The Problem of Knowledge and Data

Anna Koop

April 17, 2011

Contents

1 The Problem of Knowledge and Data

What knowledge is and how it is represented are important questions of cognition, and the debate about how to understand knowledge in biological and artificial agents is not likely to end soon. The possibilities have tantalized philosophers and scientists for centuries, but they easily becomes clouded with esoteric and untestable notions. At the same time, having a clear and defensible answer to "What might knowledge be?" is necessary if we are to design knowledgeable agents.

One reason the answer eludes us is that it is easier get lost in philosophical tangles than precisely state the question (let alone find a solution). So we build and test our artificial agents, aiming for pragmatic progress and glossing over ambiguity—to our loss. I am not arguing we should turn philosophers and forgo implementation. But a little bit of care will give us the practical benefits of functional and clear definitions.

For the purposes of this work, what I mean by 'knowledge' is the summary information used by the mind for cognition. It is general, abstract, and stable: the currency of mental activity. However represented, knowledge can be applied in many situations (especially new situations), is not tied to exact details (though it may have contextual nuance), and is reliable (and relevant) over time. This is a working definition of knowledge, intended to be inclusive.

So knowledge is summary information and we want to build agents that can use knowledge. Usually this means designing agents with some particular mental structure, a representation of some kind. Building a representation means choosing how to build it and what to build it out of. These are the questions that generally get the most attention in AI. But there is another simple question that needs to be addressed: What is it that we want to represent? Knowledge, yes, but what information is that summary meant to be about? This is the question of the *referent* of knowledge. We want an artificial agent to know something about the world it is operating in (or the world *we* are operating in) and so we tend to assume the referent should be the relevant content of the world, but here things get a little tricky. More on this objective view in a moment.

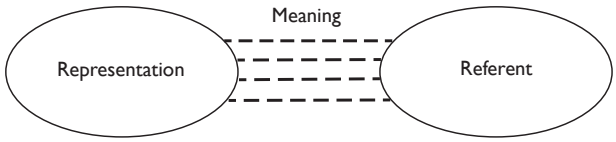


Figure 1: A visualization of the parts of knowledge.

That's knowledge, then, but what about *agent*? An agent is a decision-maker, interacting with the environment it inhabits via sensorimotor signals (see Figure 2). These signals are the mediator between the mind of the agent and the environment it is embedded in, but that doesn't explain what the mind *does* with those signals. How is sensorimotor data—those ever-changing, detailed, particular signals—connected to a stable, abstract, generally useful bit of knowledge? Biological agents integrate sensorimotor data and knowledge to create an astounding range of intelligent behaviour: We would like to do the same artificially[1]. This

[1]Functionally the same, anyway.

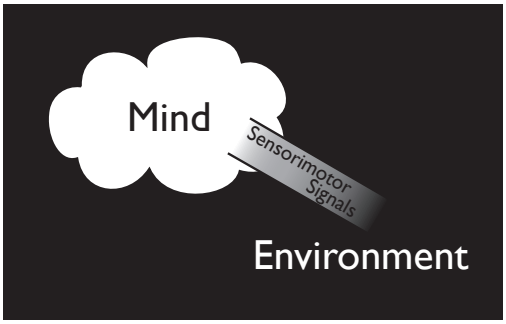


Figure 2: The architecture of the agent/environment system.

integration is the problem of knowledge and data.

Technological advances have made the problem of knowledge and data more relevant—and progress more possible—than ever before. In the early stages of AI, memory and processing resources were strikingly limited. The standard approach to knowledge representation was to concentrate on inference within a knowledge base and let human experts "write down what needs to be known" [Brachman and Levesque, 2004]. Let the knowledge engineers decide which are the important referents and how to encode them and—later—how those referents can be integrated with sensorimotor data [Nilsson, 2009]. This approach to knowledge representation, sometimes referred to as GOFAI or KR[2], is still in use today[3]. But increases in memory and processing power have increased the possibilities. The growth of machine learning has been spurred by ever-greater sources and stores of data. And recently, reinforcement learning research has begun on how the state of an agent/environment system might be represented in terms of sensorimotor data (more on that in Section 3.2).

Our ability to handle ever-more data in AI has been paralleled in cognitive science by increasing ability to observe the processes of biological minds. In the early stages of psychology, the brain was a black box, with no way of observing the activity in a living brain. Now researchers in neuroscience and psychology can avail themselves of ever-finer tools for monitoring live neurological activity. And it turns out there is activity in the parts of the brain that process sensorimotor signals even when the body accompanying that brain lies immobile in a scanning machine [Barsalou, 1999]. Furthermore, advances in comparative cognition and developmental psychology have broadened our understanding of intelligence. The evidence that non-primates and pre-linguistic children have cognitive abilities worthy of investigation opens the possibility for intelligent agents that are *not* proxies for adult, domain-expert logicians [Gallistel and King, 2009, Wasserman and Zentall, 2006, Gopnik et al., 2000]. These trends suggest it might be time to address the problem of knowledge and data directly.

1.1 Proposal overview

To make progress on the problem of knowledge and data, it will be important to understand how the integration of knowledge and data is currently addressed. I propose undertaking this

[2]For good, old-fashioned AI and Knowledge Representation.

[3]And generally what people think you mean by 'knowledge representation'. To my eternal frustration.

analysis in Section 2, as the first contribution of my thesis. In Section 2.1 I will try to define the scope of the analysis. I will be making the argument that understanding the meaning of a representation (its referent and approach to connecting representation and represented) is particularly important for the problem of knowledge and data.

The inclusiveness of my working definition of knowledge is valuable, I believe, because the problem of knowledge and data transcends representational details. A thorough survey of state-of-the-art may include control theory, GOFAI, machine learning—even comparative cognition and cognitive science. A discussion of some specific candidates for analysis is in Section 2.2. Discovering the commonalities and conflicts across these areas will be useful for understanding the core of problem.

There are, at least, to be two potential approaches to meaning. On the one hand, there is the objective view mentioned in the introduction, that knowledge is meant to accurately correspond to real things in the world. This view seems closest to our everyday use of the word—knowledge is *about* the world and its contents. The intuitiveness of objective referents certainly helps for constructing human-interpretable knowledge. On the other hand, there is the empirical view that knowledge is meant to correspond to testable data—in particular, sensorimotor data. This view provides both verifiable meaning and an intrinsic solution to the problem of knowledge and data. Between verifiable meaning and an intrinsic solution to the problem of knowledge and data, the benefits of an empirical approach could be enormous. It would certainly help for autonomous AI. Are we going to be forced to chose between understanding and autonomy? Or can we get similar benefits from different approaches? These two philosophical approaches are explored further in Section 2.3. Their different practical consequences suggest the potential in the careful analysis of referents.

Although I am arguing that clarity about knowledge and data is valuable in itself, I would like to explore implementing a knowledge representation directly. I propose a second stage to my thesis in Section 3, the development of a particular empirical representation based on current reinforcement learning research in predictive representations. What it might mean to develop a predictive knowledge representation is explained in Section 3.1, and the particular tools and relevant research are reviewed in Section 3.2. My initial ideas about how to investigate the consequences of taking the empirical and predictive stance are described in Section 3.3.

The analysis I propose in Section 2 will result in clarity about the strengths and weaknesses of current state-of-the-art knowledge representation, including the various ways the problem of knowledge and data is current addressed. The final step in my thesis will be returning to this analysis to evaluate the progress made in Stage 2.

If we can be clear about our fundamental assumptions about knowledge and data, then we can start to examine the implications of those assumptions and explore new possibilities. Knowing how and why our approaches differ, we can better understand the consequences of the differences and move towards comparisons that, rather than touting higher scores in limited test domains, illustrate the different strengths and appropriate context for different approaches. This, in turn, can help improve our understanding of the larger problem of knowledge and data, which is a critical issue for understanding minds.

