The Enacted KOAN – An Agent’s Knowledge of Agency

Justin Brody
Goucher College

Abstract
We present Knowledge Of Action Networks, which provide an enactive machine learning model for knowledge of agency in artificial intelligence. These networks, which are expected to be part of embodied intelligences existing in dynamic environments, learn to represent their environment while simultaneously learning to represent their own actions and bodies within that environment. Thus self and world are intricately coupled in their basic representations. We will also explore some of the (many) expected contributions of such networks for implementing minimal self-models, which are basic models of self-aware agents.

Keywords: Enaction, Sense of agency, Deep learning, Embodied cognition, Self

1 Introduction

In this paper we will present a machine learning model that allows an artificial agent to develop a knowledge of agency based on its actions and sensory feedback from its environment. Knowledge of agency refers to an agent’s knowledge that it is undergoing a particular action. This is generally called sense of agency [3] in the literature, but we prefer to use the term knowledge to avoid any implication that our implementation will have any qualitative feeling associated with it. Crucially, the knowledge that we will describe will be immediate in the sense of being unmediated (or at least minimally mediated), grounded in that it interacts dynamically with the “real” world (rather than being an abstract symbol), situated in that it functions as a representation of agency at a particular time and in a particular environment, and open in that it can dynamically alter in new situations (while representing the same thing.) These properties will each be discussed more fully in a later section. I see the current work as grounding many (but not all) uses of the indexical “I”; future work will integrate this into a broader model of a minimal self.

A minimal self is perhaps most simply understood as a minimal model of of an agent’s subjectivity; see [8] and [12] for two different but overlapping approaches to this idea. In [2] it was argued that an account of subjectivity is a fundamental notion in need of a good computational model and that such a model will enable progress on a number of difficult questions in cognitive science. While there is no consensual set of necessary and sufficient conditions for a minimal self-model, the broader program is predicated on endowing a model
with a knowledge of agency, knowledge of ownership, situatedness, temporal awareness and self-reference. I will argue that all but the last two are achieved (at least partially) by the current work; the final properties will be explored in future work.

The network presented here, termed a KOAN (Knowledge Of Agency Network) develops a representation of its agency by learning to predict the sensory consequences of its actions. In particular, the network will represent its environment in a hierarchical way and while it’s learning that representation will try to predict the output of one high-level feature based on the current sensory information and the action it is about to undertake. Since the prediction and representation are learned simultaneously, the result of this will be that the high level feature learned will have to represent a kind of sensory locus of action. For example, if an agent is capable of moving its body in several discrete directions, then this learning paradigm will force the representation to be a representation of the body in a way that covaries with the appropriate actions.

The model is enactive (see, e.g., [6]) in the sense that the prediction forms a simultaneous representation of self and action. Indeed, while the self is modeled in the sensory data it is modeled inherently as an actor rather than a thing. Although the action is modeled with an explicit representation, it will ultimately be represented as a dynamic interaction between self and environment.

In a review article [5], Limanowski and Blakenburg argued for using predictive coding as a way of obtaining an enactive minimal self-model. While the language we use here is different, the current model should translate well into the vocabulary of predictive coding; thus the machine learning model described here can be thought of as providing an implementation of the kind of enactive self-model they describe.

I will present the machine-learning model in the next section and follow this by a discussion of its properties and some of the potential technologies I see arising from it (many of these ideas are present in [5], but with a specific model at hand we can give somewhat more implementational detail).

The work in this paper is a piece of a larger program of implementing a minimal self-model in collaboration with Donald Perlis and his research group at the University of Maryland. In particular, the importance of the initial concepts discussed in Section 3 was pointed out to me by Perlis and many of the other ideas in this paper either directly arose from or else were inspired by conversations with him.

## 2 The Model

In this section I will define a KOAN as a general scheme that relies on an environmental encoding module and a self-action prediction module. Given an agent situated in a changing environment, we will denote the totality of its sensory input at time $t$ by $S_t$ (conceived as a vector in some high-dimensional Euclidean space). If an agent has a fixed (and finite) set of possible actions it can take, say $\alpha_1, \ldots, \alpha_n$, then we will denote the action that it does take at time $t$ by $a_t$. Our basic approach is to learn a high level representation of the environment and force that representation to pick out the effects of the agent’s actions.

To that end we will define a KOAN as consisting of two modules: a hierarchical representational network $\mathcal{H}$ (for example a deep convolutional network) and a predictive network $\mathcal{P}$ (e.g. a multi-layer perceptron). The network $\mathcal{H}$ functions to represent the environment at a given time while $\mathcal{P}$ functions to make predictions about a part that representation based on the current environment and the action taken. In order for $\mathcal{P}$ to successfully make such a prediction, the part of the representation in question must covary with the action, forcing the network to
represent its own “embodiment” in its environment. That is, the network must have an explicit representation of the parts of its environment that change predictably with each specific action, which in sufficiently complex environments should correspond to the body.

