Technical Report

UCAM-CL-TR-423
ISSN 1476-2986

Number 423

UNIVERSITY OF CAMBRIDGE

Computer Laboratory

Symbol grounding: Learning categorical and sensorimotor predictions for coordination in autonomous robots

Karl F. MacDorman

May 1997

15 JJ Thomson Avenue
Cambridge CB3 0FD
United Kingdom
phone +44 1223 763500

https://www.cl.cam.ac.uk/

This technical report is based on a dissertation submitted March 1997 by the author for the degree of Doctor of Philosophy to the University of Cambridge, Wolfson College.

Technical reports published by the University of Cambridge Computer Laboratory are freely available via the Internet:

https://www.cl.cam.ac.uk/techreports/

ISSN 1476-2986
DOI *https://doi.org/10.48456/tr-423*

Abstract

To act intelligently, agents must be able to adapt to changing behavioural possibilities. This dissertation proposes a model that enables them to do this. An agent learns sensorimotor predictions from spatiotemporal correlations in sensory projections, motor signals, and physiological variables. Currently elicited predictions constitute its model of the world.

Agents learn predictions for mapping between different sensory modalities. In one example, a robot records sensory projections as points in a multidimensional space. It coordinates hand-eye movements by using closest-point approximations to map between vision and proprioception. Thus, one modality elicits predictions more closely identifiable with another. In a different example, an agent generalizes about a car's sensorimotor relations by weighting sensorimotor variables according to their mutual influence: it learns to navigate without any *a priori* model of the car's dynamics.

With feedback from miscategorization, an agent can develop links between categorical representations and the relevant objects they distinguish. Wavelet analysis provides a neurologically plausible means of accentuating invariance that can subserve categorization. In some experiments, categorical representations, derived from inter-category invariance after wavelet analysis, proved to be efficient and accurate at distinguishing different species of mushrooms.

In a simulation of fish chemoreception, agents learn sensorimotor predictions that uncover salient invariance in their environment. Predictions are formed by quantizing a sensory subspace after each dimension has been weighted according to its impact on physiological variables. As these predictions also map from motor signals to likely changes in sensory projections, the agent can chain backwards from desired outcomes to form plans for their attainment.

Contents

Chapter 1. Symbol Systems and Symbol Grounding

Introduction

This dissertation explores how robots might learn internal symbols for representation that are grounded in sensorimotor activity. It outlines a possible path to robots that learn to interact with their environment directly—that is, without need for a programmer to anticipate what they must respond to and how they must respond to it. The approach is ecological, sensorimotor, and adaptive in the sense that it takes account of the fact that a robot's relation to its environment depends on internal adaptations that are conditioned by its particular body, internal variables, and history.

Symbol systems have proved to be very powerful at simulating essential properties of thought. Unfortunately, there have been serious problems with explaining how extrinsic properties of the world could influence the causal roles of their symbols. This dissertation seeks ways to exploit the strengths of symbol systems while overcoming some of their weaknesses, especially those related to their symbols' lack of grounding.

What is a Symbol System?

Working in such disparate fields as psychology, neuroscience, philosophy of mind, linguistics, and artificial intelligence, proponents of symbolic representation have set themselves a highly ambitious task: the explanation or simulation of thought and its physical manifestations. To this end many of them have exploited a powerful tool and metaphor. It has even been called the only game in town. It emerged in part from synergy between propositional logic and the technological marvel of our day, the digital computer. I am referring here to the symbol system. Most attempts to operationalize thinking share much in common with it. Its symbolic representations are often said to reflect a mental realm of concepts and relations which serve to represent actual objects and events. In stronger versions these representations are expressed formally in what Fodor (1975) calls a *language of thought* or *mentalese* for

short. Harnad (1990a) reconstructs from Fodor (1975), Newell (1980), Pylyshyn (1980) and others the following definition of a symbol system:

> A symbol system is (1) a set of arbitrary *physical tokens* (scratches on paper, holes on a tape, events in a digital computer, etc.) that are (2) manipulated on the basis of *explicit rules* that are (3) likewise physical tokens and *strings* of tokens. The rule-governed symbol-token manipulation is based (4) purely on the *shape* of the symbol tokens (not their "meaning"), i.e. it is purely *syntactic*, and consists of (5) *rulefully combining* and recombining symbol tokens. There are (6) primitive *atomic* symbol tokens and (7) *composite* symbol-token strings. The entire system and all its parts—the atomic tokens, the composite tokens, the syntactic manipulations (both actual and possible) and the rules—are all (8) *semantically interpretable*: The syntax can be *systematically* assigned a meaning (e.g. as standing for objects, as describing states of affairs). (p. 336)

Like Descartes' pineal gland, an intended purpose of a symbol system is to link the mental and the material (Stoutland, 1988). It is meant to offer an intermediate and functional level of explanation between physics and the beliefs, goals, and desires of folk psychology. This level has been dubbed the cognitive level. According to Fodor and Pylyshyn (1988), "any level at which states of the system are taken to encode properties of the world counts as a *cognitive* level; and no other levels do" (p. 9). Analogies have been drawn between symbol systems and computer software because a symbol system's operation is conceived of as being independent of any particular physical realization, be it a digital computer or a brain (e.g. Dennett, 1991). Because under this theory mental states can be implemented in limitless ways, mental categories do not necessarily reduce to brain categories.

A long historical development of ideas lies behind symbol symbols. Key components may be found in the work of Leibniz (1950). Inspired by Hobbes' view of all reasoning being mere calculation, in *De Arte Combinitoria* Leibniz sought to devise a logical calculus on an idea from his schooldays that all complex concepts were but combinations of a few elementary concepts. Researchers have only seriously explored the ramifications of this in the last forty years. To cite one example, Masterman (1961) developed a semantic network for machine translation at Cam-

bridge; she constructed a dictionary of 15,000 concept entries from a mere 100 primitive concept types (e.g. STUFF, THING, FOLK, BE, DO).

Some Useful Properties of Symbol Systems

Underlying the symbol system approach is the premise that cognitive processes perform their function by virtue of their attunement to the constituent structure (i.e. the syntax) of mental representations. In summarizing this position, Fodor and Pylyshyn (1988) state that a typical cognitive process transforms any mental representation satisfying a particular structural description into a mental representation satisfying a different structural description (p. 12). Additionally, symbol systems have the property that we can systematically relate the semantics of a representation to its syntax. The over-all semantic content of a complex representation is determined by its constituent structure plus the semantic content of its syntactic parts. Harnad likewise highlights the utility of symbol systems being systematic: in order that propositions may be assigned a semantic interpretation (Harnad, 1990a, p. 343). According to Fodor and Pylyshyn, constituent structure depends on the parts as well as the whole being semantically evaluable (p. 19).

They further elucidate how two ideas from computation have formed the bedrock of 'classical' theories of cognition. Both require the postulation of a syntactic level distinct from the physical and semantic level:

> The first idea is that it is possible to construct languages in which certain features of the syntactic structures of formulas correspond systematically to certain of their semantic features. Intuitively, the idea is that in such languages the syntax of a formula encodes its meaning; most especially, those aspects of its meaning that determine its role in inference. All the artificial languages that are used for logic have this property and English has it more or less. (p. 28) [The second idea] is that it is possible to devise machines whose function is the transformation of symbols, and whose operations are sensitive to the syntactical structure of the symbols that they operate upon. (p. 30, emphasis removed)

Fodor and Pylyshyn show how these two ideas straightforwardly conjoin in the notion of a symbol system:

> If, in principle, syntactic relations can be made to parallel semantic relations, and if, in principle, you can have a mechanism whose operations on formulas are sensitive to their syntax, then it may be possible to construct a *syntactically* driven machine whose state transitions satisfy *semantical* criteria of coherence. Such a machine would be just what's required for a mechanical model of semantical coherence of thought; correspondingly, the idea that the brain is such a machine *is* the foundational hypothesis of Classical cognitive science. (p. 30)

Fodor and Pylyshyn justify the reliance of classical theories on constituent structure on the grounds that it explains the *systematicity*, *productivity*, and *inferential coherence* of thought (pp. 33-49; Fodor, 1981, pp. 147-149). As they explain,

> systematicity arguments infer the internal structure of mental representations from the patent fact that nobody has a *punctate* intellectual competence (p. 40). [C]ognitive capacities always exhibit certain symmetries so that the ability to entertain a given thought implies the ability to entertain thoughts with semantically related contents (p. 3).

This ability is understood to be *intrinsic* to the system (p. 37). They take the systematicity of thought to follow from that of language:

> [J]ust as you don't find people who can understand the sentence 'John loves the girl' but not the sentence 'the girl loves John,' so too you don't find people who can *think the thought* that John loves the girl but can't think the thought that the girl loves John. Indeed, in the case of verbal organisms the systematicity of thought *follows from* the systematicity of language if you assume—as most psychologists do—that understanding a sentence involves entertaining the thought that it expresses. (p. 39)

After presenting several more illustrations of systematicity, they reiterate, "all the evidence suggests that *punctate* [i.e. unsystematic] *minds can't happen*" (p. 49).

The productivity of thought is based on the assumption that people can entertain an unbounded number of belief states. "Productivity arguments infer the internal structure of mental representations from the presumed fact that nobody has a *finite* intellectual competence" (p. 40). (As was the case with systematicity, this stance mirrors a Chomskyan argument about linguistic competence; it derives from the fact that people can understand and produce an unbounded number of sentences.) To explain the productivity of thought without recourse to constituent structure would

require an unrealistic—indeed, astronomical—number of unique internal categories (or microfeatures, see p. 23).[1]

Fodor and Pylyshyn cite two standard counter-arguments in response to the challenge that behaviour is less systematic and productive than classical theories predict. Since the theories model competence and not performance, they are not required to account for errors or limitations in human performance. Finite performance could result from "the interaction of an unbounded competence with resource constraints" (p. 34), and apparently nondeterministic behaviour could be caused by "the interaction of multiple deterministic sources" (p. 58). The former might explain, for example, why people cannot understand sentences of arbitrary complexity; the latter, why they make grammatical errors.

As for inferential coherence, Fodor and Pylyshyn draw attention to how the syntax of a mental representation connects it with its inferential role:

> The rule of existential generalization applies to formulas in virtue of their syntactic form. But the salient property that's preserved under applications of the rule is semantical: What's claimed for the transformation that the rule performs is that it is *truth* preserving. (p. 28)

They, however, make an important distinction. "The point of Classical modeling isn't to characterize human thought as supremely logical" because syntactic transformations might preserve semantic properties other than truth-value (e.g. "warrant, plausibility, heuristic value, or simply *semantic non-arbitrariness*"). Rather, "it's to show how a family of types of semantically coherent (or knowledge-dependent) reasoning is mechanically possible. Valid inference is the paradigm only in that it is the best understood member of this family." (fn. 19)

The ideas motivating symbol systems may be summarized as follows: Intentional states (thoughts, desires, and the like) are semantically interpretable (because they are *about* things). Many have found it reasonable to assume that, moreover, these states

---

[1] Chalmers (1993) among others have shown Fodor and Pylyshyn's (1988) attacks on connectionism, and exposition of its limitations, to be seriously flawed. Nevertheless, they well summarize the assumptions underlying classical cognitivist approaches.

have causal powers (most importantly, the power to cause behaviour).[2] But, as Turing (1950) made clear, a digital computer's symbols can also be semantically interpretable and causally effective. Their causal powers derive from how the symbols are instantiated in physical materials. In modern-day computers they most often depend on the physical properties of silicon, such as those influencing its electrical conductivity.

The trick is for the computer to manipulate the symbols syntactically, based on how they are realized physically, while maintaining—as much as possible—their semantic coherence. Thus, for example, given symbols that are valid propositions, the computer is to infer other symbols that are also valid propositions. It appears possible at least to set up highly constrained symbol systems that allow symbols to act causally without much (if any) loss of semantic coherence (e.g. automated theorem provers). However, for decades a nexus of problems has beset the simulation of more general kinds of intelligence. Sometimes falling under the rubric *frame problems*[3], they relate to an agent's need to take actions and predict their consequences in an ever changing world. These problems expose the difficulty in maintaining a correspondence between internal symbols and external states of affairs. These difficulties can arise even in deterministic "blocks worlds."[4]

---

[2] For an opposing viewpoint see, for example, Churchland (1986).

[3] McCarthy and Hayes (1969) originally identified the frame problem with the programmer's problem of how to set up an agent's representational form within the situational calculus so that the agent could predict that most facts about the world were unaffected by its action without recourse to frame axioms (also see Hayes, 1987). However, as a number of contributors in Ford and Pylyshyn (1996) have pointed out, the frame problem has come to be associated with more general problems of temporal reasoning. These include (but are not limited to): the prediction problem, the revision problem, the qualification problem, the control problem, Hamlet's problem (see Janlert, chap. 4), the relevance problem, the holism problem, the relevance-holism problem (see Lormand, chap. 6), the persistence problem, the temporal projection problem, and the ramification problem (see Morgenstern, chap. 8). In chapter 7, we examine the frame problem and related problems in greater detail.

[4] In fact, the original frame problem first arose in a completely deterministic environment in which the effect of every action was codified in the situation calculus.

Besides the frame problem, there is a further obstacle in using symbol systems to model intentional states. According to Fodor (1994b) a thought's inferential (i.e. causal) role cannot fix its content because the content depends at least partly on its relation to external states of affairs.[5] If this is right,

> then (some of) the intentional properties of thoughts are essentially *extrinsic*; they essentially involve mind-to-world relations. But we are still following Turing in assuming that the computational role of a mental representation is determined entirely by its *intrinsic* properties (its weight, shape, or electrical conductivity, as it might be). The puzzle is that it is, to put it mildly, hard to see how the extrinsic properties of thoughts could supervene on their intrinsic properties.... But the idea that the implementation of intentional laws *is* computational is the only serious cognitive science we've got; without it, the semantic coherence of the intentional is completely a mystery. (p. 299)

Hume (1975, 1978) had proposed that laws (or 'principles') of association structured thought—in particular, spatiotemporal contiguity, resemblance, and causation. (The latter he claimed was based on contiguity and 'constant conjunction'.) However, Fodor (1994) scoffs at the idea that such laws could attune the causal relations between mental representations to keep them true to the external world. "As everyone has known since Kant—but somehow forgets every thirty years or so—semantically coherent processes are not, in general, associative; and associative processes are not, in general, semantically coherent" (p. 296). Fodor's rejection of associationism, which he criticizes for being insensitive to structure, follows from his conception of thinking—fundamentally, in terms of structure-sensitive operations (Fodor & Pylyshyn, 1988, p. 67).

Nevertheless, a symbol system that generates propositions purely on the basis of properties intrinsic to that system (e.g. a symbol's constituent structure) cannot model how real intentional states come to be connected to the things they represent. We shall turn to this point next.

---

[5] The reasons for this will become increasingly clear in chapters 2 and 3, where we discuss category induction (see, Harnad, 1987; Schyns, Goldstone, and Thibaut, in press). Also see externalist arguments in the philosophy of language.

The Symbol Grounding Problem

To date the most successful models of higher-level intelligence have been symbolic. However, these models typically suffer from a basic limitation pertaining to how symbols are related to the external world. Harnad (1990a) has identified this limitation as the *symbol grounding problem.* Symbol systems typically rely *either* on

(1) a human designer to hardwire connections between internal symbols and external states of affairs, *or* on

(2) a human user to feed in symbols as input and then interpret and act on the symbols the system outputs.

In other words, symbol systems typically require a separate perceptual system—either a human mind or connections explicitly set up by a human mind—to map and maintain a consistent mapping between the symbol system's active representations and the states of affairs they are supposed to denote in the physical world.[6] In either case, as Harnad notes, their intentionality is parasitic on human minds. This, of course, places limits on their capacity to model both natural and artificial intelligence.

In the absence of symbol grounding, Harnad (1990a) likens the activity of a symbol system to someone who does not know Chinese (or any other language) trying to learn Chinese from a Chinese-Chinese dictionary. You want to know what a symbol refers to, so you look it up, only to find more meaningless symbols. You move from symbol to symbol without ever arriving at the actual thing that the symbol represents. Thus, to escape infinite regress, the meaning of some symbols must not be derived from other symbols.

Jakob Fries drew similar conclusions concerning scientific statements and what he called the predilection for proofs. He argued that the call for all statements to be logically justified by other statements leads to infinite regress. He concluded that the *mediate* knowledge represented in the symbols of a language must be justified by the

---

[6] Point *(8)* merely asserts that the symbols are grounded. It is unclear whether this means the symbols are required to be interpretable to the agent of which they are a part or merely to an outside observer (see Dennett, 1987)—presumably both. Devising large-scale representational systems capable of maintaining semantic coherence is difficult because of the symbol grounding problem and the frame problems.

immediate knowledge of perceptual experience. Haugeland (1985) discusses infinite regress in the derivation of symbol meaning in relation to AI programs whose only connection to the world is through textual exchanges with their users. Once again, only in the mind of their user are the symbols of these programs connected to the things they represent. In Plato's Cratylus Socrates also argued that knowledge of things cannot be derived from symbols but must be gained through an investigation of interrelations between things.[7]

As Kant pointed out in his *Critique of Pure Reason*, by itself a formal definition of a dog is not sufficient to recognize a dog when you see one. Schyns, Goldstone, and Thibaut (in press) explain this in more contemporary terms:

> there is a conceptual difficulty with the idea that (innate) theoretical knowledge constrains perceptual information: Going from theories to predict perceptual data is underconstrained. To illustrate, if a categorizer is instructed that a set of objects with an unknown complex structure is a set of hammers, an existing theory of hammer would list the components representing these objects in memory. However, unless the theory also specifies all possible perceptual appearances of these components, a segmentation procedure would still have difficulties locating the actual parts in a new object: The perceptual realization of the parts depends on the new stimulus itself. This problem is analogous to the symbol grounding problem (Harnad, 1990). (§3.3.1)

The flip side of the symbol grounding problem is what Harnad terms the *hermeneutic hall of mirrors* (Harnad, 1990b, 1990c). He notes that, not only are a symbol system's operations—and the resulting output—semantically interpretable (by point eight of the definition), but that their property of systematicity ensures their coherence.[8] As he explains, this means that our interpretation of the system's present and future output will continue to corroborate past interpretations. However, our

---

[7] "How real existence is to be studied or discovered is, I suspect, beyond you and me. We must rest content with the admission that knowledge of things is not to be derived from names. No; they must be studied and investigated in their connection with one another." (Plato, 1953, pp. 104-105)

[8] This is an *in principle* argument, which assumes that the symbol system can deliver as promised. In practice, there is still the frame problem and other problems to overcome.

interpretation plays no role in the workings of the symbol system; its operation is driven solely by the application of formal rules to representations, both of which are intrinsic to the system. We alone fix the meaning of the symbol system's output. But, as Harnad advises, there is always the danger that semantically-coherent symbol manipulation will lull the user into a false sense that the system actually knows what its talking about—in spite of the fact that the connection between its symbols and the world depends only on the user.

Harnad (1993) proposes that the frame problem is another consequence of not letting external properties of the world influence the causal role of symbols. He suggests that the reason ungrounded symbol systems suffer from the frame problem—the reason they suddenly produce a result that utterly clashes with our interpretation—is because they have only formal, syntactic constraints governing their operation. What they lack are bottom-up nonsymbolic constraints. Thus, their behaviour is hopelessly underdetermined. In the coming chapters we shall be looking at ways to build systems that are grounded in these nonsymbolic constraints.

Summary

An essential feature of intentional states is that they are about circumstances in the outside world. The fact that we can recognize this means that they are, at least to some extent, semantically evaluable. Common sense—whether it be right or wrong—tells us that intentional states also have causal powers: we act *because* of what we believe.[9]

It so happens that we can order a computer's symbols in such a way that they are both semantically evaluable and causally effective. This has lead to the development and study of symbol systems. Symbol systems are representational systems that manipulate symbols systematically—that is, according to their constituent structure—while maintaining their semantic coherence. Many cognitive scientists recognize that our mental faculties must (more or less) encompass properties of

---

[9] One example of the prevalence of this assumption is the fact that it underlies our legal system and how it apportions responsibility.

symbol systems in order to explain the systematicity, productivity, and inferential coherence of human thought.

A problem arises because intentional contents are at least partly determined by their relation to the external states of affairs they represent. But if this is the case, how can the exclusively intrinsic properties of a symbol system wholly determine the causal role of intentional contents? Somehow we need to let these extrinsic properties in; we need to let them influence the workings of the symbol system.

The symbol grounding problem is one consequence of not letting them in: the problem of causally connecting a symbol system's internal symbols to the external objects, events, and states of affairs they represent (Harnad, 1990a). For symbol systems to model embodied cognition, the meaning of a symbol cannot be derived solely from the meanings of other symbols. To let a human user's interpretation of the symbols ground them cannot explain extrinsic aspects of intentional contents because the user is standing in for the very aspects that we seek to explain. For an engineer simply to hardwire connections between internal symbols and external objects is equally unsatisfactory because of the difficulty of anticipating what new symbol categories a system might need *a priori*.

Harnad (1993) proposes that *ungrounded* symbol systems suffer from the frame problem because their operations are underconstrained: they only have formal, syntactic constraints but are completely lacking in nonsymbolic, bottom-up constraints (see chapter 7). In chapter 3 we examine his candidate solution to the symbol grounding problem: that an empirical process of category induction grounds (at least some) symbols from the bottom up (1990a, 1987).

Overview

Chapter 1. Symbol systems have proved useful for characterizing the systematicity, productivity, and inferential coherence of intentional contents. But as presently conceived they fail to explain how extrinsic properties could bear on their symbols. Their causal role is determined purely in terms of properties intrinsic to the system. Harnad (1990a) has identified the symbol grounding problem with the problem of causally connecting the system's symbols with the external objects, events, and states of affairs they are supposed to represent.

Chapter 2. Theories which propose to ground all symbols by means of the *a priori* feature detectors of an encapsulated perceptual module only beg questions: How did the detectors evolve? How do they detect new kinds of invariance? How can encapsulated sensorimotor constraints influence reasoning? Encapsulation and *a priori* feature detectors pose problems not only for understanding intentional states but also for developing clever robots. As behavioural possibilities depend on variable bodily and environmental relationships, a robot needs flexible representations that can adapt to unanticipated change.

Chapter 3. Harnad (1987, 1990a) proposes a theory of category induction: an agent empirically develops categorical representations that are causally linked with the sensory projections of the objects they distinguish. These representations exploit the invariant features of objects with the aid of feedback from miscategorization. Wavelet analysis provides one neurologically plausible way to accentuate invariance prior to categorization. Daugman (1980) has used 2-D Gabor wavelets to account for cortical cells' sensitivity to scale, localization, orientation, and quadrature phase relationships. In experiments I found that categorical representations could efficiently and accurately distinguish different species of mushrooms. The representations were derived from inter-category invariance in the input after it had been decomposed using wavelets.

Chapter 4. Intelligent robots could learn predictions from spatiotemporal correlations in sensory projections, motor signals, and internal variables in order to detect and exploit changing behavioural possibilities. Memory-based models, which learn sensorimotor relationships, have proved to be more flexible, robust, and biologically plausible than conventional models for robotic control because they are capable of adapting to bodily and environmental change without human intervention. Clocksin and Moore (1989) have shown that a robot can learn to map from stereoscopic vision to proprioception. Their approach can be made to conform with a broadly defined theory of representation based on neurophysiological predictions (Sommerhoff & MacDorman, 1994, §5) and it can be used to integrate sensory information from a wide range of sources through the mutual elicitation, maintenance, and revision of predictions.

Chapter 5. In a simulation, a robot builds up a mapping from a remote-controlled car's current perceived state and motor signals to its perceived state in the next time step. Based on limited experimentation, it extract correlations between different

sensorimotor variables and uses this information to predict its dynamics in novel states. Thus, its empirically developed model allows it to plan optimal paths through unvisited areas of its environment.

Chapter 6. A fish simulation is used to outline an empirical, bottom-up approach to grounding cognitive categories. The agent stores in a coordinate space information about its motor signals, current sensory projections, and predicted future values of internal variables. Through a process of quantization and progressive refinement, it forms categorical representations that can detect relevant sensorimotor invariance. Instantiated representations reflect the presence of particular objects and agents in the environment. The fish exploits these representations to discover and home in on food and avoid dangers.

Chapter 7. We propose three ways of moving the fish simulation towards the behavioural capacities of a symbol system without sacrificing its bottom-up grounding: *(1)* learning to map from (learned) external categories and motor signals to consequent external categories; *(2)* enhancing its mapping by removing incidental categories; and *(3)* learning to exploit constituent structure. Symbol systems usually have only formal, syntactic constraints. Harnad (1993) points out that, although these systems are underdetermined, their semantic interpretability can lead us to overinterpret them. Evidence revealing this is typically identified with the frame problem. Learned bottom-up constraints could help a robot's representations better capture the stabilities of the world by limiting the formally admissible to the empirically admissible.

Appendix A. The point of the appendix is not to propose a reinforcement-based alternative to symbol systems, since there is not enough evidence to posit such an alternative. It is rather to show that the reinforcement learning models that have been mathematically explored in AI tend to suffer from their own problems of grounding states and actions. In these systems it is normally the robot's designer who must determine how raw sensory data is mapped onto the set of possible perceived states. Additionally, the choice of action is usually artificially constrained. Appendix A proposes a method of state and action generalization based on an extension of Q-learning. It permits both states and actions to be empirically grounded in sensorimotor projections.

Chapter 2. A Review of Some Past Approaches

Introduction

In this chapter we will examine Fodor's theory of perceptual analysis. The intended purpose of his theory is to show how representations can be causally linked with the objects they represent. It is typical of many theories that are based on a fixed set of *a priori* features—for example, Biederman's (1987) Recognition-By-Components or Schank's (1972) Conceptual Dependency. These theories categorize and represent objects (events, relations, etc.) by means of nondecomposable atomic elements. As noted in the last chapter, a finite set of atomic elements and rules for their combination can be used to generate an unbounded number of representations. Moreover, they can be used to represent systematic relations among objects (see Fodor & Pylyshyn, 1988). However, fixed features come at a price:

> Fixed feature theories limit new representations to new combinations of the fixed features. Consequently, all possible categorizations are bounded by the possible combinations of the features. If a categorization requires a feature not presented in, or derivable from the feature set then the categorization cannot be learned. This is a rather limiting view of representational change. There may be occasions when features not originally present in the system are useful for distinguishing between important categories in the world that newly confront the organism. (Schyns et al., in press, §1.1)

By employing a fixed set of *a priori* features, it is possible to seal off a symbol system from the details of sensorimotor activity. This can simplify such traditionally symbolic tasks as abstract planning. Unfortunately, as we shall see, it also seals off the symbol system from the very bottom-up constraints that could help to improve planning and to circumvent the symbol grounding and frame problems.

Fodor's Perceptual Analysis

Although Fodor criticizes Hume's associationism, his language of thought hypothesis has certain points of compatibility with it. Fodor (1975) allows for the composition of complex representations from atomic symbols as the result of experience. This is analogous to how Hume conceived of complex ideas developing from simple ones. As Hume first put forth in his *Treatise of Human Nature* (1739), "The idea of a substance... is nothing but a collection of simple ideas, that are united by the imagination, and have a particular name assigned to them, by which we are able to recall, either to ourselves or others, that collection" (1978, p. 16). In his *Enquiries* Hume asserts that we can conceive of a virtuous horse because we can conceive of virtue and unite that with the familiar figure of a horse (1975, p. 19). This resembles Fodor's suggestion that the concept *airplane* may be composed of the concepts *flying* and *machine* on the basis of experience (p. 96) or Harnad's (1990a) suggestion that *zebra* may be composed of *horse* and *stripes*. It is worth noting that the system's capacity for representing airplane is intrinsic to the system; it only requires the combination of existing terms. The combination is, nevertheless, provoked by experience.

Where Harnad and Fodor differ is on the question of how symbols are to be grounded. Whereas Harnad favours an empirical approach whereby objects are rendered identifiable by associating their invariant features, Fodor favours a nativist version of mentalese whereby they trigger concepts that are already latent in the mind. This controversy, long predating mentalese, has raged in philosophy since the time of Locke and was later to infect psychology.

Both Fodor and Maze (1991) are critical of mentalese empiricism. They consider it to beg the question: if it is possible for perception to organize symbols into hypotheses like *the robin is on the lawn* by association, "where do 'robin' and 'lawn' and, for that matter, 'being on', come from? It must be obvious that the acquisition of the background knowledge presents just the same difficulties as the interpretation of current sensory information." (Maze, 1991, p. 173) As Fodor (1975) points out, "If, in short, there are elementary concepts in terms of which all the others can be specified, then only the former needs to be assumed to be unlearned" (p. 96).

Fodor proposes a nativist solution to the problem of grounding the concepts expressed in a language of thought. He posits a passive mechanism of perceptual

analysis that derives symbolic representations of real objects from sensory data (the raw unstructured readings of physical parameters). *Demons,* each sensitive to a single physical property, shriek *yes* or *no* depending on whether a hypothesis is present or absent in the environment. These demons activate innate elementary concepts which, once properly combined, are used to reason formally about the world.[10]

Part of the controversy surrounding perceptual analysis has resulted from its apparent support for the existence of a 'grandmother cell' (a neurone which fires whenever you recognize your grandmother). (This may be a misconception.) However, similar architectures have won supporters in the neurosciences. Marr (1982) proposed a theory of stereoscopic vision based on *feature detectors* analogous to Fodor's shrieking demons.[11] The results of experiments on feature recognition in monkeys (e.g., Fujita et al., 1992) have suggested to some that visual memories may be written in an alphabet of iconic figures (Stryker, 1992) not unlike Leibniz's "alphabet of human thoughts" (1951, p. 20). This interpretation contrasts with the view that recognition and recall are evoked by highly distributed brain activity (Rumelhart & Norman, 1981; Rumelhart & McClelland, 1986; Smolensky, 1988).[12] This distributed account has lead others to conclude that it may be impossible to translate between thoughts and brain activity, for example, by means of an intervening

---

[10] In characterizing perception in terms of a cacophony of demons, Fodor drew inspiration from Selfridge's (1959) demon-based pandemonium model. However, it is doubtful whether Selfridge himself would have approved of the use Fodor makes of demons. Fodor implies that the causal connection between external states of affairs and the activation of demons is set up *a priori.* Selfridge, by contrast, proposed a kind of learning: evolving a population of similarly constructed demons by means of natural selection (p. 523). Clearly, Selfridge did not conceive of perceptual categories as being *a priori.*

[11] Starting from the bottom up, he first proposed algorithms for extracting edge segments, blobs, boundaries, and orientations from a static scene and then for discriminating groups of these primitives according to their size, orientation, and spatial arrangement. These computations yielded three levels of representation: the primal, 2 $^1/_2$-D, and 3-D sketch.

[12] The difference between the notion of a grandmother cell and distributed connectionist representation may only be a matter of degree. It is quantified in Valiant's (1994) mathematical theory of representation and retrieval in the brain.

cognitive level such as a language of thought (Stern, 1991). But it is unlikely that Fodor and Pylyshyn ever saw this as the purpose of mentalese, as they argued for the autonomy of psychology and the irreducibility of intentional states.

By proposing an *a priori* method of transducing elementary concepts from sensory projections, Fodor's nativism, exemplified by Marr, does not constitute a radical departure from the empiricist tradition. Although Hume (1975), its quintessential figure, believed that our ideas were not innate, he claimed that our impressions were (p. 22).[13] And as our simple ideas, according to Hume, are just copied from and caused by these impressions, it takes only a small step to conclude, as Fodor does, that simple ideas are already latent in the mind.

While Maze and Fodor agree that symbols cannot ultimately be grounded empirically, Maze also attacks Fodor's nativist view:

> By elementary concepts he seems to mean natural kinds, which, he points out, cannot be subjected to 'definitial elimination' without loss of meaning. Thus, the innate language must be provided with terms for every natural kind which we can potentially identify, which would include, just for a start, every one of the millions of species of animals, birds, fish, insects, vegetable life and so on with which the world is stocked. (1991, p. 173)

Interpreted in this way, Fodor's nativism implies that we are hardwired to recognize specific things, like Antarctic penguins and DNA molecules, that our ancestors have never before seen. This would require a miraculous feat of evolution.[14] (However, Fodor may not adhere to such an extreme position as the one Maze extrapolates from his reading of him.)

Maze concludes that the mentalese theory of the mind must be false because

---

[13] Hume argued that ideas were not innate because the blind cannot form any idea of colours nor the deaf of sounds as long as neither had experienced the corresponding impressions (1975, p. 20). Hume's term *impression* includes affect.

[14] The anthropic principle has been invoked to justify it. The argument, imported from physics, is that if the universe were not the way it is, we would not be here to witness it. It is not a denial of evolution, but it does permit highly improbable evolutionary occurrences. In physics it is used only a last resort.

(1) its symbols would be solipsist if ungrounded and

(2) its symbols cannot be grounded, neither empirically nor innately.

Maze's first point may be arguably true. In places Fodor (1980) indeed appears to maintain that cognition is solipsist and that this is unfortunate but must be accepted (in his words, "it's tough but true"). Maze's second point remains an open question—one that can be answered empirically. In the robotic case, a preliminary chemoreceptive simulation in chapter 6 suggests that representations *can* be grounded on the basis of empirical and evolved adaptation. This is further supported by recent advances in many areas of pattern recognition (e.g. optical character and face recognition).

Nevertheless, we must grant Fodor that, since any category resulting from real-time learning could have existed already, *in principle* robots and even humans and other animals could be grounded solely by means of innate elementary categories and combinations thereof. Nevertheless, it seems unlikely that all or most of our categories are innate (see chapter 3), and it is especially unlikely that a robot engineer could second guess how to set up the right causal connections to ground a robot's categories *a priori.*

The dichotomy we have set up in this section between nativism and empiricism (i.e. arguments supporting category learning) is perhaps dated and simplistic. It has become increasingly clear that most behaviour demands some form of learning (see chapter 4), which "is often innately guided, that is, guided by information inherent in the genetic makeup of the animal. In other words, the process of learning itself is often controlled by instinct" (Gould & Marler, 1987, p. 62) Category learning must, to some extent, be driven and biased by evolved biological constraints.

The Constraints of Modularity in Robotics

Traditional academic robotics is compatible with mentalese nativism, although few in the field would exhort this philosophical stance. The robot Shakey designed at SRI is

a good (though hackneyed) example of this design methodology (Nilsson, 1984).[15] Its programmers furnished it with a stock of symbols and operator rules. The robot used the symbols to compose propositions. These propositions represented states of affairs such as relations between nearby objects. The robot scanned its environment in a perceptual stage to determine which propositions to include in its internal description of its surroundings. It might represent the presence of a box at a doorway, for example, as *at (box, doorway)*. Solving a problem involved finding a chain of operator rules whose application would transform the propositions in the robot's current state to those in its goal state. Each rule (e.g., for moving a box) had certain preconditions (like the robot being at the box) and resulted in certain additions and deletions to the robot's world description (the location of both the robot and the box having changed). The chain of operator rules served as an action plan for the robot to carry out. This sensorimotor process involved matching internal symbols against external objects so that the robot could locate and move the objects.

The trouble with this approach is that it demanded that Shakey be stocked with symbols for every elementary object and relation it could possibly be required to handle.[16] Otherwise, whenever it needed a new symbol to represent a new elementary concept, its programmer would have to add it. So long as Shakey remained within the confines of the simple environment of boxes and platforms expressly set up for it, the robot's symbols appeared to be grounded. If removed from that environment, it could not function at all. This may be owing to the fact that Shakey had to rely on its programmer to set up a causal relation between its internal symbols and the objects it detected.

It is unlikely that the adaptability of robots with this kind of architecture could approach that of most vertebrates, let alone human beings, because the programmer may not be able to set up the appropriate causal relation or anticipate the robot's

---

[15]Shakey's planning system is based on STRIPS (see Fikes & Nilsson, 1971). Brooks (1991a, b) has been one of the strongest critics of the application of symbol systems to robotics.

[16] Although more recent systems can make deductions on the basis of derived relations, they still do not learn new *elementary* relations and categories from their sensorimotor activity.

future need for an elementary category. An exception to the latter would be if we let the programmer tinker with the robot as its environment changes. But to do as much would be to include the programmer in our overall conception of what constitutes the robot. This would be unacceptable in any robotic system that is intended to explain how the mind works. This is because it introduces a homunculus, the programmer, who entirely duplicates the talents of the mind the robot's workings were meant to help explain (Dennett, 1979, p. 123). What is needed is a different sort of explanation of how an agent may become sensitized to objects (for discussion, see Smith, 1995).

Today it may seem unfair to criticize Shakey, a robot developed almost thirty years ago. However, up-to-date robotics systems like SOMASS (Malcolm, 1995), designed by researchers who write about symbol grounding, try to make a virtue of Shakey's design limitations. In SOMASS the intention was to make the clearest possible separation between the classical symbolic planning module and the behaviour-based plan execution module so that the implementation-dependent details of sensorimotor coordination would not complicate symbolic planning or limit it to a particular problem domain. In the planning module, a particular instance of an object is represented solely by types and combinations of types. In a robot with sensing equipment, presumably these types would be instantiated by feature detectors. Since types by their very nature abstract away analogue and instance-dependent detail, information crucial to sensorimotor coordination—such as information about an object's contours—is lost. This information is important for manipulating objects of varying shapes and sizes. In SOMASS, its handling is kept away from planning down in the plan execution module.

In the spirit of Fodor's *The Modularity of Mind* (1983), SOMASS tries to encapsulate planning from sensing and motor control by placing it in a separate module. The system's innovation is exhibited by how it handles the complex task of interfacing the planning and behaviour-based module. The benefit of such a scheme over earlier systems, which lack encapsulated modules, is clear: it requires much less work for the programmer to adapt the robot to a new problem domain or to a new suite of sensors and actuators. In earlier systems it often proved necessary to reprogram the system from scratch.