2 Thesis Part 1: Clarifying the strengths and weaknesses of different approaches to knowledge

The first project I propose is a broad analysis of how a variety of representational approaches within AI address the problem of knowledge and data. I will be using an inclusive definition of knowledge as the general, abstract, stable stuff of the mind. This allows me to consider a wide range of frameworks: from the logical knowledge bases of GOFAI and the dynamical models of control theorists to the state-transition models of reinforcement learning and latent variables in Bayesian networks. The explanation of scope and a brief overview of potentially-relevant representations are covered in Section 2.1 and Section 2.2. A case-study of sorts, of the practical consequences of two different ways of addressing the question of referent, is covered in Section 2.3.

2.1 Analyzing the state of the union of knowledge and data

Undertaking an inclusive analysis of knowledge representation in AI is an important and timely step for autonomous AI. Surveys of knowledge representation, while sometimes defining knowledge in inclusive terms, are overwhelming dominated by classical knowledge representation [Brachman and Levesque, 2004, Russell and Norvig, 2003].Machine learning and control theory, although they deal with the representation and application of summary information, are treated as orthogonal to KR. I believe that taking a broad view will aid in understanding core issues in knowledge representation. Implicit assumptions about the referent of knowledge and narrow focus on specific testbeds can make it difficult to compare results across fields. It is easy to be distracted by the details of particular implementations within subfields. Yet we are, at some level, working on the same problem. The growing research in reinforcement learning on predictive representations of state [Littman et al., 2002, Singh et al., 2004, Tanner et al., 2007] and relational reinforcement learning [Tadepalli et al., 2004, Diuk et al., 2008, Van Otterlo, 2009] speak to the interest in integrating knowledge and data in new ways. There have been recent advances in the integration of reinforcement learning and cognitive science [Todd et al., 2009, Ludvig et al., 2008]. These suggest that there is rising interest in inclusive analyses. It is my intention to undertake this with a broad survey of the problem of knowledge and data. A big-picture analysis of what we have learned so far will bring together contributions from disparate fields. This has the potential to improve our understanding of critical issues.

Across this broad base I will be focussing on representational choices relevant to the problem of knowledge and data: how knowledge has been integrated with sensorimotor data and how this and other issues are affected by choice of referent. Existing analyses of knowledge in AI tend to focus on representational details within subfields. Researchers concentrate on differences in representational architecture and consequently miss the significance of different assumptions about knowledge itself. The architecture of a knowledge representation and the referent of knowledge therein are separable issues. There has been little attention to what the knowledge might and should mean in a broader sense. I will be looking for clear statements about the strengths and weaknesses of the different approaches to referent and data.

2.2 Existing approaches and potential areas to investigate

My analysis is an attempt to understand many different ways of getting at abstract, general, stable representations and dealing with the problem of knowledge and data. I will be trying to identify what knowledge means in these representations. I suspect in the full analysis this will involve identifying both what the knowledge is meant to mean and what it actually does mean (the referent in theory and practice). In the full survey I am aiming for breadth on a focused question rather than depth through exhaustive comparison. At the end, I should have a clear list of the strengths and weaknesses of different choices in referent and representation, across a range of specific approaches to knowledge representation in AI.

2.2.1 KR

No survey of knowledge representation could be complete without addressing the behemoth of classical knowledge representation (here referred to as KR or GOFAI). This is, broadly speaking, the family of approaches that represent knowledge as logical statements meant to correspond to objective reality. Examples of note still actively developed include Cyc and the Semantic Web [Lenat, 1995, Berners-Lee et al., 2001]. Cyc is a huge knowledge base of 'common-sense' knowledge meant to provide the foundation for human-level AI, now developed by Cycorp [http://cyc.com, 2011]. The Semantic Web takes a crowd-sourced approach to knowledge engineering and provides the rules and structures for anyone to provide the definitions for logical content [http://www.w3.org/2001/sw/SW FAQ, 2011].

These representations are general, abstract, and stable, but tend to concentrate on reasoning to the exclusion of action. It is not that they discount the problem of knowledge and data, more that the integration of data is seen as separable from the representation of knowledge. The 'symbol grounding problem' was introduced by Harnad as the question of how symbolic knowledge has meaning at all, and is generally addressed by posing sensorimotor data as the solution [Harnad, 1990, Ziemke, 1999]. Given a massive knowledge base of human-approved knowledge, sensorimotor data is applied in various ways to give true meaning to symbols. An analysis of the ways KR has dealt with symbol grounding will most likely be key for analyzing their overall approach to the problem of knowledge and data.

Relational reinforcement learning (RRL) is a fairly young offshoot of KR, that takes the agent/environment stance of reinforcement learning but posits sensorimotor signals that correspond to objective entities [Tadepalli et al., 2004]. I believe a clear analysis of the meaning of knowledge in RRL would be helpful in understanding how relational reinforcement learning supports and subverts traditional reinforcement learning [Van Otterlo, 2009, Diuk et al., 2008].

2.2.2 Reinforcement learning

Reinforcement learning (RL) is concerned with the behaviour of agents interacting with an environment and thus consistently confronts the problem of knowledge and data [Sutton and Barto, 1998]. But the referent of knowledge for an RL agent can vary greatly.

On the control theory side, the problem of interest is monitoring and adjusting a dynamical system [Sutton and Barto, 1998]. Knowledge is represented as models of state transitions

(or in the case of partially-observable markov decision processes, state transitions and observation/action models) [Chrisman, 1992, Cassandra et al., 1995]. The referent in these representations is the true state of the system, an objective entity.

State representation (and similarly function approximation and feature discovery) is a particularly relevant aspect of reinforcement learning. The simplest view of state is that the sensory signals received from the environment correspond to unambiguous identifiers for the state of the agent/environment system. Knowledge, generally in the form of a value function (a measure of the 'goodness' of a state), refers to an empirical entity [Sutton and Barto, 1998]. Of course it quickly gets more complicated than this. When the sensory signal does not correspond to the agent/environment state, there are generally two choices for the new referent. The dominant approach is to still construct knowledge that refers to the agent/environment state, but acknowledge that the sensorimotor signal provides imperfect information about that state [Cassandra et al., 1995]. The alternative is to construct an agent-state from the sensorimotor experience of the agent, either through memory or projection [McCallum, 1995, Tanner, 2005, Rafols, 2006, McCracken and Bowling, 2006].

2.2.3 Graphical models

Graphical models have been significant in the development of modern probabilistic AI, particularly in the form of Bayesian networks [Russell and Norvig, 2003]. They have also been applied to modeling language acquisition and concept learning in cognitive science [Sanborn et al., 2006, Lucas et al., 2010]. Bayesian approaches use probabilistic networks to represent knowledge. The network represents the probabilistic relationships between the referents of the nodes, and those probabilities are updated according to available data. The probabilistic relationships are objective entities like the agent/environment state in reinforcement learning. In most Bayesian networks the nodes refer to concepts determined by domain experts[4]. Thus knowledge representation using Bayesian networks has objective referents but integrates knowledge and sensorimotor data in calculations.

2.2.4 Embodied cognition

Embodied cognition emphasizes the critical nature of being-in-the-world, positing that the mind cannot be separated from its environment [Wilson, 2002, Pezzulo, 2009]. This is obviously a field of interest for understanding the problem of knowledge and data: Embodiment insists on inherent integration with sensorimotor data. However, the referent of embodied knowledge is still often objective rather than empirical. Grush's treatise on embodiment, for example, talks about the importance of understanding activity in terms of body-awareness—the robot should know the angles of its joints [Grush, 2004]. In this case, the referent of the robot's knowledge is not inherent in the sensorimotor data. The contrast between the fierce commitment to experience and adherence to philosophical realism in this area is interesting.

