[42]. This does not coerce him into any kind of ontological commitments about the naive set theory he is using for his discussion—his proof goes

---

[40] See Gharamani, *Fundamentals of Probability*.

[41] Personal communication.

[42] See Nagel and Newman, *Godel's Proof*.

through whether one take a realist or a formalist stance on the existence of mathematical objects. I would argue that the same principle is at play here—we can hold set theory as a language of discourse and within that language make arguments about set theory as an object language.

Bibliography

Arnold, Dan. *Brains, Buddhas, and Believing: The Problem of Intentionality in Classical Buddhist and Cognitive-Scientific Philosophy of Mind*. New York: Columbia University Press, 2014.

Cox, Michael T., and Anita Raja, eds. *Metareasoning: Thinking about Thinking*. Cambridge: MIT Press, 2011.

Dennett, Daniel C. *Consciousness Explained*. London: Penguin, 1993.

Dreyfus, Georges. 'Apoha as a Naturalized Account of Concept Formation'. In *Apoha: Buddhist Nominalism and Human Cognition*, edited by Mark Siderits, Tom Tillemans and Arindam Chakrabarti, 207–27. New York: Columbia University Press, 2011.

Duckworth, Douglas. 'A Buddhist Contribution to Artificial Intelligence?' *Hualin International Journal of Buddhist Studies* 3, no. 2 (2020): 27–37.

Dunne, John D. 'Key Features of Dharmakīrti's Apoha Theory'. In *Apoha: Buddhist Nominalism and Human Cognition*, edited by Mark Siderits, Tom Tillemans and Arindam Chakrabarti, 84–108. New York: Columbia University Press, 2011.

Fodor, Jerry. 'Against Darwinism'. *Mind & Language* 23, no. 1 (February 2008): 1–24.

Gallagher, Shaun. *How the Body Shapes the Mind*. New York: Oxford University Press, 2005.

Ganeri, Jonardon. 'Apoha, Feature-Placing, and Sensory Content'. In *Apoha: Buddhist Nominalism and Human Cognition*, edited by Mark Siderits, Tom Tillemans, and Arindam Chakrabarti, 228–46. New York: Columbia University Press, 2011.

Gharamani, Saeed. *Fundamentals of Probability: With Stochastic Processes*. Boca Raton: CRC Press, 2019.

Goodman, Charles. 'Machine Learning, Plant Learning, and the Destabilization of Buddhist Psychology'. *Hualin International Journal of Buddhist Studies* 3, no. 2 (2020): 38–61.

Gunkel, David J. *Robot Rights*. Cambridge: MIT Press, 2018.

Maturana, Humberto R., and Francisco J. Varela. *The Tree of Knowledge: The Biological Roots of Human Understanding*. Boston: New Science Library/Shambhala Publications, 1987.

Nagel, Ernest, and James R. Newman. *Godel's Proof*. London: Routledge, 2012.

Noë, Alva. *Action in Perception*. Cambridge: MIT press, 2004.

Parthemore, Joel. 'The Unified Conceptual Space Theory: an enactive theory of concepts'. *Adaptive Behavior* 21, no. 3 (2013): 168–77.

Redmon, Joseph, Santosh Divvala, Ross Girshick, and Ali Farhadi. 'You Only Look Once: Unified, Real-Time Object Detection'. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 1 (2016): 779–88. DOI: 10.1109/CVPR.2016.91.

Searle, John R. 'Minds, brains, and programs'. *Behavioral and Brain Sciences* 3, no. 3 (1980): 417–24.

Siderits, Mark. 'Apohavāda, Nominalism and Resemblance Theories'. In *Studies in Buddhist Philosophy*, edited by Mark Siderits, Jan Westerhoff, and Christopher Jones, 152–60. Oxford: Oxford University Press, 2016.

———. *Buddhism as Philosophy: An Introduction*. London: Routledge, 2017.

———. 'Śrughna by Dusk'. In *Apoha: Buddhist Nominalism and Human Cognition*, edited by Mark Siderits, Tom Tillemans, and Arindam Chakrabarti, 283–304. New York: Columbia University Press, 2011.

Stoll, Joshua. 'Buddha in the Chinese Room: Empty Persons, Other Mindstreams, and the Strong AI Debate'. *Hualin International Journal of Buddhist Studies* 3, no. 2 (2020): 78–101.

Thompson, Evan. *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*. Cambridge: Belknap Press of Harvard

University Press, 2010.
Varela, Francisco, Evan Thompson, and Eleanor Rosch. *The Embodied Mind*. Cambridge: MIT press, 1991.
Waldron, William S. *The Buddhist Unconscious: The Alaya-vijnana in the Context of Indian Buddhist Thought*. New York: Routledge, 2003.
Wilson, Robert A., and Lucia Foglia. 'Embodied Cognition'. *The Stanford Encyclopedia of Philosophy* (Spring 2017 Edition). Access date June 4, 2020. https://plato.stanford.edu/archives/spr2017/entries/embodied-cognition/.