Unfortunately, the SOMASS approach is unsatisfactory from the standpoint of grounding internal representations adaptively. This is because its modularity

artificially divides planning and sensorimotor activity and makes untenable assumptions concerning their relationship. Before considering why, let us first recapitulate Fodor's (1981, 1983) apparent line of argument. According to Fodor, a symbol system is necessary to explain the productivity and systematicity of thought and language. But sensory projections do not represent the world in a form that a symbol system can process (e.g. amodal types). Thus, Fodor concludes, it is the function of perceptual analysis to provide the symbol system with representations in a form that it can handle. This form is not analogue but comprised of well-formed strings of symbols. He claims that perceptual analysis accomplishes this tokenization by filtering out *irrelevant* variability in the analogue sensory input.

In the context of SOMASS, the case against this argument is that it is improbable that the programmer can specify *in advance* precisely what kinds of variation a robot will *never* need to consider in the planning module in order for the behaviour-based module to execute, in the face of changing circumstances, intelligent movements. (This also holds if we replace the programmer with a learning or evolutionary process.) There is nothing wrong with SOMASS as a practical system set up by a programmer to function in a simple and predictable environment. But as a model of real intentional systems, it fails on the symbol grounding issue.

Encapsulation may prevent a robot from learning to detect new elementary categories and from adding new elementary symbols to its planning module. Why is this so? The purpose of having the robot learn new symbols is so that it can make plans about currently relevant—but as yet unrepresentable—features of the world. But these features depend on, and must be extracted from, analogue sensory projections and motor signals. This implies that the formation of new elementary symbols depends on an interplay between symbolic planning and (analogue) sensorimotor information. In SOMASS this information is available *only* to the behaviour-based module. Its strict division between planning and sensorimotor control places a heavy burden on the interface between them (which was not designed to be adaptive) because it must translate between sensorimotor transformations and changes in propositional descriptions of the environment. An alternative is for the robot to use

representations of sensorimotor transformations as the basis for category formation and planning about the world (see chapters 6 and 7).[17]

According to Fodor (1987), the same elementary concepts and logical syntax that can be used to define 'kosher' concepts can also be used to define any number of 'kooky' concepts (pp. 145-146). This excess of freedom and lack of stability in logical formalisms is a major cause of the frame problem (see Janlert, 1996). But (*pace* Fodor) the rules we posit for keeping kooky concepts out of our representations should be none other than the bottom-up perceptual and empirical constraints that SOMASS's planner lacks (see Harnad, 1993, Schyns et al., in press, and chapter 7).

In sum: if our intent is to develop a system that can ground its own symbols, there are at least five reasons why *not* to place an artificial rift between planning and sensorimotor coordination:

(1) It is difficult to devise an interface that can translate between analogue and propositional forms of representation without it being dependent on *a priori* symbol categories. However, if the system cannot learn new elementary categories, there may be vital sensorimotor invariance that it can neither recognize nor learn to recognize.

(2) Often an object's analogue features must enter into abstract planning as well as movement execution. These features influence not only how, for example, the robot would need to angle an object to get it through a narrow passageway (i.e. plan execution) but also whether the passageway were wide enough (i.e. planning itself: is the plan even possible?). The planner may not be able to take advantage of some opportunities because this would require it to have access to unrepresented analogue features of the environment.

(3) While, within a sufficient degree of accuracy, it may be possible to represent analogue information propositionally, this may carry prohibi-

---

[17] Similarly, Glenberg (1997) proposes that, to model its world, an agent uses analogue trajectories that develop in memory from meshed sensory projections.

tively high computational demands (see Janlert and Pylyshyn in Ford & Pylyshyn, 1996).

(4) According to Janlert (1996), to solve the frame problem one must find a representational form that preempts the need for reasoning about stabilities. Janlert believes that what the frame problem is *not* about is finding a better algorithm—one that avoids unnecessary computation (pp. 43-44). By then it is too late: you have already let empirically inadmissible concepts seep into your representational form. SOMASS's symbolic planner founders on this point because its form does not exclude propositions that should be inadmissible because of bottom-up constraints (see Harnad, 1993).

(5) It seems plausible that analogue features of particular past episodes might influence thinking without their having been typecast in advance. Higher-level thought may depend not only on the ability to draw on ever more abstract categories but also on the ability to draw on particular cases that appear superficially remote but bear an abstract relation to the matter at hand.

A main motivation behind SOMASS was to save time and money. It eliminates the need to hire someone to rewrite the robot's program from scratch for each new problem domain. A more adaptive system, however, might be able to relearn what had previously required reprogramming. Nevertheless, a learning system would still face the problem of properly integrating knowledge of a new domain with previously acquired expertise. Maintaining stability in planning across different environments remains an important issue, and there may be lessons to be learned from SOMASS—even for systems that are grounded from the bottom up in learned sensorimotor categories.

Learning Behavioural Possibilities: A Robot's World-modelling

Behavioural possibilities vary according to individual differences. Ethology provides a groundwork for expressing this (e.g. Hinde, 1987). Baron von Üxküll (1934) discussed in several publications differences among species in how they might relate to the same physical environment. It has become clearer that, although other species

may inhabit the same world that we do, they fill very different niches. An external wall that functions as a barrier for humans may be a bird's landing spot and nesting place. In psychology Gibson (1979) proposed, shortly before his death, a theory of *affordances*. It related perception to the varying behavioural possibilities that the physical world affords each species. This view, in one form or another, may be traced back at least to Tolman (1932). According to Tolman, living organisms can only know objects as potential *behavior-supports*. Tolman was a behaviourist (though far less simplistic in outlook than many of his contemporaries who spoke only in terms of reflexes). He believed it to be the psychologists duty to describe objects in terms of the specific sensorimotor adjustments they *evoke* in an organism. He contrasted psychology with physics because physicists attempt to distil from their description of an object all the particular sensorimotor conditions that could occur when an organism is faced with that object. They aim to describe objects in terms of abstract characteristics, namely, those that remain constant regardless of whether the subject is human or rodent, hungry or lustful, finned or winged. This is, of course, something psychologists can never do.

> Even physics' account of the external world is, in the last analysis, an ultimately, though very abstracted, behavioral account. For all knowledge of the universe is always strained through the behavior-needs and the behavior-possibilities of the particular organisms who are gathering that knowledge... But what outside reality may be, in and for itself, abstracted from all human behavioral needs and all human behavioral capacities, we do not, cannot, and need not know. (Tolman, 1932, pp. 430-431)

It may be possible to extend Tolman and Gibson's insights to the social domain. A number of theorists have proposed that primate intelligence has its origin in the selective demands of social interaction (Jolly, 1966; Humphrey, 1976; Cheney & Seyfarth, 1990). In a social setting, relationships and individual differences add a new level of complexity to the task of recognizing and acting on behavioural possibilities. Here we are dealing not so much with the properties of matter as the potentialities of other minds and, in particular, the ability of individuals to develop unique and productive relationships. Primates (and other vertebrates) need to develop predictions not only about the consequences of their motor activity relative to the physical world but also about their closely-coordinated interactions with particular individuals, interactions that have repercussions that may be interpreted at many levels of

abstraction. One may note, for example, that a baby girl's mother does not afford her milk in the same way that a blanket affords her warmth. Mothers have minds of their own, and the girl's own mother may sometimes afford her a whack. Nevertheless, we need not assume that interactive categories, such as those involved in guessing another person's preferences, involve category induction that is different in kind from that required for nonsocial categories.

A robot's perceptual system reflects different commitments from those of its designer's perceptual system. This is because robots do not have human bodies and human experience. Their sensors and actuators are very different from our own. What is salient to a robot may not be what is salient to a human no matter how hard we try to make it so.

This realization may sound obvious but in the AI community it has only come after years of trial and error. The point of early robotics projects was to get robots to recognize and label what are *to us* objects. More precisely, they were meant to identify, on the basis of the proximal sensory projections of distal objects, phenomena denoted by English lexemes. They would then reason about these objects and manipulate them accordingly.

Most researcher assumed that whole objects and complex features were defined in terms of a fixed set of *a priori* features.[18] Robots could then be programmed to detect those features and to use this information to identify whole objects. This decomposition of the problem mirrored decompositions found in neurophysiological models. Early AI work appeared to confirm it. This was largely because of allowances that were made in the design process. Researchers carefully crafted objects that could be

---

[18] This assumption is not limited to AI. As Schyns et al. (in press) note, "many recent approaches to categorization have continued to use stimuli that 'wear their features on their sleeves.' Clear-cut dimensions with distinct values are often used for reasons of experimental hygiene" (§1.1). In psychology these facts have created an experimental bias in support of fixed feature sets. Schyns and his colleagues argue that functionally-determined constraints on features "should be defined by the environment and not simply by the experimenter" (§2.1). "Experimental materials are more likely to promote feature creation when they are not designed with *a priori* diagnostic features, leading to obvious feature decompositions" (§4).

easily recognized by the robots that were, in turn, designed to recognize them. Thus, the correspondence between object and representation was built into the relationship between robot and environment. This is possible when both world and cognizer are the product of the same designer. Unfortunately, this made it easier for robot experiments to corroborate assumptions about fixed feature sets; their design prevented them from really putting fixed feature assumptions to the test.

Systems conceived along these lines simply did not scale up, and even their designers have temporarily put aside the hope that their robots could produce and act on a complete description of their surroundings (see Brooks, 1991b). For example, many systems now use deictic representations (e.g. *the box in view*) instead of trying to keep track of information across time about which object is which (e.g. *box #7* is at location *xyz*).[19] Some robots are able to respond to phenomena without labelling them—or at least when they do use labels they are neither designed nor intended to cover human-imposed classes of things. Nevertheless, within limits we can at least interpret the behaviour of these robots as being directed and purposeful.

Now that we have the necessary hardware to embody and test our theories about the mind, pressure has mounted to reconcile our picture of both mind and brain. Much is currently being written on the subject of hybrid systems. Indeed, many hybrid systems are under development. They graft together low-level neurally-inspired models for sensing and motor control and higher-level symbolic models of reasoning about the world. Some of these hybrid systems will fail to live up to expectations because like SOMASS they ignore the symbol grounding problem. The resulting problems will expose weaknesses in the separate and distinct theories that inspired the system's parts and, in so doing, bring the symbol grounding problem to the fore.

---

[19] For example, instead of having a moth-eating bat just keep track of the one or two moths that it can most easily catch, a traditional AI approach might have it trying to identify every moth that came within its range (e.g. as the fifth moth among eight). If it failed to identify a moth, it would issue it a name (e.g. *moth #9*) before pursuing it. (Given sonar's limited resolution, in addition to being pointless, such a design probably would not work.)

With our current understanding, probably the fastest way to get robots to perform certain practical tasks is to place hard constraints on their perceptual systems—for example, by explicitly programming them to recognize the things we need them to. However, if our aim is to develop robots that behave intelligently, philosophical insights as well as experience in academic robotics suggest that this approach is not enough. To avoid this we need first to develop a foundational understanding of embodied intelligence beneath the engineering problem. Hence, we should not focus only on short-term practical benefits. This means developing robots that work within more flexible constraints. Certain constraints will always be necessary because perceptual clusterings of salience do not simply pop out of the physical structure of an individual's sensory projections. Some form of feedback is necessary—either from natural selection, internal variables, or both—concerning the sensory data's relevance to the individual's needs and potential behaviour. The robot should have the wherewithal to learn to recognize new objects, to discover what patterns in its flux of sensorimotor input are salient. The patterns it settles towards will be *its* objects of perception as influenced by, among other things, its body, goals, and past.

If Martian scientists wished to understand how human cognitive processes work, they might start by trying to surmise what kinds of things people respond to (see Quine, 1960, 55ff; Gordon, 1992). However, if these Martians, having evolved under different conditions, were sensitized to different perceptual categories (perhaps they only see infrared light), they probably would have trouble guessing what it was that a human individual saw. Even if they could thoroughly probe a person's brain, they still would not know everything that person saw because they would not be able to see it themselves. One cannot step outside the limits of one's own perceptual system, although one may succeed in extending the limits of that system (science and technology have been of enormous benefit here). Now when the Martians turn to theorizing about humans, just like us they would be able to name the things they had names for, and with time no doubt they would invent more names. Nevertheless, some aspects of our world would fall through the gaps. No matter who is doing the theorizing, something is always lost.[20]

---

[20] This is not to deny that, in a sense, something should be lost. Theories only approximate aspects of the world that are relevant to our interests, and this is all they are meant

The limitations faced by these hypothetical Martians would not be unlike those faced by human scientists when trying to understand the workings of the adaptive systems they set up. If the wrong approach is taken in the development of robots, the limitations of the cognitive systems of robot designers could pose a more serious threat to robot adaptability than any hardware constraint. Perhaps one reason natural selection has been able to evolve marvellously complex and adaptive creatures is because it does not and cannot impose such limitations.

Summary

So far symbol systems have offered the best hope for modelling the systematicity and productivity of thought and language. However, for them to model embodied cognition or to serve as a basis for robotic activity, some of their symbols must be causally connected to the outside world. The analogue nature of sensory projections prevents symbol systems from processing them directly. As a consequence, some theorists posit a separate perceptual module that works beside the symbol system. Its purpose is to present the symbol system with representations of the world that it can process and, thus, to maintain consistency between the system's internal symbols and the world they represent. The perceptual module is supposed to do this by filtering out irrelevant variations in the sensory input—for example, by using a fixed set of feature detectors.

This approach has the benefit of encapsulating the symbol system from the messy details of sensorimotor coordination. This can make for simpler, cleaner planners that robot engineers can more readily adapt to new sensors and actuators or to new problem domains. Unfortunately, in realistic environments it may not be possible to specify in advance what kinds of sensorimotor invariance will be relevant to planning. *Encapsulation and fixed features* pose the following problems for a symbol system:

---

to do. As Tolman pointed out, if a theory reproduced the complexity of the world, it would lose its usefulness as a map "for finding one's way about from one moment of reality to the next" (Tolman, p. 425).

- They only delay explanation of how extrinsic relations help to determine the causal role of intentional states. After all, how did the fixed set of feature detectors originate and why is it now fixed?
- They may hinder a robot in learning new features and planning in terms of them.
- They prevent the analogue features of an object from playing a sometimes necessary role in the process of reasoning about the world.
- They release the symbol system's representational form from bottom-up constraints. Unfortunately, these are precisely the constraints that could have limited the form to the expression of empirically admissible concepts. Thus, they undermine our best hope for overcoming the frame problem (see Harnad, 1993; Fodor, 1987; Janlert, 1996; and chapter 7).

For agents to behave intelligently, they must be able to recognize—or learn to recognize—the changing behavioural possibilities the world affords them. These possibilities vary between individuals according to differences in body and ability and also according to the relationships that have developed among them. As we shall see in the next two chapters, learning appears to play an important role in how an agent categorizes its environment and in developing sensorimotor predictions that are responsive to bodily and environmental change.

Before technology was available to test cognitive theories in embodied systems, separate theories were developed to explain high-level thinking and low-level perception. The assumption was that they could one day function together as modules in hybrid systems. This is probably too optimistic. But the attempt to integrate these different theories will have the beneficial effect of forcing us to face the symbol grounding problem and, as a result, to rethink our theories of cognition.

Chapter 3. Category Induction for Symbol Grounding

Introduction

Psychologists have begun to uncover experimental evidence which challenges theories based on a fixed repertoire of *a priori* features. Instead, Schyns and his colleagues propose that high-level cognitive processes acting under perceptual constraints help to guide the creation of new features—for example, by means of corrective feedback from miscategorization (see Schyns et al., in press; Wisniewski & Medin, 1994; Harnad, 1987). They found that subjects tended to decompose objects into those parts that were diagnostic in categorization. With experience we learn to processes stimulus categories and dimensions separately that were originally processed together. This difference is most obvious when comparing children with adults (Smith & Kemler, 1978; Ward, 1983). Though young children may at first categorize all round object as *balls*, gradually they learn to narrow their lexical categories (Chapman, Leonard & Mervis, 1986; Clark, 1973). This evidence supports the view that our categorizations influence which features are used in representing objects.

Schyns and his colleagues (in press) explain the advantages of developing features in object concepts (§2.7):

(1) The ability to learn features that distinguish objects makes for features that are both flexible and constrained.

(2) Evidence for a proposed set of hardwired features could be explained by an equivalent set of learned features.

(3) A fixed set of *a priori* features must anticipate many potential but as yet unused categorizations. Since every learned feature must have been formed to subserve at least one categorization, a learned set should have far fewer extraneous features.

(4) Learning permits features to be tailored to the categorizations they are used to make. This lessens the need to have complex rules for making categorizations with a less specialized fixed set.

(5) Feedback from miscategorizations can cause features to be decomposed into subfeatures capable of subserving the correct categorization.

In this chapter we shall begin to explore category induction's potential to aid in the development of symbol systems that are grounded from the bottom up.

Harnad's Theory of Category Induction

The connectionism versus symbolic representation debate has seen a history of rival claims made about the adequacy of each methodology in modelling mind, brain, and behaviour.[21] Harnad (1990a) suggests, however, that a successful theory may be required to capitalize on the strengths of both. He proposes that neural networks or other statistical learning mechanisms might be able to form basic categorical representations from invariant features in the environment. The categorical representations could develop causal links with sensory projections through "the acquired internal changes that result from a history of behavioral interactions" with the distal objects they represent (p. 343). These representations could serve as grounded elementary symbols, and a symbol system could be built out of them from the bottom up.

Specifically, learning mechanisms would create iconic and categorical representations. *Iconic representations* (IRs) are analogue copies of "the proximal sensory projections of distal objects and events" (Harnad, 1990a, p. 335). They result from "an analog transformation," retaining much of "the spatiotemporal structure (i.e., physical 'shape') of the input or proximal stimulus" (1987, p. 552). "In the case of horses (and vision), they would be analogs of the many shapes that horses cast on our retinas." They allow us to discriminate between horses by "superimposing icons and registering their degree of disparity" (1990a, p. 342). Harnad (1987) notes that

> analog representations are unbounded in the sense that nothing reliably links them to a shared category except whatever natural similarities and differences they may have. But apart from such "ecological" boundaries, iconic representations would blend continuously into one another, sharing

---

[21] On the connectionist side see Smolensky (1988) and Dreyfus and Dreyfus (1988). On the symbolist side see Pinker and Prince (1988), Minsky and Papert (1988), and Fodor and Pylyshyn (1988).

> the same analog representational substrate to the degree that they shared overall physical similarities of configuration or shape. (p. 551)

Categorical representations (CRs) are "learned and innate feature detectors that pick out the invariant features of objects and event categories from their sensory projections" (p. 335). "They are icons that have been selectively filtered to preserve only some of the features of the shape of the sensory projection: those that reliably distinguish members from nonmembers of a category." They allow us, for example, to identify a horse as a horse and not "a mule or a donkey (or a giraffe, or a stone)" (p. 342). As Harnad explains,

> successful categorization depends on finding the critical features on the basis of which reliable, correct performance can occur. These will depend, not on the inherent "features" of any particular instance (there are an infinity of them), but on the context: the range of confusable alternatives involved, the specific contrasts to be made, the invariant features that will reliably subserve successful categorization. And because ranges can change (and instantiation and categorization are never-ending processes), all categories and the features on which they are based will always remain provisional and approximate. (1987, p. 540)

Harnad points out that, at higher levels in perceptual processing, categorical representations may demonstrate rule-like properties: for example, a test for *one-leggedness* could sort many instances of animals and trees by counting their number of supporting structures (as detected by lower-level processing) and registering *animal* if there are at least two supports, *plant* otherwise. The categories employed by this test (e.g. *one* and *leg*) are grounded either directly in the sensory projections (e.g. by category learning) or recursively in simpler categories which, at base, are themselves grounded directly. "As anomalies were introduced, the rule could be elaborated so as to tighten the approximation in accordance with the new contingencies." As categorical representations are only accountable to the data encountered so far, "better and better approximations are all one can ever expect. No 'essence' of a tree or of an animal could ever be captured by a process such as this." (p. 539) Nevertheless, since categorical representations encode only those invariant features of instances sufficient for categorization, they are not extensional but intensional, capturing abstract properties, rules, and relations (p. 556).

Harnad's suggestion that the old analogue versus digital distinction be replaced by a digitalization continuum may prove highly fruitful for adaptive robotics. He proposed that the more invertible a transformation is the more analogue (and less digital) it is (p. 560, fn. 4). If, for example, by inverting an $X \rightarrow Y$ transformation, we can perfectly recreate X from Y, then the transformation is completely analogue. If, however, after the $X \rightarrow Y$ transformation, all of Y's spatiotemporal structure has been abstracted away[22] and no degree of inversion is possible, then the transformation is completely digital.

The digitalization continuum introduces thought-provoking middle ground to both the connectionism and imagery debates. .Certain mental operations appear to be constrained by some (but not all) of the analogue features of objects. It is possible that, in a person's cognitive economy, they are represented with some, but not all, of their analogue detail stripped away. They would thus rest somewhere midway in the analogue-to-digital conversion process, coming after quantization and dimensionality reduction but before complete symbolization (see, again, fn. 4). As we have seen from our discussion of SOMASS, for many tasks (especially those involving, for example, spatial reasoning), symbolic planning would appear to require recourse to just this kind of representation. It may be able to strike the right balance between the expressive freedom of symbolic representation and the avoidance of reasoning about stabilities offered by certain kinds of analogue representation (see Janlert, 1996).

Perceptual analysis is usually thought of as setting up a type-token relationship so that an instance of a particular type activates a corresponding symbol. In this way the sensory projections of a robin might instantiate the *robin* symbol. Methods of learning this kind of input-output mapping divide according to those that are *supervised* and, hence, require at least some information about what their correct output should be and those that are *unsupervised* and, hence, require no additional information.[23]

---

[22] For example, if Y is a representation manipulated purely according to the formal properties of its lawfully combined arbitrary symbols (as opposed to its analogue shape).

[23] Examples of unsupervised learning include statistical cluster analysis, vector quantization (Swaszek, 1985), Kohonen nets (Kohonen, 1984), competitive learning (Rumelhart

Unsupervised learning methods form categories solely on the basis of spatiotemporal correlations in their input—that is, without attaching any value(s) to it. If the proximal projections of distal objects are the only source of input, these methods can uncover no more than clusters of patterning in this data.[24] It is unlikely, however, that all objects we perceive can be discriminated on the basis of fortuitous correlations (e.g. discontinuities) in their sensory projections. As Harnad (1987) writes, although this

> is a significant simplifying factor for some categorization problems, it by no means represents a general solution to the problem of category acquisition. For example, the problem of perceiving "object constancy" [the continuance of unobserved objects] under spatial transformations still requires the selective detection of invariants. Any domain in which instances vary continuously is a potential problem. So is any domain in which the variation, though discrete, is so complex, subtle, or confusable — i.e., *underdetermined* — as to necessitate selective search and filtering. Finally, there is the domain of abstract "objects," whose instances vary along conceptual rather than physical dimensions. (p. 561, fn. 13)

Nature is full of mimicry, and often abstract objects like the edible and inedible must be discriminated on the basis of fine-grained differences, which unsupervised learning is unlikely to distinguish from noise.

In addition, more often than not, the correlations on which discriminations are based are *sensorimotor* in character. Even the ability to passively recognize objects largely develops by means of sensorimotor activity (see chapter 4). Furthermore, our internal reactions determine what is worth looking at—what patterns we ought to discriminate—as do the selective pressures behind them. It is by sensitizing themselves to these patterns that cognitive systems acquire perceptual categories. Hence,

---

& Zipser, 1985), and adaptive resonance theory (Grossberg, 1988). Examples of learning from a teacher include the backpropagation of errors (Rumelhart, Hinton & Williams, 1986) and the generalized delta rule, and from reinforcements include temporal differencing (Sutton, 1988) and Q-learning (Watkins & Dayan, 1992).

[24] Of course, if the input is augmented to include evaluative dimensions, the clusters become more than mere clumps of physical similarity. But, in this case, the unsupervised learning algorithm is being used as a form of supervised learning.

there is little reason to believe that clusters of sensory patterning from distal objects would coincide with the boundaries demarcating all the things we perceive as objects.

Let us assume, for point of illustration, that a neural network learned to produce a particular output given a particular sensory projection as input. One of the sources of this projection might be a robin at the window. For the network to make it possible for the robin to trigger a particular output, it would have to be able to ignore irrelevant variations in the projection: for example, variations related to lighting, viewing perspective, or the bird's posture and position. In the absence of additional feedback (e.g. from motor signals, internal reactions, and other mechanism that result from natural selection), this is not easy. Harnad brings to light this weakness of unsupervised learning in his criticism of Tversky:

> nondirected or 'ad lib' similarity... seems unlikely to explain how we categorize. Categorization is an *imposed* rather than an ad lib task. Hence the relevant dimensions of similarity must be found and selected by active processing guided by feedback from the consequences of *mis*categorization.... In nontrivial (i.e., confusable, underdetermined) categorization problems the solution is not obvious in the precategorical (ad lib; unsupervised) similarity structure. (1987, p. 561, fn. 15)

Indeed, it is because unsupervised methods accentuate the gross physical features of sensory projections that, in so doing, they are likely to filter out the kinds of potentially crucial details required for making life or death discriminations (e.g. edible versus poisonous). By contrast, supervised learning methods, which may exploit various reinforcements (e.g. physiological reactions) or knowledge acquired from a teacher, can make these kinds of discriminations. We shall examine this more fully in chapter 6 and appendix A.

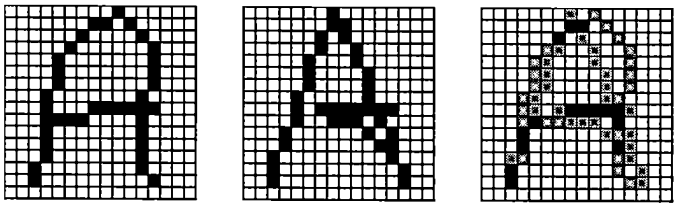
Vanishing Intersections Revisited

As Harnad (1990a) has pointed out, category induction has had many critics over the years; they have denied that categorical representations, capturing invariant aspects of sensory data, can be learned and revised empirically:

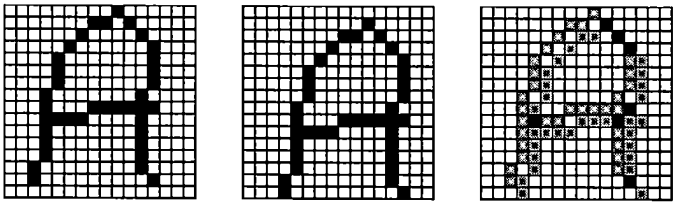
> It has been claimed that one cannot find invariant features in the sensory projections because they simply do not exist: The intersection of all the projections of the members of a category such as "horse" is empty. The

> British empiricists have been criticized for thinking otherwise.... The problem of vanishing intersections (together with Chomsky's "poverty of the stimulus argument") has even been cited by thinkers such as Fodor as a justification for extreme nativism. (p. 344)

Let us consider a simple illustration of vanishing intersections at the level of the raw optical character input to a computer. The first two columns display two hand-written instances of the letter *A;* the third column their superimposition (with points of intersection in black):



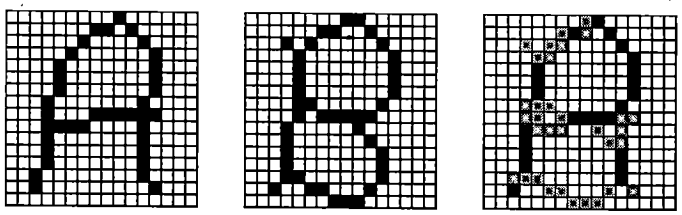
Though perfectly scaled and aligned, less than forty percent of the pixels in the first letter's image match the second's, and vice versa. If we repeated this process, intersecting the fifteen remaining pixels with some other varied instances, none may remain.[25] Below we see that even identical characters, slightly misaligned, may yield few points of intersection (less than ten percent):



This lack of overlap would be even more pronounced at higher resolutions. Categorization is further complicated by the fact that an instance of one category may have more points in common with an instance of another category than with other instances of its own. The *A* below shares sixty-two percent of its points with the *B*:

---

[25] For another simple demonstration of this, imagine superimposing two images of a horse on a computer screen and then blackening out all pixels whose colours differ at corresponding points in the two images. One need only repeat this process a few times with different pictures of horses to see the screen turn completely black.



In a project I helped to supervise, images of handwritten characters were normalized for scale and translation, detilted, decomposed using Gabor wavelets (see the next section), and then reprojected into a space of lower dimensionality using the K-L transform.[26] A neural network, which adjusted its weights by means of the RPROP algorithm, found invariant features in the output of this process. It outperformed undergraduate students at categorizing highly-degraded handwritten characters.

If one assumes that the intersections are taken solely at the level of raw sensory projections, then for many domains intersections may well vanish. However, extreme nativism does not solve the problem of vanishing intersections; *a priori* feature detectors would still need to categorize particular instances of types on the basis of invariance identifiable in the sensory projections of those instances. The only difference is that, under extreme nativism, there would be no need to *learn* how to do this. This is because the features detectors would already have evolved specifically for it.

Arguably, intersections do occur but at higher levels of abstraction. Horses do not exist as recognizable entities at the lowest levels of visual processing. At this level invariance may facilitate the detection of simpler and more universally applicable

---

[26] The Karhunen-Loeve (K-L) or Hotelling transform is an analogue, orthonormal transformation of an input vector. The vector is rotated by multiplying the transpose of the eigenvectors of the covariance matrix with it. These eigenvectors are the principal components. Thus, the axes of its new dimensions correspond to a set of orthogonal axes along which the original data varies maximally (see Haykin, 1994). Dimensions in the new space that provide little information may be discarded. Although principal components analysis and the K-L transform are never mentioned in *Numerical Recipes* (Press et al., 1992), the book provides routines for performing almost all the necessary calculations. After subtracting the mean from the data points, I used a chunk of code from `pearsn` (§14.5) to calculate the correlations for the correlation matrix. Since the correlation matrix is symmetric, I was able to use `tred2` to tridiagonalize it, before calculating the eigenvectors and eigenvalues with `tqli`, the QL algorithm (§11.2).

categories, and here it may be appropriate to speak in terms of Marr's detectors of edge segments, blobs, boundaries, and orientations. Higher levels are sensitive to horse-indicative invariance in lower levels, and no doubt this invariance is integrated from multiple sensorimotor modalities. Yet it may be more fruitful to think in terms of *feature selection* than feature detection. The global 'interpretation' of a figure may place constraints on lower-level processing thus permitting a categorization to have multiple equilibria as the Necker cube, Jastrow's (1900) duck-rabbit, and Rubin's (1915) vase-face figures exemplify (see Wittgenstein, 1958, further examples in Gregory, 1970, and the Gestalt literature).

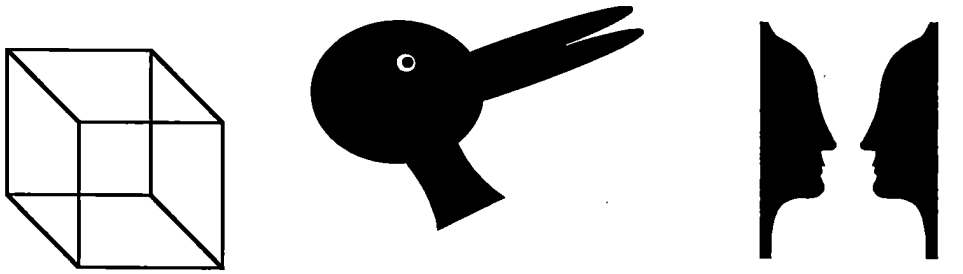


Fig. 2.1: The Necker cube, Jastrow's duck-rabbit, and Rubin's vase-face exemplify how a higher-level global interpretation may lead to the selection of different features at lower levels.

To exploit physical invariance in sensorimotor projections, the *same* invariant features need not be present for every instance of a category. Any one of a disjunctive set of invariant features, sampled from any number of sensory modalities, can serve to indicate sensorimotor invariance at a more abstract level. And it is at more abstract levels that invariance is most significant. Either consciously or nonconsciously, an individual needs to be able to recognize *abstract* invariance—what something is, the ways that individual can interact with it, and the likely outcome of those interactions. At more concrete levels *variance* can be crucial—not the abstract properties an instance shares with other members of a class but its particular shape, weight, and so on.

This is where Harnad's digitalization continuum comes in handy (1987, p. 560, fn. 9). Since we need not conceive of representations as entirely analogue or entirely digital (i.e. categorical or symbolic), we can make use of representations that have

undergone various degrees of digitalization to exploit simultaneously both their instance particular, variant features and their invariant, formal features—and features in between. In the present discussion, one category may be referred to as being more abstract than another simply because it is further removed from the physical structure of the sensory projection. We might expect this to imply that later brain processing is involved and that its recognition is more likely to depend on learned predictions than hardwired detectors. Indeed, sometimes even the distinction between iconic and categorical representations may be blurred if, to some degree, a representation can serve both iconic and categorical roles.

Wavelet Multiresolution Analysis for Category Induction

Studies of peripheral sensory processing in the brain have supported the view that sensory input undergoes considerable filtering, reprojection, dimensionality reduction (see Kaas, 1995), and low-level discrimination (e.g. edge detection) before being categorized at anything like a semantically penetrable level.[27] For no sense is this more true than vision. In this section, we examine some of the reasons for 'pre-categorical' processing with the aid of a specific model: the discrete wavelet transform (DWT). The DWT accentuates variations in sensory projections at varying scales and localizations. As Schyns and his colleagues (in press) note (though without mention of wavelets),

> Many psychophysical and computational models are converging on the observation that perception operates simultaneously at multiple spatial scales and that the coarser scales are often sufficient for effective processing of complex pictures (e.g., Burt & Adelson, 1983; De Valois & De Valois, 1990; Maar, 1982; Schyns & Olivia, 1994; Watt, 1987; Witkin, 1986). Multi-scale representations suggest that the input stimuli are discretized at different scales, possibly using scale-specific feature repertoires.... Large features may be registered without being composed out of smaller features, and small features may sometimes be created by decomposing larger features. (§3.1)

---

[27] This observation does not undermine the view that semantically penetrable categories take part in feature learning.

After briefly introducing how the DWT works, we apply it, in four sets of experiments, to the task of learning visual categorization. Not only is wavelet analysis compatible with Harnad's (1987) theory of category induction but preliminary results involving its application have revealed the theory's utility relative to other approaches.

Although wavelets have developed from work in mathematics, engineering, and the physical sciences (especially signal analysis, image processing, approximation theory, and physics), their promise for modelling early visual processing has recently come to light. In 1980 Daugman (1980) proposed that a parameterized family of two-dimensional Gabor filters (see Gabor, 1946) could offer a suitable model of the anisotropic receptive field profiles of single neurones in several areas of the primary visual cortex. Van Essen (1979), among others, has found these neurones selectively respond to the 2-D location, orientation, spatial frequency (size), symmetry, colour, motion, and stereoscopic depth of visual stimuli. The 2-D Gabor filters are able to account for the selective tuning of simple cells (i.e. linear neurones) for characteristic scale (spatial frequency), localization, orientation, and quadrature phase relationships (Daugman, 1988). Daugman (1985) has shown by chi-squared tests that this family of elementary functions fits the profiles of ninety-seven percent of simple cells in the cat visual cortex (based on measurements reported in Jones & Palmer, 1987).

How the Discrete Wavelet Transform Works

The fast Fourier transform (FFT) is typically used to decompose a signal (e.g. a soundwave, image, or video clip) into a series of coefficients, each of which describes the signal at one frequency only. (The FFT has an inverse, since the sum of the functions resulting from the multiplication of the coefficients by their respective sines and cosines approximates the original signal.) To take a biological analogue, the cochlea performs the same kind of mapping nonalgorithmically by virtue of its spiral shape.[28] Wavelets provide localization in characteristic scale (roughly analogous to frequency), but unlike sines and cosines they also provide localization in space.

---

[28] The Fourier transform and its inverse also provide a model of the rapid adaptation of surface markings in tropical flatfish to patterns at various spatial frequencies in their background (Ramachandran et al., 1996).

Thus, the discrete wavelet transform (DWT) is useful for analysing many kinds of bounded, discontinuous, aperiodic signals that are ill-suited to the Fourier domain—for example, an image with many sharp edges. Like the FFT, the DWT is quick to compute (linear-time complexity), and the sparsity of its output—most coefficients are negligible or zero—makes it ideal for data compression and fast numerical solutions.

In the 1-D case, the DWT takes a data vector, whose length is an integer power of two, and transforms it into a numerically different vector of the same length.[29] The transform decomposes the data vector into a particular wavelet basis (e.g. Haar, Daubechies, Gabor). A basis is a minimal set of vectors that can generate any possible data vector in the vector space through linear combination. Since wavelets are a set of linearly independent functions spanning a vector space, a wavelet basis is a minimal spanning set of functions that are also linearly independent. The number of possible wavelet bases is infinite. Difference among them usually reflect a trade-off between compactness and smoothness. Smooth bases provide higher numerical accuracy, while compact bases are better for data with sharp discontinuities.

The original data vector may be reconstructed from the output vector by multiplying each output coefficient by its respective wavelet basis function and summing up the result. Thus, DWT effects an analogue-to-analogue transformation because the input is recoverable from the output to within the numerical accuracy of the computing elements (be they neurones or a floating point chip, see Harnad, 1987). The method of recovery is the discrete wavelet inverse transform, and it has been used to restore many kinds of data after decompression, including images and video.