Specifically, let us assume that $H$ is a hierarchical model that at time $t$ extracts basic features from $S_t$ and then extract features of those features, and so on iteratively. Formally, we assume that $H$ transforms $S_t$ into a sequence of representations $S^0_t(=S_t), S^1_t, \ldots S^m_t$ where each $S^i_{t+1}$ is a set of features $\{S^i_{t+1,1}, \ldots, S^i_{t+1,l_i}\}$ that can be used to describe the features of $S^i_t$. For example, in a convolutional network the features of $S^1_t$ might represent the result of detecting various kinds of edges in the input image while $S^2_t$ might then represent combinations of those edges, etc. Ideally of such a representation might correspond to our own higher level concepts about a sensory input (for example, we might detect “person” based on detecting various shapes in a particular configuration, while those shapes are based on detecting various types of edges and shadings, etc.) Let us denote the first feature at the highest level of the hierarchy by $\sigma_t$ (thus $\sigma_t := S^m_t$); then our goal will be for $\sigma_t$ become a representation of the agent itself in the environment. \footnote{An interesting alternative representation would be to attempt a prediction of $\sigma_{t+1}$ based only on $\sigma_t a_t$ (rather than using all of $S_t$; the disadvantage here is that one would not be able to account for context-dependent consequence of actions (e.g. bouncing right when moving left). The advantage would be that using the whole environment may make incorporate too much in the represented self. In some ways this may be good, since there is experimental evidence from Haggard (discussed later) that agency can be attributed to surprising parts of the environment.}

This will be accomplished by using the network $P$. We will assume that $P$ takes as input at time $t$ the vector $S_t$ along with a vector-representation of the action $a_t$. It will then output a vector $\pi_t$ which will be its prediction of what $\sigma_{t+1}$ will be.

To train our model, we will assume that $H$ computes the representations on the basis of learnable parameters (at time $t$) $\theta_t$ and that learning proceeds by minimizing a loss function $L_H(S_t; \theta_t)$. If our predictive network is parameterized by $\psi$, then we will define a new loss function $L_P(S_t S_{t+1}; \psi \theta_{t+1})$ that measures the predictive error as some kind of difference between $\pi_t$ and $\sigma_{t+1}$. The total loss at time $t + 1$ would then be the sum of these:

$$L_{t+1} = L_H(S_t; \theta_t) + L_P(S_t S_{t+1}; \psi \theta_{t+1})$$ \hspace{1cm} (1)
The simplest choice for $L_P$ might be the mean-squared error, so that equation (1) becomes

$$L_{t+1} = L_H(S_t; \theta_t) + \eta \| \sigma_{t+1} - \pi_t \|^2$$

where $\eta$ is a meta-parameter that controls the balance between trying to learn good features of the environment and trying to make good predictions. It is worth noting that a priori this function runs the risk of having both $\pi_t$ and $\sigma_{t+1}$ be constant, but that since this should effectively decrease the value of $\sigma_{t+1}$ as a high level feature of $H$, this shouldn’t happen in practice.

Currently experimentation is using stochastic gradient descent to minimize this loss; given reasonably well-behaved choices of $L_P$ and $L_H$ faster convergence should be possible.

### 3 Network Properties and Potential Capabilities

One of the main purposes of this network is to provide for a grounded representation of the self. If we view the network as representing itself and its action, then this network has a number of desirable properties that would seem to match our human self-representations which are not present in classical logic-based agents.

In the first place, the self-representation of a KOAN is immediate. There is a long history in the phenomenological philosophical tradition (see, e.g. [13]) of noting that self-awareness is *pre-reflective*. That is, our knowledge of our selves does not come about after some of conceptual/deductive reflection – it is rather given immediately to experience. In the same way, once a self-representation has been learned, the self-representation in the highest level of $H$ will be immediately given at any moment – it arises from the structure of the network rather than from doing additional computations on the input.

In a similar way, our self-representation is grounded. I mean by this that the network’s representation of itself is not just an abstract symbol, but is based on real experiences of itself and tied in an explicitly causal manner to its own actions. Thus it avoids the philosophical (and practical) difficulties posed by the symbol grounding problem [11]. Further, Don Perlis has argued that a good grounded self-symbol can potentially serve as an anchor for grounding meanings more generally [7].

Finally, such a symbol is situated and open. The representation of an action in the hierarchical network will not be a fixed entity, but will rather be instantiated by a different value for the feature in each given situation, and potentially in new situations as well. This is true even while the fact that the contents of the feature are a self-representation remains constant. Thus we have a constant symbol (sense) with a kind of shifting reference. In particular, this formulation shows a path toward allowing for the indexical “I” in artificial systems. This is a difficult problem (if not impossible) for systems that do not use situated symbols, as described eloquently in the work of John Perry [9, 10].