The most extreme stance within embodied cognition is the rejection of mind-without-physical-body [Wilson, 2002]. In AI, this means only developing robotic agents, since software-based agents are considered incapable of cognition. I am interested in investigating this view to understand why the sensorimotor signals of a software agent are seen as categorically

[4]My overall impression, given cited works. Specifics will, of course, be necessary in the actual analysis

different (regardless of the complexity of the environment) from the sensorimotor signals of a physically realized agent. The softer version of this stance, that the physical realization of a robot profoundly changes its representational requirements, seems to me to illustrate the need to analyze the problem of knowledge and data [Tedrake et al., 2004].

In behaviour-based robotics, it is the data that is the focus and the knowledge that gets postponed, in direct contrast to standard KR. This is most famously seen in Brooks' 'use the world as its own model' rallying cry [Brooks, 1991]. It has led to widespread research in behaviour-based robotics. A brief analysis of these and other anti-represention approaches could prove interesting.

Mapping and navigation for robotic systems is another potential area for analysis. Machine learning techniques such as SLAM (Simultaneous Localization and Mapping) build bird's-eye maps to assist the robot's autonomous navigation [Dissanayake et al., 2001]. This representation has objective referents (the map is meant to refer to physical space) built from objective sensorimotor models (possibly learned from data, but meant to represent the physics behind the sensorimotor signals), adjusted with empirical data. This way of integrating knowledge and data is arguably the standard in machine learning overall.

2.2.5 Machine learning

Supervised learning does not usually come to mind when considering approaches to knowledge representation, but it is about summary information: a regression or classification function. Supervised learning is all about correspondence with the labels—by definition. Objective, then. But within supervised learning we talk about the problems of overfitting and underfitting, which can be seen as wanting to correspond with the true function generating the data, or as wanting to improve accuracy on future test data. Similar problem, opposite referent. The simplicity with which the problem of supervised learning can be defined might make it useful to relate to the distinctions I make in referential approaches.

Unsupervised learning and manifold discovery are two areas of machine learning I think could be interesting to analyze as approaches to knowledge representation, although they infrequently deal with temporal data. Unsupervised learning and manifold discovery are both about uncovering patterns in data. This can again be understood in both an objective and empirical sense. They are often presented as ways of uncovering underlying systems (objective), but can also be seen as representing patterns within the data without needing deeper referents (empirical). Including these methods in my overview of knowledge representation could prove helpful.

2.3 Potential referents

Let us turn now to looking at the objective/empirical distinction in more detail. There are two immediate candidates for the referent of knowledge: the world itself and the sensorimotor signals by which we perceive the world. Artificial intelligence has generally been dominated by an objective approach. Knowledge is facts about the world—physical objects, state-transition matrices, classification functions. Objective knowledge defines knowledge as being about the real, external world. Knowing about coffee cups is knowing something about the physical entity in the environment. Empirical knowledge defines knowledge as being about

patterns in sensorimotor data. Knowing about coffee cups is recognizing something about this particular moment in time[5]. These two main approaches to the referent of knowledge are described in more detail in the following sections, with an preliminary overview of their strengths and weaknesses.

2.3.1 Objective referent: Knowledge about real entities

What could be more stable, abstract, and general than a true understanding of the world itself, its objects and relations and laws? Knowledge should extend beyond individual circumstance into the commons, not be constrained by the experience of a particular mind. Surely it must do so at least partly by referencing this shared, objective space. An intelligent agent may misunderstand or be unaware of what is really out there, but it is the objectively real physical and metaphysical things that form the basis of truth for an objective approach. It seems intuitive, so let us explore the implications.

True Essence: Objective reality is general and abstract. Having objective reality as the basis of knowledge seems to get at the core of the general abstractions we want our agent's knowledge representation to have. If our definitions and models correspond to the true essence of the wider world, then by construction they have abstracted away incidental details. When we want our agents to have 'justified, true belief' about the facts of the world, our referent should be that objective reality. Not even Descartes' demon can hide the fact that what we want to represent is reality, uncertain though our access to that reality may be.

This objective approach brings up the problem of definition in a broader sense. The definition of reality is unknown and our best understanding of it shifts. Are there four elements or 118[6]? Is smallpox caused by bad blood or a virus? General relativity or quantum mechanics? An objective stance defines knowledge as real and true things that may be only viewed through shadows on the cave wall, mediated by our senses and never truly knowable.

Reliable Communication: Objective reality provides a stable basis for communication. An old philosophical conundrum: how do you know that what I mean by red is what you mean by red? Our desire for stable knowledge that transcends the individual seems to tautologically require objective data: how could knowledge about subjective experience be shared? Your experience of red may be an ineffable mystery, a personal quirk of your own consciousness[7], but at least we can share the real things between us.

Language is thus a favoured example for the objective approach. Words may have nuance, connotation, sense and tense, but they also have definition, denotation, and reference to something external and shared. They are used as if they refer to real entities[8]. If I ask you to pass the cup of coffee, we must have some shared understanding of what cups of coffee are and which thing in the world I want you to move. According to the objective representation we will understand each other because our representations have the same referent. As long

[5]Real or imagined: empirical knowledge does allow for offline reasoning [Silver, 2009]

[6]or more

[7]Consciousness being a can of worms I will leave closed, if possible.

[8]concrete or abstract

your representation and my representation are both refering to the same coffee cup, we should get along just fine.

'Parasitic' meaning: Objectivity requires intermediary channels. Knowledge about objective reality, whatever the structure, requires a mediator between reality and the representation. Reality is a black box (see Figure 2). Sensorimotor data is generally volunteered as the channel between mental representations and reality for objective representations. This may require a relatively simple mapping, as when the observation signal and agent/environment state fortuitously align (see Section 2.2.2). Or it may be more difficult to define, requiring sophisticated supervised learning to connect referent and data. Grounding objective knowledge requires not only a method for relating the objective representation to the sensorimotor data, but also an understanding of how the sensorimotor data arises from the real world. Sensor and movement models, conceptual labels, state-transition matrices must be constructed.

Thus the second option for connecting objective reality and an internal representation: oracular intervention. We humans can provide the labels and the concepts, the objects and laws. For a supervised learning algorithm, this means providing human-labeled data. For a logical knowledge base, the inferences can be translated back into human-readable terms and checked by humans. The human mind provides the link between internal representations and external reality. It may be this method of grounding makes the meaning of the representation parasitic on our own minds, but perhaps Harnad and Searle were both unduly concerned by this symbiosis between humans and machine [Searle, 1980, Harnad, 1990]. Having other agents be an intrinsic part of epistemology may not be unique to artificial knowledge[9].

Existing research suggests, however, that requiring an intermediary for the grounding and verification of knowledge makes artificial autonomy difficult. The reliance human effort to construct knowledge has been a long-standing issue in classical knowledge representations, slowing development and creating problems with brittle representations [Brachman and Levesque, 2004, Taylor et al., 2007]. Providing labelled data for machine learning can be expensive and time-consuming. The intermediary required by objective representation comes at a price.

2.3.2 Empirical referent: Knowledge about sensorimotor signals.

Sensorimotor signals are uninterpreted, noisy, ridiculously fine-grained, constantly changing, and wholly dependent on an individual. Yet they are the only sure thing in an intelligent agent's world. All I have available to construct knowledge and determine my behaviour is the contents of my mind and the sensorimotor signals of the moment. This does not mean that I am trapped in a solipsistic delusion. Sensorimotor signals are generated by the agent/environment system, not by the agent alone, and the environment exists whether as our universe, a computer game, the Matrix, or Dual Earth. But whatever the system is, only the sensorimotor signals are directly accessible to an agent's mind.