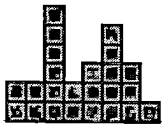
A rigorous introduction to wavelets would require a lengthy mathematical exposition (see Chui, 1992; Daubechies, 1992). However, a simple example will suffice to glean a basic understanding of how they work. Consider this data vector

---

[29] Notes and exceptions: *(1)* We shall later see that, from the 1-D case, it is easy to generalize to matrices (e.g. for images) and tensors (e.g. for video). *(2)* Data vectors whose length is not an integer power of two may be padded before applying the DWT. *(3)* For the box basis, the DWT *does* result in a numerically identical vector. We use the box basis only for purposes of explanation.

$$(2, 2, 6, 2, 3, 5, 3, 1)$$

and its graphical representation



It may be decomposed into an equivalent set of coefficients using the box basis, as shown in the first column of Figure 2.2 below:

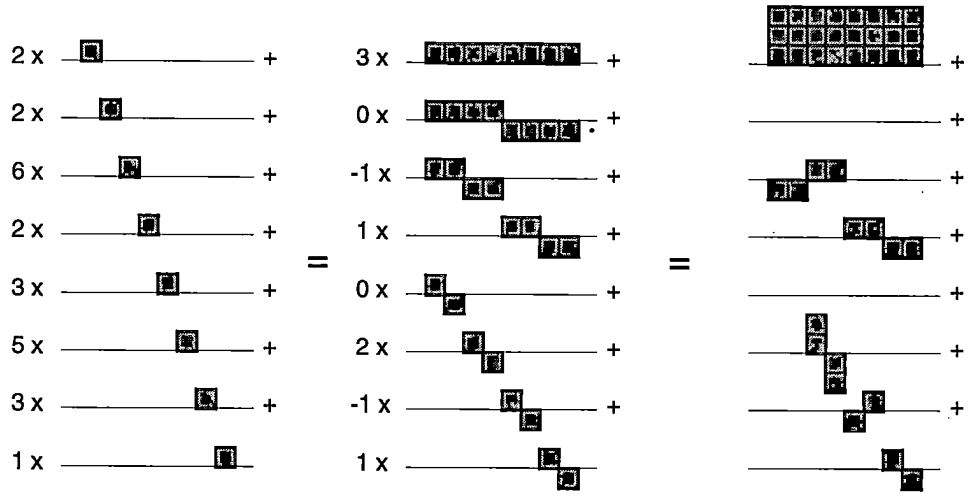


Fig. 2.2

When multiplied out, these box functions all sum up to the original input vector. If we treat the input vector as a piecewise-constant function on the half-open interval [0, 1), we can generate the eight box functions (also called Haar scaling functions)

$$\phi_0^3, \phi_1^3, \phi_2^3, \phi_3^3, \phi_4^3, \phi_5^3, \phi_6^3, \phi_7^3$$

from the function

$$\phi_i^j(x) \leftarrow \phi(2^j x - i) \qquad i = 0, \ldots 2^j - 1$$

where

$$\phi(x) \leftarrow \begin{cases} 1 & \text{for } 0 \le x < 1 \\ 0 & \text{otherwise} \end{cases}$$

The data vector may, however, be decomposed into a different basis—for example, the Haar (1910) basis—as shown in the centre column of Figure 2.2. This is the simplest and most compact basis. Its first function (Figure 2.2, top, centre) is

called the mother function. (Other bases often have more than one.) The mother function for the Haar basis is the box function

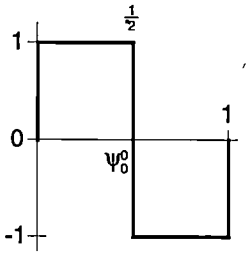
$$\phi_0^0$$

Notice that the mother function is completely smooth (flat) for the interval [0,1). Its matching coefficient (3, in our example) is called the mother-function coefficient. When the mother function and its coefficient are multiplied together, the result is an average for the entire vector (Figure 2.2, right column, top).

The remaining seven functions (which follow in sequence under the mother function in the centre column) are called wavelets. We can generate the seven Haar wavelets in Figure 2.2

$$\psi_0^0, \psi_0^1, \psi_1^1, \psi_0^2, \psi_1^2, \psi_2^2, \psi_3^2$$

by scaling and translating the wavelet



given by the function

$$\psi_i^j(x) \leftarrow \psi(2^j x - i) \qquad i = 0,\ldots 2^j - 1$$

where

$$\psi(x) \leftarrow \begin{cases} 1 & \text{for } 0 \le x < \frac{1}{2} \\ -1 & \text{for } \frac{1}{2} \le x < 1 \\ 0 & \text{otherwise} \end{cases}$$

As an alternative, we can generate Haar wavelets directly from the box function

$$\psi(x) = \phi(2x) - \phi(2x-1)$$

Unlike the mother function, wavelets do not respond to data vectors that are flat in their response area (hence, their integral is always zero). Their response is positive if the left-hand side of their response area is greater than the right-hand side; negative if less than. The wavelets' matching coefficients (0, -1, 1, 0, 2, -1, 1, in Figure 2.2) are

called wavelet coefficients. These detail coefficients[30] pick out variations in the data vector at various scales and translations. By multiplying each wavelet by its respective coefficient, we get the third column in Figure 2.2, which sums up to the original data vector. As is true of any Haar basis, all the basis functions in the second column

$$\phi_0^0, \psi_0^0, \psi_0^1, \psi_1^1, \psi_0^2, \psi_1^2, \psi_2^2, \psi_3^2$$

are orthogonal to each other. Not all bases share this property (e.g. Gabor does not). When they do not it rather complicates the calculation of the DWT.

We obtain the wavelet coefficients for the Haar DWT by means of a filter bank. A filter bank calculates $n/2$ pairwise averages for a vector of length n and $n/2$ pairwise differences. This makes it both a smoothing and a detailing filter (it is called a quadrature mirror filter in signal processing parlance). The detail coefficients can be used to recover the information that has been averaged away. We apply the filter bank first to the entire data vector, then to the resulting $n/2$ ($n/4$, $n/8$,...) smoothed coefficients (the pairwise averages), recursively, until—in the case of the Haar basis—only one smooth coefficient remains. This is the mother function coefficient, the average for the entire data vector. This is sometimes called a pyramidal algorithm.

We are now prepared to calculate the wavelet coefficients for the input vector in the above example (final values are in italics). We begin with the input vector

2 2 6 2 3 5 3 1

For each pair of coefficients, we calculate the average

2 4 4 2

and half the difference:

0 2 -1 1

Then, for each pair of averages, we calculate the average

3 3

and half the difference:

---

[30] Wavelet coefficients are sometimes called detail coefficients. This is a useful distinction for the following reason: As a kind of shorthand, the term *wavelet coefficients* is often applied to all the coefficients collectively. Thus, the term detail coefficients makes clear that we are excluding the (smooth) mother-function coefficient(s).

-1 1

Finally, for the remaining pair of averages, we calculate the average

3

and half the difference:

0

Collecting the coefficients, we have

3 0 -1 1 0 2 -1 1

A *C* function for the Haar DWT is given as follows:

```
void hdwt (double *vector, int length)
{
    double *temp;
    int    n, mid_pt, i, twice_i;

    if ((temp = (double *) malloc (length * sizeof (double))) == NULL) {
        fprintf (stderr, "Error in hwd(): unable to allocate enough memory.\n");
        exit (EXIT_FAILURE);
    }
    for (n = length, mid_pt = length/2; n >= 2; n = mid_pt, mid_pt /= 2) {
        for (i = twice_i = 0; i < mid_pt; i++, twice_i += 2) {
            temp[i]       =  (vector[twice_i] + vector[twice_i+1])/2;
            temp[i+mid_pt] =  (vector[twice_i] - vector[twice_i+1])/2;
        }
        for (i = 0; i < n; i++)
            vector[i] = temp[i];
    }
}
```

The first half of the program does no more than declare the variables and allocate space in memory for a temporary array used to hold the smooth coefficients. The outer `for` loop recursively subdivides the vector. (For a vector of length eight, it sets the length to 8 and the mid point to 4 on the first iteration, 4 and 2 on the second, 2 and 1 on the third.) For each pair of coefficients in the vector, the first inner `for` loop accumulates the averages (to the left of the mid-point) and the differences (from the mid-point rightward). The second inner `for` loop simply copies the smooth coefficients from the temporary array back into vector.

It is often useful to use a normalized basis (for example, when a standard measure of error for all coefficients is required). We may calculate the DWT for a normalized Haar basis by dividing by $\sqrt{2^j}$ each output coefficient (calculated by `hdwt`) whose matching wavelet has the superscript *j*. (The normalization can also be incorporated into the DWT itself. I have found both implementations to give the same results.)

The 2-D DWT generalizes neatly from the 1-D case. The standard approach is to apply the above function sequentially to every row of the data matrix and then to every column. To do this, the following *C* function calls `hdwt`:

```
void hdwt_2d (double **matrix, int x_size, int y_size)
{
    double *temp;
    int    i, j, max_size = x_size < y_size ? y_size : x_size;

    if ((temp = (double *) malloc (max_size * sizeof (double))) == NULL) {
        fprintf (stderr, "Error in hwd(): unable to allocate enough memory.\n");
        exit (EXIT_FAILURE);
    }

    for (j = 0; j < y_size; j++) {
        for (i = 0; i < x_size; i++)
            temp[i] = matrix[i][j];
        hdwt (temp, x_size);
        for (i = 0; i < x_size; i++)
            matrix[i][j] = temp[i];
    }
    for (i = 0; i < x_size; i++) {
        for (j = 0; j < y_size; j++)
            temp[j] = matrix[i][j];
        hdwt (temp, y_size);
        for (j = 0; j < y_size; j++)
            matrix[i][j] = temp[i];
    }
}
```

The first outer `for` loop indexes through the rows. Its inner `for` loop copies each row into a temporary array. It then computes the 1-D Haar DWT on that array and copies the results back into the matrix (overwriting the original values). The same process is repeated in the second outer `for` loop for the columns. To avoid all the copying, I combined `hdwt` and `hdwt_2d`.

Some Categorization Experiments using Wavelets

The 2-D DWT may be used to prepare images for categorization and category induction. Unlike random patterns, images of the natural world are highly redundant (in information-theoretic terms). It is possible to predict the values for a set of pixels on the basis of another, possibly remote, set. Examples of this self-correlation are locally similar values for colour or luminance, edge continuance, and uniformity of texture. Wavelets provide a means of exploiting these correlations to varying extents. As we have noted, wavelets selectively respond to variations in the input, which they describe in a hierarchy of resolutions and localizations. The larger the variation, the larger (in absolute value) the coefficients for the wavelets that respond in that region; the smaller the variation, the smaller the coefficients. It is precisely because images

are highly redundant that most wavelets show little or no response.[31] I took advantage of this feature in a simple experiment of sorting mushrooms. The matrix positions of twenty significant coefficients (out of a total of 4096) provided a highly compact signature from which to extract invariant features.

In the experiments I took greyscale snapshots of sixty mushrooms with a video camera—fifteen from each of four species: shiitake, chestnut, oyster, and closed cup. I took the images under good, though somewhat varying, indoor lighting conditions with a greyish cardboard background, from about the same angle (15°–25°) and distance (12–15 cm). The shiitake and especially the oyster mushrooms were highly asymmetrical, so I decided to digitize the mushrooms at random rotations to prevent the categorization program from exploiting these asymmetries. I roughly centred each mushroom in its image but made no attempt to keep uniform its distance to the edges of the image. (In object recognition research, it is common to scale the object horizontally or vertically in order to fit it to the frame of the image.) The images measured 64-by-64 pixels, with 256 shades of grey, varying from black (0) to white (255). They were in PGM (P5) format—a binary pixel-by-pixel representation. Figure 2.3 displays two mushrooms from each species.

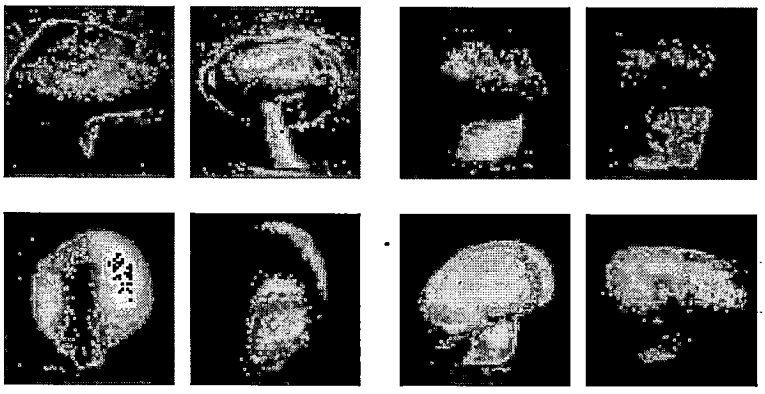


Fig. 2.3: Shown above are pairs of mushrooms chosen at random from each species in the data set: (from top left) shiitake, chestnut, oyster, and

[31] This is why, with no observable loss of detail, image compression programs can compress an image at ratios sometimes exceeding 20:1. They truncate or quantize the small coefficients and then recode the sparse matrix in a compact form (e.g. using Huffman and run-length coding).

closed cup. These images were among those used in the experiments (64x64 pixels, 256 greys).

I performed four sets of experiments. The criterion of correctness for the experiments was to categorize each mushroom with the other members of its species.[32] The purpose of the first set—the baseline experiment—was to gain some insight into what degree of categorization accuracy I should expect when such considerations as memory requirements and computational efficiency are ignored. The second set—the first of the signature-based experiments—was intended to find out whether the performance of signature-based categorization could approach that of the brute force method taken in the baseline experiment. The third set—the prototype experiments—was designed to appraise the effectiveness of using an 'average mushroom' from each species for categorization. The fourth set—the experiments with categorical representations—were meant to test whether four categorical representations, one for each species, could successfully categorize the mushrooms. The program developed these representations by filtering in within-category invariance while filtering out between-category invariance. This proved to be the most promising approach so far attempted.

The baseline experiment. In the first experiment, the program performed a normalized Haar discrete wavelet transform (DWT) on all sixty images. For purposes of comparison we may conceive of each of the sixty resulting 64-by-64 matrices as a 4096-dimensional vector. The program simply calculated the Euclidean distance between all the vectors; a mushroom's best match was determined according to whichever mushroom's vector was closest to its own. The program categorized ninety-seven percent (58/60) of the mushrooms correctly (i.e. with other mushrooms of the same species). However, the program's memory and computational requirements were formidable. Sixty double-precision (10 byte) images occupied nearly 2.5 million bytes of memory. To compare just two matrices required 4096 multiplica-

---

[32] In these categorization experiments, we need not presuppose either language or a teacher. The category induction could have been based simply on a physiological response. (To me this became all too clear when I discovered in the course of the experiments that I have an allergy to shiitake mushrooms.)

tions; to compare every pair required almost 7.25 million multiplications (4096 times 1770). This method was not promising to scale up well.

The first signature-based experiments: categorization from past instances. In this set of experiments, the program converted each 40960-byte matrix (4096 pixels times its 10 byte double-precision representation) to a signature after performing the DWT on the image. The signature is simply an ordered list of the signs and 1-D positions of the k largest detailed coefficients in the matrix. (For the purposes of this routine, the matrix was treated just as a long 1-D array.) Only 13 bits were required for each entry in the list (12 for the position and 1 for the sign). Diminishing returns usually set in for signatures with more than 20 entries and, as the number of entries got into the hundreds, performance declined. The memory requirement for a signature of length 20 was only 260 bits (20 times 13) or 33 bytes. Thus, for sixty matrices the total memory requirements were 1980 bytes, less than a tenth of a percent of the 2.5 million bytes required for storing the matrices.

To create a signature, the program first obtained the value of the kth largest coefficient (in absolute value) in the matrix. It calculated it by creating a heap (see Sedgewick, 1988) from the first k coefficients in the matrix, with the smallest at the root of the heap, and then replacing the smallest coefficient in the heap with any larger coefficients found in the remainder of the matrix. This is the most efficient method when k is much smaller than n, where n is the size of the matrix (i.e. 4096). It takes only order $n \log_2 n$ operations. The program then used the value of the kth largest coefficient to find and extract the signs and 1-D positions of the k largest coefficients in the matrix. To save on the number of operations required to compare two signatures, the program first backed through the matrix finding the negative coefficients that were largest in absolute value and then passed forward through it finding the largest positive ones. A sample signature is

-1793 -576 -387 -322 -256 -196 -128 -67 -65 -64 -4 -2 1 3 6 192 195 323 448 896

The minus signs do not indicate negative 1-D positions (positions range in value from 0 to 4095) but positions whose coefficients are negative. The program compared pairs of signatures by counting matches. (A match occurred only if both the sign and position were the same.) If there was a match, the program incremented the indexes in

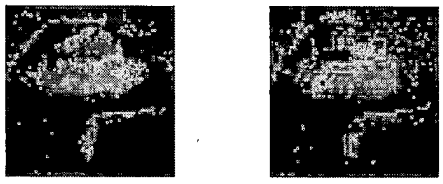
both lists to the next entry in each; if there was a mismatch, it would increment whichever index pointed to the lower value.

To appraise the performance of this scheme, I used two measures. The first was the same as the one used in the baseline experiment except that the program counted matching sign-position entries in signature pairs rather than computing distances between vectors. For the second measure, the program selected one mushroom from each species in turn without replacements. If, in the first turn, it selected a shiitake mushroom and, in the next, a chestnut, it would have no choice but to match the chestnut with the shiitake since it would have no other signatures to match it against. As the program accumulated more instances of different species, its performance improved, but under neither scheme did its performance ever approach that of the baseline experiment.

I repeated the experiments using the Haar basis without normalization and my implementation of the Daubechies 4-tap basis.[33] Some of my results are summarized in Table 2.1. The percentage of correct matches listed in the table is based on signature matching alone. In other experiments I factored in a comparison of the image

---

[33] Press et al. (1992, chap. 13.10) elucidate the theory behind Daubechies wavelet bases. Daubechies 4 is a highly compact basis, though less compact than Haar. The code for the Daubechies 4 DWT is no longer than that for the normalized Haar DWT but it runs slower, largely because it requires twice as many multiplications. In general, Daubechies 4 is much better than the Haar bases for image compression. Both of these images were produced from the same snapshot from just under 5% of their detail coefficients:



The image on the left used the Daubechies 4 basis, while the image on the right used the Haar basis (notice the blocky artefacts). Nevertheless, for categorization the normalized Haar basis performed at least as well as Daubechies 4. This can probably be put down to the fact that the highly compact (and, hence, unsmooth) Haar basis is very sensitive to changes in luminance (e.g. edges) and that this factor is more critical in categorization than numerical accuracy.

pair's average luminance.[34] Although this improved some statistical measures, it had little influence on the number of correct matches. Based on these and other statistical measures, I came to conclude that the normalized Haar basis performed marginally better than Daubechies 4, which performed somewhat better than the Haar basis without normalization. While one would expect high error rates at the beginning of learning, if we compare rates after learning, it is clear that the best accuracy obtained by the signature-base method—ninety percent—fell significantly short of the ninety-seven percent accuracy of the baseline experiment.

Wavelet basis for the DWT	*Length of signature*	*Percentage correct after learning*	*Percentage correct while learning*
Haar	20	85% (51/60)	72% (43/60)
Haar	60	80% (48/60)	67% (40/60)
Normalized Haar	20	90% (54/60)	77% (46/60)
Normalized Haar	60	90% (54/60)	77% (46/60)
Daubechies 4	20	85% (51/60)	80% (48/60)
Daubechies 4	60	88% (53/60)	68% (41/60)

Table 2.1

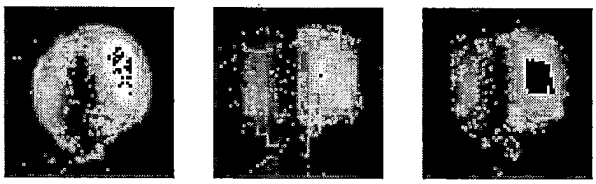


Fig. 2.4: The image on the left resulted from performing the normalized Haar inverse transform on a complete 64-by-64 matrix (40960 bytes). The centre image resulted from performing the normalized Haar inverse transform on a matrix generated from the mother function coefficient and a signature of length 60 (108 bytes). The image on the right resulted from the same process as the second, only using the Daubechies 4 inverse

[34] This figure was derived from the mother function coefficients. For the Haar basis, there is only one mother function coefficient, and it is the average luminance. For the Daubechies 4 basis, there are four mother function coefficients for the 2-D transform; two for the 1-D transform.

transform. Although the image in the centre and on the left were produced using only a fourth of a percent as many bits as the first, the mushrooms are still discernible. Note that even though the image produced using the Daubechies 4 basis appears clearer than the one produced with the normalized Haar, the normalized Haar basis proved to perform slightly better overall than the Daubechies 4 basis.

The prototype experiments. A prototype is the most typical member of a category. For example, a robin might be the most typical bird—certainly more typical than a goose or an ostrich (see Rosch née Heider, 1971). From previous experiments, it had already become clear that basing categorization on the most average mushroom projection of each species would give poor results, so I decided to base it on a prototype-like representation—an average of the projections of all the mushrooms of a category. Comparing a mushroom to, say, the shiitake representation was equivalent to finding out how many matches it would have on average with all the instances from that species. The accuracy of this method was relatively low: only 70% correct for signatures of length 20, 73% for signatures of length 60. These poor results are owing to the fact that oddly shaped mushrooms belonging to a given category may sometimes better match the average representation of a category other than their own. Such results are to be expected as long as each category has only one prototype. Prototypes fail at categorizing disjunctive sets for the similar reasons.[35]

Experiments with categorical representations. The basic idea for these experiments was to develop one categorical representation for each species of mushroom. The representation was meant to capture as much within-category invariance and as little between-species invariance as possible (see Harnad, 1987). The program derived each categorical representation from the signatures of ten mushrooms of the same species. Learning was then turned off, and the program assessed not only how accurately the four representations categorized the four sets of ten mushrooms from which they were derived but also the remaining twenty (hitherto unseen) mushrooms. I attempted various schemes which placed different weight on the importance of

---

[35] Minsky and Papert, 1988, discussed this limitation in relation to single-layer perceptrons and showed why it prevented them from learning the logical exclusive-or (XOR).

including within-category invariance in each representation relative to excluding between-category invariance.

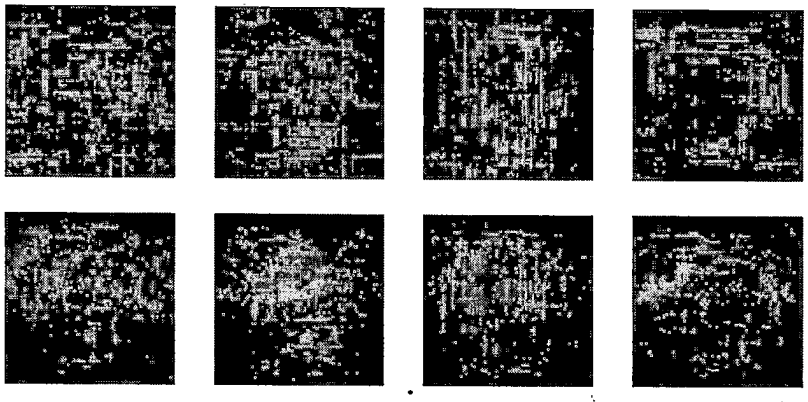


Fig. 2.5: Here is a first attempt at displaying within-category invariance derived from ten mushrooms of each species: (columns, from left to right) shiitake, chestnut, oyster, and closed cup. The normalized Haar inverse transform was used for the top row; Daubechies 4 for the bottom row.

The scheme which struck the best balance was to include a sign-position entry in a categorical representation's signature if the total number of instances matching that entry from the same category exceeded one-fifth the total number matching for the other three categories combined. (This scheme would need to be adjusted for increases in the total number of categories.) The program correctly categorized *all* of the forty mushrooms from which the categorical representations were derived, thus showing better accuracy than the distant metric used in the baseline experiment. It miscategorized only one of the twenty novel mushrooms—a 95% success rate. There were some variations among the signature lengths for the different species (106 for closed cup, 146 for shiitake, 148 for oyster, and 155 for chestnut). Only 902 bytes (7215 bits) of memory were required to store the four categorical representations. Not only did this scheme prove to be more accurate than that of the baseline experiment, but it required less than one twentieth of one percent the memory and far less computation.

There are, however, potential problems with this scheme. Recall back to Harnad's test of whether something is a plant or animal depending on how many supporting

structures (i.e. legs) its has (1987, p. 539). Let us consider an even simpler problem. Our agent is presented with images revealing a leg on either the left or right-hand side, or on both sides, or neither, and it must categorize each image as either one-legged or not one-legged. (This is a version of the XOR problem, see Minsky & Papert, 1988.) The agent will not be able to develop a pair of categorical representations to discriminate instances as either one-legged or not one-legged because images of two legs are essentially linear combinations of images of legs on either the right or left-hand side. Since ONE-LEGGEDNESS is equivalent to (LEG-ON-LEFT *and not* LEG-ON-RIGHT) *or* (LEG-ON-RIGHT *and not* LEG-ON-LEFT), it requires a two-tier logical structure and the lower-level categorical representations LEG-ON-RIGHT and LEG-ON-LEFT.

So we need to address

(1) how the robot can determine when it needs to form additional categorical representations, and

(2) how to decide which instances to use in deriving them.

One suggestion might be to use feedback from miscategorization (Harnad, 1987). Miscategorization can occur for many reasons, and it is not necessarily a symptom of the above problem. However, miscategorization in conjunction with a low-learning rate may indicate a need for additional representations and logical structure.

A similarity measure may be useful in deciding which instances to use in developing new lower-level categorical representations. This would help to ensure that within-category invariance is high relative to between-category invariance, so the categories are fairly stable. However, in determining similarity this measure should probably include more than just the physical shape of the sensory projections. As we shall discuss in chapter 6, it may be useful to include similarities in internal (e.g. physiological) feedback, so that the invariance being picked up includes more of the factors that are relevant to the agent. If the new lower-level categorical representations still fail to improve learning, the similarity measure can be progressively tightened. This would narrow the range of instances from which each categorical representation could be derived and widen the number of possible representations that could be formed. Thus, it would prompt the agent to form new categorical representations until the learning rate improves. More advanced techniques could use cluster analysis in place of a similarity measure.

Summary and Conclusion

The balance of evidence may now be tipping in support of theories that permit the formation of (completely) novel features. There are a number of advantages to being able to learn new features rather than having to rely on a fixed set. Schyns and his colleagues (in press) propose that features are formed through the interaction of top-down functional and bottom-up perceptual constraints. Feedback from miscategorization plays a key role in initiating and guiding their development.

Harnad (1987, 1990a) proposes that categorical representations could develop causal links with the sensory projections of objects during behavioural interactions. Categorization feedback informs the inductive process by which the representations are sensitized to sensory features. The learning process is aimed at developing *intensional* representations that are capable of detecting those features—and only those features—that have, in previous sensory projections, reliably distinguished positive and negative instances of a category. Thus, they can uncover abstract properties, rules, and relations. When anomalies lead to miscategorization, a categorical representation tightens its approximation. Which features can best subserve categorization depend on the scope of confusable alternatives. Harnad's approach has clear advantages over unsupervised learning. The latter must rely on fortuitous patterning and discontinuities in the sensory projections and, thus, can pass over spatial details that may be critical to an organism's activity.

Harnad (1990a) has indicated that categorical representations can act as grounded symbols. Thus, they might form the basic building blocks from which a symbol system could be developed from the bottom up.

Often it may not be possible to discover a category's invariant features by intersecting the sensory projections of its instances: the intersection of all the instances might be empty. But it would be wrong to conclude from this that category induction is impossible:

(1) Intersections may appear after the sensory input has been reprojected into a new domain by means of an analogue transformation like the fast Fourier transform or the discrete wavelet transform.

(2) Intersections may also appear after the input has been processed by low-level feature detectors or after a brief interplay between bottom-up and top-down constraints.

(3) Category induction does not depend on every instance of a category sharing the same invariant features. Different clusters of instances may share different clusters of invariant features. The presence of features associated with any one of a disjunctive set of clusters is enough to subserve categorization.

(4) An organism may first learn categorical representations for more basic and easily learned categories. These representations may then bootstrap the learning of more complex or abstract categories that are expressed in terms of them.

The discrete wavelet transform provides a means of exploiting the high degree of redundancy or self-correlation typical of the sensory projections of natural environments. Wavelet coefficients can pick out variations in the sensory projections at various scales and translations. Daugman (1980) has devised a family of 2-D Gabor filters that accurately model the receptive field profiles of simple cells in the primary visual cortex. These wavelet filters account for the cells' selective tuning for scale, localization, orientation, and quadrature phase relationships.

In a series of experiments, I prepared images of mushrooms for categorization by writing a program to decompose them using the Haar and Daubechies 4 wavelet bases. The program then learned to categorize the mushrooms by means of a number of different methods. Some were brute force, others used compact and easily compared signatures. Some were inspired by prototype theories, others tried to find closest matches between individual instances. One technique, based on Harnad's theory of category induction, proved to be significantly more accurate at categorization than the other approaches. It was also the most efficient in terms of computation and memory requirements. Category representations were derived from within-category invariance with between-category invariance filtered out. After exposure to only ten mushrooms from each species, the program was able to classify new mushrooms with ninety-five percent accuracy.

Chapter 4. Learning Sensorimotor Predictions

Introduction

This chapter considers the shortcomings of nonadaptive approaches to grounding perceptual categories. It attempts to go beyond the minimal requirement that they be causally linked with actual objects, events, and relations by developing a fuller, more ecological notion of grounding. At their foundations, perceptual categories are seen as part of an organism's sensorimotor loop with its environment and, through an empirical process of adaptation, they mediate its activity.

In developing an adaptive approach to grounding perceptual categories and, by extension, all internal representation, this chapter advances and elaborates a theory of sensorimotor predictions. It also explores computational techniques for implementing predictions. An agent learns predictions from spatiotemporal correlations in sensory projections, and these predictions serve as a basis for how it models its environment. Sensorimotor predictions are not to be confused with response probabilities, which may be learned without requiring a model of how actions affect the environment or internal variables (see appendix A). Neither are sensorimotor predictions to be confused with hypothesis testing. Nevertheless, much risk can be avoided by testing hypotheses against these predictions before taking action.

It would be difficult for proponents of symbolic representation to explain its evolutionary origin without granting that it, at some point, must have conferred on its bearer selective advantage.[36] Symbolic representation can offer no benefit, however,

---

[36] The argument has been put forward that a language system developed like a spandrel in a cathedral, that is, because of the demands of surrounding architecture and not because of anything that it contributed in its own right. This argument could also be extended to symbolic representation. Yet considering the complexity of both language systems and symbol systems and the lack of empirical evidence that surrounding architecture would have created a selective demand for them, it seem unlikely that either could be spandrels. A comparison of the anatomy of the human mouth and throat with that of our simian

unless symbolic categories capture features of the surrounding world that are relevant to an organism behaving so as to enhance its reproductive success. Therefore, to achieve minimal biological grounding, an internal category must stand in some relation to sensorimotor coordination.

Genes are passed on when an organism is sufficiently able to discriminate sensorimotor patterning relevant to its reproductive success. The essential kind of invariance in this patterning, the kind that motivates an organism at least to behave as though it categorizes different instances, with their disparate sensory projections, as similar, is that those different instances can be dealt with in similar ways to produce similar physiological effects.

This chapter is structured according to a line of argument that runs as follows. Unaugmented, fixed feature detectors are inadequate to ground perceptual categories adaptively in sensorimotor coordination. It is not enough for internal indicators to be causally linked to outside states of affairs. They must also capture sensorimotor relationships. By capturing these relationships they enable an intelligent agent to anticipate likely consequences on the basis of its sensory projections and motor signals. These relationships involve the learning of mappings that relate information from various sources and sensory modalities.

Since sensorimotor relationships vary according to both body and environment, they must reflect individual differences. Environmentally-mediated processes of growth, damage, regeneration, and decay change an organism's sensorimotor dynamics in unforseeable ways. This suggests not only that an organism learns predictions in order to anticipate the likely consequences of potential motor signals but also that it alters these predictions in light of bodily and environmental changes.

To take this into account, we develop a high-level theory of sensorimotor representation. It is based on the empirical development of sensorimotor predictions.

---

ancestors reveals that our vocal track is anything but a spandrel. It is good for speech and not much else (see Lieberman et al., 1992). To make room for a rounded tongue and enhanced voice box, our teeth are crowded, and we have a longer, more vulnerable neck. It is much likelier for humans to die from choking on food or from abscesses caused by impacted molars. The fact that speech came at a price supports the thesis that it must have had selective advantage.

Although the theory does not state how to implement systems that develop predictions, it may be used to appraise prospective implementations. We do precisely that in our examination of a memory-based model that has been used to learn sensorimotor mappings. The appraisal not only shows the implementation to be promising but also demonstrates that, after adding a few extensions, it can be brought into conformance with the theory. I am currently supervising a project which incorporates my extensions while expanding the system's performance. Finally, we reconsider the implications of the theory for the simulation of intelligent behaviour and find a rich vein for future work.

Evidence from Vision for Sensorimotor Integration

The brain's processing of visual input is highly complex, and much occurs even before this information is integrated with sensorimotor information from other modalities. Nevertheless, sensorimotor integration must come at some point in order to ground perceptual categories in sensorimotor activity. An appreciation for evolution would lead us to believe that this grounding comes prior to and serves as the foundation for more abstract, systematic modes of behaviour. Unfortunately, an emphasis on vision and its modality-specific processing in isolation from both other sensory modalities and motor coordination has lead to potentially misleading analyses of the nature of visual processing. This has biased computational approaches towards a reliance on fixed feature detection as opposed to empirical adaptation. But eventually we must confront the issue of grounding, and it would be a mistake to postpone its consideration until after we obtained a thorough understanding of early stages of visual processing. This is in part because there is *no* stage of visual processing whose evolution has not been influenced by the fact that an organism's perceptual categories are grounded in its sensorimotor activity. In our study of physiological structures, we should not lose sight of their function in biological adaptation. Hence, a deeper level of analysis is called for than just considering input-output relations.

Vision's dependence on coordination is well acknowledged. The manner and degree of this dependence has been contested by neuroscientists for more than 40 years—ever since it became clear that whether the brain attributes a movement on the retinal image to the viewer or to an object depends on whether the motion was self-

induced (see Gyr et al., 1979).[37] It is only by taking self-induced movements into account, for example, that the brain can maintain the stability of the visual scene. Kohler (1964) showed that such movements were necessary for subjects to adapt to seeing the world through reversing prisms. The importance of movement is illustrated by the fact that if the eye is held focused at a fixed point in a static scene, once neurones have habituated the viewer even fails to see. Distance perception may depend more heavily on movement parallax than binocular disparity, and evidence suggests that infants become sensitive to it first (Slater, 1989). Unfortunately, Marr's stereoscopic theory of vision echoes perceptual analysis in its neglect of sensorimotor coordination. There has, nevertheless, been a growing realization in the 1990s that computational models need to be extended to account for the role of action in vision (Ballard, 1991; Blake & Yuille, 1992).

A consideration of the drawbacks in Marr's attempt at perceptual analysis will illustrate the importance of relating internal categories to sensorimotor coordination. Marr proposed that different neurones in a network could be attuned to varying degrees of binocular disparity in the retinal image. In this manner a network could extract a viewer's distance from a perceived object. However, to be grounded, this distance measure would need to be related to the viewer's behaviour, such as the movements necessary to grasp the object. Otherwise, the measure could not serve any behavioural ends nor could a selective process be involved directly in its evolution.

This observation is acutely evident with body variables. For example, although there is a correspondence between the topography of the retina and arrays of cells in various cortical regions, the output from these cells does not capture any spatial relationship. To accomplish this, further representations would be required to relate the retinal image to the individual's motor activities such as the eye movements that would be required to foveate a peripheral object. To be effectual and, therefore, evolutionarily explicable, symbols must be grounded in sensorimotor categories that are relevant to coordination. Otherwise, they are as useless as a cockpit furnished

---

[37] Under the right circumstances, of course, the brain can be easily fooled. For example, a person seated in a train compartment at a station may sometimes look at an adjacent train that is moving and think that his or her cabin is moving instead.

with unlabelled and undimensionalized instruments in the hands of a pilot lacking instructions as to their bearing on the flight of the aircraft.

Research Caveats

The methodology of the neurosciences can lead theorists to overestimate the importance of feature detection and neglect the brain's essential role in setting up sensorimotor mappings. This is hardly surprising considering the primitive tools we have for exploring the brain. And so, for example, they resort to probing neurones with electrodes and trying to find out what kinds of stimuli can make them fire.

In devising intelligent agents we run up against limits in our own perception with its particular epistemological commitments. When an agent is confronted with an object, we may be inclined to infer what it detects on the basis of our own perceptions. We usually imagine that it 'sees' pretty much what we do. In ordinary life this might seem like a reasonable assumption. A life-time of consensual endorsements within a language community backs up the inferences we make concerning consistencies in the appearance of objects. Unfortunately, inferences based on consensual endorsements within our own language community may be of little use when analysing the workings of intelligent agents. This is because these inferences take for granted that the agent, like us, is human and partakes in the same language and culture.

It would be a mistake, for example, to assume that a resemblance between the source of an agent's sensory projections and the patterns of neural firings in one of its cortical regions implies that the agent sees what we see or can respond to what we can. It is not the resemblance that matters. If the sensory inputs to a cortical region were all shifted spatially, the resemblance would remain but the organism would not be able to handle the object until it had adapted to this change. Furthermore, if it were possible to displace neurones spatially without interfering with their interconnectivity, the resemblance could be obscured without affecting the behaviour of the organism. Whatever form these patterns take, that form must be ecologically mediated. It depends on the relationship between the organism and its environment as set up by its particular body and sense organs.