I also contend that a KOAN can provide support for a number of cognitively important capabilities, many of them tied to other properties of a minimal self-model. It is worth noting that while I believe a KOAN is an effective implementation of the predictive coding framework

\[ \text{This description is usually in the context of self-awareness conceived as consciousness rather than bodily-self awareness. In embodied traditions, where the body is an integral part of such awareness, a bodily-self awareness would arguably be a crucial component of such self-awareness.} \]

\[ \text{While one could argue that passing the sensory input forward through the network is itself a kind of extra computation, it seems reasonable to assume that any perceptual system with a hierarchical representations would have all levels of such representation present at any given moment as a “baseline” for computation.} \]
for a sense of agency described by Limanowski and Blakenburg, I do not believe that the latter by itself entails a minimal self-model. In particular, in [2], Don Perlis, Michael Cox and I argued that such a model will necessarily have an immediate access to its own processing (as self-awareness) in real-time.

Capabilities that the KOAN model seems likely to provide for include a knowledge of ownership associated with bodily selves, the capability for multimodal integration (as discussed in [5]), a fluid sense of agency described in P. Haggard’s experimental work [4], the learning of affordances, corollary discharge and real-time error detection and correction.

Knowledge of ownership refers to an agent’s awareness of its body as its own. There are some subtleties involved in what constitutes ownership, but a reasonable first stipulation might be that whatever an agent controls is “owned” by that agent. Thus all moveable body parts would be considered owned by an agent, as would its voice and tactile sensation. Some thought reveals control may be sufficient for ownership but is not necessary, since if it were I would not own my thoughts or autonomic nervous system. There does, however, seem to be some kind of relation between such phenomena and those that are controlled by me (my autonomic nervous system is contained in the body that I move volitionally; my thoughts may not be completely volitional but are at least partially so.) It thus seems plausible that knowledge of ownership can be grounded in the kind of knowledge of agency described here. In particular, a KOAN can pick out what the parts of the environment under its control by simply looking at a kind of inverse image of $\sigma_t$. That is, once an effective representation of self is established we can take those parts of $S_t$ that are represented in $\sigma_t$ as being controlled by the agent.

Multimodal integration refers to the fact that agency is not usually experienced in a single modality but across modalities. For example, my sense of pressing the keys on my keyboard draws on the visual feedback of my fingers moving the keys, the tactile and proprioceptive feelings of pressing the keys and the sounds that the keyboard makes as I interact with it. These do not give rise to three separate agencies, but rather to a single unified sense of agency that draws on all of these sensory modalities as contributions. Representing agency multimodally is straightforward in our model—we simply let $S_t$ denote the totality of sensory input from all modalities. Such a representation gives rise to a number of interesting questions.

One immediate question is the role of various modalities in contributing to agency. For example, if I close my eyes and continue to type I still derive a sense of agency from the sensory input I receive from other modalities. It is an empirical question whether this sense of agency is any way weaker than it would be with all senses engaged or if different modalities contribute differently. More generally one might ask whether a sense of agency is a binary or graded phenomenon. It is similarly an empirical question of how a trained KOAN model would respond to this question; presumably if its training disagreed with the empirical evidence about humans, the model parameters could be modified to agree with human behavior.

Another interesting avenue of investigation would be to examine the role of certain kinds of prior knowledge in agency and see how they could be implemented in a KOAN. For example, it seems plausible that a lack of visual input will have a smaller effect on my sense of agency if I know that my vision is impaired (and can thus amplify other modalities) than if I don’t.

In a somewhat similar vein, we note experiments of Patrick Haggard in which subjects learned a sense of agency from pressing a button with their right hands while simultaneously having their left hands moved involuntarily by a machine [4]. Haggard found that these subjects also developed a sense of agency associated solely with the involuntary motion. Such a derived sense of agency would naturally be modelled by our networks. Indeed, since our predictive network looks for statistical regularities (rather than, e.g. causal connections) this will be their

---

4As before, the literature refers to this as a sense of ownership.
default behavior. As such they model (and perhaps hint at an explanation for) an interesting phenomenon in human psychology.

Further, these networks may allow for the natural learning of affordances. Indeed, I hypothesize (and am in the process of testing empirically) that an agent which is forced to learn a representation of itself will naturally be pushed to learning meaningful representations of its environment. In particular, an agent with a robust self-representation should more easily learn representations for different ways an action can effect an environment and representation for what parts of an environment can effect it.

Another use of these networks would be to subtract the expected sensory output of an action from the received sensory input. This mechanism is known as corollary discharge in humans, and is thought to be the mechanism at play in the phenomenon of one’s being unable to tickle oneself [1] – the tactile input from your own touch is subtracted (at least partially) from the total tactile input so that self-touch is not felt as strongly.

Finally, we can use these networks to perform real-time error detection and correction (as is done with “visual servoing” in robotics). Here the expected sensory feedback can be compared with the received feedback, and corrective action can be taken if they don’t match. For example, if an agent intended to move its arm five centimeters to the right but found that it only moved three centimeters, it can continue its motion for longer, continue it with more force, or else try to engage in some reasoning to determine what’s going on and what an appropriate corrective would be.

References