This empirical approach is both an intensely embodied and a radically disembodied view of cognition. It is intensely embodied in that the physics of the agent/environment

[9]Cf. Wittgenstein's argument that meaning requires community [Wittgenstein et al., 2009]

system will completely alter the sensorimotor streams. A robot with no light sensors cannot see colours. A bipedal robot modelled after a wooden toy can more easily learn to walk [Tedrake et al., 2004]. Each has its own particular way of sensing and its own particular channels for acting, and its particulars are dependent on its body and the environment it inhabits. But the sensorimotor signals are disembodied in the sense that knowledge of sensorimotor signals need not define the physics of the body or environment, and certainly a physical body is not required. We may be in a world where matter is a strange probabilistic beast and energy can be reified. We may be in a world that was dreamed up last Tuesday, with all our memories and historic records intact but false. It doesn't matter. We can never know, because everything we know is mediated by our senses. It is the statements about the sensorimotor data that have knowledge content, not statements about the objective system.

In an empirical representation, the referent itself is available to the agent. Knowledge is grounded in accessible data and parasitic on no one. Verifying knowledge becomes a simple matter of comparing the contents of the representation to the available data. This is the scientific approach: string theory and loop quantum gravity can spin tales about whatever theorized entities they like. Their veracity always comes back to the testable predictions they pose. But some difficulties remain.

Unmediated, noisy data: Sensorimotor signals are uninterpreted. Sensorimotor signals should not be confused with perceptions and actions in casual language. The agent/environment interface is a statement about an information boundary, not a physical one. Sensorimotor signals are the uninterpreted, place-coded bits going into and out of a mind. In a robot this could be the current to the motors and the pixels from the sensors: bits as raw as we can make them. Or it could be motor signals programmed to be a particular direction and speed, combined with vision signals that are the output of an edge-detector operating on a captured frame. In either case the signals are still—*to the agent*—uninterpreted data coming and going in the memory space. They may be sensorimotor signals that we, the programmers, design to have a certain referent to makes our lives easier, but that meaning is in our representation, not the agent's.

Nor should the certainty of sensorimotor data be confused with correspondence with reality. The empirical understanding of meaning does away with such correspondence. Hallucinations exist, to say nothing of phantom limbs or neuronal misfirings. Sensorimotor signals are occasionally inconsistent and frequently imprecise. They are generated by the complex interactions between agent and environment and transmitted through lossy channels. Although we hope and believe these systems are ultimately predictable (and have evidence that biological minds manage reasonably well), the way an intelligent agent makes sense of the signals may or may not correspond to how the signals are generated by the environment. Thus the distinction between objective and empirical representation: If knowledge is about reality, then it is accurate when it corresponds to reality. If knowledge is about empirical data, then it is accurate when it corresponds to that data, regardless of its correspondence to objective reality.

The Devil in the Details: Sensorimotor signals provide great detail and no summary. The abundance of minute detail found in sensorimotor data can be overwhelming.

For all agents, sensorimotor signals travel on different channels, communicating through different modalities. So an empirical representation has rich variety to draw on, but no inherent summaries. Not only can you not see the forest for the trees, you can't see the trees for the flashes of green and the roughness of bark and the sting of pine-scented wind. There is gloriously personal detail in the sensorimotor data, but it is all detail.

The agent might have various modalities and detailed inputs and outputs, but there is an embodied consistency in the location of the signals. Although we can remap the touch sensors on the tongue to the vision processor of the brain, the mind only adapts if that mapping has reliable temporal consistency [Doidge, 2010]. Observation bit 7 may be meant to correspond to the presence of a block or the smell of cut grass, but its values are determined by the embedded system. The consistent mapping and existence of some system hiding in the black box of the environment allows an empirical representation to map the patterns in the wide stream of signals. Patterns in the data allow the possibility for stable, general summaries of what is inherently specific and changeable.

Egocentric cognition: Sensorimotor signals are inherently subjective. The sensorimotor data stream is not only ridiculously detailed, it is particular to a given agent at a given moment in time. Basing knowledge in such subjective data would seem to invite disaster for our goal of general, stable, and abstract knowledge. But as place-coding provides a potential basis for abstract patterns, temporal context provides a source of stability.

All the intelligent agents we know of are trapped in time. They access the past only by imperfect memory, predict the future with varying degrees of uncertainty, and can act only in the now, that instantaneous moment. It seems problematic to ask for the creation of stable knowledge from a transient existence. But the moment in which knowledge needs to be used is a moment with its own sensorimotor, temporal, and internal context, it is a moment in the timeseries of the agent/environment system. The temporality of experience gives an empirical representation context that is missing in disembodied, dissociated inputs. The context comes for free, the stable patterns have to be uncovered.

2.4 Referential advantages and disadvantages transcend representational detail

An objective representation lends itself to general, abstract, and stable content, which we have taken as the definitive characteristics of knowledge. Objective knowledge seems to get at the heart of the abstract concepts humans are used to talking about. At the same time, knowledge about objective reality seems to require an external source of data for grounding and verification.

An empirical representation has an easier task in grounding and verification, being about internally accessible data. This provides a practical advantage for autonomous knowledge representation. However, constructing general, abstract and stable content from the ephemera of sensorimotor signals might be difficult.

A thorough analysis of the different strengths and weaknesses common to referential and representational choices should support these conclusions. Or it may contradict, which would be interesting in itself. But it will be necessary to analyze a broad range of existing

representational approaches to find the consistencies (if any) of fundamental attitudes to knowledge and data and reference.

The analysis will provide a reference point for further research in knowledge representation. With a clear list of strengths and weaknesses, I can have a measure of progress that does not depend only on my latest score in Tetris or Blocks World or Go. It also provides direction for the extension of predictive representations to general knowledge representation. These two features of the analysis will be invaluable for the experiments I propose next.

3 Thesis Part 2: Investigating an empirical approach to knowledge

The second stage of my thesis is to develop an empirical knowledge representation. In particular, I wish to extend existing work on predictive representations towards more general, abstract, and stable knowledge. Preliminary work on predictive representations has shown surprising potential for defining knowledge in terms of experience, but examples are scattered and no unified look at the representational possibilities exists. I would like to continue this work, developing the use of predictions for conceptual knowledge that uses an empirical representation but approaches the kinds of knowledge that come naturally to objective representations.

Designing an empirical knowledge representation leads to certain definitional choices about the role of mind, knowledge, and sensorimotor data. These are described in Section 3.1. Previous research in reinforcement learning and predictive representations has provided groundwork for this project, defining and investigating several tools for abstraction. I briefly cover these in Section 3.2. I intend to take advantage of my reinforcement learning background through the extension of these tools and development of simple test environments. My initial ideas for testing are described in Section 3.3. In designing these simple worlds I hope to draw inspiration from comparative cognition. The study of intelligence across species provides fertile ground for the investigation of artificial general intelligence.

3.1 Preliminary choices for an empirical representation

Working within an empirical approach and developing a predictive representation means making particular choices about knowledge and data. In such a representation, predictions are the threads that weave knowledge from data. Knowledge is constructed from and refers to information actually available to the intelligent agent. Both the referent and the representation are empirical. Predictions are the key mechanism for defining and anchoring knowledge in concrete signals.

The empirical role of the mind is to predict and control sensorimotor signals. An empirical choice for the role of the mind is to have it deal directly with the sensorimotor data stream. Reasoning and planning are grounded in sensorimotor data and they cannot be separated from the dynamical context the mind inhabits. Planning can occur offline in an empirical representation, but it does so by projecting sensorimotor experience [Silver, 2009, Sutton et al., 2007]. In my proposed work, I will be concentrating on the prediction of sensorimotor signals in particular.