A Theory for Learning Sensorimotor Predictions

The aim of this dissertation is to propose a possible method of modelling the development of behaviourally-relevant categories. When an agent is using an internal representation, what makes that representation a representation is not the fact that it looks like one to us (i.e. that we can interpret it) but that it can function as one for the agent. Hence, we propose that sensorimotor representations form the foundation for our model of cognition. Sensorimotor representations denote adaptations that take place within an individual on the basis of past sensorimotor projections (and evolution), and they depend on, among other things, the individual's particular body, internal reactions, and life history. As a working definition, we shall say that representation is *about* something (e.g. an object, event, or relation) to the extent that those sensorimotor processes of the agent that need to be appropriately related to that thing can be appropriately related to its representation. By using this criteria to qualify it as a representation of that thing, we avoid the danger posed by our ability to interpret symbols as standing for things even if, to the agent, they have no causal connection (i.e. grounding, see Harnad, 1990b, on the hermeneutic hall of mirrors).

There is likely to be a close relationship between how representations function and how they develop and change. With this in mind, Sommerhoff and MacDorman (1994, §5) proposed an approach to modelling sensorimotor relationships based on neurophysiological predictions. They also discussed evidence for these predictions from brain research and suggested one way they might be represented neurally. Their approach shall be corrected and, from the standpoint of behavioural simulation, further developed here.

An individual learns predictions (within the constraints of biological adaptation) on the basis of spatial and temporal correlations in sensorimotor projections and internal variables. These predictions need not be conscious. They may concern, for example, the effect of self-induced movement on sensory projections from internal and external sensory receptors. An individual may also learn them passively. They may, therefore, concern likely trends in internal physiological and external sensory projections that can occur either without motor involvement or independently of it. An individual's anticipation of consequences is always contingent on that individual acting in particular environmental circumstances. Therefore, sensorimotor projections

(as well as internal processes like thinking, remembering, and feeling) elicit, revise, and sustain an individual's learned predictions, and only a subset of them will be active at any given moment. They are in this sense conditional predictions.

Once activated, predictions prepare the individual for the kinds of sensorimotor projections that typically ensue. They do this by expediting *anticipatory responses*. Unexpected sensorimotor projections initiate *orienting responses*. They stimulate the revision of old predictions and the development of new ones in order to account for the unexpected projections. Since sensorimotor predictions are learned by means of sensorimotor projections from actual bodily interactions, they are able to model the relationship between individual and environment. Other individuals are an important part of that environment and certainly the most complex (see Cowley & MacDorman, 1995).

We shall next consider how this theory can be adaptively used to ground an *a priori* symbol system. Memory-based techniques in robots provide one possible route to its implementation.

Grounding an *A Priori* Symbol System

The propositions of a symbol system resemble the disembodied knowledge found in books (see Harnad, 1993). This contrasts with an individual's knowledge which is full of internal variables and relates to body and personal history. As with the words in a book, the symbols of a symbol system are ungrounded insofar as they have been *abstracted* from their relation to a particular person. Moreover, the symbols manipulated by most AI programs are not part of an autonomous system: they are not embedded in an agent that can cope with the world on its own (e.g. a robot). Hence, they are not required to mediate between sensory projections and motor signals. This is because the program has neither. The symbols the program generates are only intended for human consumption, and their meaning is determined by a user's interpretation (Harnad, 1990c). Although these programs cannot interact with the world through bodies under their direct control, their output influences their user's behaviour; and people have successfully applied them to a variety of endeavours from mineral exploration to medical diagnosis.

A robot, by contrast, must rely on its own sensors and actuators. If internal symbols can be functionally analogous to intentional states (as Fodor argues), in order to relate these symbols to the outside world, they ought to be abstracted from and grounded in the robot's own sensorimotor coordination. If we consider the difference between an agent acting on the basis of phenomena it has itself detected or instead acting on the basis of a description of that phenomena (perhaps one produced by its own perceptual system), intuitively, the difference may seem trivial. This is because we straightforwardly perceive a world populated by objects, and much of the brain activity involved in recognizing and responding to them is introspectively opaque. Being subconscious it can only be inferred.[38] We further take for granted that our similar bodies, sense organs, language, and culture make it possible for us to agree on what constitutes an object. But far from being trivial, grounding symbols in behaviour has proved unexpectedly difficult for robot designers. For example, it appears to be much easier to automate the task of deciding what chess moves to make (the 'intellectual' part) than physically moving chess pieces about the board.

Few programs, including programs that play chess, could be said to have the clear semantics of a symbol system. Internal symbols (in the computer science sense) often do not stand for external states of affairs, and indeed their semantics can be rather convoluted (Smith, forthcoming). Nevertheless, by Harnad's eight criteria (1990a, p. 336 and chapter 1), we may give a symbol-system interpretation to significant portions of a chess-playing program. The game has explicit rules for the abstract movement of pieces (as opposed to their physical movement on a board). The manipulation of internal representations is purely syntactic, and a significant subset of these representations may be interpreted semantically—for example, as standing for chess pieces, board positions, and chess moves.

To illustrate how learning can provide a well-adapted sensorimotor grounding to the symbols shared by a chess program and a piece locating system, let us consider in more detail an aspect of the hand-eye coordination task of moving chess pieces. A robot must move its six-jointed arm so that its hand is positioned to grasp a piece. For eyes the robot has two cameras, both trained on the chess board. If either its

---

[38] It has been noted that we cannot explain how we recognized someone's face with the same certainty that we can explain, for example, how we solved a mathematics problem.

hand or a chess piece becomes visible, feature detectors activate a proposition that contains the corresponding symbol and its location in camera-based coordinates. For example, the proposition *at (rook, cam$_1$loc (497, 325), cam$_2$loc (890, 194))* may signify that there is a castle centred at the position $x = 497$, $y = 325$ in the first camera's image plane and $x = 890$, $y = 194$ in the second camera's image plane. To give the symbol *rook* a sensorimotor grounding, the robot must be able to pick up the castle and move it across the board. The robot could move its hand into position by constantly monitoring the visual consequences of small changes in its joint angles, and using this negative feedback to correct for errors in the manner of a servo-mechanism. This method would be comparatively slow, however. If only the robot were able to know which joint angles corresponded to the object's visual location, it could make a beeline through the coordinate space of joint angles, thus positioning its hand in one swift and graceful movement.

Technically, the robot needs to be able to make a *coordinate space transformation* from points in the 4-D visual coordinate space[39] of the camera's image planes to points in the 6-D proprioceptive coordinate space of the robot arm's joint angles. In humans some neuroscientists have assigned this task to the cerebellum. Pellionisz and Llinás (1979) conjectured that Purkinje cells in the cerebellum were responsible for making this kind of coordinate space transformation by acting as a metric tensor.

In robotics the standard approach has been to compute the transformation according to a mathematical function that has been analytically derived from a set of equations describing the geometry of the robot's cameras and the kinematics of its arm (Paul, 1981). This kind of *a priori* mapping leads to robots whose behaviour lacks resilience to mishaps, such as a joint sticking or a camera being knocked askew. Omohundro (1990a) contrasts this rigidity with the adaptability found in nature:

> Biology must deal with the control of limbs which change in size and shape during an organism's lifetime. An important component of the biological solution is to build up the mapping between sensory domains by learning. A baby flails its arms about and sees the visual consequences of

---

[39] In the literature these coordinate spaces are also referred to as vector spaces, state spaces, and phase spaces. Sometimes more general non-Euclidean spaces are discussed such as dissimilarity spaces which have no fixed dimensionality.

> its motor actions. Such an approach can be much more robust than the analytical one. (p. 310)

Neither hard-wired transformations nor fixed motor patterns are flexible enough to provide a sensorimotor mapping for bodies that grow and change in often unpredictable ways. Following any significant alteration in kinematics, perseverative changes in a creature's nervous system are needed to effect adaptations in sensorimotor mappings—at least in performing complex or learned movements.

Considering the speed at which neural signals travel, servo-mechanical response is not fast enough to account for the fluidity of human movement. Taking a swing at a golf ball and many other kinds of movement are simply too rapid to rely on it. This is especially true of ballistic movements like throwing a ball (see chap. 9 of Carpenter, 1990). Were these movements to exploit any kind of feedback, it could *only* come from an internal mapping. This is because external feedback concerning where the ball has landed relative to the goal arrives after control of the ball has been relinquished. As Jeannerod (1994) makes clear, these mappings must be of considerable complexity to explain the resilience of human performance to environmental perturbations:

> The corrections in movement trajectory and/or kinematics in response to these perturbations can be so fast (usually within less than 100 ms) that they cannot be due to programming another movement based on feedback error detection. Instead, they have to rely on an open-loop adjustment of the ongoing program. This suggests that the central representation must be a "dynamic" structure, in the sense that it permanently monitors movement-related signals (e.g., proprioceptive) and compares them with the ongoing efferent commands. Any deviation arising from this comparison would immediately trigger corrections (see Prablanc et al. 1979; Jeannerod 1990; and for a computer model using the same principle, Bullock and Grossberg 1988). (p. 200)

Considering the relatively slow speed of neural pathways and the need to adapt to unpredictable sensory and kinematic change, sensory feedback and innate motor patterns probably play a relatively minor role in grounding perceptual categories as compared to patterns of behaviour developed in light of sensorimotor experience. We should not be surprised if all perceptual information is, to some extent, filtered through learned sensorimotor mappings. In this case sensorimotor learning must be a prerequisite for any kind of passive learning.

Evidence from patients with brain lesions supports a somewhat hierarchical framework for motor control in which the basal ganglia affects the initiation and perhaps the strategic planning of movements, the cerebellum their execution with reference to sensorimotor feedback, and the motor cortex the transformation from desired limb positions to motor signals governing muscular forces (Carpenter, 1990, pp. 300, 293-294). People who lack a working cerebellum cannot make rapid well-coordinated movements and must apparently rely on constant conscious monitoring of small motions.

These findings support Marr's (1969) hypothesis that, in learning a skill, the cerebellum automates consciously-controlled movements. His hypothesis is that, at first, neural signals from higher centres pass along climbing fibres through Purkinje cells to the motor cortex. Meanwhile, the sensorimotor context creates patterns of activation across parallel fibres. Gradually, presynaptic and post-synaptic coincidences in the firings of parallel fibres and Purkinje cells result in the strengthening of connecting synapses, so that the performance of movements no longer requires direct conscious control. Nonetheless, neurophysiological evidence for these mechanisms is inconclusive, and in some ways Marr's explanation may be too modular: it downplays the fact that knowledge of sensorimotor relationships must be consciously available in sensorimotor planning, perceptual recall, and dreaming.

Memory-based Implementations for Learning Sensorimotor Mappings

There is a broad class of algorithms for learning nonlinear mappings like those needed for sensorimotor coordination including, for example, feed-forward neural networks that learn connection weights by back-propagation (Rumelhart & McClelland, 1986). Indeed, connectionism has already been applied to learning in sensorimotor domains (e.g., Mell, 1988). However, Omohundro (1987, 1990a) has instead advocated the application of closest-point (i.e. geometric) learning algorithms to sensorimotor problems. They are easier to analyse and, for this kind of application, learn many orders of magnitude more quickly and accurately (at least when implemented on ordinary computers, also see Gross & Wagner, 1996). Methods abound for finding closest points (see Bentley, 1975, Friedman et al., 1977; Sproull, 1991; Tamminen, 1982; Ramasubramanian & Paliwal, 1992; Micó, Oncina & Vidal, 1994). There are even efficient methods of finding closest points in dissimilarity spaces where, as is the

case for phoneme recognition, the dimensionality of the instance and the categorical representation are likely to differ and the computational cost of comparing them is high. For simplicity's sake, however, discussion in this dissertation will be limited to Euclidean distance measures in ordinary vector spaces.

Clocksin and Moore (1989) successfully applied a closest-point algorithm to learning stereoscopic hand-eye coordination like that required for moving chess pieces. Clutching a pen light (to simplify the recognition of its hand), their robot moved its arm about. It recorded in a composite coordinate space the locations it visited in its visual and proprioceptive subspaces. If nearby points in its visual subspace had been reached by more than one set of joint angles, it would only remember the set that resulted in the least contortion to its arm. During a 'dreaming' period, the robot was disconnected from its arm and recorded further points by making linear interpolations from neighbouring points. To approximate the joint angles corresponding to a new visual location, the robot consulted the joint angles for the closest point recorded in the visual part of its coordinate space. Thus, previous sensory projections acted as categorical representations to divide up the coordinate space into Voronoi hyperpolyhedra (a special class of multidimensional convex hulls where the hyperplane boundaries that quantize the coordinate space are equidistant from the closest two points). If an object were recognized, the image plane coordinates of its location would fall into one of these hulls, and the corresponding representation would serve as the robot's prediction about the movements required to approach it. In a comparatively short period of time the robot learned to move its hand to within a centimetre of a given goal point without the use of an algorithm for either stereoscopic vision (like that proposed by Marr) or arm kinematics.

Even more significantly, coordinate space learning methods (of which the closest-point method is but one example) are highly general. There are only two assumptions implicit in the above model, and both of them generalize to other geometric problems. The first is *continuity* which indicates that nearby points in the domain of a mapping are also nearby in the range. The second is *smoothness* which indicates that the mapping can be approximated by local interpolation. The shortcoming of the current chess example is that a chess board shares little in common with the complexity of the real world. However, the learning techniques used here generalize to more intricate environments. As Clocksin and Moore note, the same class of techniques could have

been used if the robot were controlling wings or fins or viewing the world through reversing prisms.

Improving the Implementation

Sommerhoff and MacDorman's high-level description of how an individual could develop predictions (beginning with representations about sensorimotor relationships) does not immediately suggest a particular implementation. Nevertheless, we may appraise the suitability of an implementation in terms of a description such as theirs. We shall presently consider Clocksin and Moore's hand-eye coordinating robot. The suitability of their implementation will depend on many things that must, of course, be left out of the high-level description, for example, what computing elements are available and the intended sensorimotor domain. Unsurprisingly, implementations that model qualitatively different kinds of patterning (e.g. acoustic, proprioceptive) will require some degree of differentiation and modularity. Nevertheless, given some of the broad adaptive features of biological systems considered in this chapter, we may use Sommerhoff and MacDorman's description to appraise Clocksin and Moore's implementation and to suggest how it may be improved.

We can do this by assigning the functional particulars of the implementation an *intertheoretic interpretation* in terms of the description. In this way, we can evaluate the plausibility of the implementation vis-à-vis the description. Indeed, this process can be applied—in addition to implementations—to representational frameworks such as Fodor's language of thought. The more detailed the description, the more deficiencies it can potentially expose in a theory or implementation, so ideally it should be very detailed (containing as much low-level information as possible), and cognitive scientists should all be able to agree on it. This is, unfortunately, not realistic given how controversial representational theories in this field are. In what follows we shall analyse, in the context of our chess game, Clocksin and Moore's implementation of hand-eye coordination in terms of Sommerhoff and MacDorman's high-level account of how an individual represents sensorimotor relationships (§§3, 5, 1994).

The proposition about the rook's location on the robot's image planes qualifies as a sensorimotor representation because the sensorimotor processes that need to be appropriately related to the manipulation of the rook are appropriately related to the

proposition. Specifically, the processes that determine where to move the robot hand rely on the proposition's image plane location values for the rook. Also, the robot derives lasting benefit from sensing the visual and proprioceptive consequences of its arm movements insofar as they leave their mark in the form of corresponding sensorimotor predictions about the movements required to reach visual locations. Indeed, the core of the robot's behaviourally-grounded representations consists of these learned predictions. Hence, Clocksin and Moore's implementation certainly reflects the spirit of Sommerhoff and MacDorman's high-level description and, up to a point, conforms with the description. The robot, however, does not exhibit *orienting responses* to unexpected sensorimotor projections. They are generally accompanied by a shift in attention from ongoing activity to the unanticipated perception and, according to the present approach, lead to the formation of new predictions or the modification of insufficiently accurate predictions.

To account for orienting responses, the robot must be able to shift emphasis between learning, correcting, and using its coordinate space model as circumstances dictate. If the robot's predictions are sufficiently accurate for it to position its hand so as to move a piece, the robot need not augment or modify them. If not, the robot can, in the final stages of positioning its hand, resort to positioning it servo-mechanically by exploiting negative feedback from vision. (This appears to be what a person does in extending an arm to place a finger on a spot. The arm movement slows right before the spot is reached.) Once suitably positioned, the robot can then store in its coordinate space the new visual and proprioceptive position of its hand so as to fill in the gap in its sensorimotor model. It is also important that the robot be able to distinguish between a lack of predictions and predictions that are no longer sufficiently accurate because of bodily or environmental change. A lack of predictions can be corrected simply by further exploration so that predictions may be developed from previously-unencountered sensorimotor projections. However, insufficiently accurate predictions must be modified, amended, or forgotten. If the robot's arm or cameras are knocked out of alignment, this may necessitate replacing all predictions related to hand-eye coordination, and it may be more efficient simply to forget them and to start afresh by relearning the mapping.

Once amended Clocksin and Moore's implementation conforms to Sommerhoff and MacDorman's high-level description at least insofar as it concerns the robot's

sensorimotor coordination in mapping from visual locations to corresponding arm movements. (Piece recognition is another matter.) All the points in the robot's coordinate space constitute learned conditional predictions. Active predictions—those that have been elicited, revised, or sustained by sensorimotor projections—prepare the robot with anticipatory responses for sensorimotor projections that are likely to follow. In the present example, the recognition of a particular chess piece triggers the anticipatory response of the robot's arm movement; in turn, the arm movement triggers the anticipatory response of grasping the rook.[40] Unexpected projections initiate orienting responses in order to learn new predictions or modify insufficiently accurate predictions. If the arm is not positioned close enough to the rook, the robot must resort to positioning it servo-mechanically and then augment or modify its predictions.

I am supervising a project, currently at the simulation state, which extends the above enhancements to Clocksin and Moore's model. Instead of simply moving its arm to a point in space, the robot learns to predict the future position of a ball based on its position when it came into view of both cameras' image planes and again its position some milliseconds later. The arm moves a basket to that position to catch the ball.

Implications and Extensions

Reconceptualizing Visual Processing

One important feature of Clocksin and Moore's memory-based approach to hand-eye coordination is that no intermediary form of representation was necessary for the robot to map from vision to proprioception. The robot was able to integrate information from these two modalities without, for example, computing a depth map from binocular disparity. From our earlier discussion on sensorimotor integration, we

---

[40] These responses are rather simple and uninteresting because they do not involve interaction. Nevertheless, predictions may elicit anticipatory responses when two or more agents must coordinate their activity—for example, if one must catch a ball thrown by another during a game.

note that a depth map by itself is only grounded insofar as it serves (or could potentially serve) to coordinate the robot's activity. It would have required considerably more computation than Clocksin and Moore's memory-based approach, and it is unclear what it could have contributed because values for depths would still need to be mapped onto arm movements. The mapping is more straightforwardly performed directly from the image planes of the two cameras. In using a depth map, if the depth values were to be represented in some objective unit of measure, the robot would need to have a means of calibrating itself accordingly. A depth map would also probably make it harder for the robot to adapt to unanticipated change. Even were it to choose its units subjectively for its own convenience, if its cameras were knocked out of alignment, it would be difficult for it to compensate for the change in its camera geometry. Incorrect values would require it to relearn the sensorimotor mapping from scratch, thus offering no advantage over Clocksin and Moore's approach.

This is not to deny that intermediary forms of representation do at some point become useful or even necessary. However, there is little point in developing representational models without first considering how they are to mediate between sensation and action. Otherwise, we might develop models that, in practical robotics problems, would prove unworkable or superfluous.

Computational approaches often conceive of visual processing as operating on a static scene. The task is to mark out and label objects according to their abstract class (e.g. *chair*, *table*, *cup*). This information can then be converted into a symbolic description of the objects and their relations to each other which, in turn, can be used to make inferences. Contact between the visual systems and objects is limited to a stationary image. It is impossible to touch, hear, smell, taste, or in any way handle objects. These constraints and the totally noninteractive conceptualization of the problem naturally lead to an undue emphasis on *a priori* feature detectors. This is because, with such limited sensorimotor contact, it is virtually impossible for the visual system to discover what might constitute an object. But when people recognize objects they implicitly recognize the behavioural and physiological possibilities they afford. This is because multimodal interactions with objects enable us to distinguish them from their surroundings and, with time, to discover what they do and do not afford (see chapter 6).

A tendency to introspect verbally about what is in a scene and then to try to link up the symbols (words) used to describe it with the objects they stand for perhaps tempts researchers to take a short cut. It may tempt them to rely on *a priori* feature detectors and overlook the low-level sensorimotor dimension that may underpin much of recognition. Although it is patently true that adult humans can look at a static scene (e.g. a photograph) and recognize objects without moving in that scene, this ability may depend on much prior learning. At base this learning may need to be a sensori*motor* kind. The development of our powers of recognition may be closely tied to the relationship between movement and perception. The next subsection explores one way to conceive of this happening.

Multimodal Integration Facilitates Abstract Perception

In the hand-eye coordination example, sensorimotor projections from vision elicit predictions about proprioception. This kind of mapping is also possible within the same modality. By the coordinate space methods described, for example, a robot can learn to map between movement parallax and binocular disparity, or binocular disparity and vergence, or vergence and changes in apparent size, etc., or any combination of these. Information from these sources can also be used to set up predictions about the movements required to reach an object. Bodily movements can likewise be used to set up predictions related to vision. As mentioned at the beginning of the chapter, this has been observed and studied extensively in humans. For example, a person walking without vision from one location to another automatically updates the relative direction of surrounding objects (see Rieser, Guth & Hill, 1986). From past experience an individual develops a feeling about which way to turn.

This kind of behaviour can be modelled in robotics. A robot can learn to update, among other things, predictions about the direction it would have to turn to reach objects solely on the basis of proprioceptive information. In this way, a robot can use one source of sensory information to elicit predictions that are more directly related to a second source and, thereby, to compensate for a lack of information from that second source. In cases like this, the predictions would probably be less accurate; however, coordinate space models at least offer graceful degradation in predictive accuracy when fewer dimensions are available. (In the simplest closest-point model, this means fewer terms in the Euclidean distance measure.)

As mappings are learned between and within different sensory modalities, it is possible to see how predictions can be elicited by ever more indirect means of sensorimotor contact. For example, a robot might initially be obliged to follow the contours of objects with its hands and eyes to discern their shape and how they behave. Gradually, sensorimotor predictions could develop from it having touched and foveated points along the contours of objects. These predictions map the topography of the image plane to eye saccades so that it is no longer necessary to follow the contours of an object with hand or eye to elicit predictions about its shape. Instead its shape is recognizable from the still image. Eventually, sensorimotor predictions develop not only through self-induced movement but also by passively watching things happen.

If infants develop the ability to recognize objects passively only after developing rich predictive maps integrating the multimodal sensory consequences of motor actions, this may suggest that robots should do the same. In this way they could learn perceptual categories that can ground their symbols. Object recognition would certainly be more robust if the robot had access to such maps: If the robot could not categorize an object with certainty by passively looking at it, it could try foveating along its edges or moving around it. This activity would probably lead to the confirmation or violation of its active predictions and would sustain and elicit further predictions. There should be little doubt that having integrating maps can contribute to recognition. The most adaptable (and perhaps easiest) way to have them is to learn them. This learning may be best modelled in terms of the development and revision of sensorimotor predictions. An understanding of the process by which robots can learn these predictions may help to diminish their reliance on *a priori* feature detectors.

Multimodal Integration Facilitates an Object-Centred Frame of Reference

Predictions about how bodily movements transform a viewpoint-dependent representation of an object (or configuration of objects) can in fact serve as a viewpoint-independent representation of that object. *By themselves*, the sensory projections of an object as viewed from a particular perspective can say little about either the projections of that object as viewed from another perspective or how motor actions could be related to that object. However, a viewpoint-dependent representation *plus* predictions about how motor actions transform its projections make it possible to

predict the appearance of the object from other angles and distances. Thus, they together provide a viewpoint-independent representation.[41] From this we can see that sensorimotor predictions concerning how motor actions transform an object's sensory projections indeed facilitate the formation of an object-centred (allocentric) frame of reference from a viewer-centred (egocentric) frame of reference (see Feldman, 1985, for background). This is because predictions concerning transformations caused by self-induced movements can be used to compensate for those movements, for example, in visualizing activity from an object-centred perspective.

The Need for More Realistic Environments

So far we have only increased the adaptability of the robot's hand-eye coordination. Yet although the robot can now adaptively adjust to changes in its visual geometry and arm kinematics, its feature detecting mechanisms do not improve with practice. (This could be remedied, for example, through the learning of categorical representations as described in the last chapter.) Moreover, the robot can only classify objects as belonging to a finite number of specified categories. This is no disadvantage in playing chess because there are only six kinds of pieces and two colours. But the real world is more complex. To adapt to environmental change, it may be necessary to become sensitized to new kinds of sensorimotor patterning.

Plainly, there are a number of ways in which playing chess is not like getting about in a natural environment. For one, the fundamental objects of chess (its abstract pieces, board positions, legal moves, etc.)—on which all chess players agree—are not representative of all perceptual categories.[42] Perceptual categories probably vary somewhat between individuals—and certainly between different species. I have suggested this be explained in terms of their reliance on sensorimotor mappings developed in the course of a particular individual's life.

---

[41] This representation need not rely on any intermediary form of representation, such as a representation of the object's shape in allocentric (object-centred) coordinates.

[42] Chess players may, of course, disagree about other matters such as whether the pieces are situated in a particular strategic arrangement, the significance of that arrangement, and what term should best describe it.

There are other limitations about the domain of chess that, unlike physical environments, make the game amenable to symbol systems. Unlike real environments, chess only requires the recognition of a few objects which are known in advance: six kinds of chess pieces of two different colours, and their position on the chess board. Likewise, abstract piece movement is constrained by the rules of play which do not change. Furthermore, a chess program can get away with being solipsistic. A human can directly interpret the meaning of its output so the program does not need to connect its symbols to actual pieces and board positions (see Harnad, 1990c). However, even in a highly constrained domain like chess, if a chess program had to recognize and move actual chess pieces, an *a priori* grounding of the kind provided by feature detectors may not be flexible enough to recognize all the different kinds of chess sets with which it might potentially be confronted—and this would be a good place to start applying the methods of the last section.

Chapter 6 is intended to illustrate an adaptive approach to discovering classes of objects. The next chapter extends the approach taken here to learning sensorimotor predictions that can be applied to the planning of action sequences.

Conclusions

Human beings exhibit a general capacity to adapt: we can learn to eat with chopsticks, ride a horse, play tennis, or feel our way through darkness with a cane (from Bateson, 1979), without need for the genes of our ancestors to have been selected specifically for these purposes.

In pondering the nature of representation, theorists often consider sensory modalities in isolation. By overlooking sensorimotor integration, they place an undue emphasis on peripheral sensory processing (which, because of its proximity to the sense organ, is specific to that modality alone). This results in theories that are too optimistic about what peripheral sensory processing can do. This bias may also lead to theories that exaggerate the importance of *a priori* feature detectors. This is perhaps because, with impoverished data from only one sensory domain, it is very difficult to discover relevant features empirically. Multimodal interaction with objects enables the perceptual system to sensitize itself simultaneously to the

sensorimotor projections of the objects and to the behavioural and physiological possibilities they afford.

Through the empirical learning of conditional sensorimotor predictions, it seems likely that one may simulate many aspects of intelligent behaviour. According to the predictive theory put forward, active predictions form the nexus of an organism's grounded sensorimotor representations. They relate motor signals to their sensory (including physiological) consequences. Since they are learned empirically, unlike *a priori* feature detectors they reflect bodily and environmental change. It is necessary for an organism to learn an internal mapping of this kind to explain the adaptability and fluidity of coordinated movement. As Jeannerod (1994) has shown, the use of negative feedback in closed-loop control requires too much time to explain the speed with which people can adjust their movements to environmental perturbations. Open-loop control requires access to a mapping—a mapping which to some extent must be learned empirically. According to Jeannerod (1994) this "central representation" would need to have a dynamic structure in order to compare proprioceptive and motor signals and respond in accordance with their changing relationship.

Predictions may allow an organism not only to anticipate the effects of its actions but also to model mentally its relationship to its surroundings. The role of *a priori* feature detectors, however vital it may be, is probably at the periphery of this process.

It is relatively easy to simulate an organism's learning of sensorimotor predictions by means of any number of algorithms for learning nonlinear mappings including those used by artificial neural networks (e.g. backprop). Especially robust is the closest-point technique used by Clocksin and Moore (1989) and Omohundro (1990a). By exploiting the geometric properties of continuity and smoothness, this learning algorithm can be applied to any sensory domain.

Sensorimotor categories are grounded in coordinated activity. There is some evidence to suggest that these categories precede and, to some extent, subserve abstract reasoning. Wason (1981) tests have demonstrated that people can more easily solve many kinds of abstract problems once they have been redescribed in concrete settings. This suggests that a sensorimotor context can aid deduction. Could even logical reasoning partly depend on the elicitation of sensorimotor predictions?

(Some research has indicated that, insofar as thinking is logical, learning is important: sociocultural institutions like schools are influential, see Scribner & Cole, 1981, pp. 126-128; Wertsch, 1991.) If in thinking we exploit sensorimotor transformations that are analogue in nature and, for the most part, learned, then thinking would appear to depend on more than what could be explained by an *a priori* symbol system.

This chapter proposes a theory which claims that learned predictions can model the integration of multimodal sensorimotor information. Early in cognitive development predictions may integrate information related to the same sensory source or modality. Later they may integrate information from various sources and modalities. This is possible because earlier predictions provide scaffolding for later ones to develop. In this way predictions initially identified with one sensory source or modality come to elicit predictions related to other sources and modalities. As more abstract and indirect mappings are built up, predictions may be elicited by ever more indirect means of sensorimotor contact. Thus, our sophisticated ability to recognize objects passively may in part be rooted in active sensorimotor learning. If this can be shown then it would confirm that all perceptual categories do not need to be triggered by *a priori* feature detectors.

Chapter 5. Adaptive Sensorimotor Route Planning

Introduction

In this chapter we examine how an autonomous robot can learn about its sensorimotor dynamics from experimentation. As in the preceding chapter, the robot empirically learned predictions about the sensory consequences of motor signals, and these predictions served to model its relation to its world. However, in the last chapter we focused on how the robot could learn a mapping between sensory modalities. We specifically considered how Clocksin and Moore's robot learned proprioceptive predictions about visual positions. The control part was straightforward and only depended on this mapping: the robot moved its arm into position by making a beeline through its proprioceptive coordinate space.

In many domains movement is not so simple. For example, it may be necessary to compute paths that circumnavigate obstructions. (This is usually called *route planning*. Of course, there does not need to be anything consciously thought out about this planning.) In this chapter we shall consider how a robot can learn sensorimotor predictions about the dynamics of a remote controlled car and use these predictions to plan sequences of movements in cluttered environments. In experiments with a simulated car, the robot developed generalizations about its movements, and used this information to successfully plan paths through uncharted areas of its state space. In the next chapter, we will consider more abstract forms of planning (as well as how a robot can learn to recognize objects without advance knowledge).

We first begin by considering rudimentary models of learning in which a robot only takes a single action before evaluating the success of its behaviour. To be more precise, given its current sensory projections from external sensors **e** (e_1, e_2,... e_n), it produces motor signals **m** at one time step and then appraises the outcome **o** at the following time step. In this simple example **e**, **m**, and **o** are not vectors but scalars, and together (*e, m, o*) determine a point **p** in a *sensory-motor-outcome* coordinate space.

We first consider learning by trial and error and then examine how the number of trials may be minimized by exploiting consistencies: for example, the closest-point method can be effective if sensory projections and motor signals that are closer in the sensory-motor subspace are more likely to lead to outcomes that are more similar in the outcome subspace than sensory projections and motor signals that are further apart. We shall also investigate the criteria the robot may use to select motor signals and the trade-off between taking the time to enhance its sensorimotor predictions (which model its relationship to its surroundings) and making immediate use of them. That is, the trade off between exploration and exploitation. We then explore how the robot could cope with a nondeterministic environment and how apparent nondeterminism can actually be caused by an inadequate suite of sensors or the inability to recognize an underlying regularity.

Finally we examine how the robot may properly scale the dimensions of its coordinate space—that is, the range of values produced by each sensor—in order to improve behavioural prediction, to reduce noise, to eliminate behaviourally redundant or irrelevant sensors, and to uncover regularities that can be used to make predictions about the consequences of motor signals—even when the robot is in a novel perceived state, having never detected physically similar sensory projections before. The capacity to make accurate predictions by making generalizations about sensory projections that are apparently very different equips the robot with sufficient knowledge to successfully plan out sequences of motor signals without the need of having previously tested each sensory-motor combination.

A Preliminary Example

In an introductory example, Clocksin and Moore (1989) explored the use of an adaptive model for solving the task of getting a car safely across a street while another car is approaching. A common solution would be to apply standard control theory in order to devise a mathematical model of the dynamics of the system including the performance of the cars involved. The usefulness of the model, however, would depend on how accurately this idealization represented the real world. It would, therefore, be sensitive to problems of measurement and could not adapt to changes in the environment or the functioning of its sensors or mechanics. Clocksin and Moore

sought rather to design a robot which does not require an *a priori* model but can adapt to its environment regardless of variations in the car's sensing or performance.

The robot is given the velocity and location of the car on the main road when it first appears. These numbers index the fixed amount of acceleration it will use. If there is a collision, the value that was indexed is randomly replaced with a new value. Eventually the robot learns to cross the road. The robot achieves somewhat better performance with smoothing—that is, resetting untested or unsuccessful indexed entries to successful neighbouring entries. Still better performance was achieved by using the "Boxes" method (Michie & Chambers, 1968) where the likelihood that an action is selected is proportional to its estimated probability of success.

The robot learned slowly because it can only determine the effect of its motor signals in light of previous trials that were very similar. If the nature of the robot's environment is such that, for any given motor signal, the mapping from the sensory-motor subspace to the outcome subspace is fairly flat and continuous, nearby points in coordinate space are likely to produce similar results. The closest-point method (Friedman et al., 1977) exploits this underlying consistency: the robot dispatches motor signals that were successful for a nearby point in coordinate space (Clocksin & Moore, 1989; Moore, 1990; Omohundro, 1990a). Of course, this method will be less effective in regions characterised by sharp boundaries, because the same signals will produce very different behaviours for nearby points. Moore proposes the use of a multidimensional tree (Bentley, 1975) to record the previous attempts of the robot, because it facilitates the rapid look up of a point's nearest neighbour. Its complexity is order $2^k \log_2 n$ where k is the number of dimensions and n the number of past trials. The exponential k term posed no problem in experiments using up to twenty dimensions.

In a practical system we would want the robot to be able to modify the speed of the car while it is passing through the intersection so that it could adjust to changes in the speed of the other vehicle. It might also be useful for the car to know how much time it is taking to pass through the intersection or the amount of fuel it is consuming, in order to optimize its behaviour. Means of addressing these issues in situations where sequences of motor signals are required will be discussed later in this chapter.

Should the gazelle stop to check if the rustling in the bushes is a predator or should it immediately flee? Living creatures are able to balance the need for updating and embellishing their model of the world with the need for exploiting it, because those that could not perished with their genes. The question of exploration versus exploitation is similar to the *n*-arm-bandit problem (Jervis, 1992). Should one risk losing money in order to be more certain that one is pulling the arm of the slot machine with the highest expected pay-off? Of course, animals developed the ability also to balance the priorities of maximizing long-term pay-off with those of short-term pay-off. If a squirrel squanders all of its time surveying its habitat, it may starve before it can take advantage of its efforts. People have used decision theory to balance the cost of gaining more information with its utility (this has been called *bounded rationality*, see Russell, 1991). For example, we would like the robot to be able to determine based on the goals we set out for it whether it should explore boundaries between regions where different behaviours are predicted or avoid them. The robot should be able to determine, for example, whether to weight newer information more heavily than older information.

Moore (1990, chap. 8.11) uses a probability distribution to ensure that previously successful motor signals (and those most similar to them) would be attempted more often than other motor signals, although all would have some chance of being tried.[43] The best candidates are those reliably predicted to be successful. The next best are those unreliably predicted to be successful. The worst are reliably predicted to be unsuccessful. To illustrate let us again assume that our robot is trying to guide a car across the street given that another car is approaching with a certain speed and position. Furthermore, let us assume that the robot may choose among ten actions: accelerating the car between one and ten feet per second. If the car had safely crossed

---

[43] Moore called these motor signals *actions*. This is an unfortunate use of language because it blurs an important distinction. Motor signals cannot be equated with physical movement because, although they influence it, they do not determine it. The environment also comes into play. For example, when animals copulate, the behaviour of one is not the direct result of its motor signals but depends in part on the movement of its mate. It is also important to distinguish physical movement from perceived behaviour. Much of what animals do goes unobserved because of the limits of our own sense organs and the way we categorize and conceptualize their input (Gordon, 1992).

the intersection with accelerations two and nine already, but was unsuccessful with four, and if the robot is trying first to maximize safety and then minimize time across, it would be more likely that the robot would attempt accelerating eight or ten feet per second rather than three or five. Of course, the robot would face more complicated choices in a real driving situation, because there are other dimensions to control such as steering.

What if the driver of the other car sometimes spotted the robot controlled car in time to avert an accident? Unmodified the performance of the closest-point algorithm would be impaired by this external stochastic process. There would be a region of coordinate space in which a motor signal may or may not lead to an accident. Because the success or failure of a point in this region depends on the other driver's behaviour, the fact that a motor signal had been successful when the car was in a nearby location is a poor indicator of whether it will be successful again.

Ideally, instead of marking these points as successes or failures, they would be marked as potentially being either. If success and failure are represented by real values, the expected outcome of taking a given action in this region could be represented by an average of the utility of the n closest points to the current sensory-motor point being considered, weighted according to their proximity to it. Hence, the expected utility of the outcome would be somewhere between the values for success and failure. If the robot always tried to choose the best action possible, it would then learn to avoid this area. However, there is quite a difference between a point that is consistently mediocre and one that could be successful but is potentially fatal. An animal's reproductive success, for example, might sometimes be enhanced if it acted to ensure subsistence rather than taking unnecessary risks in the hopes of greater gains. At other times the opposite might be true. Therefore, it may also be useful for the robot to calculate something like a weighted standard deviation. This information could be exploited by a well-adapted control strategy in light of changing circumstances.