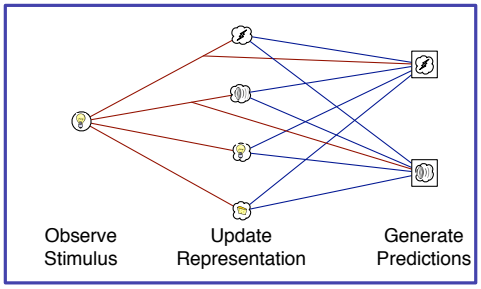
Empirical knowledge is about patterns within sensorimotor data. Empirical knowledge is about the signals rather than the system. The agent is operating within a system that generates sensorimotor signals, presumably a principled system that is possible to predict. But although we may have guesses about the entities contained within the environment and the physics governing their behaviour, empirical knowledge cannot be directly *about* those

things. All hypotheses are mediated by and measured against sensorimotor data. Thus it is appropriate that knowledge refer to and be built from that data.

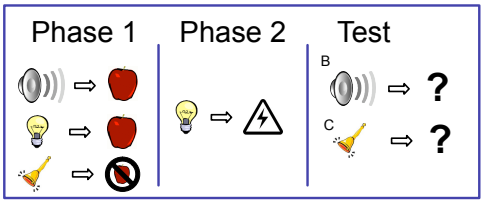
Predictions about specified behaviours form the statements of empirical knowledge. Conceptual knowledge is created out of predictions about the outcomes of specified behaviours. Predictions are themselves signals—their value changes as sensorimotor signals change and as the agent thinks. But predictions can be about complex patterns in the agent's representation or future sensorimotor experience and can be conditioned on general and temporally extended ways of behaving. This creates general and stable signals from detailed and shifting data.

3.2 Specific tools for building abstraction

Previous work in the area of predictive representations provides some interesting results and useful tools for developing empirical knowledge. These results, although arising largely from state-representation research (and animal learning models), suggest the possibility of using prediction as the basis of abstraction. Developing a predictive framework specifically for empirical knowledge seems a reasonable next step.



(a) Representation for mediated association



(b) Acquired Equivalence preparation.

Figure 3: The representation used to model mediated association effects. The stimuli on each timestep are used to update the internal representation, which is then used to calculate the prediction features. In acquired equivalence experiments the shock effect is generalized to unrelated stimuli based on previous experiments where the two stimuli had similar rewards. See [Ludvig et al., 2008] for details.

3.2.1 Prediction: Grounded association and classification.

Predictions are representational units that represent the expected future value[10] of another specific representation bit. Elliot Ludvig and I developed a model of a particular aspect of animal learning that illustrated how predictions can link disparate stimuli. Mediated association is when two unrelated stimuli, such as a tone and a click, become associated through shared consequence (see [Ludvig et al., 2008] for more detail). In our model, the stimuli were related by the prediction units within the representation, and learning new associations in the model showed similar transfer as in experimental data for rats and pigeons.

[10] Or other estimable statistic

This is related to what Zentall calls associative concept learning [Zentall et al., 2008], and is related the humans the ability to link arbitrary symbols (such as words) to varied stimuli. An earlier study by Rafols et al. showed how predictions can form a practical basis for generalizing over distinct environment states, by classifying according to 'predictively equivalent classes' [Rafols et al., 2005]. Both these prior works hint at how predictions provide a possible mechanism for abstraction.

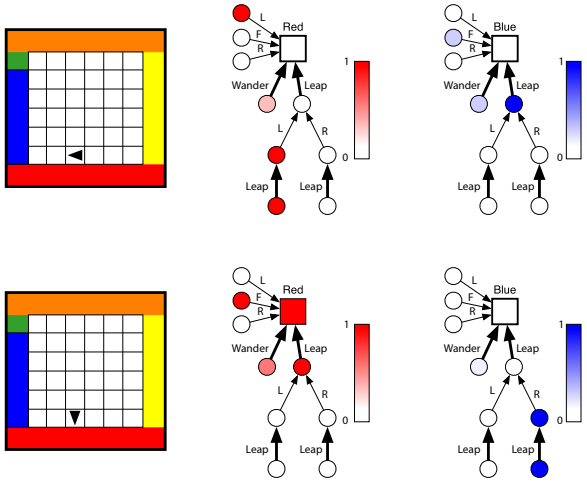


Figure 4: The compass world, on the left showing the current state of the environment and on the right showing part of the agent's internal representation. The squares represent the observation value of the particular colour bit at that time point (so the red square is filled in when the agent is directly facing a red wall). The round shapes represent predictions which as about the representation bit they are pointing to after taking the action the associated arrow is annotated with. The degree of shading illustrates the predicted expected value on the current time step. See [Sutton et al., 2006] for more detail.

3.2.2 Option-Based Prediction: Abstracting over time.

Predictions can be conditioned on specific actions, including extended ways of behaving that in the reinforcement learning literature are known as *options* [Sutton et al., 1999]. This opens up the possibilities for representing notions that span more than a moment of time. In the Compass World, Sutton et al. demonstrated how a predictive representation (specifically, a temporal-difference network with option-conditioned predictions) can allow an agent to maintain knowledge of its allocentric orientation even while receiving uninformative observations [Sutton et al., 2006]. The agent could represent 'which wall it was facing' even while its immediate observations showed only empty white space, and could maintain this representation indefinitely. Furthermore, because the predictions were conditioned on options, it was possible for the agent to transfer learning to larger worlds, even worlds that were too large to learn from scratch [Rafols, 2006]

Figure 5: An older child is surprised when the ball is no longer visible when the barrier is removed. A younger child is not. This experiment is used to test (and illustrate) knowledge of the persistence of objects. See [Koop, 2008] for more details.

3.2.3 Temporal Coherence: Stability from patterns in predictions.

In my Master's thesis I began an investigation into empirical knowledge representation, and developed the idea of temporal coherence as relevant for abstracting sensorimotor data [Koop, 2008]. Temporal coherence is, as you would expect, the tendency for a signal to be consistent (or not) over local time spans. Objective reality seems to have relatively high temporal coherence: objects stay where you put them, seasons slowly wane, people rarely change. Empirical data comes in a noisy, chattering stream of signals. Identifying the temporal coherence of particular bits seems to be relevant for the development of stable knowledge.

As part of that work, I examined the idea of an empirical object concept. The transient retinal input as our eyes flit around the room does not translate directly to the apparent stable persistence of objects. The objective view of persistence: I know the ball is there even when I can't see it. My thesis developed the empirical equivalent to this objective knowledge: I predict that a search procedure will lead to ball-experience even when immediate ball-experience is absent. The prediction signal has temporal coherence from signals with low temporal coherence.

3.2.4 Empirical concepts: Units of empirical representation

Knowing that the coffee cup is on the table in an objective knowledge representation means being (actually and justifiably) aware that out there in the physical world the coffee-cup-object has a particular spatial relationship to the table-object. In an empirical, predictive knowledge representation, on the other hand, it means knowing that if I move my eyes around a certain way I can set off the visual-coffee-cup-experience, and moving my eyes a certain way after that will lead to the visual-table-experience. Coffee-cup-experience and coffee-cup-object are different entities. They can be used in very similar ways, and may even have similar representational structure, but they have different referents, differences with

practical implications.

Empirical concepts are ways of splitting up moments of time, not objects in the environment or features of the environment. They may pick out specific sensorimotor patterns or they may pick out patterns of prediction, but we have experience-of-X and timesteps-on-which-Y concepts. Not things that can be defined in terms of universal characteristics. Objects are eddies in the experiential stream, not physical entities existing within the objective environment. Knowing about chairs has more to do with timesteps-that-afford-sitting than a 3D model of a prototypical object. Thus, the concepts that come out of this determined empirical approach will not have exact parallels with traditional concepts. We will not be grounding objective concepts in sensorimotor data à la Harnad [Harnad, 1990], but developing a new understanding of empirical concepts.

The most sophisticated tests, whether large hadron collider or spoken question, are still mediated by our senses. We have theories about the physical world. Theories about, for example, objects and solidity and boundaries and gravity are supported every time that you sit in a chair and don't slip through it to the floor. But the chair-ness of the chair is not directly perceived (if chair-ness exists at all). The theory is tested by the sensorimotor signals it predicts. In an empirical representation, theories about the system underlying the environment side are encouraged: compact ways of predicting that unending stream better allow the mind to predict and control the sensorimotor stream. Helping the mind do its job is the point of knowledge. Concepts in an empirical representation may exist as mental representations [Margolis and Laurence, 2006], but will be different than the concepts of more traditional representation. Part of the research endeavour I propose will be exploring how well these prediction-based concepts can fill the role of traditional concepts.

3.3 Experimental framework

I would like to have a clear way to investigate and illustrate the capabilities of my predictive representation. I will be aiming to develop experiments that are small and informative rather than grand-challenges that prove the worth of one approach over another. In this I hope to draw inspiration from the field of comparative cognition. Experiments in comparative cognition tend to be designed to disambiguate "mere" association from "actual" learning, or test different kinds of cognitive abilities and conceptual distinctions. They also, being largely focused on non-linguistic species, provide an interesting model for testing the abilities of an agent whose brain we can scan but whose thoughts we can't fathom. Thus, the first set of experiments I propose will be in a simple gridworld, whose complexity I can scale slowly according to the conceptual questions that arise.

3.3.1 The agent/environment interface

I am coming at this research from the a reinforcement learning perspective, which poses the problem of intelligence as that of an agent interacting with an environment. The sensorimotor signals are the stream of information passing between the agent and environment; The agent is the part of the system that can make decisions; The environment is everything else. This may include the agent's body, the physically external environment or other agents, but it

can all be understood as the complicated black-box system generating and responding to the sensorimotor stream.

In my proposed experiments I will be designing the environment and programming it, so it will not be a black-box to me. This makes it important to ensure there is a way of illustration the differences in the objective and empirical approaches. I will be able to look inside the environment system and likely provide objective views on its operation, but I will be emphasizing the subjective experience of the agent, at least through a simultaneous display of its sensorimotor experience and internal state. Since I am investigating empirical knowledge representation, I want to be able to switch between the traditional objective view, where the agent can be seen from a bird's-eye-view, to a subjective view, with a reasonable display of the realtime sensorimotor signals the agent is experiencing, and an fMRI-inspired view of elements of the agent's internal representation, so that predictions can be singled out and their fluctuation observed in realtime. The agent's access to the environment is always and only via the stream of sensorimotor inputs and outputs it receives and gives. The empirical representation embraces this view and I hope to build an experimental interface that supports this.

3.3.2 Comparative cognition: Guidance for disambiguating knowledge

Experiments in comparative cognition are carefully designed to disambiguate one model of animal intelligence from another, to elicit behaviours that arise from particular knowledge or skills. In artificial systems we know what model the intelligent agent is using and have complete control over its behaviour. But we don't always know what particular knowledge or skills might be supported by the knowledge representation we have given it.

There are several ways in which comparative and development cognition overlap intriguingly with artificial intelligence. There is the intelligence controversy: as various capabilities are claimed as distinctly human (until discovered not to be), so various grand challenges are decided not to really require intelligence (once an AI has mastered the task). The moving goal posts of human cognition compared to animal and artificial provide incentive in both fields to be clear about our claims. Besides this, the emphasis in comparative cognition on crossing or disregarding language barriers[11] predisposes researchers to be tuned to the problem of knowledge and data. For example, Zentall and Wasserman's book on animal cognition defines cognition as planning with and making sense of sensorimotor data [Zentall et al., 2008] while Margolis and Laurence's survey of concepts in human cognition defines the role of the mind in objective terms [Margolis and Laurence, 2006]. The necessities of exploring the cognitive activity in pre-linguistic children and animals tends to alleviate selection bias towards intelligent behaviour.

Investigating experiments in comparative cognition that have been developed to probe the representation of particular kinds of concepts could be particularly useful for my research project. Zentall provides a nice definition of different kinds of conceptual learning, from simple association to detection of meta-relationships, with associated experiments [Zentall et al., 2008]. These experiments could prove enlightening for testing knowledge representation systems Caching experiments in scrub jays have demonstrated that

[11] mainly the absence of language

birds can remember details about cached food, apply new knowledge of food spoilage to previously cached food, and adjust behaviour according to the presence of other watching birds [Gallistel and King, 2009]. All these abilities were apparently once considered[12] beyond the scope of non-primates. The experiment design used to support the generality of their knowledge could be helpful in my investigation. Similar experiments could be used in a gridworld for illustrating object-like and other-mind awareness in an artificial framework. There are a wealth of experiments for animal navigation that provide intriguing comparisons to traditional for robot navigation [Tarsitano, 2006, Smith and Litchfield, 2010, Wasserman and Zentall, 2006]. SLAM (simultaneous localization and mapping) and other robotic techniques rely on building an objective-view map [Dissanayake et al., 2001]. This may be the exception rather than the norm for biological agents

Inspiration from comparative cognition, taken together with the results of the analysis I propose in Section 2 should help in the development of compelling, yet simple, tests of artificial systems. I think developing illustrations of cognitive difference, may ultimately be more fruitful for general artificial intelligence than the current standard approach. Grand challenges are met and then declared improperly won. Scores on limited test domains are vulnerable to overtuning and testing bias. It may be that pursuing better test scores in particular domains will never suffice for measuring progress on the vague, yet important, problem of knowledge representation.

3.3.3 Gridworld test domains: Clear illustration and development.

The illustrative environment I intend to build will have the physics of a gridworld (discrete motion that can be blocked or impeded by features of the environment) with an egocentric point-mass agent (local observation signals and minimal physics). The particular agent I am considering can rotate left or right and move forward or backward. It observes the grid squares in front and beside it as bits: the front bits correlate to the grids in front, and the side bits are the logical OR of the squares on each side of the agent. Thus we have a foveal focus in front, peripheral vision in the sides, and a blind spot behind. The agent can additionally take an action to observe something about the square it is currently on, and take an action which has no effect. I am proposing this action/observation space as minimal-yet-interesting. The observation space in particular will have signals that vary how much information they provide, which I suspect will be a helpful extension to the bit-to-bit gridworlds in usual predictive representations.

There is a conundrum for knowledge representation in toy worlds: that perfect, specific knowledge does accurately describe what we want our agents to represent. A gridworld is completely described by the state list, transition function, and reward function on the environment side; and the policy on the agent side[13]. If you know those functions, you know exactly everything there is to know about the world, at least according to the objective understanding of meaning. And if this specific world is the only one you will ever operate in, this is reasonable. However, perfect knowledge of a toy example is not generally considered impressive, in objective or empirical terms.

[12] And I hear there's some hold-outs

[13] Also starting distributions state-space and such might need to be defined.

I believe that staying in a simple test domain but shifting to emphasize the empirical data might provide a new perspective on this issue. Walking through potential objective and empirical representations in these simple worlds should test this theory. It will be illuminating to see how empirical and objective knowledge differ in these environments.

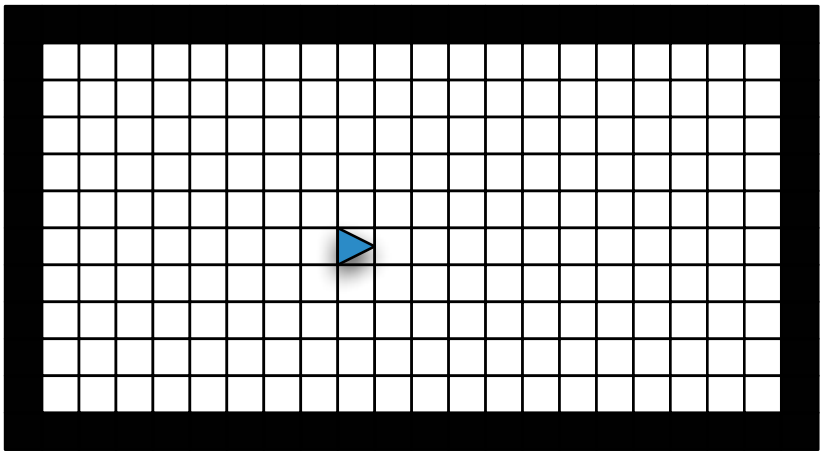


Figure 6: The simplest of the test gridworld environments: an empty, rectangular room. It provides a simple illustration of the difference between allocentric and egocentric knowledge, the issue of state aliasing and symmetry, and the ability of predictions to move a sensorimotor representation beyond a solely reactive agent.