One difficulty with any method of smoothing the coordinate space is that the robot could be smoothing out important detail. How does it distinguish a very convoluted area of coordinate space from a nondeterministic one? This might be revealed by an attempt at optimizing the weighting of the average, trying, for example,

steeper Gaussians. Another consideration is that different areas of the coordinate space could require different amounts of smoothing.

Apparent nondeterminism can actually be caused by an inadequate sensor suite or the inability to recognize an underlying pattern. It may therefore be possible to reduce it by incorporating another sensor input. In the street crossing problem, for example, the effect of the behaviour of the other driver may not be genuinely nondeterministic. If the robot were able to sense that the other driver had braked, the nondeterminism would disappear.

Given the current sensory projections, there are many methods of predicting more accurately the outcome of motor signals beyond deferring to the closest point in the sensory-motor subspace. The nature of the sensorimotor domain will determine the relative merits of each method. Prediction can be improved by scaling the dimensions of coordinate space so that the proximity of a pair of points in the sensory-motor subspace better indicates their proximity in the outcome subspace. Scaling may also be necessary to put the ranges of relevant variables within the operating ranges of computing elements (for example, neurones). Not only may different dimensions be scaled relative to each other but different regions within a dimension may also be scaled relative to each other. The next chapter discusses how local linear interpolation can perform both scaling and interpolation based on nearby points in the coordinate space. This approach is very flexible and can be used after or in place of global scaling.

A robot's sensory projections may be determined by information from a number of sensors. Sometimes a sensor may be providing information that is superfluous because it is redundant or because it is unlikely to have any bearing on behaviour. It may also be necessary to discount a malfunctioning sensor if it is producing more noise than useful information. Discovering the best scaling of dimensions means knowing which sensor inputs should be weighted more heavily. Of course, the computational cost of finding a closest point will increase with each new dimension added to the coordinate space and will depend on which closest-point method is being used. (For example, although the cost of adding new points to the coordinate space is very low for a multidimensional tree, growing logarithmically, it is very high for adding new dimensions, growing exponentially, see Friedman et al., 1977). If a sensor is not contributing any useful information, it would be best to ignore it all together.

However, if it contributes some information, it may be worth considering whether the information it provides warrants the added computation.

Let us now formalize this. The closest-point method would perform optimally if the distance between nearby points in the sensory-motor subspace were proportional to their corresponding distance in the outcome subspace. Intuitively, this means that, the more similar the robot's sensory projections and motor signals (after scaling), the more similar the outcome. Let $\mathbf{s}$ represent a point in the sensory-motor subspace of dimensionality $|S|$, where $\mathbf{s} = (s_1, s_2, \ldots s_{|S|})$, and o represents a point in the outcome subspace O, where $\mathbf{o} = (o_1, o_2, \ldots o_{|O|})$. Then for all points i, j where $i \neq j$ and $|\mathbf{o}_i - \mathbf{o}_j| < \chi$ (sufficiently close to be considered *nearby*), we would like $|\mathbf{s}_i - \mathbf{s}_j| \propto |\mathbf{o}_i - \mathbf{o}_j|$. Unfortunately, it is highly unlikely that this is attainable. We can, however, scale the dimensions of the sensory-motor subspace with a weight vector $\mathbf{w}$ to minimize the overall disproportionately between the subspace S and the subspace O. We, therefore, wish to solve for $\mathbf{w}$ such that

$$\min \sum_{i=1}^{n} \sum_{j=i+1 \& |\mathbf{o}_i - \mathbf{o}_j| < \chi}^{n} \left(\sqrt{\sum_{d=1}^{|S|} (\mathbf{w}_d (\mathbf{s}_{i,d} - \mathbf{s}_{j,d}))^2} - \sqrt{\sum_{d=1}^{|O|} (\mathbf{o}_{i,d} - \mathbf{o}_{j,d})^2} \right)$$

This equation can be minimized using Powell's method (Brent, 1973, chap. 7) or the downhill simplex method (Nelder & Mead, 1965; Press et al., 1992).

If a sensory input dimension d provides no indication of the robot's likely behaviour, $\mathbf{w}_d$ should be close to zero. In this case we can drop the dth dimension. We may also wish to drop it if it is very small relative to the other components of $\mathbf{w}$. We may compute some ε below which the value of the information contributed is not sufficient to justify the added complexity of using the dimension. If two dimensions provide exactly the same information, removing one will double the weighting of the other. By selectively removing dimensions it is possible to discover what minimal set offers adequate predictive power. A less computationally intensive approach which should offer reasonable performance for most applications is to eliminate sensory dimensions that are either not correlated or very weakly correlated with the outcome. Dimensions can then be scaled using linear least-squares minimization. Dimensions that most strongly influence the outcome will have larger coefficients. As the same variation in distance in these dimensions will have a larger likely impact on the

outcome than in other dimensions, closest-point prediction can be improved by using these coefficients to weight the distance function so that distances are extended in these dimensions relative to less influential dimensions (see the next chapter for a more detailed explanation).

Route-planning Experiments

Discovering the Building Blocks of Plans

Our sensory projections are in constant change, and one moment is never quite like the next. Nevertheless, we and other vertebrates are able to develop from experience predictions about the consequences of behaviour. This feat would require an astronomical amount of time were our brains not able to generalize. In the last section we discussed how a robot could improve the accuracy of its predictions about the outcome of its motor signals. In the remainder of this chapter, we will consider how a robot can develop predictions that specifically concern the effect its motor signals have, from moment to moment, on its sensory projections from external sensors. These predictions can be used to model transitions in the robot's perceived state. In experiments in which a robot learns to control a simulated car, we find that prediction can be improved by scaling the dimensions of a sensorimotor coordinate space so as to enhance the predictive accuracy of the closest-point mapping (at time t) **perceived state**$_t$×**motor signals**$_t$→**perceived state**$_{t+1}$. This permits the robot to form generalizations that leverage on consistent sensorimotor regularities.

We shall also consider how learned sensorimotor predictions can replace an *a priori* mathematical model in domains where it is necessary to plan a sequence of actions. We shall be looking specifically at the task of route planning. In experiments with controlling a simulated car, the robot used these predictions to plan, among other things, a three-point turn. This is an example of the robot finding the best route when constraints have been placed on its mobility. The robot needed to be able to predict the consequences of its motor signals for many intermediate states on its way to a goal state. This is useful because there will be times when it is not feasible for the robot to test in advance these unexplored regions of its coordinate space before

planning a course through them. In the experiments, the robot avoided this by forming generalizations on the basis of past trials.

My aim in the route-planning experiments was clear: We would like the robot to be able to form, on the basis of a minimal number of trials, generalizations that are sufficiently accurate for it to be able to plan with them. For example, if the car is at a certain location with a certain orientation, speed, and steering position, the robot should be able to predict approximately where it will be located in the next interval if it applies a certain force to the brake, accelerator, or steering wheel—even if it has never visited that location before. The robot should be able to discover that the current speed of the car, the amount of pressure placed on the accelerator, and the gradient of the street largely determine its speed in the next interval. (If the remote controlled car is kept to a flat surface—as was the case in the experiments—then the gradient can also be discounted.)

Of course, the car could conceivably encounter a street where its past generalizations do not apply (perhaps the street is covered with ice or it has a gradient while the roads it was trained on did not). In this case its previous predictions would need to be modified, augmented, or forgotten and replaced by new predictions that were conditional on the relevant environmental change as sensed. Also if the dynamics of the car itself change (say, if the tires go bald), old predictions will fail to forecast changes in its sensory projections, and this will signal the need for corrective measures. The experiments were set up in two stages. In the first half, the robot experimented moving the car to develop a sensorimotor model. In the second half, it used the model to plan. Ideally, however, these steps should operate in tandem—and this is a good place to begin future research.

The same method of scaling dimensions that we used before to increase the predictive power of the closest-point method were used here to extract sensorimotor regularities to plan in novel circumstances. We may characterize an environment which requires a sequence of motor signals as **perceived state**$_t$×**motor signals**$_t$→**perceived state**$_{t+1}$. Therefore, an **outcome** can simply be thought of as a state transition **perceived state**$_{t+1}$ – **perceived state**$_t$. The robot computed a set of weight vectors $\{\mathbf{w}_1,\ldots \mathbf{w}_d,\ldots \mathbf{w}_{|O|}\}$ for each of the possible $|O|$ outcome dimensions. Since, in the experiments, the location and orientation did not affect the car's speed in the

next interval (obstacles notwithstanding), after scaling the weighted vector associated with the *change in speed* outcome had a value at or very near zero for its two components that scale the location and orientation dimensions of the coordinate space. Therefore, to plan the robot was able to apply what it has learned about regulating speed at other regions of the coordinate space—those associated with different locations and orientations—to unfamiliar areas which must be crossed on the way to its goal. This form of inductive generalization permitted more accurate predictions about the consequences of motor signals with far fewer trials.

Using Dynamic Programming for Optimal Plans

In this section we will consider models that have been used previously in various domains for path planning given an *a priori* mathematical model of the robot's dynamics. The findings of my experiments with a remote-controlled car show that, after a period of experimentation, planning may be performed using acquired sensori-motor predictions and that this period may be greatly reduced by scaling the dimensions by the methods discussed above. The position, orientation, velocity, acceleration and lateral acceleration of the remote-controlled car were originally to be sensed by means of an overhead camera with the outline of the car marked out in reflective tape. Obstructions (e.g. the edges of the road) were marked out in the same way. Unfortunately, I had to move the experiment to computer simulation because the particular car I was using was too quick to control well and because the controller's crystal gradually began to detune after I had soldered together a computer-interfacing box.

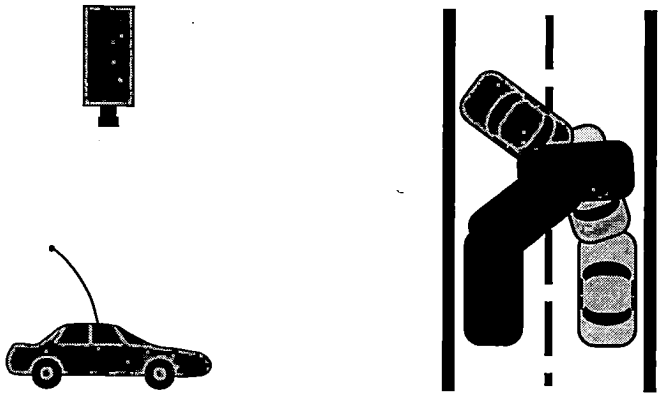


Fig. 5.1: First, the robot collected data in a sensorimotor coordinate space about how its actions affect the position, orientation, and velocity of the remote-controlled car. It then rescaled the dimensions of the coordinate space to eliminate uninformative dimensions. (This is a kind of inductive generalization.) For example, having experimented with the car on a flat surface, the robot was able to ignore the position dimensions in predicting its velocity in the next time step. Second, the robot exploited its sensorimotor model to form optimal plans by means of dynamic programming. Dynamic programming involved backward chaining from the goal to the initial state while maintaining only optimal subpaths in a quantized coordinate space.

A brute force way to find an optimal path might have been to begin with the initial state and search by hypothetically trying various accelerations and turning torques until the goal is reached. In a depth-first search, provisions would have to be made for the robot to keep the car from returning to a position it had visited before, thereby getting itself caught in a potentially infinite loop. Moreover, when the robot does find a path to the goal, there is no guarantee that it will be the most efficient path. These problems can be avoided with a breadth-first search (or a depth-first search with progressive deepening). There are numerous problems with this method, not the least of which is exponential growth in the size of the search tree as its depth increases.

One way of limiting this growth is to eliminate all suboptimal subpaths, and hence all suboptimal paths. For example, if there are several ways of getting from one intermediate point to another, the robot should only retain the path which spans the two points in the least amount of time (fewest time steps). (If several do, it should choose only one path to keep.) Now, when it is ready to search further, it will have eliminated all but an optimal subpath. Of course, it is unlikely that two different

plans would take us to exactly the same point, so it would be impossible to get much leverage out of employing this technique by itself. That is why it is important to make these calculations with some degree of coarseness. The planning coordinate space may be divided into hypercubes, with all points that fall within the same hypercube treated as the same point.

In the last subsection the robot used a sensorimotor coordinate space as a predictive model for the consequences of its actions. Here it uses an additional data structure, a quantized coordinate space, for dynamic programming. It fills in hypercubes of the new planning coordinate space with values indicating transitions in the state of the car based on its actions by virtue of the now scaled sensorimotor coordinate space.

In the experiments, the above technique proved sufficient for planning a number of simple manoeuvres. Some of them, like three-point turns, required the car to at first move away from the goal. Thus, they could not have been solved by servoing or by simple forms of means-end analysis. All the solutions were optimal to within the accuracy of the coordinate-space quantization.

So far we have only discussed forward chaining. However, the search algorithm I implemented branched out from the goal, trying all possible actions in reverse. This is the same as asking what action could produce the outcome of being in the goal state (or, in our coordinate space representation, the goal hypercube). The robot used its sensorimotor model to determine which hypercube in the coordinate space the car must be in in order to get the goal. Using this method, it was possible to backward chain to all hypercubes leading directly to the goal, and then to all hypercubes which could lead to those hypercubes that were one step from the goal, and so on, until the initial state of the car was finally reached. (The algorithm could have forward chained just as easily. As we shall discuss shortly, a combination of forward and backward chaining can lead to huge efficiency gains.)

Of course, the algorithm had to ignore unacceptable possibilities, as when backward chaining took the car outside the coordinate space or on to an obstacle. At the start of planning, all hypercubes were initialized to infinity. When the search algorithm reached a hypercube, this figure was replaced with the number of time steps it took for the robot to reach it. It immediately pruned away multiple paths by only

keeping track of the first action to reach a given hypercube. The algorithm then set a pointer in the newly reached hypercube to the originating hypercube. Once the algorithm reached the initial state, it simply followed the pointers forward from hypercube to hypercube until it reached the goal, accumulating a list of actions along the way. This list constitutes the robot's plan for getting from its initial state to the goal.

If we wished to forward chain from the initial state to the goal state, we would have followed a similar procedure: For every hypercube the car visits in its coordinate space (which includes its position, orientation, acceleration, and lateral acceleration), we can first store the forward and lateral acceleration responsible for taking it there. If we begin this process from the initial state then, once the goal is found, it is easy to trace the optimal path back to that state. Because the algorithm only stores optimal subpaths, it will be the only possible path between them.

Obstructions were marked out as inaccessible car positions in the coordinate space regardless of the values for other dimensions (acceleration, lateral acceleration, etc). One advantage to forward chaining is that, if all accessible paths in the coordinate space are charted from a starting point, it is a trivial calculation to find the optimal path to any accessible hypercube. If we have not yet selected a goal, we can use the forward search to give us information about what regions of the coordinate space are accessible and how efficiently the car can reach them. This information can then help the robot pick its goal. The method used in the actual experiments exploited the opposite advantage conferred by backward chaining. Since the search was conducted in reverse, starting from the goal, the robot simultaneously discovered how to reach the goal from any arbitrary initial hypercube. Of course, both methods could have been combined. But for the simple problem of getting from one initial state to one goal state, this would be redundant. It results in roughly double the computation.

If we are only concerned about computing the optimal path from one initial position to one goal position, there is a method for greatly pruning the search: we can search forward from the initial position and backward from the goal position simultaneously. When the two searches meet at the same hypercube, we know we have found an optimal path. This effectively transforms a problem of order k^n computations to a problem of order $2k^{n/2}$ computations, where k is the number of dimensions and n the length of the path.

Even when using dynamic programming (Larson, 1978; Smith, 1991), as the size, resolution, or number of dimensions of the coordinate space increase, the amount of search required still grows exponentially. However, dynamic programming can often be used to extend the horizon beyond which search is no longer feasible. Although the curse of dimensionality is not avoided, for many applications, it may be sufficiently postponed so as to make them tractable.

There are a number of ways to represent the quantized coordinate space for planning. In the above experiments, the robot used a very large multidimensional array. The array was indexed according to the location, orientation, steering position, and speed of the car. This information was used to look up the optimal forward or lateral acceleration to get to another hypercube. The quantization was sufficiently fine so that there was always enough forward and/or lateral acceleration to get the car out of the current hypercube. Were this not the case, the time interval would need to be extended to lengthen these paths. This is of particular importance with the dynamic programming of variable resolution tries because there is great variation in the side of the hypercubes (Moore, 1991; also see Moore & Atkeson, 1995).

The disadvantage of using large arrays is that the data structure could be much larger than necessary, if, for example, much of the space is inaccessible. This was not the case in the above experiments. It could also be slow if the array is stored on pages of virtual memory that are frequently being swapped in from disk and out to it.

An alternative is to store points in a kind of fast lookup table (i.e. a hash table) that places points that are sufficiently close together in the same location. The overhead of using a hash table should not be significantly greater than that of an array: because the access time for both is also $O(n)$ (polynomial time).[44] Although we may need to dynamically increase the size of a hash table during the process of calculating a solution, we would also need to dynamically increase the size of the multidimensional array if we wanted to increase the size or resolution of the problem space. A

---

[44] For example, n-dimensional arrays are often represented by one-dimensional arrays internally. This means that every time the array is accessed, each index must be multiplied by a constant and summed to compute the offset for the one-dimensional array. It is also possible to represent n-dimensional arrays as an $n - 1$ dimensional array of pointers, but the access time is still $O(n)$.

special function would be needed to access the array (with its own overhead) in order to make these changes in size and resolution possible. Hashing has the added convenience that one only need store the locations that one uses: not only are inaccessible locations never stored, but neither are out of the way locations. By this I mean that once the robot forms a plan in which the goal is reached, all the points that lie beyond the goal need not be stored.

Adapting and Optimizing Previously Successful Plans

Given a starting state and a goal state, dynamic programming can be used to determine the optimum path a car may follow. The question arises how one might best cope with a similar but not identical situation. Dynamic programming is more efficient than an exhaustive search, but even so it may require too much computation to be effective for real time situations if the state space is very large. Therefore, it may be useful to store plans for future use. These plans could either be indexed based on the perceived features of the world weighted according to their relevance, or all plans could be executed simultaneously and those which fail because they are not legal, safe, or possible under the current circumstances could then be rejected. Depending on the cost of failure, sometimes it might make the most sense simply to try out a plan and then revise it when it does not work. If we are working under the constraints of a sequential machine, in order to increase the efficiency of the implementation, the indexing system could rank the plans according to which had been most successful in similar circumstances. They could then be executed in the mental model until one can be found which sufficed.

But what if no plan can be found that describes a path that fits perfectly? A new path may be either calculated using dynamic programming or else interpolated or extrapolated from the paths described by one or more existing plans.

If a nearly sufficient plan is found, one method of reducing the computation required to find a better plan would be to apply dynamic programming only to region of coordinate space within a certain distance of the path suggested by the original plan. Under this scheme if the new path at some point is as far from the original path as is permitted, then it is unlikely to be optimal. So if the new plan is not good

enough, the process may be repeated, only with dynamic programming being applied to the region about this new path.

Moore describes a similar method in which dynamic programming is applied to a coordinate space at a coarse resolution and an attempt is made to find an optimal path (Moore, 1991). Then the resolution of the area about this path is increased, and dynamic programming is again applied to the space. The difference is that rather than limiting computation by applying dynamic programming to only the area around the best mental trajectory computed so far, it is applied, at an increased resolution, to the region of all the previous trajectories attempted.

This method of searching for a better plan in the area about a plan that already works is a form of hill climbing. It is only guaranteed to find a solution that is locally optimal. If we begin by starting with a three-point turn, we may find a three-point turn that better fits the circumstances. However, should there be enough space for the more efficient U-turn, it is unlikely that this method would find it. Unfortunately, Moore's method suffers from the same limitation. When there is a choice between two very different alternatives, increasing the resolution about one would not necessarily uncover another—and perhaps more efficient—path through the multidimensional coordinate space, because the resolution of far away points is hardly increased.

Rather than indexing and adapting a path in its entirety, it may be more feasible to modify only the part of the path which is not successful. As mentioned earlier, we can dynamic program from the initial position forwards and from the goal position backwards until they touch. If most of the path does not require correction, we can dynamic program only those segments that do. If we imagine, for example, that points along our original path are lettered from A at the initial position to Z at the goal position, and only the segment E to H and Q to T are unacceptable, then the segment A to E can be copied into the new path, E can be set to the starting position and H to the goal. This can be dynamic programmed and the results can be copied into the new path. The segment from H to Q can likewise be copied over. Then Q can be set to the starting position and T to the goal and so on. Even if there are many smaller segments to dynamic program, there will be less to compute than if we dynamic programmed a new path from A to Z. Of course, this path is not guaranteed to be the best if we are trying to optimize for such things as time to the completion of the task.

We can combine this method with that of limiting the region about the path to be dynamic programmed.

What Plans Make Good Generalizations?

There is, of course, a trade off between generating, storing, and indexing a large number of plans and finding a few prototypes or generalizations that can be easily adapted to many situations. A three-point turn can be looked upon as a particular kind of generalization. Is it necessary, however, to have many different three point turns available to solve this class of problems, or is one sufficient? If only a single or a small number of prototypes are necessary, then what qualities would we want them to exhibit?

Earlier I discussed how the process of hill climbing may be used to find a locally optimal plan for a three point turn if we have already been given the plan for an adequate three-point turn. It would be ideal if we could start with a description of an optimal three-point turn, but unfortunately that is not possible because the minima will shift with small changes in, say, the width of the road. If from a plan for any three-point turn, we would find the optimal three-point turn solution, then we do not need to store multiple generalizations of the same plan. However, we would need to store a generalization of a U-turn if the locally best solution that would be found with it could also be the globally best solution.

We can imagine a landscape of peaks and valleys for a given problem where the lowest point in a particular valley is a local minimum. It would be fortuitous if we only needed one prototype for each valley. The trouble is that, as the particulars of the problem change such as the width of the road, not only can new local minima arise but also the positions of existing valleys can shift. Therefore, two prototype plans may uncover two different local minima or the same local minimum depending on the circumstances.

Stretching and Distorting Plans

When learning handwriting, a child first tries to trace or copy examples of individual letters. Eventually, the child no longer needs to consciously imagine the letters, but develops a personal style of writing. Were we to program a robot to write longhand,

we might represent each letter as an array of motor torques for each interval. Of course, we can write more quickly or more neatly as need be, so the length of the interval should likewise vary. Because certain letters like 'o' leave us further above the line than other letters like 'e,' adjustments would have to be made to connect letters in different sequences.

Even though we have a sense of the location of our writing based on our joint positions, because we may tend to drift up or down when we do not watch where we are writing, some form of closed loop control may be useful. This was used by Vogel (1992) to monitor the gait of a bipedal robot. The ease with which we write letters of arbitrary sizes implies that we do not have to relearn the alphabet in order to make these adjustments. Our robot should be able to increase torques linearly to produce this effect without any additional learning. Of course, when we move from the fine motor hand coordination required for writing on paper to the gross motor arm coordination required for writing on a chalk board, people often require a period of adjustment. This fact should not be surprising: the larger writing can no longer be accomplished by means of linear distortion of torques. In our computer model of hand-eye coordination, when it is no longer possible simply to increase the torques of the motors controlling the hand joints, we will have to relearn the torques for the arm joints.

Our capacity for distorting handwriting (e.g. bending it around corners or shrinking it) has been revealed in many a postcard where the sender has had more to say than space permits. Relating this back to the example of the three-point turn, we see that the robot must also be able to fit its plan for various driving manoeuvres, such as passing, turning off, or following the bend of the road. Just as when we write on a postcard and are forced to bend, distort, or compress letters to fit our words into the limited space, so may the constraints of the problem be used to distort a plan that has a reasonably close fit. The robot, therefore, needs to adjust these manoeuvres to the space available and to perform tracking using servos or closed-loop control. Even if stretching and shrinking plans to fit the circumstance is not adequate, it may at least suggest an area of coordinate space to be explored by dynamic programming.

Summary

The experiments discussed in this chapter show how, from limited experimentation, it is possible to learn predictions about the influence of self-induced movement on sensory projections and to use these predictions for route planning. In one simulation I used dynamic programming to compute a three-point turn for a car that had limited space to manoeuvre. This could not have been achieved servo-mechanically because the solution requires the car at first to move away from the goal.[45] Thus, the model is based on planning with empirically learned predictions about self-induced transitions between sensory projections.

Planning approaches, whether based on learned predictions or symbolic rules, require heuristics to set computational limits on search. This problem has been the subject of intense investigation and is not the main focus of this dissertation. Nevertheless, we have seen that there are approaches which can reduce the need for computation, especially if we do not require a globally optimal solution.

---

[45] A servo mechanism would simply have reversed the car towards the goal. At this point the car would be stuck—unable to move anywhere because any movement would take it further away from the goal.

Chapter 6. CSM Representations: A Grounded Foundation

Introduction

In chapter 3, we explored a few possible models for learning to recognize objects. A preliminary model which combined some of the strengths of wavelet analysis and Harnad's (1987) theory of category induction proved successful in experiments with classifying mushrooms. In chapter 4, we began with a pre-existing chess playing symbol system and a pre-existing piece locating perceptual module and then developed a method of grounding their shared symbols adaptively in sensorimotor coordination. We found it useful to develop the concept of sensorimotor predictions. These predictions can support adaptive mappings between different sensory modalities. By modelling sensorimotor relationships, they provide a basis for an agent to decide how to act. In chapter 5, we demonstrated how a robot can use sensorimotor predictions to plan optimal routes.

In this chapter, we explore how an agent can inductively learn behaviourally and physiologically relevant categorical-sensorimotor (CSM) representations. It can use the same CSM representation both for recognizing an object and for modelling the consequences of its actions directed at that object—simultaneously in terms of its actions' environmental and internal (i.e. physiological) consequences. Thus, these CSM representations extend the concept of sensorimotor predictions: it is possible for the predictions to not only refer to sensorimotor mappings but also real world objects—or, to be more precise, their correlates in the sensorimotor projections. Furthermore, we show in our fish simulation that these predictions can be learned inductively.

As research moves in the direction of learning perceptual or sensorimotor categories, to some extent, we may be able to wean our models off what may be an over-reliance on *a priori* feature detectors. No doubt, such detectors can contribute to the recognition of low-level features and even certain whole objects (namely, those, like faces and hands, that are of enduring significance to the proliferation of the genotype across evolutionary time scales). However, at some point category

induction probably must come into play, especially in learning to recognize objects that entail novel sensorimotor patterning or combinations of features. This sensorimotor patterning will not necessarily be semantically penetrable. Like the wavelet coefficients of chapter 3, it may be no more than a novel configuration of frequency and localization invariance in the sensorimotor projections.

So, in this chapter, we shall focus on how an agent could learn behaviourally and physiologically relevant categories and use them to predict the consequences of its actions. Specifically, we shall consider category induction in a simulation of a fish that moves about a simple aquatic environment. By chapter's end, we will be able to start considering how its learned categories could serve as the elementary symbols of a symbol system, and how a symbol system might develop from them in a bottom-up fashion (see Harnad, 1990a). But first let me introduce the representational model that we shall use.

A Coordinate Space Model for Multimodal Sensorimotor Integration

Of the five senses, perceptual research has tended to concentrate on the last to evolve: vision. This sense appears to rely on highly refined special-purpose processes for recognition. Unfortunately, this fact has tended to distract attention away from the importance of learning to integrate multimodal sensorimotor information. The purpose of the following predictive model is to demonstrate how it is possible to directly map sensorimotor information from more basic senses like smell and touch to cognitive categories that an agent can use not only for reactive behaviour but also higher-level planning. It is likely to be possible to adapt similar principles to vision, for example, by using more complex feedback loops, multiresolution analysis (see chapter 3) or, for bootstrapping purposes, evolved feature detectors and sensorimotor pathways. But to focus on that now would be a distraction.

We may represent the sensory stimulation that an agent receives from *external* sources as a point $\mathbf{e}$ $(e_1, e_2, \ldots e_n)$ in a sensory subspace E. The stimulation may arrive either directly from its sensors or pass through an intervening preprocessing

stage.[46] Likewise, we may represent the sensory stimulation that an agent receives from *internal* sources as a point **i** in a sensory subspace *I*. In a very simple agent, the point may be determined by a single scalar value such as a reinforcement from a reward function (see appendix A). In a more complex agent, it may be determined by numerous sensed physiological variables (e.g. thirst resulting from dehydration) and other internal reinforcement feedback loops as well as higher-level cognitive variables. (For simplicity, these shall be referred to collectively as internal variables.) Finally, we may represent the motor signals the agent sends to its actuators as a point **m** in a motor subspace *M*.

The agent can represent the consequences of a self-induced action by storing points in a *sensorimotor coordinate space* composed of its motor subspace and its sensory subspaces before and after carrying out the action. We assume for the moment discrete time steps. We further assume that the sensory effects of an action are available by the next time step. Thus, at time t, ($\mathbf{m}_t$, $\mathbf{e}_t$, $\mathbf{e}_{t+1}$, $\mathbf{i}_t$, $\mathbf{i}_{t+1}$) determine the location of the point $\mathbf{p}_t$ in a composite coordinate space. The internal changes $\Delta\mathbf{i}_t$ correlated with ($\mathbf{m}_t$, $\mathbf{i}_t$, $\mathbf{e}_t$) are simply $\mathbf{i}_{t+1}$ - $\mathbf{i}_t$, and the external changes $\Delta\mathbf{e}_t$ are $\mathbf{e}_{t+1}$ - $\mathbf{e}_t$.

As long as there is some degree of consistency across time and space in what the agent senses, it can exploit the points stored in its sensorimotor coordinate space to predict the consequences of its actions. Thus, the past is brought to bear on the present. Once again, this may be achieved by the closest-point method. Thus, the agent may approximate the likely effect of a motor signal **m** on its sensory stimulation ($\mathbf{i}_t$, $\mathbf{e}_t$) by finding the point $\mathbf{p}_x$ ($\mathbf{m}_x$, $\mathbf{i}_x$, $\mathbf{e}_x$, $\mathbf{i}_{x+1}$, $\mathbf{e}_{x+1}$) such that ($\mathbf{m}_x$, $\mathbf{i}_x$, $\mathbf{e}_x$) is the closest point to ($\mathbf{m}$, $\mathbf{i}_t$, $\mathbf{e}_t$) in that subspace. The predicted values for sensory stimulation in the next time step are ($\mathbf{i}_{x+1}$, $\mathbf{e}_{x+1}$). It follows that the predicted change in the agent's internal sensory stimulation is $\mathbf{i}_{x+1}$ - $\mathbf{i}_t$; and the predicted change in the agent's external sensory stimulation is $\mathbf{e}_{x+1}$ - $\mathbf{e}_t$. Thus, points contained within the sensorimotor coordinate space may serve as a basis for predictions about the physiological and ecological consequences of possible actions. The agent may select the action it deems

---

[46] Preprocessing in Clocksin and Moore's penlight-holding robot, for example, reduced two large matrices of light intensity values from the two camera's image planes to a point in a visual coordinate space of just four dimensions.

best on the basis of these predictions or, in order to fill in for gaps in the model, an action with relatively unknown consequences.[47]

The closest-point approach upholds the property of *graceful degradation* found in many kinds of artificial neural networks. For example, if closest point selection is based on a Euclidean distance measure, estimates of distances will only be somewhat less accurate if the squared difference between the current external sensory projection and all previously recorded external sensory points ($\mathbf{e}_1$ to $\mathbf{e}_t$) cannot be calculated for every relevant dimension. However, this only affects the estimate if contribution from those dimensions would have resulted in the selection of a different point, and even a different point may be acceptable. Just as someone who normally performs a task sighted may still be able to accomplish it blindfolded (e.g. with touch and proprioception), so may an agent using the this model despite there being missing or incomplete values for one or more dimensions of coordinate space (e.g. input from one or more sensory modalities or sources of sensory information).

The closest-point method can be used to calculate many other kinds of approximate mappings such as those required to fill in for missing sensory data (e.g. the blind spot). This is done by finding, on the basis of available dimensions, the closest point in the external sensory subspace. The values for the missing dimensions can be estimated from the corresponding dimensions of that point.

Unlike neural networks, the closest-point method does not compromise semantic penetrability. To an outside observer (with the appropriate sense organs or sensing equipment), the semantics of the system should be relatively clear. By peering into the coordinate space we can always explain why a certain prediction was made by tracing it to the particular sensorimotor projections in which it is grounded. Of course, this does not mean that an observer will always be able to see clearly what the agent is responding to. What the agent can respond to will depend on what it senses and, therefore, its particular sensors and how it appraises their output in terms of its learned predictions.

---

[47] It is possible to make more accurate predictions from less data by interpolating values on the basis of nearby points. We shall examine this method in the next section.

It would be useful for the agent to be able to learn the most accurate sensorimotor model possible in the least amount of time. Ideally, the closest point to the agent's current sensory stimulation in coordinate space will indeed provide a prediction that offers the best prediction. But just what should the agent be trying to predict? That depends on what we want the agent to do. In nature selective pressures should favour models that offer a clear choice between those actions that are likely to enhance reproductive success and those that are not. As an animal cannot relive its life to see what works best, it can only estimate indirectly what action is most likely to enhance its reproductive success. The expected future proximity of essential physiological variables to their optimum provides a measure of its expected future well being (and, as we shall discuss later in appendix A, it can be estimated using Q-learning or other methods of learning from delayed rewards). Internal variables can indicate, among other things, an organism's physiological state. Precisely how these internal variables are related to its bodily condition and how they are best weighted can be left for evolution to determine, though to some extent they may be fine-tuned empirically. According to this model then, evolution influences perception and behaviour not only by mean of evolved adaptations but also indirectly by means of intermediate internal variables.

There are numerous ways in which an agent can improve its model once it has indicators (e.g. internal variables) in terms of which it can evaluate its model's predictive power. Statistical methods can be used to estimate the linear or nonlinear correlation between the indicators and the other sensorimotor dimensions so that the other dimensions can be weighted according to their likely impact on the indicators. (This method can also be used to eliminate irrelevant sensorimotor dimensions, see chapter 5.) However, often their relative influence will vary from one region of the coordinate space to another. To illustrate this point, we may consider a two-segmented arm monitored by a camera that is fixed perpendicular to the plane in which the arm has mobility. A change in the angle between the base and the first segment of the arm will have a much greater influence on the location, in camera-based coordinates, of the arm's end when the arm is extended than when it is contracted. (You may easily demonstrate this by moving your own arm.)

In this arm example, a better method of approximation would be local linear interpolation (see Omohundro, 1987). We are trying to approximate a mapping from

the visual to the proprioceptive subspace, so that we can approximate the proprioceptive coordinates of some new point **p** given only its visual coordinates. If the proprioceptive subspace has k dimensions, we need $k + 1$ nearby points in the visual subspace in order to interpolate (or extrapolate) the proprioceptive coordinates of **p**. An added restriction is that all points not fall on the same hyperplane of dimensionality $k - 1$. A weighted sum of vectors emanating from one of the $k + 1$ points to the k remaining points describes the location of **p** in the visual subspace, which, of course, is known from the camera's image plane. This provides a system of k linear equations with k unknowns. A straight-forward algorithmic solution provides k weights (see Press et al., 1992, chap. 2). The $k + 1$ points in visual subspace correspond to $k + 1$ points in the proprioceptive subspace. Hence, the k vectors in terms of which the new point was originally described in the visual subspace have k corresponding vectors in the proprioceptive subspace. The weights may be applied to these vectors, providing an estimate of the new point in the proprioceptive subspace. Figure 6.1 illustrates this technique in two dimensions.

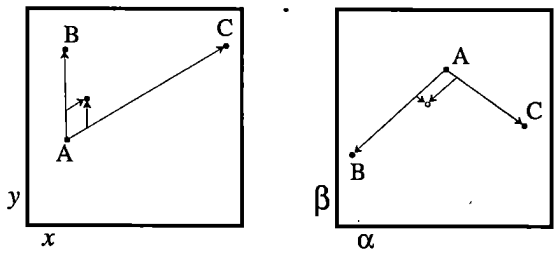


Fig. 6.1: For ease of illustration, let us consider two dimensions only. From past experimentation, the robot has determined that the points *A*, *B*, and *C* in the visual *xy*-subspace correspond to matching points in proprioceptive $\alpha\beta$-subspace. Given enough points, the robot can perform a (globally) nonlinear mapping between the two subspaces by means of local linear interpolation. For example, the robot predicts a new location (the unfilled point) in the proprioceptive subspace from a known location in visual subspace by representing the known location in terms of ratios between vectors connecting its neighbouring three points and using these ratios to interpolate or extrapolate its approximate location in the proprioceptive subspace.

Gross and Wagner (1996) have applied interpolation to learning sensorimotor mappings using a type of *k*-d tree (also see Omohundro, 1987, 1990a). Their results show massive gains in efficiency over other learning methods like backprop. In one experiment with learning the inverse kinematics of a robot arm, building the *k*-d tree required less than one hundred thousandth the time that training the network did. However, the post-training memory requirements for *k*-d trees are significantly higher than for neural networks.

Approximating an input-output mapping for a point **p** based on the $k + 1$ closest points can lead to discontinuities when **p** falls outside their simplex (i.e. the convex hull whose vertices are the $k + 1$ points). Here **p**'s mapping is extrapolated from the $k + 1$ closest points, while points lying on just the other side of a Voronoi hyperplane are interpolated by a somewhat different set of points—hence, the discontinuity. This is not a crucial performance issue for the robot applications we have considered so far, since slight discontinuities within a margin of error will generally go unnoticed.

To ensure that a mapping has no discontinuities, one may triangulate using the vertices of the simplex enclosing **p**. The question is how to choose which simplex to use. The Delaunay criteria, which takes advantage of the fact that vertices of the simplex also specify a circle (in 2-D), sphere (in 3-D), or hyper-sphere (in 4-D and above), selects the simplex which contains only **p** and no other neighbours. This provides a bias in favour of smaller, more equilateral simplexes whose predictive accuracy is likely to be higher than that of long, skinny simplexes in which significant errors may arise because of nonlinearity in the long dimension. The worst case error of the Delaunay triangulation has been proven to be lower than that of any other triangulation method (Omohundro, 1990b).

Unfortunately, the Delaunay triangulation can be time consuming when there are more than a few dimensions because, as the number of dimensions increases, the points closest to **p** are less likely to be the vertices of the simplex enclosing it (as specified by the Delaunay criteria). For learning mappings which involve between ten and a hundred dimensions, Omohundro suggests using normalized Gaussian blending functions to blend affine or quadratic maps and bumptrees for the fast look-up of segments relevant to the query (personal communication, also see Omohundro, 1991, 1989).

The Fish Experiments: Forming Relevant Categories by Adaptive Quantization

In this section, I give a description of a program I wrote to simulate fish chemoreception. I did not intend the model to accurately depict the behaviour of fish but to illustrate how an agent can learn categories that pick out physiologically and behaviourally relevant invariance (e.g. edible and inedible objects) and to use these categories to predict the consequences of actions.

Staying in constant motion the fish swims about its shallow (2-D) pond. As the fish sways back and forth, its receptors sample the concentrations of different chemicals in the water.[48] It controls its behaviour through two decision-making systems, one which controls its oral cavity and the other which controls its navigation. The former is only activated when the fish bumps into something, perhaps another fish or a piece of debris. When this happens, the fish selects one of four motor responses. It *(1)* ignores it, *(2)* ingests it, *(3)* takes it up in its mouth, or *(4)* empty the contents of its mouth.[49] Its choice of action at time t is represented by a point $\mathbf{m}_t$ in the motor subspace of its *sensorimotor coordinate space* ($\mathbf{m}_t$, $\mathbf{e}_t$, $\mathbf{e}_{t+\delta}$, $\mathbf{i}_t$, $\mathbf{i}_{t+\delta}$, $\Delta\mathbf{i}_t$). The levels of stimulation to its different kinds of chemoreceptors from waterborne substances determine its external sensory projection. Before performing an action they are represented by the point $\mathbf{e}_t$ and the point $\mathbf{e}_{t+\delta}$ afterwards. Before performing an action, its essential physiological variables are represented by the point $\mathbf{i}_t$. After the results of the action have had time to take effect, it is represented by the

---

[48] Many vertebrates exploit olfaction for feeding, navigation, reproduction, and the avoidance of predators (Stoddart, 1980). Evidently even birds like the petrel assess the direction of an odour source by swaying their heads from side to side (p. 192). Fish swim up the gradient of chemicals emitted by food to its source. They make use of klinotaxis, successive comparisons of concentrations, and tropotaxis, simultaneous comparisons among distinct chemosensory organs (Jones, 1992, p. 292). Although fish have natural preferences for certain odours and tastes, experience can change their response to chemical cues (Jones, 1992, p. 300).

[49] Clearly, this scheme is simplistic. Some form of action generalization is also called for so that an intelligent agent can control actuators with real-valued motor signals (see appendix A).

point $\mathbf{i}_{t+\varepsilon}$. We shall assume that the dimensions of $\mathbf{i}_t$ and $\mathbf{i}_{t+\varepsilon}$ have already been weighted so that the fish's health is roughly proportional to the proximity of $\mathbf{i}_t$ to a known optimum. The change in the fish's health that is correlated with the action taken and the current external sensory projections ($\mathbf{m}_t$, $\mathbf{e}_t$) is represented by the point $\Delta\mathbf{i}_t = \mathbf{i}_{t+\varepsilon} - \mathbf{i}_t$.

The Integrated Learning of Perceptual Categories and Act-Outcome Predictions

The fish learns predictions about the effects of its motor actions on its physiological variables by recording unexpected sensorimotor projections in its coordinate space. In a new situation the fish approximately predicts the effect of alternative actions on its physiological state by finding the closest point in sensorimotor coordinate space given its current levels of chemoreceptive stimulation. (We shall delay for a moment discussion of what we mean by *closest*.) Therefore, in this example, points corresponding to unexpected sensorimotor projections help to form representations that are at once categorical and sensorimotor. The CSM representations quantize the coordinate space into convex hulls to set up predictions about the consequences of future actions.

How do we quantify what makes a sensorimotor projection unexpected in determining when it is appropriate for the fish to learn a corresponding prediction (i.e. CSM representation)? If we are to be very strict then any occurrence that deviates from the fish's current predictions is unexpected. In this case (except in the unlikely event that its predictions were spot on), the fish should simply commit all its experiences to memory. Economy speaks against this, because the fish would be learning detail that is of marginal utility. We note that the fish's receptors are of limited accuracy and also that some of the stimulation they receive would not be coming from whatever the fish were currently trying to distinguish. It would be a waste of the fish's precious resources to learn what is for the most part noise in the system. There is another reason not to remember all past sensorimotor projections. The better able the fish is to treat behaviourally equivalent things as the same, the better able it is to survive and reproduce. This implies that the fish should not only ignore extraneous detail but also take determined measures to filter out any form of misleading information. I opted for this approach in programming the simulation.

In a complex and rapidly changing world, there are certain things worth discriminating rather finely. If two different species of fish, for example, produce nearly identical sensory projections ($\mathbf{e}_x \approx \mathbf{e}_y$) but radically different physiological changes when ingested ($\Delta\mathbf{i}_x \neq \Delta\mathbf{i}_y$)—if, for example, one is poisonous—, then priority must be given to detecting any telltale differences in their sensory projections. Thus, $\mathbf{e}_x$ and $\mathbf{e}_y$ should map to different categories, which we may write as

$$c\,(\mathbf{e}_x) \neq c\,(\mathbf{e}_y)$$

However, if two different species of fish produce somewhat different sensory projections ($\mathbf{e}_x \neq \mathbf{e}_z$) but nearly identical physiological changes ($\Delta\mathbf{i}_x \approx \Delta\mathbf{i}_y$), then they can be lumped together in the same behaviourally equivalent category.[50] Thus,

$$c\,(\mathbf{e}_x) = c\,(\mathbf{e}_y)$$

In the simulation I tried to capture this intuition by first globally weighting the dimensions of the coordinate space so that each variable received proper emphasis. If, for example, slight changes in a physiological variable spell life or death for the fish but the level of stimulation to a particular kind of chemoreceptor is relatively unindicative of what is in its environment, then the weighting had to reflect this. The weighting was performed by multiple linear least-squares. Here we are trying to predict the change in the fish's overall health $|\,\Delta\mathbf{i}_t\,|$ in terms of its sensorimotor projection $\mathbf{p} = (\mathbf{m}_t, \mathbf{e}_t, \mathbf{e}_{t+\delta}, \mathbf{i}_t, \mathbf{i}_{t+\varepsilon}, \Delta\mathbf{i}_t)$. If we have a set of points $\mathbf{p}_1, \mathbf{p}_2, \ldots \mathbf{p}_r$ corresponding to r categorical-sensorimotor representations, each k dimensional, then we are looking for a weight vector $\mathbf{w} = w_1, w_2, \ldots w_k$ that minimizes the linear squared error of the following set of linear algebraic equations:

---

[50] This has a parallel in comparative linguistics. A language may lump together many different kinds of objects if the speakers of the language deem them to be of like importance (see Malinowski, 1923).

$$w_1 p_{1,1} + w_2 p_{1,2} + \ldots + w_k p_{1,k} = |\Delta \mathbf{i}_1|$$
$$w_1 p_{2,1} + w_2 p_{2,2} + \ldots + w_k p_{2,k} = |\Delta \mathbf{i}_2|$$
$$\vdots$$
$$w_1 p_{r,1} + w_2 p_{r,2} + \ldots + w_k p_{r,k} = |\Delta \mathbf{i}_r|$$

The r equations relate the k unknowns w_1, w_2,... w_k. Of course, the p's and **i**'s were known from the CSM representations which were learned from past sensorimotor projections. (The method of calculation I used will be discussed below.)

Once the coordinate space had been properly weighted, it was quantized. The granularity of the quantization varied according to the different distance thresholds I tried out. The threshold determines when an 'experience' is sufficiently different from nearby CSM representations to warrant the formation of a new representation (i.e. category). I discovered the most advantageous granularity at which to perform this quantization by trial and error.[51] The categories thus formed reflected similarities among sensory projections in their likely internal (i.e. physiological) and behavioural outcomes. This, of course, goes beyond the detection of structural similarities (i.e. spatiotemporal correlations) in the projections themselves.

As mentioned earlier, ideally the coordinate space should be quantized so as to minimize the misleading effect of chemicals from more distant sources in discriminating what is at hand. I approached this problem in the following manner. Note that each relevant category is determined by the placement of its CSM representation and those of surrounding categories. Therefore, the collection of all these representations quantized the coordinate space. At first, each sufficiently distinct sensorimotor projection acted as a representation. However, when a second sensorimotor projection fell in the same category, the representation was refined by intersecting it with the new sensorimotor projection and resetting it to the result (see Figure 6.2). In our specific example, for each kind of chemoreceptor (that is, each dimension), the future

---

[51] The granularity is roughly analogous to vigilance in adaptive resonance theory (Grossberg, 1988) or the number of units whose weights are to be adjusted in competitive learning (Rumelhart & Zipser, 1985) and, indeed, we see that the quantization itself need not have been performed by the closest-point approach but could have been performed by these or other methods of unsupervised learning.

value of the representation was assigned the minimum of its present value and the value of the new sensorimotor projection.

This is compatible with Harnad's insight that the purpose of a categorical representation is to "reliably distinguish members from nonmembers of a category" (1990a, p. 342, also see 1987). Therefore, the agent did not use prototypical representations. Rather, it learned categorical-sensorimotor representations in order to distil from sensorimotor projections the invariance that was shared by members of a category but not shared by nonmembers.

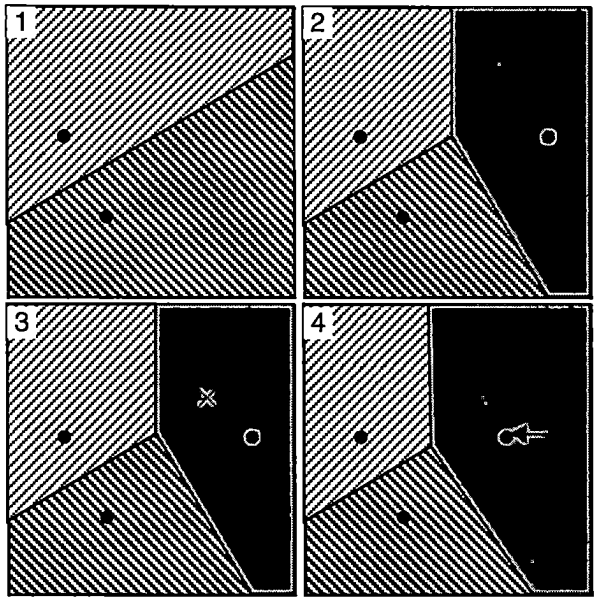


Fig. 6.2: *(1)* Two CSM representations quantize two dimensions of the sensorimotor coordinate space. *(2)* A sufficiently dissimilar sensorimotor projection forms a new representation. *(3)* A sensorimotor projection (×) is categorized by a representation. *(4)* It is intersected with the representation thus refining it.

The Navigational System

Apart from the decision making system for controlling its oral cavity, the fish has a second decision making system for navigation. The two systems are autonomous except for the fact that the second system exploits the categories formed by the first. Based on the CSM representations that the oral cavity decision making system had

formed so far, the navigational system ascertained which corresponding sources of chemicals are present in its vicinity. It did this by trying to account for the level of stimulation to the fish's different kinds of chemoreceptors in terms of its CSM representations by means of a multiple linear least squares technique that I had modified to the peculiarities of this application (see Figure 6.3).

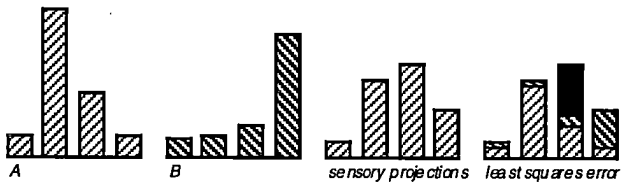


Fig. 6.3: This simplified illustration shows how the agent used least-squares error to account for a sensory projection in terms of the known categorical-sensorimotor representations *A* and *B*. Both *A* and to a lesser degree *B* appeared to be present. However, there appeared also to be an unexplained source of the third chemical transmitter, shown here in black. Potentially the fish could have actively sought to uncover this unknown chemical source.

The l linear algebraic equations below relate the r unknowns $a_1, a_2, \ldots a_r$. The weight coefficients $a_1, a_2, \ldots a_r$ determine the likely contribution of the l-dimensional external sensory subspace $\mathbf{e}_1, \mathbf{e}_2, \ldots \mathbf{e}_r$ of each of the k-dimensional CSM representation $\mathbf{p}_1, \mathbf{p}_2, \ldots \mathbf{p}_r$ to the current external sensory projection $\mathbf{e}_t$:

$$
\begin{aligned}
a_1 e_{1,1} + a_2 e_{2,1} + \ldots + a_r e_{r,1} &= e_{t,1} \\
a_1 e_{1,2} + a_2 e_{2,2} + \ldots + a_r e_{r,2} &= e_{t,2} \\
&\vdots \\
a_1 e_{1,l} + a_2 e_{2,l} + \ldots + a_r e_{r,l} &= e_{t,l}
\end{aligned}
$$

Notice that the weight coefficients $a_1, a_2, \ldots a_r$ are being put to a different use from the weight coefficients $w_1, w_2, \ldots w_k$ for the equations discussed earlier. In the former we were scaling the k dimensions of the r CSM representations relative to each other. Each representation's overall influence remained fixed and equivalent. Thus, we applied multiple least squares to the matrix of points $\mathbf{p}_1, \mathbf{p}_2, \ldots \mathbf{p}_r$. Here we are scaling the r CSM representations and *not* their individual dimensions. Thus, we apply

multiple least squares to the transpose of the matrix of points $\mathbf{e}_1$, $\mathbf{e}_2$,... $\mathbf{e}_r$. These points are determined by the l-dimensional external sensory subspace of the k-dimensional CSM representations.

Although this is not how least squares is typically used, it effectively accounts for the current sensory projection in terms of known CSM representations. We must make two non-standard restrictions on our how we calculate least squares. None of the weight coefficients are permitted to be negative. This is, of course, because a substance cannot have a negative presence in the water. Also, the weighted sum of the representations cannot exceed the sensory projection for any one chemical transmitter.

I implemented an algorithm to perform the least squares computations by singular value decomposition (SVD). For a detailed explanation of this technique and proofs for its efficacy, see Golub and Reinsch (1971, they invented it) and Forsyth, Malcolm, and Moler (1977). Further details may be found in Lawson and Hanson (1974) and Press et al. (1992). The following discussion will make plain why it was necessary to use SVD to solve the above two sets of equations instead of, for example, Gaussian elimination, LU decomposition, or QR decomposition. (The latter is probably the technique most commonly found in statistical packages.)

In the simulation, we start off without any CSM representations, so the number of dimensions in both the sensorimotor coordinate space and in the external sensory subspace exceeded the number of representations. When sufficient CSM representations have been learned, their number will exceed the number of dimensions. This means that the first set of equations started out being underdetermined: there were ambiguous combinations of the weight coefficients w_1, w_2,... w_k that scale the dimensions. It then, however, became overdetermined when the number of representations exceeded the number of dimensions. As soon as there was more than one CSM representation, the second set of equations became underdetermined: there were ambiguous combinations of the weight coefficients a_1, a_2,... a_r that scale the representations.

Unlike other methods, SVD cannot fail and has been proved never to fail (see Golub & Reinsch, 1971; Forsyth, Malcolm & Moler, 1977). If the equations are overdetermined, SVD calculates the best possible least-squares approximation. If they

are underdetermined, it calculates the smallest weight coefficients in the least squares sense. These results are unaffected by the slight modifications I have made in the algorithm when applying it to the second set of equations. Other methods can fail because the matrices are singular (because of row degeneracy, column degeneracy, or both) or because they are numerically so close to being singular that, because of round-off error, for the purpose of calculation it is as if they were.

For example, when Gaussian elimination encounters a zero pivot, it can yield infinities that unstably cancel each other out; when it encounters a pivot that is very close to zero, it can yield large magnitudes that do likewise. SVD avoids this. The way SVD decomposes the standard deviations ensures their mutual independence (i.e. that all of them are uncorrelated). This makes it possible to sum them up in the root-mean-square fashion. If any of their singular valued denominators is zero or very close to zero, its reciprocal is set to zero instead of infinity or some large magnitude. Thus, a zero multiple is added to the fitted weight coefficients. When the basis functions are degenerate in the fit, this averts the problem of adding some huge multiple that is a linear combination of them.

SVD requires an additional matrix and is purported to be significantly slower than other methods of calculating least squares. Its speed and memory requirements, however, posed no problem in the simulation: the fish could be seen to interact in its environment in real time. The calculations for the above set of equations required less than a second on a moderately loaded Alpha, even with several dozen CSM representations of several hundred dimensions.

Of the CSM representations whose corresponding sources are apparently present (above a certain threshold), the fish selected the one which could potentially, given the most appropriate action, bring its current physiological variables closest to their optimum. To climb the gradient of chemicals associated with that CSM representation, it used klinotaxis (the successive comparison of concentrations; see Jones, 1992) as follows. We assume that the concentrations of chemicals will in general exponentially increase as the fish swims directly towards their source. As the fish moves ahead, it may either veer to the left or right. It will continue veering in the same direction as long as the concentrations accounted for by the selected CSM representation are accelerating. Otherwise, it will try veering in the other direction. In simula-

tion, the fish very quickly homes in on the desired source, especially if it is not obscured by intervening sources that emit a similar set of chemicals.

Conclusion and Summary

In the fish example we have shown that perceptual information can be expressed *directly* in terms of sensorimotor categories. However, by *sensorimotor* we must now include *internal* feedback from motor signals and physiological variables. The simulated fish learned to interact with its environment so as to optimize its physiological variables. It learned for example, what to eat and what to avoid eating. From the raw sensory projections of its different kinds of chemoreceptors it developed CSM representations that progressively better approximate behaviourally similar aspects of its environment. Based on the predictions it developed about the consequences of its actions from its past experiences, it followed a course of action that kept its essential variables within their normal ranges for survival.

The sensorimotor coordinate space is limited to only one level, and our quantization of it is based on the structure of the coordinate space, or specifically, on the distribution of the CSM representations within it. Nevertheless, once sensory projections have been scaled with respect to alternative motor actions and their likely physiological consequences, the structure thus uncovered can no longer be equated merely with patterns in external sensory projections. This is because, in certain cases, the same category will be activated by structurally dissimilar sensory projections and, in other cases, different categories will be activated by structurally similar sensory projections. The former occurs when the same action produces a similar physiological effect; the latter when structurally minor details can be exploited to discriminate things that appear similar but produce widely different physiological effects.

In extending this model, there are many avenues that may be pursued. Animals become social early on in evolution and must compete for mates and distinguish friend from foe. The navigational decision making system could be enlarged so that the fish not only homes in on fish that it can profitably interact with but also avoids predators. Also further variables analogous to hormone levels could be added to its sensorimotor coordinate space. To keep them in their normal ranges, actions governing courtship and mating behaviour would need be taken.

Chapter 7. Learning a Symbol System from the Bottom Up

Introduction

Let us now tie together various threads from the previous four chapters. In chapter 3 we devised a program to learn categorical representations inductively in order to distinguish different kinds of mushrooms. In chapter 4 we developed a robot that learned sensorimotor representations to map between different sensory modalities, and in chapter 5 we developed a robot simulation that learned and used sensorimotor representations to plan optimal routes through its environment. However, in both chapters 4 and 5 we presupposed feature detectors—in chapter 4, to recognize chess pieces and, in chapter 5, to recognize a car and its obstacles.

In chapter 6, we took a first step toward developing a single form of representation that serves a dual purpose. The agent uses it both for categorizing (and learning to categorize) salient objects in its environment *and* for modelling the consequences of its actions (or nonactions) vis-à-vis those objects. In addition to learning to distinguish objects (at time t) according to their sensory projections ($\mathbf{e}_t$) and physiological effects ($\mathbf{i}_{t+\varepsilon}$), the simulated fish also learned from past interactions to predict how performing different actions ($\mathbf{m}_t$) on those objects would affect its physiological state ($\mathbf{i}_{t+\varepsilon}$).

In this chapter we consider ways of extending the fish's model so that it can plan in terms of learned representations about the world. The ability to represent how actions transform external states of affairs offers a distinct advantage over, for example, reinforcement learning, which can only represent internal goals. It lets the agent plan in novel situations, applying knowledge acquired in the pursuit of previous goals. But the more thoroughly our model articulates the structure of the world, the more difficult it is to keep it efficiently updated. Thus, we end the chapter by discussing the frame problem and appraising its relationship to our representational model.

Developing CSM Representations into a Symbol System

At the one extreme, we could improve the performance of our fish simulation by taking a hybrid approach. We could design a symbol system and graft it on top of the fish's learned categories. To obtain a description of the current state of the world, a symbol would be instantiated if its corresponding perceptual category is estimated to be present in its surroundings above a certain threshold. This provides a listing of all the chemical producers within a certain region near the fish. It may be calculated in the manner described in chapter 6: by accounting for the current sensory projections in terms of the fish's discovered CSM representations by least-squares minimization. The symbol system part would then need to be programmed with goals and operator rules for their attainment. For example, if the fish had the goal of taking care of its young, then it might have an operator rule telling it that, among other things, bringing certain kinds of food to its offspring would help it to achieve this goal. There would be other rules telling it how to get the food: for example, by searching for it, taking it into its mouth, searching for its young, and then spewing out the food. The symbol system could be implemented using a symbolic planner like STRIPS, SNLP, or SOAR.

What would be unsatisfying about this approach is that the system would still require domain-specific programming. To adopt it would be to quit when the fish is only part of the way toward developing a grounded symbol system. The fish has gone through the trouble of learning its own perceptual categories but then the programmer dictates how it should plan with them.

At the opposite extreme from the hybrid approach, we could extend the simulation so that it could exploit an advanced form of reinforcement learning—one that allows it to learn profitable patterns of behaviour across time. Since the fish has already learned relevant categories that it is able to exploit, it is in a good position to assemble simple interactions into more complex patterns of behaviour. A method akin to learning from delayed rewards would provide an adequate means of doing this—for example, by the temporal differencing of Sutton's adaptive critics (1988) or Watkin's Q-learning (1989). The fish's state would be determined according to $c\,(\mathbf{e}_t)$: the categorization of its sensory projections when it bumps into something. If in these states the fish happened to perform actions in an order that produced some kind of internal reward, then this reward would be credited across the last few states (or

state-action pairs). Eventually, it would learn to produce rewardable sequences of actions in the correct circumstances. So, for example, it would learn to take the right food to its offspring. Within the constraints of how we define state and after a sufficiently long period of trial and error, the fish would be behaving as if it had an *a priori* symbol system.

As an enhancement, instead of having a reward system based on scalar Q-values, finer discriminations are possible by using Q-vectors (which I propose in appendix A). The fish can empirically learn which physiological parameters determine the dimensions of these vectors and how they are weighted relative to each other (e.g. by the methods used in chapters 5 and 6, in Omohundro, 1989, 1991, or by a genetic algorithm). In reverting to a form of reinforcement learning, however, we would lose much. The symbol categories that the fish had learned to individuate would be folded together into states, since states are what form the basis of reinforcement learning. The fish would also not be able to model transformations in its external world—for example, as based on changes in its active set of perceptual categories.

At present the fish uses the coordinate space to map from external sensory projections and desirable internal physiological changes to appropriate motor actions. It is worth noting that the fish's CSM representations provide some information that our current model has yet to exploit. As stated in chapter 6, five subspaces ($\mathbf{m}_t$, $\mathbf{e}_t$, $\mathbf{e}_{t+\delta}$, $\mathbf{i}_t$, $\mathbf{i}_{t+\varepsilon}$, $\Delta\mathbf{i}_t$) determine each point $\mathbf{p}$ in the fish's sensorimotor coordinate space. On the basis of its sensory projections, the fish is already able to activate a salient group of representations for learned categories. However, it has neglected to use the external sensory projection $\mathbf{e}_{t+\delta}$. This projection indicates the levels of stimulation to the fish's different kinds of chemoreceptors after the effects of its action have had time to propagate. Thus, $\mathbf{e}_{t+\delta}$ could allow it to predict how its motor signals transform the external categories it has learned to detect. If the fish were only to exploit this information, it would be able to start doing a very basic kind of STRIPS-style planning. In this chapter we consider ways of extending the fish's model so that it can perform more abstract kinds of planning.

In chapter 2, we briefly introduced the robot Shakey. Shakey used the STRIPS planner, which is based on symbolic representations and *a priori* operator rules. In STRIPS a list of predicates describes the world's current state, and another list the

world's goal state. Each operator rule consists of a list of preconditions, an add list, and a delete list. The operator rule's preconditions state under what conditions the robot may apply the rule. For example, to pick up a ball, the robot might need to be next to the ball and empty-handed. The add list states what becomes true when the robot performs an action—for example, that it is holding the ball. These predicates are added to the robot's description of the world's current state. The delete list states what is no longer true and, therefore, what needs to be deleted from the description. For example, that the robot's hand is no longer empty.

In a hypothetical extension to the fish's model, we may characterize changes in its environment in terms of transformations on its active perceptual categories. To simplify our discussion, we will overlook the fact that the fish has two separate control systems—one for its oral cavity, the other for navigation. Given that $\mathbf{e}_t$ is its current sensory projection, let $c\ (\mathbf{e}_t)$ be the list of perceptual categories that the fish can distinguish in $\mathbf{e}_t$ because of its past learning. Thus, the fish's model may be characterized by

$$f\ (\mathbf{m}_t, c\ (\mathbf{e}_t)) \rightarrow c\ (\mathbf{e}_{t+\delta})$$

It may be useful to compare and contrast this model with STRIPS. The combined effect of STRIPS's add and delete lists are roughly analogous to the fish's potential transformations on its perceptual categories. The fish also has the advantage that it has actually learned its perceptual categories and transformations so, unlike STRIPS, it is not reliant on a programmer to set up its model of the world.

Since the coordinate space tells the fish what effect its actions are likely to have on succeeding perceptual categories, the mapping f would serve as a rudimentary model of the world allowing the fish to plan sequences of actions it has never attempted before. Therefore, if it has a certain projection as its goal, it can backward chain to the currently active perceptual categories in the same manner that a symbolic planner backward chains on operator rules. The fish can then attempt to perform the chain in its correct order. Each motor action in the chain has the precondition of the fish being exposed to a external sensory projection that (once categorized) matches the planned action.

Of course, in a nondeterministic environment, provisions would need to be made for the fish to retry actions or take other corrective measures when the action first

attempted fails to produce the desired change in c ($\mathbf{e}_{t+\delta}$). (We may need to introduce closed loop control.) We must also take care in determining the fish's goals. Presumably, a goal would be the performance of an action under conditions in which the fish's physiological state is likely to improve. Planning comes in because that action may not be immediately available. It may require a series of intervening actions.

But as yet the notion of precondition in the fish's model is crude. It is based on a learned mapping, which is just a simple state transformation. By folding categories together into states the fish is, in fact, sacrificing representational power. If, given the same set of active perceptual categories, the fish's motor signals **m** were previously successful in making the desired transformation, then it can try to repeat the action here. But this sort of precondition is rather too inflexible. What if more or fewer perceptual categories happen to be active? The fish would have no basis for determining what effect its action would have. How does the fish learn which perceptual categories are essential and which incidental to the performance of an action? In other words, when I twist a doorknob, how do I know that the colour of the sky and a myriad of other things do not determine whether the door will open?

Here we have hit upon a nexus of related problems. Clearly the fish needs to be able to inductively generalize on its transformations. How to know that the generalizations are correct is a fundamental problem for inductive learning and has been much examined in the philosophy of science. I think the short answer is that we cannot definitively know that all our generalizations are correct. Even if it were possible to test them in every identifiable situation, that would not prevent some unnoticed factor from invalidating at least some of them in the meantime. The best we can hope for is that the world has enough stability for our agents to get by. We neither need nor can expect perfect models of the world.

Models are approximations, and we can only expect them to be answerable to the data thus far encountered (Harnad, 1987). The approximate nature of models can, in fact, be an advantage, since it is necessary to strike a balance between inadequate models and models that are burdensomely complete (see Weld, 1991; Janlert, 1996). In science at least we generally prefer simpler models—models that better predict the data without overfitting it (Lakatos, 1976). Tolman (1932) observed this fact in the domain of scientific models:

> All science necessarily presents, it seems to us, but a map and picture of reality. If it were to present reality in its whole concreteness, science would be not a map but a complete replica of reality. And then it would lose its usefulness. For it would have to cover as many pages as does life itself; it would no longer serve as a brief and a handbook. One of the first requisites of a science is, in short, that it be a map, i.e., a short-hand for finding one's way about from one moment of reality to the next—that it be a symbolic compendium by means of which to predict and control. (pp. 424-425)

Similar desiderata probably apply to cognitive models. The need to balance representational adequacy and economy arises because there is no complete solution to the qualification problem (McCarthy, 1980): it is impossible to enumerate and test every condition that could block the predicted effect of an action.

In the first proposed extension, the fish treats the set of categories like a single state and learns a mapping from state to state:

$$f\,(\mathbf{m}_t,\, c\,(\mathbf{e}_t)) \rightarrow c\,(\mathbf{e}_{t+\delta})$$

Unfortunately, there will be a large number of states, so it would take a long time for the fish to learn all the salient state transitions. To some extent, it is possible to estimate state transitions from similar past transitions (see appendix A on state generalization). This technique is unlikely to work well here though, since necessary and incidental categories are as yet neither comparable nor distinguishable. The fish could use the model f as it stands. The model is just not very clever because it would require of the fish more trials than we human beings would need to make good inferences. It is doubtful that the model's limitations would really become apparent in the fish's simple environment as long as steps were taken to limit the fish's sensing to a handful of very close objects. (This is easily accomplished by using a high threshold and, in this way, exploiting the rapid drop off in concentrations that comes with the inverse squares law.) But a representational system that functions only because of the insensitivity of the fish's sense organs or the simplicity of its environment will not scale up.

A second extension to the fish's model would be to reduce the number of extraneous categories included in its CSM representations. The fish's predictions would no longer be based on simple state transitions, mapping from the set of all

detected categories at time t to the set of all detected categories at time $t + \delta$. Instead, it would map from one subset of detected categories—the subset necessary to enable the transition—to another subset of detected categories—the subset necessary to enable other transitions.

As yet perceptual categories are either active or inactive, but there is no notion of them standing in relation to one another in some abstract way. This makes for cognitive capacities that are rather primitive in a phylogenetic sense. As Fodor and Pylyshyn (1988) note, intelligent behaviour depends not only on learning to detect relations between objects but also being able to apply this knowledge in a systematic way:

> It is not, however, plausible that only the minds of verbal organisms are systematic. Think what it would mean for this to be the case. It would have to be quite usual to find, for example, animals capable of representing the state of affairs *a*R*b*, but incapable of representing the state of affairs *b*R*a*. Such animals would be, as it were, *a*R*b* sighted but *b*R*a* blind since, presumably, the representational capacities of its mind affect not just what an organism can think, but also what it can perceive. In consequence, such animals would be able to learn to respond selectively to *a*R*b* situations but quite *un*able to learn to respond selectively to *b*R*a* situations. (pp. 40-41)

To move towards truly symbol system-like behaviour would require a third extension: the fish would need to be able to sensitize itself to constituent structure at the perceptuo-cogntive level (see chapter 1).

The Frame Problem

As an agent performs actions, most aspects of the world do not change. The original frame problem asked: How can an agent take account of this fact *without* recourse to frame axioms? McCarthy and Hayes (1969) first uncovered the problem within the situation calculus. They tried to represent nonchanges explicitly with frame axioms but discovered that, for complex problems, the number of required frame axioms would far outstrip the computational power of any computer. Besides the fact that frame axioms are impractical in terms of computer memory and processing time, they

demand of the programmer something close to omniscience. The programmer must second guess every conceivable nonchange that could occur.

Since programmers are not omniscient (and neither are their programs), bizarre things tended to happen in programs that use frame axioms. Everything might appear all right until the agent leaves the room. Then suddenly the room disappears (see Harnad, 1993). The programmer forgot to explicitly represent this nonchange. The programmer could add another frame axiom, but this kind of tinkering never seemed to be enough. It often introduced new problems. Sometimes things *do* change and in unexpected ways. But frame axioms cannot cope when several things happen at once or when the effects of actions take time to propagate. This is a problem because, although opening a tin of beans does not cause the Prime Minister to resign, the Prime Minister could resign while you are opening a tin of beans.

Researchers first noted the frame problem in deductive systems because that is where it is most apparent. It was less obvious in (mainly) procedural systems (like STRIPS) which used means-end reduction for planning instead of logic—at least until researchers tried to expand these systems to handle conditional actions, concurrent actions (as in the example where the Prime Minister resigned), and actions with side-effects (like a cup moving with its saucer, see Janlert, 1987, p. 20-21; Morgenstern, 1996, p. 116). However, when it became clear that simply abandoning logical formalism would not remedy the problem of representing change, researchers identified the *general* frame problem, that is, "the problem of finding a representational form permitting a changing, complex world to be efficiently and adequately represented" (Janlert, 1987, pp. 7-8).

> The frame problem is a representational problem, a problem of epistemology, of how to maintain an internal symbolic map of a (changing) world, not a problem of heuristics, of how to puzzle over and engage in the world. Even if the agent refrains from planning and interacting, simply maintaining its representations in step with the represented world, there will generally be a frame problem. (Janlert, 1996, p. 43)

This last sentence makes clear the relation between the general frame problem and the symbol grounding problem. The symbol grounding problem asks how to causally *connect* symbols to the things they represent; the frame problem asks how to *maintain* that causal connection.

Now if we assume that, to produce anything close to a human-level of performance, a robot must at least learn some of its symbol categories (Harnad, 1987, 1989, 1990a, 1993; Schyns et al., in press; and chapter 3), then it stands to reason that the same mechanisms involved in category induction may also be involved in maintaining the connection between its symbols and the things they represent. In other words, it is likely that the maintenance of symbol-object correspondences is part of the ongoing symbol grounding process. Unfortunately, in discussing the frame problem researchers often presuppose that there is either no symbol grounding problem or that the problem has already been solved. The grounding of most (if not all) systems designed to address the frame problem is not a real grounding: it is either parasitic on a human user's interpretation (as is the case with disembodied systems) or it relies on hardwired connections usually based on a few *a priori* feature detectors (see Harnad 1990a; chapters 1 and 2).

This is why it is useful to recast the frame problem in terms of insights drawn from the symbol grounding problem and the hermeneutic hall of mirrors (Harnad, 1990a, 1990b):

> At any point, a symbol system has only dealt with a small amount of data (relative to human-scale performance). That's why such systems are often called "toy" systems. Toy performance, relative to human-scale performance, is highly underdetermined... Yet in projecting a systematic (usually natural-language) interpretation onto such a toy, one is at the same time overinterpreting it (typically overinterpreting it mentalistically, in terms of what it "knows," "thinks," "means"). And, in my view, a "frame" problem arises every time we run up against evidence that we have exceeded the limits of that underdetermined toy; evidence that we are overinterptreting it—and have been all along. (Harnad, 1993, ¶11)

If, to escape the frame problem, we find that a symbol system must be grounded in the external states of affairs it represents, this would be another strong piece of evidence indicating that the semantic coherence of intentional states must depend at least partly on extrinsic properties (see chapter 1).

The frame problem is associated—and sometimes confused—with other related problems (see chapters in Ford & Pylyshyn, 1996):

book-keeping problem	how most effectively to keep track of hypothetical situations and plans
control problem	how to choose a suitable action without considering unrelated facts and inferences
Hamlet's problem	how to decide what to contemplate on and when to stop thinking and start doing
holism problem	how to make salient inferences without precluding the possibility that what has been inferred may depend on virtually any proximal condition
indexing problem	how to find the most promising heuristic for a given problem
interfacing problem	how to interface heuristics and representational forms that are specialized for different domains in solving problems that cut across those domains
persistence problem	how to ignore unchanged facts
prediction problem	how to formulate representations that stay applicable
qualification problem	how to take adequate yet parsimonious account of conditions that could prevent the predicted effect of an action
ramification problem	how to predict change across time including the delayed and indirect effects of actions (compare with the *temporal projection problem*)
relevance problem	see the *control problem*
relevance-holism	how to predict what is likely to pertain to a particular goal without exploring too many possibilities
revision problem	how to resolve conflicting facts and integrate unexpected information
temporal projection	how to predict change across time

A number of computer scientists (e.g. Hayes and McDermott) have criticized philosophers (e.g. Haugeland and Fetzer) for trying to broaden or change the definition of the frame problem. Their consensus is that the frame problem is a problem *for the robot designer* or programmer: "The problem is not an epistemic one of [the robots ascertaining what changes and does not change]—it is a representational one of how to express the information we want our robots to use" (Hayes, 1991, p. 72).