The first environment is a simple rectangle, shown in Figure 6. Even in this empty room there are interesting knowledge issues to explore. One is the difference between a perfect specific representation and a general representation. The usual approach in reinforcement learning to an environment such as this would be to build a state model that allows the agent to disambiguate each square and orientation, and learn the transition probabilities for each action. This is fine, but transferring that knowledge to a room even one unit larger becomes problematic, although empirical representations tend to fare better than objective in this regard [Rafols, 2006, Tanner et al., 2007]. Treating the environment state as the most important thing you can know about the world ties us to specific worlds, and misses the strengths of general knowledge representation.

At the same time, a general knowledge representation is not capable of the detail that a representation based on agent/state is. There are symmetries in the world in Figure 6 that cannot be broken by any sensorimotor signals. The objective referents then contain strictly more information than the empirical. But if no sensorimotor data of any past, present or future can disambiguate the two corners, then the corners cannot be disambiguated in any representation unless oracular knowledge is provided. This distinction, sometimes overlooked in applications of reinforcement learning, is one reason I have argued understanding what we want to represent is just as important as deciding how to represent it. The match-up of sensorimotor data to objective referents cannot happen in this simple world. Objectively this world has north, south, east, and west walls. The empirical representation data only of long and short walls, to my right or to my left.

I envision a series of simple additions to this gridworld that allow for new conceptual distinctions. This gridworld can be extended to include fixed or moving blocks, stochastic actions and observations, and movable shapes and other agents. I can then explore how the predictive representation distinguishes (or fails to distinguish) these additions. Periodically moving blocks explode the state space in the objective representation and provide an illustration of how timing might be represented objectively and empirically. Movable and randomly moving blocks provide a simple illustration of causality: how does an empirical representation distinguish between blocks-that-can-be-moved, blocks-that-are-fixed, and blocks-that-move-of-their-own-accord? Different modes of navigation can be explored when landmarks are introduced to the environment. Finally, it might even be possible to explore the representation of theory-of-mind[14] in a simple environment with multiple agents. Theory of mind experiments in pigeons and scrub jays have reasonably simple setups, and mimicking those explorations in a simple gridworld may not be beyond consideration.

[14]Opponent modelling, one might say. If one worked in games.

4 Goals and Contributions Review

The first contribution I am aiming for is a clear analysis of how a broad range of approaches to knowledge representations address the problem of knowledge and data. I hope to precisely describe the possible choices of referent and have an analysis of the strengths and weaknesses that come naturally to each. A clear understanding of the problem of knowledge and data and a survey of the representational approaches will provide a measure for progress on the second portion of my thesis.

At the conclusion of the second portion of the thesis, I hope to have provided both proof-of-concept and compelling examples for the use of prediction for interesting, abstract knowledge representation. I sill use the analysis of the first portion to measure progress in the second. I hope to provide a clear understanding of what comes easily to a predictive, empirical representation, and what may be difficult at this point in its development.

4.1 Proposed Timeline and Milestones

Spring/Summer 2011: Analysis

Analysis completed: survey of representational approaches, noting their representational philosophy, uncovering strengths and weaknesses of representational choices, including of referent and epistemic meaning.

Paper on the problem of knowledge and data.

Fall 2011: Implementing empirical knowledge

Development of a gridworld that allows for a variety of concepts relating to the strengths and weaknesses previously identified.

Programming a predictive representation for initial gridworld tests, answers to some questions: what kinds of predictions are necessary to represent identified concepts? How much has to be provided to the representation and how much can be easily learned? Is the representation sensitive to parameter tuning or relatively stable? What kinds of predictions can be transferred to new environments?

Collaborating with an expert in comparative cognition for the development of rigorous tests in the programmed environment.

Winter 2012: Testing the new framework

Testing the capabilities of predictive representation in selected experiments. Investigating the scalability of the algorithm developed so far.

Analysis of progress with respect to strengths and weaknesses determined earlier.

Paper on the empirical knowledge representation developed: advances and difficulties so far.

Spring/Summer 2012: Defence

Final Write-up and Defence

Glossary

abstract summarized to an ideal; containing only necessary information.

agent a decision-making entity capable of learning and sensing. Embedded in a causal system: i.e. decisions have effect.

cognition general mental activity; thinking, reasoning, planning, making decisions, remembering, learning, imagining, sensing, perceiving, emoting.

comparative cognition the study of intelligence across species, the differences and similarities in mental processes.

data units of information.

detailed containing particular information.

empirical defined in terms of testable data.

empirical representation a knowledge representation whose knowledge has with a referent that can be directly tested.

general applicable to more than one situation or context.

GOFAI Good, old-fashioned AI; see KR.

grounding connecting a representation to its referent.

knowledge summary information used by the mind for cognition.

knowledge representation summarized information stored in particular form. Also, the study of the storage of summarized information.

KR the classical approach to knowledge representation in artificial intelligence, still pursued today, that holds that knowledge representations should be constructed from symbolic propositional statements.

meaning tthe relationship between a representation and its referent. Also: the actual or intended referent.

mind the centre of cognition for an intelligent agent. Not necessarily bound by brain or body.

objective external to and independent of an agent.

objective representation a knowledge representation with objective referents for the knowledge therein.

particular with characteristics determined by or applicable to a specific context.

referent the thing knowledge is about; what knowledge is meant to refer to. Also, the target of a representation; what the representation is representing.

representation a structure that attempts to capture the essence of some thing; a representation of something's essence in usable form.

sensorimotor experience the sensorimotor signals of a particular agent over time.

sensorimotor signals the data sent to and from a mind that can be understand as sensory input and control (or motor) output.

signals dynamic carriers of information across an information boundary that have particular values at any moment.

stable staying consistent over time.

subjective particular to an agent.

temporal coherence the tendency of a signal to be stable over local time spans.

the problem of knowledge and data the question of how mental representations (and knowledge in general) can be integrated with sensorimotor signals.

verification evaluating the accuracy of knowledge or a knowledge representation.

References

[Barsalou, 1999] Barsalou, L. W. (1999). Perceptual symbol systems. *The Behavioral and brain sciences*, 22(4):577–609; discussion 610–60.

[Berners-Lee et al., 2001] Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The Semantic Web. *Scientific American*, 284(5):34–43.

[Brachman and Levesque, 2004] Brachman, R. J. and Levesque, H. (2004). *Knowledge Representation and Reasoning*. Morgan Kaufmann Series in Artificial Intelligence. Morgan Kaufmann.

[Brooks, 1991] Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence Journal*, 47:139–160.

[Cassandra et al., 1995] Cassandra, A. R., Kaelbling, L. P., and Littman, M. L. (1995). Acting Optimally in Partially Observable Stochastic Domains. In *Proceedings of the National Conference on Artificial Intelligence*, pages 1023–1023.

[Chrisman, 1992] Chrisman, L. (1992). Reinforcement Learning with Perceptual Aliasing: The Perceptual Distinctions Approach. In *Proceedings of the National Conference on Artificial Intelligence*, pages 183–183.