> The frame problem is not about the content of the representation as such; the "physics" of the world is the concern of the prediction problem. The frame problem concerns the choice of a form in which the physics can be stably expressed, reflecting a categorization of the world that minimizes change. The right form, the specific form that catches the stabilities of the world, is, of course, dependent on the world, and in that sense, the frame problem does not have one solution, but at least as many as there are worlds.... the designer will have to look to the specific world. (Janlert, 1996, p. 43)

Of course, until we have robots that can evolve their own representational forms, the frame problem will be a problem for the designer. However, from the standpoint of the symbol grounding problem, it may be misleading to focus a designer's attention on finding the right representational form for stably expressing the physics of a particular domain. If our ambition is to build systems that could eventually approach a human level of performance—or even if it is only to model intentional processes which appear to operate under both intrinsic and extrinsic constraints—then our first priority should be to find representational forms that allow symbols to be empirically and adaptively grounded from the bottom up. We should solve the symbol grounding problem first and then worry about the frame problem if it is not solved in the process.

If we focus on solving the frame problem for different domains, we will only later be faced with an enormous interfacing problem when we try to simulate general intelligence by integrating domain-specific abilities. Janlert claims that the frame problem has as many solutions as there are worlds. Dreyfus and Dreyfus (1988) might cite this as an example of a researcher confusing the concept of a universe with that of world:

> A set of interrelated facts may constitute a *universe*, like the physical universe, but it does not constitute a *world*. The latter, like the world of

> business, the world of theatre, or the world of the physics,... are not related like isolable physical systems to the larger systems they *compose* but rather are local elaborations of a whole that they *presuppose*. Microworlds are not worlds but isolated meaningless domains, and it has gradually become clear that there is no way they could be combined and extended to arrive at the world of everyday life. (Dreyfus & Dreyfus, 1988, in Boden, 1990, pp. 324-325)

Although this prognosis is perhaps premature, Dreyfus and Dreyfus have identified a problem with integrating different ungrounded formalisms.

Progress in specialized domains will no doubt appear to slow if, instead of devising domain-specific symbol systems, we start to build systems that are fully grounded in robotic capacities. However, the development of grounded systems may ultimately provide the best means of integrating domain-specific expertise. For example, if all expertise is rooted in categorical-sensorimotor representations, these representations might be able to provide a common basis—and, hence, a common interface—for the different kinds of abstract models that are developed from them.

Harnad's (1993) proposed solution to the frame problem is to ground symbolic capacities in robotic capacities:

> An ungrounded symbol system has only one set of constraints: purely formal, syntactic ones, operating rulefully on the arbitrary shapes of the symbol tokens. A grounded symbol system would have a second set of constraints, bottom-up ones, causally influencing its internal symbols and symbol combinations, constraints from the internal, nonsymbolic machinery underlying its robotic capacities, especially categorization (Harnad 1987, 1992; Harnad et al. 1991), which is what would allow the system to pick out what its symbols are about without the mediation of external interpretation. (¶8)

Since the frame problem has already been identified with unconstrained logical formalisms, it is easy to see why bottom-up constraints might be crucial in finessing it: As Janlert (1996) notes, "*Because* logic can be used to represent an extremely wide range of circumstances, *because* so much is logically admissible, which is not empirically admissible, logic representations suffer from the frame problem" (p. 45). Bottom-up constraints, grounded in learned (and evolved) categories and sensorimotor predictions, could help to confine the logically admissible to the empirically admissi-

ble. And, to link with our earlier point, they could also help to ensure that the various abstract domain-specific models an agent develops remain compatible.

Schyns and his colleagues (in press) have censured inductive techniques that embrace the expressive power of symbolic formalisms while disregarding the importance of perceptual and empirical bottom-up constraints (§3.2). These techniques underconstrain the feature space because their symbolic form permits features that are arbitrarily complex and because they are oblivious to perceptual factors that influence the creation of features in humans (e.g. topology, proximity, contiguity, global coherence). They have argued that learned features can more flexibly characterize internal representations in terms of relatively *raw* stimulus properties—prior to any interpretation or symbolization.

Janlert states that "at the bottom of the frame problem lies a tension between freedom on the one hand and stability and simplicity on the other" (p. 45). He may be right in arguing that with analogue representations—or *pictures* as he calls them—you can have all three. However, if any lesson is to be learned from the symbol grounding problem, it is that stability and simplicity should be rooted not in programmer-imposed pictures but in bottom-up analogue constraints.

Whether the fish simulation suffers—or *can* suffer—from the frame problem depends on your perspective. We might deem any system to be suffering from the frame problem if it cannot meet certain performance criteria. This is the position Morgenstern (1996) takes. She criticizes STRIPS, despite its clarity and efficiency, because it uses the oft applauded strategy of letting sleeping dogs lie: She notes that the possibility of unknown concurrent actions renders false the assumption that things do not change unless they are explicitly declared to change (p. 116). As it stands, the fish simulation is not able to plan about the world. But if it were successfully extended in the three ways proposed in the last section, this would tie it to the *sleeping dogs* strategy. Furthermore, though capable of planning, the fish would still not be able to reason deductively.[52]

---

[52] STRIPS had been enhanced to include synchronic deduction in addition to its diachronic planning.

However, there is another sense in which the fish simulation, even with extensions, clearly does *not* suffer from the frame problem. It does not need frame axioms or anything like them to represent the fact that most aspects of its environment do not change as it performs actions. This is the power of the sleeping dogs strategy. Furthermore, unlike STRIPS and many other systems, its symbol categories are inductively grounded in the objects they represent.

Summary

In chapter 6 we developed a simulation in which a fish acted in accordance with learned categorical-sensorimotor representations. In this chapter we proposed three ways of extending its predictive model:

(1) The fish could learn to map from (learned) external categories and motor signals to consequent external categories.

(2) It could develop a more parsimonious and useful mapping by removing incidental categories from its representations of predicted changes in its external categories.

(3) It could learn to represent relations between objects by means of constituent structure.

If these extensions proved to be possible, then the fish's performance would be roughly on par with STRIPS's diachronic planner and, additionally, the fish's symbol categories would be grounded in their aquatic environment.

There are many ways for our fish to represent its world, but some are clearly better than others. The frame problem concerns finding the right form of representation. It should mirror the stabilities of the world so that a robot can efficiently maintain a causal connection between its internal symbols and external states of affairs. Our fish simulation is not faced with the frame problem but that has much to do with the fact that its representations are impoverished by human standards.

Harnad (1993) has astutely pointed out that, in order to solve the frame problem—at least for robots whose performance is to approach our own—we have to solve the symbol grounding problem first. The reasons are clear. Before finding a form of representation that helps maintain causal connections between symbols and

objects, we have to think about how those connections could have come to be; we have to think about how the robot's internal symbols could be grounded. Symbol grounding is likely to involve processes of category induction and sensorimotor learning. If, as Harnad suggests, the frame problem is an instance of the symbol grounding problem, then these same processes will be crucial to solving both.

The trouble with most approaches to the frame problem is that they are based exclusively on symbolic formalisms. Since many more concepts are formally admissible than are empirically admissible, it is easy for a system with only formal constraints to get bogged down trying to sort out the absurd and implausible. Harnad (1993) suggests that bottom-up constraints could help a grounded symbol system to overcome the frame problem.

Thus, we should focus our efforts on finding forms of representation that allow robots to ground their symbols empirically. Categorical and sensorimotor representations could provide a common basis on which a robot could develop more abstract models. Perhaps this common basis could make it easier to interface between the robot's various specialized models when it comes time to apply them to a problem that cuts across several domains.

Chapter 8. Conclusion

Summary of Contributions

Every human body is unique as is every personal history. This places each individual in a unique ecological relation, and it is from this vantage that we develop an understanding of our surroundings and each other. We come to this understanding largely through experience. A theory of representation, irrespective of its form, should be able to take these facts into account if it is to explain how our extrinsic relation to the world can causally influence our intensional states. My doctoral findings suggest that, by learning sensorimotor predictions, it may be possible to simulate intelligent behaviour that is grounded in physiology and experience.

In considering several experiments with actual robots as well as creatures that exist only in computer simulations, we explored ways in which agents can learn sensorimotor predictions about the consequences of their behaviour on the basis of past interactions and how they can profitably use these predictions to guide their future activity. Preliminary results suggest that learning sensorimotor predictions can enable robots to perform such fundamental cognitive tasks as

- learning categorical representations that capture invariant features of objects (chapter 3)
- learning sensory mappings for hand-eye coordination (chapter 4)
- learning sensorimotor mappings for planning movement through cluttered environments (chapter 5)
- discovering salient features of their environment (chapters 6 and appendix A)
- learning how to act on the basis of sensorimotor predictions (chapters 6 and appendix A)

The dissertation has also argued that the development of predictions can enable robots to perform such complex cognitive tasks as

- making plans in terms of sensorimotor categories (chapter 7).

There is sufficient evidence to conclude that an adaptive ecological alternative to the methodological solipsism of symbol systems (see Fodor, 1980) not only exists in principle but can be used in simulations with currently available technology. Inroads have also been made in showing that this alternative can be used to simulate verbal behaviour, the first and last bastion of the symbol system.

Discussion

Not Particularist Enough?

Brooks (1991a, b) and his colleagues have built robots that can perform certain tasks not only without *a priori* symbolic representation but also without centralized representation of any kind (e.g. a coordinate space). In fact, some of them avoid any learning, planning, or search. The performance of these robots appears to degrade more gracefully in the face of environmental change than that of symbolic approaches.

This is an important demonstration. Yet it is one of limited significance for our purposes because Brooks' approach does not permit a robot to adjust, without human intervention, to either a radically different environment or a radically different suite of sensors and actuators. With a predictive model, a high degree of adaptability is possible. By contrast, Brooks' robots perform well in a new domain only after a long and painstaking process in which the engineer has designed, tested, and redesigned every layer of their architecture. For example, if Brooks wanted his robot Herbert to collect soda cans underwater instead of in office spaces, he would not only need to give it a very different body but also a completely reworked controller. But if we wanted to use a closest-point method to model the dynamics of a remote-controlled boat instead of a remote-controlled car (see chapter 5), there is nothing fundamental about the predictive model that would need to be changed, although the robot would be guiding a different vehicle with different actuators across a different medium. Also, from the perspective of the robot engineer and user, robots that can adapt to new tasks with minimal tinkering will be superior to those that cannot.

However, a predictive model might appear behaviouristic, especially when compared to Brooks, because it concentrates on adaptive capacities shared by

different species and because it characterizes the different kinds of intelligence that these species exhibit in terms of the same theoretical construct: learned sensorimotor predictions. In this respect it is similar to learning theory and unlike ethology, a field which studies species and individuals in terms of their particular niches and relationships.

This dissertation's consideration of how differences develop between individuals and between species may be faulted for being too narrowly focused and empirical. We have discussed how sensorimotor predictions may be learned that take account of an individual's particular body, sense organs, physiological reactions, and life history. But we have neglected to discuss how evolutionary and developmental adaptations have shaped and differentiated (both ontogenetically and phylogenetically) the physiological processes involved in empirical adaptation. Just as Brooks and his colleagues have very specifically adapted robots to their environments, so too has natural selection very specifically adapted animals to theirs. Cognitive abilities are not species independent but reflect niche-specialization not accounted for by the proposed predictive model.[53]

The aim of this dissertation is certainly not to rehabilitate learning theory. It rejects the premise that operant conditioning shapes thinking with feedback from only sensory stimuli and a scalar-valued reward function. Learning to 'map' the world is valuable for predicting outcomes in novel situations. The dissertation's aim is only to draw attention to the following points:

- Empirical adaptation is necessary for even the most basic kinds of behaviour: indeed, any kind of behaviour that requires an agent to exploit sensorimotor relations that change in unforeseeable ways.
- It is useful to characterize empirical adaptation in terms of the learning of sensorimotor predictions—predictions that are particular to the individual.

---

[53] That is why intelligent behaviour cannot be mapped onto a phylogenetic ladder with humans at the top. At certain cognitive tests pigeons can outperform humans (see citations in McFarland & Bösser, 1993, pp. 9 and 14).

- Their learning can be simulated computationally.
- Not only can a predictive model cope with environmental and somatic variations, it can also serve as a foundation for modelling higher-level cognitive processes.

The results of this dissertation support the conclusion that a predictive model is sufficiently powerful to supplant competing solipsist models in many domains.

Admittedly, individual and species differences are even more important than the present predictive model suggests. Evolved species-specific specialization and neuroanatomical development are crucial in facilitating well-adapted behaviour. Different individuals have different brains—brains that are the product of different developmental processes (Edelman, 1992)—and different species form qualitatively different kinds of predictions with greater or lesser alacrity. But aspects of the behaviour of all vertebrates are amenable to a characterization in terms of developed predictions. Undoubtedly, there will be great variations in how predictions manifest themselves in the behaviour and physiology of individual organisms.

Symbolic representation, by contrast, tends not only to discount these kinds of variations—the kind resulting from evolved neurophysiological adaptations—but also variations resulting from differences in body and experience. What I have tried to show is that representation must at least be adapted to these latter, and that learned sensorimotor predictions are a suitable form of representation for doing this.

Future Work

Neuroscientific Applications. Closest-point techniques were chosen to implement the learning of sensorimotor predictions because they have a clearer semantics than neural networks and learn more efficiently in the domains considered (see Omohundro, 1987, 1990a). However, predictions may prove to be more that just a convenient way of describing and simulating behaviour. Neuroscientists may wish to ascertain whether the brain has structures that indeed function as predictions. If it does, they may wish to isolate them, find out how they develop, and model them computationally. Sommerhoff and MacDorman (1994, §5) point to one possible neural implementation.

Future research could be directed towards simulating how perseverative changes in networks of neurones could account for the learning of predictions.

Analysing and Improving Other Models. Chapter 4 introduced a technique of *intertheoretic interpretation* for analysing a theory or implementation in terms of a high-level cognitive description. Researchers may apply this technique destructively: to critique nonadaptive theories. They may also apply it constructively: to enhance an implementation and, in doing so, to find a new application for the predictive model.

Multimodal Integration for Perception. Even when information is available to only one sensory modality, we are usually still able to make out objects. Nevertheless, evidence suggests that this ability often develops *because of* multimodal interactions (see chapter 4; Kohler, 1964). A potential line of research is the construction of robots that can learn to map between, for example, tactile contact, movement parallax, binocular disparity, vergence, changes in apparent size, and motor signals. The robot might begin by following the contours of an object with its hand while foveating the point of contact. Once it has learned predictions that map from the topography of the retina to hand positions, the object's appearance can begin to elicit predictions about its shape. This would enable a robot to recognize the object—and the behavioural possibilities it affords—without hands on contact. Methods such as these, that make use of multimodal feedback at various levels of abstraction, may be able to finesse the problems of object constancy and vanishing intersections *without* bootstrapping from evolved feature detectors.

Probabilistic Planning. We need to gain more insight into how an agent can cope with uncertainty and, if possible, eliminate it. To eliminate it, it must learn to detect signs that are correlated with unexpected outcomes. There are certain challenges that need to be addressed in applying a closest-point approach to a nondeterministic environment. If motor signals only occasionally produce a desired effect, it may be worth repeating an action until the effect is achieved. It is possible to implement probabilistic predictions that base estimates on the ratio of successes to failures among nearby sensorimotor representations. But there may in fact be no nondeterminism—just a very convoluted coordinate space. In this case, the above method of estimating probabilities would actually be wiping out critical sensory detail. Future research might explore how best to balance these opposing demands.

Abstract Planning. This dissertation has suggested that abstract planning is not so abstract: it is probably rooted in learned sensorimotor predictions (see chapters 6 and 7). These predictions relate sensory projections either to consequent sensory projections or to (expected discounted values for) internal variables (see chapters 6 and 7 and appendix A). No attempt has yet been made to integrate these approaches in a simulation. In traditional programs, planning is motivated toward some external goal. The problem with these systems is that they require the programmer to specify in advance what the goals can be. At the other extreme, reinforcement learning does not make use of any explicit goals. The disadvantage of this approach is that it cannot map relations between external objects or events: there are only states. Being able to map relations between objects and events can be extremely useful. We should perhaps look for an integrated approach that permits an agent to plan in terms of predictions that are about both its external world and its internal variables.

Action Generalization. Fixing in advance an agent's choice of action places serious constraints on its adaptability (see appendix A). It would be worthwhile to reimplement the simulation in chapter 6 to allow the fish to use real-valued motor signals to manipulate its mouth and fins. We might wish to pursue the model in a reinforcement learning paradigm. In this case it must be able to reinforce negatively an action sequence that leads to the attainment of a bad goal (e.g. searching for and eating poisonous food) while reinforcing *positively* successful subsequences (e.g. quickly getting the food into its mouth). How to do this with discounted rewards is an open question, since short-term and long-term goals fade into one another.

Appendix A. Grounding State and Action in Reinforcement Learning

Introduction

In the past decade interest in reinforcement learning has renewed. Approaches based on symbolic representation have not lived up to expectations, and AI researchers have become more apt to question whether the simulation of formal reasoning provides insight into even human minds. Attention has been refocused on the full embodiment of simpler forms of behaviour. However, the intent of reexamining reinforcement learning has not been to revive behaviourism. The idea that complex behaviour develops from a *tabula rasa* by means of simple conditioning and reinforcement learning has long been put to rest (see e.g. Chomsky's 1959 critique of Skinner, 1957).[54] Rather the intent is to devise computational models that better reflect the adaptability of complex organisms.

Although no one in AI claims that reinforcement learning could supplant symbolic representation outright, on the surface it at least appears less reliant on programming. Some researchers hope that reinforcement learning might one day complement symbolic approaches in a hybrid system. Reinforcement learning has recently benefited from the further development and application of a class of algorithms for learning from delayed rewards. They allow an agent's choice of action to be revised not only according to immediate rewards but also rewards predicted to occur in the future on the basis of its past history.[55] Disenchantment among some with symbolic AI has coincided with recent progress in other once maligned research fronts like neural networks—fronts which often lay claim to being more neurologically plausible.

---

[54] Even when reinforcement learning offers the best explanation of behaviour, its efficient application relies heavily on an evolved domain-specific reward function.

[55] The fact that an action is correlated with higher future rewards does not mean that they resulted from that action. Nevertheless, it appears possible to learn intelligent behaviour without solving this problem, sometimes called the credit assignment problem.

Many people in AI are looking for new paradigms to rally behind—paradigms that at least on the surface appear more biological and adaptive.

However, in appraising the adaptability of our artefacts we must confront questions concerning what they are capable of sensing and responding to and how they come to do this. For an agent to be truly intelligent it will need to be able to cope with qualitatively new kinds of things, and so the distinctions it draws about the world will need to be flexible and not entirely predetermined by its programming.

The State Grounding Problem

Regardless of whether symbolic planning or reinforcement learning is used to control an autonomous robot, to act effectively the robot must causally connect internal representations to external conditions. In symbolic planning, these representations are composed of symbols that stand for objects, events, and relations. In reinforcement learning, they stand for *states*. The *state grounding problem* concerns identifying a robot's state from spatiotemporal correlations in its sensorimotor projections (viz. the input from its sensors and motor signals).

Reinforcement learning has been proven to converge to an optimal action policy (as $t \rightarrow \infty$) so long as the Markov property applies. For this to be true, the identification of the robot's current state must provide enough information for the learning algorithm to determine what action has the highest expected reward. We must assume the following in order for the property to hold:

- The robot has a state identification mechanism that is able to determine its current state.
- The state identification mechanism has no source of input apart from the robot's sensors and motor signals.

Unfortunately, it is difficult to delimit precisely what state is. When applying reinforcement learning to a simple physical system, it is often sufficient to characterize state by a complete, properly quantized description of that system (e.g. the positions of its various parts). If we are concerned with building robots that can operate in more natural environments, such detailed information is not available. A robot's state should not be confused with its current sensorimotor projections,

although it may draw on current and previous projections. We can only say that, to guarantee convergence, a state must distinguish everything necessary for the robot to select the best action.

In practice it is generally a programmer who must, on a case by case basis, hand-code a state identification mechanism that is appropriate to an agent's particular domain, and much foresight and ingenuity are required to ensure that states differentiate the domain's relevant aspects. For an autonomous robot, this usually entails the construction of a separate module for perceptual analysis whose job it is to detect preselected features in the raw sensory data.

However, the reliability of reinforcement learning, unlike symbolic planning, does not depend on a correspondence being maintained between external objects, events, and relations and their internal counterparts. This is because it does not make use of symbolic reference except insofar as each state is associated with a particular set of external factors. Therefore, the perceptual analysis required by robot learners differs from that required by robot planners in that, in order to identify states, it may be unnecessary to identify the particular objects, events, and relations that constitute those states. This may mean that, in domains where reinforcement learning is tenable, it is possible to supplant the symbol grounding problem (Harnad, 1990a; MacDorman, 1997) with the perhaps less complex state grounding problem.

Nevertheless, the need to connect internal categories with external conditions remains, and there are several good reasons why not to leave its accomplishment to the robot designer:

(1) The designer must posit the existence of some fixed set of features that are relevant to the robot and that can be specified in advance.

(2) It would require the designer to introspect about which features are relevant to the robot. This is not an easy task because the designer's body places that individual in a substantially different sensorimotor relation to the world from that of the robot. It is doubtful the designer could specify even the set of features relevant to him or her. (Brooks, 1991b, and Clocksin, 1995, discuss related problems with the introspective approach.)

(3) It would require the robot to rely on environmental features specified by the designer in advance. This places the robot at the mercy of the designer's ability to second-guess what features it will need to detect in order to behave intelligently. The robot could not cope with an environment that changed in ways not anticipated by the designer's choice of features.

(4) A major impetus for learning empirically is to minimize the need to make assumptions about the nature of the environment. If the designer must specify features in advance, to some extent, this lessens any advantage of having a robot learn to act on the basis of past rewards rather than in accordance with an *a priori* model.

(5) There is a trade-off between having too few states to distinguish potentially relevant environmental features and having too many to learn efficiently. Although needs may change in unpredictable ways, this trade-off must be set in advance. Deciding how many states are sufficient for reinforcement learning to learn adequate responses is a black art.

(6) It is costly to employ a programmer to set up a state identification mechanism for each new kind of robot or environment.

(7) A model solely dependant on *a priori* states can offer no satisfying explanation of how animals discover what environmental features are relevant to successful behaviour. We have criticized planning with purely *a priori* symbols as unbiological because natural selection could not have anticipated the need for specific symbol categories to represent objects, events, and relations never experienced by an individual's genetic ancestors (see e.g. Maze, 1991; MacDorman, 1997; chapter 2).[56] However, the same argument applies to state categories. Natural selection is no better able to anticipate the need for state categories to

---

[56] Rather it is likely that natural selection has resulted in the evolution of cognitive mechanisms capable of category induction within the lifespan of the animal, see Harnad 1987 and chapter 3.

> represent environmental circumstances never experienced by an individual's genetic ancestors than it is symbol categories. This is clearly because such an adaptation is unlikely to have evolved and to have spread throughout a population until after the need for it has already arisen.

It is clear, however, that at least many vertebrates are able to learn to recognize new patterns of sensorimotor invariance and to exploit them. There is no need to wait for evolutionary pressures to result in the evolution of specific adaptations. An important area of investigation is how mechanisms could evolve that would permit an animal to sensitize its responses to the behavioural possibilities afforded it by its environment. It would appear that even a digger wasp—whose behaviour is largely driven by instinct—can do this; it memorizes landmarks that help it find its ways back to its nest.[57] The mechanisms that permit this are, of course, constrained by biology, and even in far more adaptive creatures, neurological evidence suggests that evolved feature detectors may be sensitive to common low-level types of invariance. What could not be fixed by a creatures genes, however, are feature detectors to recognize all potentially recognizable patterns of invariance. Presumably many of these patterns must be learned, although to some extent bootstrapping from evolved low-level detectors may minimize the need for experimentation.

---

[57] Although the digger wasp appears able to recognize novel objects—pine cones, coloured blocks, etc.—which implies learning of invariant features, it appears poor at adapting its behaviour to new conditions. When it retrieves a caterpillar, it will inspect its hole, before pulling the caterpillar down. If, however, the experimenter pulls the caterpillar away from the opening while the wasp is inspecting the hole, it will drag the caterpillar back to the opening and inspect the hole again. This process may be repeated indefinitely. This example suggests that the ability to learn to recognize relevant invariance may be more fundamental from an evolutionary perspective than the ability to experiment with various behaviour patterns and to learn to select in the future those that had been more successful.

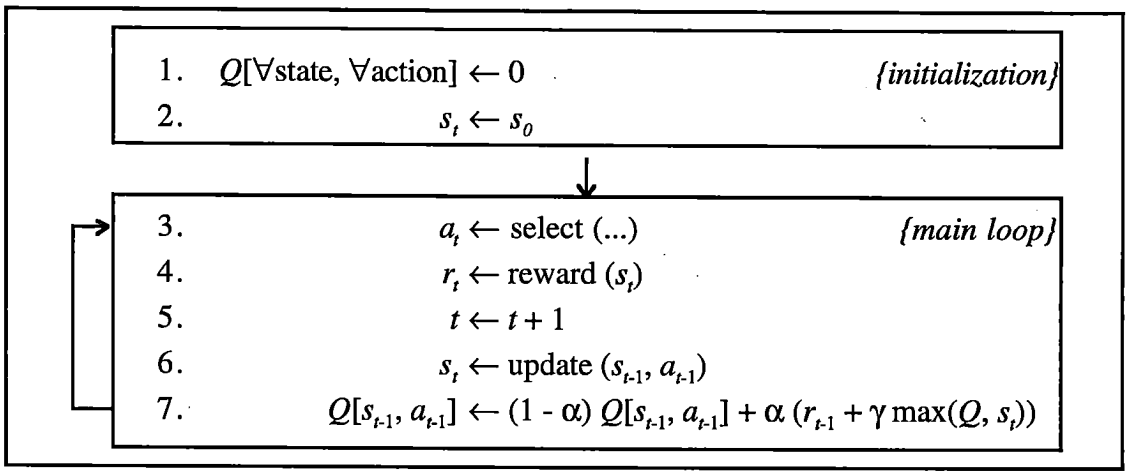
Grounding States with Q-learning

An intelligent agent needs to be responsive to relevant kinds of environmental invariance. Different paradigms conceptualize the distinctions an agent needs to make in different ways. Reinforcement learning uses states, and an agent must be able to make those distinctions that are necessary to maximize reward. What techniques are appropriate for discovering (or better individuating) states depends on whether the action policy is learned on the basis of immediate or delayed rewards. It also depends on whether the method used to learn the policy requires a model of state transitions and rewards. Since our aim is to propose a means for an agent to learn grounded internal states that are not known in advance, it may be better for the robot to use a learning method that does not require a state transition and reward model. This is because this model would normally require the robot to already be able to identify the states it was meant to discover. (However, there are methods of state generalization that may allow the robot to circumvent this limitation.)

If the aim is to discover relevant states, there are certain advantages to having a model to consult. It may aid in detecting a hidden state or equivalence between states. This information is helpful in creating new grounded internal states or eliminating redundant ones. For example, if we assume a deterministic environment, we deduce an anomaly as soon as a state x is found to lead to two (or more) different states y and z, given the same action. Either y and z are in fact the same state (and, hence, need to be subsumed under the same category) or else x indicates more than one underlying state (and, hence, needs to be split). However, there are many trade-offs to consider when both an action policy and a model of state transitions and rewards must be learned. Also, since, for any given action, transitions may occur between any two states, it is possible that the data structure for the transition probabilities would need to be proportional to the number of possible actions times the square of the number of states. For the sake of simplicity, we shall examine a method of learning from delayed rewards that does not require a state transition and reward model.

Watkins proposed a technique called Q-learning based on "incremental dynamic programming by Monte-Carlo method" (1989, p. 81). Instead of performing experiments on an internal model, an agent using one-step Q-learning learns an action policy solely by trying out different actions in different states and calculating the

expected discounted reward associated with them. Hence, the data structure only contains discounted rewards and is proportional to the number of states times the number of actions. Watkins proved that Q will converge to the optimal policy Q* as the number of trials approaches infinity (pp. 220-228). This holds in both deterministic and nondeterministic environments as long as they are stable (in the sense of state transition probabilities remaining fixed). Once learning has converged, the optimal policy is followed simply by picking the action with the highest expected discounted reward for the current state. The Q-learning algorithm is given below:



(1) First initialize (usually to zero) the array Q of discounted rewards for each possible state-action pair and *(2)* set the current state to the initial state. Then repeat the following steps. *(3)* Select an action. Q-learning does not specify how to do this; the convergence proof holds even if it is performed randomly. Sutton (1990a, 1990b) proposed a stochastic procedure based on the Boltzmann distribution. It ensures that the more promising an action is the more likely it is to be selected but that even actions with a low expected discounted reward have some chance of being selected. A temperature value controls the degree of randomness in the selection process so that the higher the temperature, the more likely a less promising action is to be selected. *(4)* An external process determines the reward the agent receives for being in the current state. *(5)* The clock ticks. *(6)* Environmental feedback determines the current state. In deterministic environments, it depends solely on the previous state and the action the agent took. *(7)* The expected discounted reward of the action taken in the previous state is modified according to the reward received in that state and the

expected discounted reward of taking the most promising action in the current state. 0.9 is a typical value for γ and α decreases gradually throughout the learning process.

Quantizing Sensory Projections

Uncovering hidden state is essential to solving the state grounding problem. Hidden state refers to information that could help an agent behave in a more appropriate manner were the agent only able to use it. State may be hidden because it is not directly obtainable from current sensory projections. This is especially a problem for agents that learn exclusively from immediate rewards.

It may be necessary for an agent to take a particular action (or sequence of actions) to create circumstances that will enable it to receive a reward some time in the future. If an agent's state is based solely on its *current* sensory projections and it learns exclusively from *immediate* rewards, it is impossible for it to learn which action to take no matter how many times it produces or fails to produce the necessary action. What if each state takes into account the recent past? If the agent's state is based on its last n sensory projections (a kind of short-term memory), then the number of states will grow exponentially not only with the number of bits in each sensory projection but also with the number of projections. This impedes learning by decreasing the likelihood of the agent being in the same state twice. Even so, an agent that learns exclusively from immediate rewards is still only able to detect a correlation between an action and a reward if both occur no more than n clock ticks apart. Methods of learning from delayed rewards like Q-learning permit an agent to learn what action to take in order to maximize its expected discounted rewards without need for states that take account of anything more than what is available in the current sensory projections.

Oddly enough, state may be hidden from an agent even when it is potentially obtainable from its current sensory projections. This is because, in order to calculate Q-values for an optimal policy, an agent must be able to try out different actions in

the same state repeatedly.[58] But if there are as many states as there are possible sensory projections and the projections are of a reasonably high resolution, learning is nearly impossible because seldom will the same state be entered twice. Nevertheless, since only a portion of the information in the sensory projections will have any bearing on choosing an appropriate action, states do not need to capture every possible distinction in the sensory projections. They only need to capture those relevant to maximizing expected discounted reward.

It is sometimes argued that environments where there is an inherent nondeterminism diminish the potential for uncovering hidden state:

> How can we know if there is hidden state in a nondeterministic world? The answer is probably that we cannot. However, if we assume that the environments we are dealing with have certain useful properties of predictability, it may be appropriate to postulate the existence of hidden state when it is not possible to find a low-entropy operator description for certain salient postconditions (Kaebling, 1993, p. 156).

It is clearly impossible for an agent to improve its action policy in environments that behave completely randomly. This is because there is nothing that can be learned—rewards from past actions say nothing about the relative utility of current alternatives—and because there is nothing to learn—no policy for selecting actions could be better than any other policy. However, all interesting environments offer some degree of predictability. I shall argue that, with only a few *a priori* assumptions about an environment's "properties of predictability," it is possible to discover and exploit hidden state.

What we need to determine is not how Q-learning can in principle discover hidden state in nondeterministic environments. As long as state transition probabilities do not change it can, given a sufficient, perhaps infinite amount of time, discover state that is both hidden in the sense of not being directly obtainable from current sensory projections and in the sense of being buried in a mass of sensory data. What we need to determine is how it can do this efficiently.

---

[58] Later, we can relax this requirement by allowing the agent to approximate Q-values based on the Q-values of similar states—that is, those that are nearby by in a coordinate space.

It may be possible to approximate the Q-values for a newly encountered state from Q-values for previously encountered states if those states are similar enough to the new state (e.g. close enough to it in a coordinate space). For example, for a set of neighbouring states, each dimension of the coordinate space of sensory projections may be weighted according to its effectiveness at predicting Q-values, and then the influence of the states may be weighted according to their proximity to the current state. In chapter 6 we have already discussed various methods of interpolating (e.g. by Delaunay triangulation or using normalized Gaussians to blend quadratic or affine maps). They may be used to approximate the current state's Q-values from the set of neighbouring states. Atkeson has used a proximity-weighted least-squares error minimization to approximate Q-values (see Atkeson, 1990; Atkeson & Shaal, 1995). Triangulation, blending, or error minimization all provide what is essentially a continuous mapping from sensory projections to Q-values. They permit Q-values for an infinite number of states to be approximated based on the results of a finite number of state-action trials.[59]

However, for learning to take place, Q-values must also be updated. It is not immediately clear how to do this with the above schemes. Past states are represented as points in a coordinate space on $\Re^n$, their number is potentially infinite[60] and it is unlikely that the same state would be visited more than once. Thus, if we do not modify how we update Q-values, learning is impossible. We could quantize the coordinate space, but then we would need to determine an appropriate granularity for this, which might not be uniform. Perhaps a better alternative is the following: when Q-values are updated, average the change into the Q-values for nearby past trials in coordinate space, weighting the change according to the proximity of the nearby point to the one that is being updated.

The accuracy of Q-values computed in the above manner will depend on, among other things, how continuous the mapping is between states and Q-values (given the

---

[59] Instead of making use of a closest-point approach, state generalization may be achieved by other methods; for example, Tesauro, 1992, successfully used neural networks with back propagation for state generalization in a backgammon player.

[60] Or at least it is astronomical with the limiting factor being the floating-point accuracy of the representation.

same action) and how much the mapping will change with the passage of time. A vast number of trials and a relatively stable environment may still be required to obtain good performance.

Grounding Actions

It is important to note that actions also need to be grounded. This is often ignored in the literature. Problems are chosen, with pole balancing being the archetypal example, where there are as few as two available actions (Michie & Chambers, 1968). However, when one realizes that, for example, a simple pair of opposing hydraulics will produce the same movement for an infinite number of pairs of forces as long as their difference remains constant, the problem of action generalization becomes clear. The above state generalization method may be applied to actions as well. Instead of having an agent chose between a finite number of actions, Q-values may be stored in a coordinate space whose dimensions include both real-valued sensory inputs and real-valued motor outputs. (For work on approximating real-valued motor outputs for Q-learning see Baird & Klopf, 1993).

Q-Learning As a Model

In some respects Q-learning is not intuitively appealing. In reinforcement learning experiments, maze-running rats are rewarded for following the path that leads to a piece of food. However, evidence suggests that as they run the maze they also learn geometric relations between locations; the rats may be forming a kind of mental map (Tolman, 1932). For example, if the reward is moved, they show improved performance at following the new path. An agent using Q-learning may be able to navigate its way to the reward after a sufficient period of trial and error, but it will not have any internal representation of the geometrical relationships between intervening locations. Having a representation of this kind would save it much trial and error if the experimenter changes the rat's initial location or the location of the reward. Agents possessing internal models can have the benefit of having their hypotheses die in their stead (to quote Popper, see Dennett, 1994, 1996). This is one reason why this dissertation has pursued a path toward the development of systems that can represent objects and relations among objects.

A variant of Q-learning may be used to model the effect of potential actions on internal variables. The agent would use a vector of Q-values—a Q-vector—in place of a single Q-value. Each Q-value in the vector signifies the agent's expected discounted reward as it relates to one or more physiological variables—P-variables, for short. For example, a Q-value related to 'hunger' satisfaction may be derived from P-variables for blood sugar level and the state of digestive processes; a 'stuffed' Q-value may be derived from these and other P-variables.

The calculation of Q-values from P-variables is not straight-forward in part because the agent needs to optimize P-variables, not maximize them. For example, an animal should not eat if it is already bursting at the seams. Facts like these make the relationship between P-variables and Q-values nonlinear. Hence, the calculation of the relative urgency of ameliorating a P-variable is also nonlinear. In addition, it is dynamically influenced by other P-variables.

Despite these complications, it is worth noting that even with the simple Q-value strategy the agent's physiological condition needs to be included in the notion of state. Otherwise, learning would be made very difficult because what appeared to be the performance of the same action in the same state would sometimes be rewarded (e.g. when the agent happened to be hungry) and at other times not (e.g. when the agent happened to be full). As with Q-vectors, the simple Q-value strategy requires a reward function that, when maximized, disallows overfeeding. So an overall reward criteria—for example, an indicator of general health or reproductive fitness—must likewise take into account the nonlinear way it relates to P-variables.

Q-vectors have the potential to provide an agent with a much richer model than the simple Q-value strategy because they can map changes in internal conditions based of the effects actions have in particular states. However, unlike with symbolic planners (see chapter 7), goals cannot be related to particular external objects because those objects are nowhere represented explicitly. In the simplest scheme, the Q-vectors may be mapped on to a single Q-value that indicates the best action for the agent to take in a given state. If the agent is very thirsty and in the past it has been able to reach potable liquids from its present state, the corresponding Q-value will predominate. Thus, it can obtain maximal reward for pursuing a course of action that is predicted to increase its intake of potable liquids. If the agent is instead very hungry, a different Q-value will predominate, and it can obtain maximal reward by

pursuing a course of action that is predicted to increase its intake of calories. This is no different from the way ordinary Q-learning works.

There are a number of reasons to use Q-vectors in place of Q-values. Since an animal cannot relive its life to test the effectiveness of different strategies at fulfilling the long-term goals of survival and reproduction, its behaviour must be rewarded for achieving intermediate goals like hunger satisfaction. The motivation to strive for these intermediate goals will, of necessity, be largely determined by natural selection. And it is likely to be easier to evolve mechanisms to set relative priorities between competing intermediate goals than to assign rewards directly to actions. Furthermore, Q-vectors permit an agent to tune the relative weighting among its vector's component Q-values on the fly and without necessitating more learning. With the simple Q-value strategy, if the reward function changes, Q-values must be relearned from scratch.