[Dissanayake et al., 2001] Dissanayake, M., Newman, P., Clark, S., Durrant-Whyte, H., and Csorba, M. (2001). A solution to the simultaneous localization and map building (SLAM) problem. *IEEE Transactions on Robotics and Automation*, 17(3):229–241.

[Diuk et al., 2008] Diuk, C., Cohen, A., and Littman, M. (2008). An object-oriented representation for efficient reinforcement learning. In *Proceedings of the 25th international conference on Machine learning*, pages 240–247. ACM.

[Doidge, 2010] Doidge, N. (2010). *The Brain That Changes Itself: Stories of Personal Triumph from the Frontiers of Brain Science*. Scribe Publications.

[Gallistel and King, 2009] Gallistel, C. R. and King, A. P. (2009). *Memory and the Computational Brain: Why Cognitive Science Will Transform Neuroscience*. John Wiley and Sons.

[Gopnik et al., 2000] Gopnik, A., Meltzoff, A. N., and Kuhl, P. K. (2000). *The scientist in the crib: what early learning tells us about the mind*. Perennial.

[Grush, 2004] Grush, R. (2004). The emulation theory of representation : Motor control , imagery , and perception. *Behavioral and Brain Sciences*, pages 377–442.

[Harnad, 1990] Harnad, S. (1990). The Symbol Grounding Problem. *Physica D*, 42:335–346.

[http://cyc.com, 2011] http://cyc.com (2011). Cycorp, Inc.

[http://www.w3.org/2001/sw/SW FAQ, 2011] http://www.w3.org/2001/sw/SW FAQ (2011). W3C Semantic Web FAQ.

[Koop, 2008] Koop, A. (2008). *Investigating Experience: Temporal Coherence and Empirical Knowledge Representation*. Master's thesis, University of Alberta.

[Lenat, 1995] Lenat, D. B. (1995). CYC: a large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.

[Littman et al., 2002] Littman, M. L., Sutton, R. S., and Singh, S. (2002). Predictive representations of state. In *Advances in neural information processing systems*, volume 2, pages 1555–1562.

[Ludvig et al., 2008] Ludvig, E. A., Sutton, R. S., and Kehoe, E. J. (2008). Stimulus representation and the timing of reward-prediction errors in models of the dopamine system. *Neural computation*, 20(12):3034–54.

[Margolis and Laurence, 2006] Margolis, E. and Laurence, S. (2006). Concepts. *Stanford Encyclopedia of Philosophy*.

[McCallum, 1995] McCallum, A. K. (1995). *Reinforcement Learning with Selective Perception and Hidden State*. PhD thesis, University of Rochester, Rochester, New York.

[McCracken and Bowling, 2006] McCracken, P. and Bowling, M. (2006). Online Discovery and Learning of Predictive State Representations. In Weiss, Y., Schölkopf, B., and Platt, J., editors, *Advances in Neural Information Processing Systems 18*, pages 875–882, Cambridge, MA. MIT Press.

[Nilsson, 2009] Nilsson, N. (2009). *The quest for artificial intelligence*. Cambridge University Press New York, NY, USA, web version edition.

[Pezzulo, 2009] Pezzulo, G. (2009). Grounding Procedural and Declarative Knowledge in Sensorimotor Anticipation. *Mind and Language*, pages 1–28.

[Rafols, 2006] Rafols, E. J. (2006). *Temporal Abstraction in Temporal-Difference Networks*. Master of science, University of Alberta.

[Rafols et al., 2005] Rafols, E. J., Ring, M. B., Sutton, R. S., and Tanner, B. (2005). Using predictive representations to improve generalization in reinforcement learning. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence*, pages 835–840.

[Russell and Norvig, 2003] Russell, S. J. and Norvig, P. (2003). *Artificial intelligence: a modern approach*. Prentice Hall.

[Searle, 1980] Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3):417–424.

[Silver, 2009] Silver, D. (2009). *Reinforcement Learning and Simulation-Based Search*. Doctor of philosophy, University of Alberta.

[Singh et al., 2004] Singh, S., James, M. R., and Rudary, M. R. (2004). Predictive State Representations: A New Theory for Modeling Dynamical Systems. In *Uncertainty in Artificial Intelligence: Proceedings of the Twentieth Conference (UAI)*, pages 512–519.

[Smith and Litchfield, 2010] Smith, B. P. and Litchfield, C. A. (2010). How well do dingoes, Canis dingo, perform on the detour task? *Animal Behaviour*, 80(1):155–162.

[Sutton and Barto, 1998] Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. MIT Press.

[Sutton et al., 2007] Sutton, R. S., Koop, A., and Silver, D. (2007). On the Role of Tracking in Stationary Environments. In *Proceedings of the 24th International Conference on Machine Learning (ICML 2007)*, pages 878–886, Corvallis, OR.

[Sutton et al., 1999] Sutton, R. S., Precup, D., and Singh, S. (1999). Between MDPs and Semi-MDPs: A Framework for Temporal Abstraction in Reinforcement Learning. *Artificial Intelligence*, 112(1-2).

[Sutton et al., 2006] Sutton, R. S., Rafols, E. J., and Koop, A. (2006). Temporal abstraction in temporal-difference networks. In *Advances in Neural Information Processing Systems 18 (NIPS 2005)*, Vancouver, BC.

[Tadepalli et al., 2004] Tadepalli, P., Givan, R., and Driessens, K. (2004). Relational reinforcement learning: An overview. In *Proceedings of the ICML 04 workshop on Relational Reinforcement Learning*, volume 4, pages 1–9.

[Tanner, 2005] Tanner, B. (2005). *Temporal-Difference Networks*. Master of science, University of Alberta.

[Tanner et al., 2007] Tanner, B., Bulitko, V., Koop, A., and Paduraru, C. (2007). Grounding abstractions in predictive state representations. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI 2007)*, pages 1077–1082, Hyderabad, India.

[Tarsitano, 2006] Tarsitano, M. (2006). Route selection by a jumping spider (Portia labiata) during the locomotory phase of a detour. *Animal Behaviour*, 72(6):1437–1442.

[Taylor et al., 2007] Taylor, M. E., Matuszek, C., Klimt, B., and Witbrock, M. (2007). Autonomous Classification of Knowledge into an Ontology. *Artificial Intelligence*, (May).

[Tedrake et al., 2004] Tedrake, R., Zhang, T., and Seung, H. (2004). Stochastic policy gradient reinforcement learning on a simple 3D biped. In *Proceedings of the 10th International Conference on Intelligent Robots and Systems, 2004 (IROS 2004)*, volume 3, pages 2849–2854. IEEE.

[Todd et al., 2009] Todd, M., Niv, Y., and Cohen, J. (2009). Learning to use working memory in partially observable environments through dopaminergic reinforcement. In *Neural information processing systems*, pages 1689–1696.

[Van Otterlo, 2009] Van Otterlo, M. (2009). *The logic of adaptive behavior*. IOS Press, Amsterdam.

[Wasserman and Zentall, 2006] Wasserman, E. A. and Zentall, T. R. (2006). *Comparative cognition: experimental explorations of animal intelligence*. Oxford University Press.

[Wilson, 2002] Wilson, M. (2002). Six views of embodied cognition. *Psychonomic bulletin & review*, 9(4):625–636.

[Wittgenstein et al., 2009] Wittgenstein, L., Hacker, P. M. S., and Schulte, J. (2009). *Philosophical investigations*. John Wiley and Sons.

[Zentall et al., 2008] Zentall, T. R., Wasserman, E. a., Lazareva, O. F., Thompson, R. K. R., and Rattermann, M. J. (2008). Concept Learning in Animals. *Comparative Cognition & Behavior Reviews*, 3:13 –45.

[Ziemke, 1999] Ziemke, T. (1999). Rethinking Grounding Approaches to Grounding. *Cognitive Science*.