Perhaps most importantly, Q-vectors are far more likely to finesse the state grounding problem than ordinary Q-values. Any practical model demands a way of handling state generalization for reasons stated in the last section. Q-vectors provide a much richer source of information to draw on in deciding how to generalize from the sensory projections. They make it far easier for an agent to determine how similar a pair of states are because, unlike with the simple Q-value strategy, they reveal the internal changes that result from taking actions in those states. These changes provide a kind of signature that the agent can use to distinguish different states. A multidimensional signature is much more effective at discriminating states than a single scalar Q-value—as is the case with simple Q-learning—because it can discriminate states according to rewards attributed to actions that cause particular effects (and not just a change in overall health). This is one reason why, in the fish experiments of chapter 6, I opted for a multidimensional representation of internal state.

Summary

For an autonomous robot to learn what action to take in a given state by means of reinforcements, it must be able to detect relevant variations in its environment solely on the basis of spatiotemporal correlations in its sensorimotor projections. Thus, those sensorimotor correlations to which it is sensitized determine its current state.

The problem of determining how to map from a robot's current and past sensorimotor projections to its current state is the state grounding problem.

Its solution is generally left to the robot designer who must hand-code a state identification mechanism for the particular robot in question. This approach is problematic for reasons such as the following:

- The designer, whose sensorimotor relation to the world differs substantially from that of the robot, is required to determine which spatio-temporal correlations in the robot's sensorimotor projections would indicate environmental variations relevant to the robot's appropriate choice of action.
- Since the designer must set up a state identification mechanism that fixes in advance what kinds of correlations are detectable, the robot cannot adapt to environments that change in ways not captured by the correlations the mechanism detects.

A designer can often make a simple reinforcement learning domain tractable both by specifying how sensorimotor information is to be partitioned into states and by limiting a robot's choice of possible motor signals. It is, however, unlikely that this approach would work in complex environments. Some method of state-action generalization is necessary to provide an empirical method of grounding both states and actions in sensorimotor projections. As a correct action or series of actions may be necessary for a robot to receive a reward some time in the future, it also makes sense to adopt a method that permits learning from delayed rewards. If a robot using reinforcement learning is to discover state for itself, Q-learning offers a method of learning from delayed rewards that has the advantage of not requiring a state transition and reward model. Normally, this model would require the robot to already be able to identify the states we would wish it to discover.

Q-values for an infinite number of state-action pairs may be approximated from a finite number of state-action trials. This may be accomplished by weighting each dimension of a distance metric for the coordinate space of state-action values according to its effectiveness at predicting Q-values. The distance measure may then be used to find the n-closest points to the robot's current state. The Q-values themselves may be approximated, for a given action, by applying proximity-weighted

least-squares error minimization to those points, or by a number of other methods. In order to keep the Q-values for past trials up-to-date, when Q-values for a given trial are updated, the change is spread over nearby points as well, weighted according to their proximity to the updated point. The accuracy of Q-values calculated in this way will depend on the number and distribution of past state-action trials, the smoothness and continuity of the mapping from state-action pairs to Q-values, and the stability of the environment.

Q-vectors may be used in the same straight-forward way that Q-values are. However, they also open the door to a number of enhancements, such as the tuning of an agent's priorities without relearning. Q-vectors can be used to improve state generalization (and in a manner wholly autonomous from learning or control). This is because they reflect physiological changes that result from performing certain actions in certain states. This added information can be used to better individuate and weight the various dimensions of states and actions during generalization. In sum, Q-vectors, state and action generalization, and learning from delayed rewards can work together to make state grounding possible within the reinforcement learning paradigm.

References

Anderson, J. R. (1983). *The architecture of cognition.* Cambridge, MA: Harvard University Press.

Argyle, M., Salter, V., Nicholson, H., Williams, M. & Burgess, P. (1970). The communication of inferior and superior attitudes by verbal and non-verbal signals. *British Journal of Social and Clinical Psychology, 9,* 222-231.

Ashby, W. R. (1952). *Design for a brain.* London: Chapman & Hall.

Atkeson, C. G. & Shaal, S. (1995). Memory-based neural networks for robot learning. *Neurocomputing, 9*(3), 243-269.

Atkeson, C. G. (1990). Memory-based approaches to approximating continuous functions. In *The Sixth Yale Workshop on Adaptive and Learning Systems.*

Auer, P. and di Luzio, A., Eds. (1992). *The Contextualization of Language.* Amsterdam: John Benjamins.

Austin, J. L. (1962). *How to do things with words.* Oxford: Clarendon Press.

Ballard, D. H. (1991). Animate vision. *Artificial Intelligence, 48*(1), 57-86.

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology.* Cambridge: Cambridge University Press.

Bateson, G. (1979). *Mind and nature: A necessary unity.* New York: Ballantine.

Bateson, P. P. G. (1988). Biological evolution of cooperation and trust. In P. Gambetta (Ed.), *Trust: Making and breaking cooperative relations.* Oxford: Blackwell.

Bentley, J. L. (1975). Multidimensional binary search trees used for associative searching. *Communications of the ACM, 18*(9), 509-517.

Berlin, B. & Kay, P. (1969). *Basic color terms: Their universality and evolution.* Berkeley: University of California Press.

Bickhard, M. H. (1980). *Cognition, convention, and communication.* New York: Praeger.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review, 94*, 115-147.

Blake, A. & Yuille, A. (1992). *Active vision.* Cambridge, MA: MIT Press.

Bloomfield, L. (1933). *Language.* New York: Henry Holt.

Boden, M. A. (Ed.) (1990). *The philosophy of artificial intelligence.* Oxford: Oxford University Press.

Brent, R. P. (1973). *Algorithms for minimization without derivatives.* Englewood Cliffs, NJ: Prentice-Hall.

Brooks, R. A. (1991a). Intelligence without reason. In *IJCAI-91: Proceedings of the Twelfth International Joint Conference on Artificial Intelligence*, Sidney, Australia (Vol. 1), pp. 569-595. San Mateo, CA: Morgan Kaufmann.

Brooks, R. A. (1991b). Intelligence without representation. *Artificial Intelligence, 47*, 139-159.

Bullock, D. & Grossberg, S. (1988). Neural dynamics of planned arm movements. Emergent invariants and speed-accuracy properties during trajectory formation. *Psychological Review, 95*, 49-90.

Burt, P. & Adelson, E. H. (1983). The Laplacian pyramid as a compact image code. *IEEE Transactions on Communication, 31*, 532-540.

Campbell, R. L. & Bickhard, M. H. (1986). *Knowing levels and development stages.* Basel: Karger.

Carpenter, R. H. S. (1990). *Neurophysiology* (2nd ed.). London: Edward Arnold.

Chapman, K. L., Leonard, L. B. & Mervis, C. B. (1986). The effect of feedback on young children's inappropriate word usage. *Journal of Child Language, 13*, 101-107.

Cheney, D. L. & Seyfarth, R. M. (1990). *How monkeys see the world: Inside the mind of another species.* Chicago: University of Chicago Press.

Chomsky, N. (1957). Review of B. F. Skinner's *Verbal Behavior. Language, 35*, 26-58.

Chomsky, N. (1965). *Aspects of the theory of syntax.* Cambridge, MA: MIT Press.

Chomsky, N. (1986). *Knowledge and language: Its nature, origin, and use*. New York: Praeger.

Chui, C. K. (1992). *An introduction to wavelets.* San Diego: Academic Press.

Churchland, P. S. (1986). Neurophilosophy: Toward a unified science of the mind/brain. Cambridge, MA: MIT Press.

Clark, E. V. (1973). What's in a world? On the child's acquisition of semantics in his first language. In T. E. Moore (Ed.), *Cognitive development and the acquisition of language* (pp. 65-110). New York: Academic Press.

Clocksin, W. F. & Moore, A. W. (1989). Experiments in adaptive state-space robotics. *Proceedings of the 7th Conference of the Society for Artificial Intelligence and Simulation of Behaviour*, pp. 115-125.

Clocksin, W. F. (1995). Knowledge representation and myth. In J. Cornwell (Ed.), *Nature's imagination: The frontiers of scientific vision.* Oxford: Oxford University Press.

Collins, A. M. & Loftus, E. F. (1975). A spreading activation theory of semantic processing. *Psychological Review, 82*, 407-428.

Collins, A. M. & Quillian, M. R. (1969). Retrieval time for semantic memory. *Journal of Verbal Learning and Verbal Behavior, 8*, 240-247.

Conrad, C. (1972). Cognitive economy in semantic memory. *Journal of Experimental Psychology, 92*(2), 149-154.

Cowley, S. J. & MacDorman, K. F. (1995). Simulating conversations: The communion game. *AI & Society, 9*(3), 116-137.

Cowley, S. J. (1994). Conversational functions of rhythmical patterning: A behavioural perspective. *Language and Communication, 14*, 353- 376.

Cowley, S. J. (1996). Conversation, co-operation and vertebrate communication. *Semiotica.*

Cowley, S. J. (in prep.). Applications of the English perfect: A behavioural view.

Daubechies, I. (1992). *Ten lectures on wavelets* (CBMS/NFS series in applied mathematics). Philadelphia, PA: Society for Industrial and Applied Mathematics.

Daugman, J. G. (1980). Two-dimensional spectral analysis of cortical receptive field profiles. *Vision Research, 20*, 847-856.

Daugman, J. G. (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America, 2*(7), 1160-1169.

Daugman, J. G. (1988). Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 36*(7), 1169-1179.

De Valois, R. L. & De Valois, K. K. (1990). *Spatial Vision.* New York: Oxford University Press.

Decety, J. & Michel, F. (1989). Comparative analysis of actual and mental movement times in two graphical tasks. *Brain and Cognition, 11*, 87-97.

Dennett, D. C. (1979). *Brainstorms*. Hassocks: Harvester Press.

Dennett, D. C. (1987). *The intentional stance*. Cambridge, MA: MIT Press.

Dennett, D. C. (1991). *Consciousness explained.* Boston: Little, Brown.

Dennett, D. C. (1994). Language and intelligence. In J. Khalfa (Ed.), *What is intelligence?* Cambridge: Cambridge University Press.

Dennett, D. C. (1996). *Kinds of minds: Towards an understanding of consciousness.* London: Weidenfeld & Nicolson.

Dowes, R. M. & Kramer, E. (1966). A proximity analysis of vocally expressed emotion. *Perceptual and Motor Skills, 22*, 571-574.

Dreyfus, H. & Haugeland, J. (1978). Husserl and Heidegger: Philosophy's last stand. In M. Murray (Ed.), *Heidegger and modern philosophy*. New Haven, CT and London: Yale University Press.

Dreyfus, H. L. & Dreyfus, S. E. (1988). Making a mind versus modeling the brain: Artificial intelligence back at a branchpoint. *Daedalus, 117*(1), 15-43.

Dreyfus, H. L. (1971). *What computers can't do*. New York: Harper & Row.

Dreyfus, H. L. (1992). *What computers still can't do*. New York: Harper & Row.

Edelman, G. (1992). *Bright air, brilliant fire: On the matter of the mind.* London:

Penguin.

Erickson, F. & Shultz, J. (1981). *The counselor as gatekeeper*. New York: Academic Press.

Feldman, J. A. & Ballard, D. H. (1982). Connectionist models and their properties. *Cognitive Science, 6*, 205-254.

Feldman, J. A. (1985). Four frames suffice: A provisional model of vision and space. *Behavioral and Brain Sciences, 8*(2), 265-288.

Fikes, R. & Nilsson, N. (1971). STRIPS: A new approach to the application of theorem proving to problem solving. *Artificial Intelligence, 2*(3-4), 189-208.

Fodor, J. A. & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition, 28*, 3-71.

Fodor, J. A. (1975). *The language of thought*. Cambridge, MA: Harvard University Press.

Fodor, J. A. (1980). Methodological solipsism considered as a research strategy in cognitive psychology. *Behavioral and Brain Sciences, 3*(63), 63-110.

Fodor, J. A. (1981). How direct is visual perception?: Some reflections on Gibson's "ecological approach". *Cognition, 9*, 139-196.

Fodor, J. A. (1981). *Representations: Philosophical essays on the foundations of cognitive science*. Brighton, UK: Harvester.

Fodor, J. A. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.

Fodor, J. A. (1987). Modules, frames, fridgeons, sleeping dogs, and the music of the spheres. In Z. W. Pylyshyn (Ed.), *The robot's dilemma: The frame problem in artificial intelligence*. Norwood, NJ: Ablex.

Fodor, J. A. (1994a). *The Elm and the expert: Mentalese and its semantics*. Cambridge, MA: MIT Press.

Fodor, J. A. (1994b). In S. Guttenplan (Ed.), *A companion to the philosophy of mind*. Oxford: Blackwell.

Ford, K. M. & Hayes, P. J. (Eds.) (1991). *Reasoning agents in a dynamic world: The frame problem*. Greenwich, CT: JAI Press.

Ford, K. M. & Pylyshyn, Z. W. (1996). *The robot's dilemma revisited: The frame problem in artificial intelligence*. Norwood, NJ: Ablex.

Forsythe, G. E., Malcolm, M. A. & Moler, C. B. (1977). *Computer methods for mathematical computations*. Englewood Cliffs, NJ: Prentice-Hall.

Friedman, J. H., Bentley, J. L. & Finkel, R. (1977). An algorithm for finding best matches in logarithmic expected time. *ACM Transactions in Mathematical Software*, *3*(3), 209-226.

Fujita, I., Tanaka, K., Ito, M. & Cheng, K. (1992). Columns of visual features of objects in monkey inferotemporal cortex. *Nature*, *360*(6402), 343-346.

Funnell, E. (1995). Objects and properties: A study of the breakdown of semantic memory. *Memory*, *3* (3/4), 497-518.

Gabor, D. (1946). Theory of communication. *Journal of the Institute of Electrical Engineers*, *93*(22), 429-457.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston, MA: Houghton Mifflin.

Glenberg, A. (1997). What Memory Is For. *Behavioral and Brain Sciences.*

Golub, G. H. & Reinsch, C. (1971). Chapter I.10. In J. H. Wilkinson & C. Reinsch (Eds.), *Handbook for Automatic Computation*. New York: Springer-Verlag.

Goody, J. & Watt, I. (1968). The consequences of literacy. In J. Goody (Ed.), *Literacy in traditional societies*. New York: Cambridge University Press.

Gordon, D. M. (1992). Wittgenstein and ant watching. *Biology and Philosophy*, *7*, 13-25.

Gould, S. J. & Marler, P. (1987). Learning by instinct. *Scientific American*, *256*(1), 62-73.

Greenfeld, P. J. (1986). What is grey, brown, pink, and sometimes purple: The range of wild cards in color terms. *American Anthropologist*, *88*(4), 908-916.

Gregory, R. L. (1970). *The intelligent eye*. London: Weidenfeld & Nicolson.

Gross, E. M. & Wagner, D. (1996). *k*-d trees and Delaunay-based linear interpolation for function learning: A comparison to neural networks with error backpropaga-

tion. *IEEE Transactions on Control Systems Technology, 4*(6), 649-653.

Grossberg, S. (1988). *Neural networks and natural intelligence*. Cambridge, MA: MIT Press.

Gumperz, J. J. (1982). *Discourse strategies*. Cambridge: Cambridge University Press.

Gyr, J., Wiley, R. & Henry, A. (1979). Motor sensory feedback and geometry of visual space: A replication. *Behavioral and Brain Sciences, 2*, 59-64.

Haar, A. (1910). Zur Theorie der orthogonalen Funktionen-Systeme. *Mathematische Annalen, 69*, 331-371.

Haas, W. (1968). The theory of translation. In G. H. R. Parkinson (Ed.), *The theory of meaning*. Oxford: Oxford University Press.

Halliday, M. A. K. (1975). *Learning how to mean: Explorations in the development of language*. London: Edward Arnold.

Harnad, S. (1987). Category induction and representation. In S. Harnad (Ed.), *Categorical perception: The groundwork of cognition*. Cambridge: Cambridge University Press.

Harnad, S. (1989). Minds, machines and Searle. *Journal of Experimental and Theoretical AI, 1*, 5-25.

Harnad, S. (1990a). The symbol grounding problem. *Physica D, 42*(1-3), 335-346.

Harnad, S. (1990b). Against computational hermeneutics. *Social Epistemology, 4,* 167-172.

Harnad, S. (1990c). Lost in the hermeneutic hall of mirrors. Invited commentary on: Michael Dyer: Minds, Machines, Searle and Harnad. *Journal of Experimental and Theoretical Artificial Intelligence, 2,* 321-327.

Harnad, S. (1993). Problems, problems: The frame problem as a symptom of the symbol grounding problem. *Psycholoquy, 4*(34). frame-problem.11.

Haugeland, J. (1985). *Artificial intelligence: The very idea*. Cambridge, MA: MIT Press.

Hayes, P. J. (1987). What the frame problem is and isn't. In Z. W. Pylyshyn (Ed.), *The robot's dilemma: The frame problem in artificial intelligence*. Norwood, NJ: Ablex.

Hayes, P. J. (1991). Commentary on "The frame problem: Artificial intelligence meets David Hume." In K. M. Ford & P. J. Hayes (Eds.), *Reasoning agents in a dynamic world: The frame problem*. Greenwich, CT: JAI Press.

Haykin, S. S. (1994). *Neural networks: A comprehensive foundation*. New York: Macmillian.

Heidegger, M. (1962). *Being and time*. New York: Harper & Row.

Heider, E. R. (1971). 'Focal' color areas and the development of color names. *Developmental Psychology, 4*, 447-455.

Hinde, R. A. (1987). *Individuals, relationships and culture: Links between ethology and the social sciences*. Cambridge: Cambridge University Press.

Hume, D. (1975). *Enquiries concerning human understanding and concerning the principles of morals*. Oxford: Oxford University Press.

Hume, D. (1978). *A treatise of human nature*. Oxford: Oxford University Press.

Humphrey, N. K. (1976). The social function of intellect. In P. P. G. Bateson and R. A. Hinde (Eds.), *Growing points in ethology*. Cambridge: Cambridge University Press.

Janlert, L.-E. (1996). Modeling change—The frame problem. In K. M. Ford & Z. W. Pylyshyn (Eds.), *The robot's dilemma revisited: The frame problem in artificial intelligence*. Norwood, NJ: Ablex.

Janlert, L.-E. (1996). The frame problem: Freedom or stability? With pictures we can have both. In K. M. Ford & Z. W. Pylyshyn, *The robot's dilemma revisited: The frame problem in artificial intelligence*. Norwood, NJ: Ablex.

Jastrow, J. (1900). *Fact and fable in psychology*. Boston: Houghton Mifflin.

Jeannerod, M. (1990). The representation of the goal of an action and its role in the control of goal-directed movements. In E. L. Schwartz (Ed.), *Computational neuroscience*, Cambridge, MA: MIT Press, pp. 352-368.

Jeannerod, M. (1994). The represented brain: Neural correlates of motor intention and imagery. *Behavioral and Brain Sciences, 17*(2), pp. 187-202.

Jervis, T. T. (1992). *A Baysian multiple-model approach to the two-armed-bandit*

problem. Unpublished paper. Engineering Department, Cambridge University.

Jespersen, O. (1922). *Language: Its nature, development and origin*. London: George Allen and Unwin.

Jespersen, O. (1924). *The philosophy of grammar*. London: George Allen & Unwin.

Jolly, A. (1966). Lemur social behavior and primate intelligence. *Science, 153*, 501-506.

Jones, J. & Palmer, L. (1987). An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology, 58*, 1233-1258.

Jones, K. A. (1992). Food search behaviour in fish and the use of chemical lures in commercial and sports fishing. In T. J. Hara (Ed.), *Fish Chemoreception*. London: Chapman & Hall.

Kaas, J. H. (1995). The reorganization of sensory and motor maps in adult mammals. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences*. Cambridge, MA: MIT Press.

Katz, J. J. & Fodor, J. A. (1963). The structure of semantic theory. *Language, 39*(2), 170-211.

Katz, J. J. (1964). Mentalism in linguistics. *Language, 40*(2), 124-137.

Kohler, I. (1964). The formation and transformation of the perceptual world (Fiss, trans.). *Psychological Issues, 3*, 1-173.

Kohonen, T. (1984). *Self-organization and associative memory*. Berlin: Springer-Verlag.

Kosslyn, S. M. (1994). *The resolution of the imagery debate*. Cambridge, MA: MIT Press.

Krebs, J. R. & Dawkins, R. (1984). Animal signals: Mind reading and manipulation. In J. R. Krebs and N. B. Davies (Eds.), *Behavioural ecology: An evolutionary approach* (2nd ed). Oxford: Blackwell.

Kueker, D. & Smith, C. (Eds.) (1996). *Learning and geometry: Computational approaches*. Boston: Birkhauser.

Laird, J. E., Newell, A. & Rosenbloom, P. S. (1987). Soar: An architecture for general intelligence. *Artificial Intelligence, 33*(1), 1-64.

Landauer, R. K. & Freedman, J. L. (1968). Information retrieval for long-term memory: Category size and recognition time. *Journal of Verbal Learning and Verbal Behavior. 7*, 291-295.

Landauer, T. K. (1962). Rate of implicit speech. *Perceptual and Motor Skills, 15*, 646.

Larson, R. E. (1978). *Principles of dynamic programming*. New York: Marcel Dekker.

Laver, J. (1993). *Principles of phonetics*. Cambridge: Cambridge University Press.

Lawson, C. L. & Hanson, R. J. (1974). *Solving least squares problems*. Englewood Cliffs, NJ: Prentice-Hall.

Leibniz, G. (1951). De arte combinatoria. In *Selections*. New York: Schribner.

Lenarcic, J. & Ravani, B. (1994). *Advances in robot kinematics and computational geometry*. Dordrecht & Boston: Kluwer.

Levinson, S. C. (1995). Interactional biases in human thinking. In E. N. Goody (Ed.), *Social intelligence and interaction: Expressions and implications of the social biases in human intelligence.* Cambridge: Cambridge University Press.

Lieberman, P., Laitman, J. T., Reidenberg, J. S., Gannon, P. J. (1992). The anatomy, physiology, acoustics and perception of speech: Essential elements in analysis of the evolution of human speech. *Journal of Human Evolution, 23*(6), 447-467.

Linell, P. (1982). *The written language bias in linguistics*. Linköping: University of Linköping.

Locke, J. L. (1993). *The child's path to spoken language.* Cambridge, MA: Harvard University Press.

Luria, A. R. & Vygotsky, L. S. (1992). *Ape, primitive man, and child: Essays in the history of behavior*. New York: Harvester.

Luria, A. R. (1976). *Cognitive development: Its cultural and social foundations* (trans.). Cambridge, MA: Harvard University Press.

Lyons, J. (1977). *Semantics* (Vols. 1-2). Cambridge: Cambridge University Press.

MacDorman, K. F. (1997). How to ground symbols adaptively. In S. O'Nuallain, P.

McKevitt & E. MacAogain, *Readings in computation, content and consciousness*. Amsterdam: John Benjamins. Based on a paper presented at the Cognitive Science Workshop of AISB-95.

Malcolm, C. M. (1995). The SOMASS system: A hybrid symbolic and behaviour-based system to plan and execute assemblies by robot. In *Hybrid Problems, Hybrid Solutions*, J. Hallam, et al. (Eds.), pp. 157-168. Oxford: ISO Press.

Malinowski, B. (1923). The problem of meaning in primitive languages. In C. K. Ogden and I. A. Richards, *The meaning of meaning*. London: Routledge.

Mangan, B. B. (1993). Taking phenomenology seriously: The fringe and its implications for cognitive research. *Consciousness and Cognition, 2*(2), 89-108.

Marcken, C. G. de (1996). *Unsupervised language acquisition.* Doctoral dissertation. Massachusetts Institute of Technology, Cambridge, MA.

Marr, D. (1969). A theory of cerebellar cortex. *Journal of Physiology, 202*, 437-470.

Marr, D. (1982). *Vision.* New York: Freeman.

Masterman, M. (1961). Semantic message detection for machine translation using an interlingua. In *Proceedings of the International Conference on Machine Translation*, pp. 438-475.

Matthews, P. H. (1981). *Syntax*. Cambridge: Cambridge University Press.

Maze, J. R. (1991). Representationalism, realism and the redundancy of 'mentalese'. *Theory & Psychology, 1*(2), 163-185.

McCarthy, J. & Hayes, P. J. (1969). Some philosophical problems from the standpoint of artificial intelligence. In B. Meltzer & D. Michie (Eds.), *Machine Intelligence* (Vol. 4, pp. 463-502). Edinburgh: University of Edinburgh Press.

McCarthy, J. (1980). Circumscription—A form of non-monotonic reasoning. *Artificial Intelligence, 13*, 27-39.

McFarland, D. & Bösser, T. (1993). *Intelligent behavior in animals and robots*. Cambridge, MA: MIT Press.

Mell, B. W. (1988). Building and using mental models in a sensory-motor domain: A connectionist approach. In *Proceedings of the Fifth International Conference on*

Machine Learning, 207-213.

Meyer, D. (1970). On the representation and retrieval of stored semantic information. *Cognitive Psychology, 1*, 242-300.

Michie, D. & Chambers, R. (1968). Boxes: An experiment in adaptive control. In E. Dale & D. Michie (Eds.), *Machine Intelligence 2*, pp. 137-152. Edinburgh: Oliver & Boyd.

Micó, M. L., Oncina, J. & Vidal, E. (1994). A new version of the Nearest-Neighbour Approximating and Eliminating Search Algorithm (AESA) with linear preprocessing time and memory requirements. *Pattern Recognition Letters, 15*, 9-17.

Miller, W. T., Sutton, R. S. and Werbos P. J. (Eds.). (1990). *Neural networks for control*. Cambridge, MA: MIT Press.

Minsky, M. (1975). A framework for representing knowledge. In P. Winston (Ed.), *The Psychology of Computer Vision.* New York: McGraw-Hill.

Minsky, M. L. & Papert, S. A. (1988). *Perceptrons* (expanded ed.), Cambridge, MA: MIT Press.

Moore, A. W. & Atkeson, C. G. (1995). The parti-game algorithm for variable resolution reinforcement learning in multidimensional state-spaces. *Machine Learning, 21*(3), 199-233.

Moore, A. W. (1990). *Efficient memory-based learning for robot control.* Technical Report 209. Computer Laboratory, Cambridge University.

Moore, A. W. (1991). Variable resolution dynamic programming: Efficiently learning action maps in multivariate real-valued state-spaces. In L. Birnbaum & G. Collins (Eds.), *Machine Learning: Proceedings of the Eighth International Workshop*. San Mateo, CA: Morgan Kaufmann.

Muhlhauser, P. & Harré, R. (1990). *Pronouns and people: The linguistic construction of social and personal identity*. Oxford: Basil Blackwell.

Nagel, T. (1974). What is it like to be a bat? *Philosophical Review, 83*, 435-351.

Nelder, J. A. & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal, 7*, 308-313.

Newell, A. & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice Hall.

Newell, A. & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Newell, A. (1980). Physical symbol systems. *Cognitive Science*, *4*, 135-183.

Nilsson, N. J. (1984). Shakey the robot. Technical Report No. 323. SRI AI Center, Menlo Park, CA.

Norvig, P. (1987). *A Unified theory of inference for text understanding* (doctoral dissertation). Technical Report UCB/CSD 87/339. University of California, Berkeley.

Olson, D. R. (1977). From utterance to text: The bias of language in speech and writing. *Harvard Educational Review*, *47*(3), 257-281.

Omohundro, S. M. (1987). Efficient algorithms with neural network behavior. Technical Report UIUCDCS-R-1331. Department of Computer Science, University of Illinois at Urbana Champaign.

Omohundro, S. M. (1989). *Five balltree construction algorithms*. Technical Report TR-89-063. International Computer Science Institute, Berkeley, California.

Omohundro, S. M. (1990a). Geometric learning algorithms. *Physica D*, *42*(1-3), 307-321.

Omohundro, S. M. (1990b). *The Delaunay triangulation and function learning*. Technical Report TR-90-001. International Computer Science Institute, Berkeley, California.

Omohundro, S. M. (1991). Bumptrees for efficient function, constraint, and classification learning. In R. P. Lippmann, J. E. Moody, and D. S. Touretzky, *Advances in neural information processing systems 3*. San Mateo, CA: Morgan Kaufmann.

Paul, R. P. (1981). *Robot manipulators, mathematics, programming and control*. Cambridge, MA: MIT Press.

Pellionisz, A. & Llinás, R. (1979). Brain modeling by tensor network theory and computer simulation. The cerebellum: Distributed processor for predictive coordi-

nation. *Neuroscience*, *4*, 323-348.

Pfeifer, R. & Leuzinger-Bohleber, M. (1986). Applications of cognitive science methods to psychoanalysis: A case study and some theory. *International Review of Psycho-analysis*, *13*(2), 221-240.

Piaget, J. & Inhelder, B. (1969). *The psychology of the child* (trans.). New York: Basic Books.

Piaget, J. (1959). *The language and thought of the child* (trans.). London: Routledge.

Piaget, J. (1969). *Judgement and reasoning in the child* (trans.). London: Routledge.

Plato. (1953). *The dialogues of Plato: Cratylus*, (Vol. 3), B. Jowett (trans). Oxford: Oxford University Press.

Prablanc, C., Echallier, J. F., Komilis, E. & Jeannerod, M. (1979). Optimal response of eye and hand motor systems in pointing at a visual target. I. Spatio-temporal characteristics of eye and hand movements and their relationships when varying the amount of visual information. *Biological cybernetics*, *35*, 113-124.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., Flannery, B. P. (1992). *Numerical recipes in C: The art of scientific programming* (2nd ed.). Cambridge: Cambridge University Press.

Prince, A. & Pinker, S. (1988). On language and connectionism: Analysis of a parallel distributed-processing model of language-acquisition. *Cognition*, *28*(1-2), 73-193.

Pylyshyn, Z. (1980). Computation and cognition: Issues in the foundations of cognitive science. *Behavioral and Brain Sciences*, *3*, 111-169.

Pylyshyn, Z. W. (Ed.) (1987). *The robot's dilemma: The frame problem in artificial intelligence*. Norwood, NJ: Ablex.

Quillian, R. M. (1968). Semantic memory. In M. A. Minsky (Ed.), *Semantic information processing*. Cambridge, MA: MIT press.

Quine, W. V. O. (1960). *Word and object*. Cambridge, MA: MIT Press.

Ramachandran, V. S., Tyler, C. W., Gregory, R. L., Rogers-Ramachandran, D., Duensing, S., Pillsbury, C., Ramachandran, C. (1996). Rapid adaptive camouflage in tropical flounders. *Nature*, *379*(6568), 815-818.

Ramasubramanian, V. & Paliwal, K. K. (1992). An efficient approximation-elimination algorithm for fast nearest-neighbour search based on a spherical distance coordinate formulation. *Pattern Recognition Letters, 13*, 471-480.

Reade, C. (1989). *Elements of functional programming*. Wokingham: Addison-Wesley.

Reddy, M. (1979). The conduit metaphor: A case of frame conflict in our language about language. In A. Ortony (Ed.), *Metaphor and thought*. Cambridge: Cambridge University Press.

Rieser, J. J., Guth, D. A. & Hill, E. W. (1986). Sensitivity to perceptive structure while walking without vision. *Perception, 15*, 173-188.

Rose, S. (1995). The rise of neurogenetic determinism. *Nature, 373,* 180-182.

Rosenfield, I. (1992). *The strange, familiar, and forgotten: An anatomy of consciousness*. New York: Knopf.

Rubin, E. (1915/1958). Figure and Ground (trans.). In D. C. Beardslee & M. Wertheimer (Eds.), *Readings in perception*. Princeton: Van Nostrand.

Rumelhart, D. E. & McClelland J. L. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition, Volume 1: Foundations*. Cambridge, MA: MIT Press.

Rumelhart, D. E. & Norman, D. A. (1981). A comparison of models. In G. E. Hinton & J. A. Anderson (Eds.), *Parallel models of associative memory*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Rumelhart, D. E. & Zipser, D. (1985). Feature discovery by competitive learning. *Cognitive Science, 9*(1), 75-112.

Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986). Learning internal representations by error propagation. In Rumelhart & McClelland (1986).

Rumelhart, D. E., Lindsay, P. H. & Norman, D. A. (1972). A process model for long-term memory. In E. Tulving & W. Donaldson (Eds.), *Organization and memory*. New York: Academic Press.

Russell, S. J. (1991). *Do the right thing: Studies in limited rationality*. Cambridge, MA:

MIT Press.

Sandholm, T. W. & Crites, R. H. (1995). Multiagent reinforcement learning in the iterated prisoner's dilemma. University of Massachusetts at Amherst, Computer Science Department.

Saugstad, P. (1989). *Language: A theory of its structure and use*. Oslo: Solum Forlag.

Saussure, F. de (1916/1959). *Course in General Linguistics* (trans.). London: Peter Owen.

Savage-Rumbaugh, E. S. (1986). *Ape language: From conditioned response to symbol*. Oxford: Oxford University Press.

Schaeffer, B. & Wallace, R. (1969). Semantic similarity and the comparison of word meanings. *Journal of Experimental Psychology, 82*, 343-346.

Schank, R. C. & Abelson, R. P. (1977). *Scripts, plan, goals and understanding: An inquiry into human knowledge structures*. Hillsdale, NJ: Erlbaum.

Schank, R. C. (1972). Conceptual dependency: A theory of natural language understanding. *Cognitive Psychology, 3*(4), 552-631.

Schank, R. C. (1973). Identification of conceptualizations underlying natural language. In R. C. Schank & K. M. Colby (Eds.), *Computer models of thought and language*. San Francisco: Freeman.

Schank, R. C. (1982). *Dynamic memory*. New York: Cambridge University Press.

Scheff, T. J. (1990). *Microsociology: Discourse, emotion, and social structure*. Chicago: University of Chicago Press.

Scherer, K. R., Koivumaki, J. & Rosenthal, R. (1972). Minimal cues in the vocal communication of affect: Judging emotions from content-masked speech. *Journal of Psycholinguistic Research, 1*, 269-285.

Schyns, P. G. & Olivia, A. (1994). From blobs to boundary edges: Evidence for time and scale dependent scene recognition. *Psychological Science, 5*, 195-200.

Schyns, P. G., Goldstone, R. L., & Thibaut, J.-P. (in press). The development of features in object concepts. *Behavioral and Brain Sciences*.

Scribner, S. & Cole, M. (1981). *The psychology of literacy*. Cambridge, MA: Harvard

University Press.

Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge: Cambridge University Press.

Sedgewick, R. (1988). *Algorithms* (2nd ed.). Reading, MA: Addison-Wesley.

Selfridge, O. G. (1959). Pandemonium: A paradigm for learning. *National Physical Laboratory, Symposium No 10: Mechanisation of Thought Processes*, Her Majesty's Stationery Office, London, Vol. 1, pp. 513-531.

Selz, O. (1913). *Über die Gesetze des geordneten Denkverlaufs*. Stuttgart: Spemann.

Seyfarth, R. M., Cheney, D. L. & Marler, P. (1980). Vervet monkey alarm calls: Semantic communication in a freeranging primate. *Animal Behaviour, 28*, 1070-1094.

Shanon, B. (1993). *The representational and the presentational: An essay on cognition and the study of the mind.* London: Harvester.

Shapiro, S. C. (1971). A net structure for semantic information storage, deduction, and retrieval. In *IJCAI-71: Proceedings of the Second International Joint Conference on Artificial Intelligence*. San Mateo, CA: Morgan Kaufmann.

Slater, A. (1989). Visual memory and perception in early infancy. In A. Slater & G. Bremner (Eds.), *Infant development*. London: Lawrence Erlbaum.

Smith, B. C. (1995). *On the origin of objects.* Cambridge, MA: MIT Press.

Smith, B. C. (forthcoming). *The middle distance (Vols. 1-5).*

Smith, D. K. (1991). *Dynamic programming: A practical introduction.* New York: Ellis Horwood.

Smith, E. E., Schoben, E. J. & Rips, L. J. (1974). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review, 81*(3), 214-241.

Smith, L. B. & Kemler, D. G. (1978). Levels of experienced dimensionality in children and adults. *Cognitive Psychology, 10*, 502-532.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences, 11*, 1-74.

Sommerhoff, G. & MacDorman, K. F. (1994). An account of consciousness in physical and functional terms: A target for research in the neurosciences. *Integrative Physiological and Behavioral Science*, *29*(2), 151-181.

Sowa, J. F. (1992). Semantic networks. In S. C. Shapiro (Ed.), *Encyclopedia of artificial intelligence* (2nd ed.). New York: Wiley.

Sowa, J. F. (Ed.) (1991). *Principles of semantic networks: Explorations in the representation of knowledge*. San Mateo, CA: Morgan-Kaufmann.

Sproull, R. F. (1991). Refinements to nearest-neighbor searching in *k*-d trees. *Algorithmica*, *6*(4), 579-589.

Stern, D. G. (1991). Models of memory: Wittgenstein and cognitive science. *Philosophical Psychology*, *4*(2), 203-218.

Stern, D. N. (1977). *The first relationship: Infant and mother.* London: Fontana.

Stoddart, M. D. (1980). *The ecology of vertebrate olfaction.* London: Chapman & Hall.

Stoutland, F. (1988). On not being a behaviourist. In L. Hertzberg & J. Pietarinen (Eds.), *Perspectives on human conduct.* Leiden: E. J. Brill.

Struhsaker, T. T. (1967). Auditory communication among vervet monkeys (*Cercopithecus aethiops*). In S. A. Altmann (Ed.), *Social communication among primates*. Chicago: University of Chicago Press.

Stryker, M. P. (1992). Elements of visual perception. *Nature*, *360*(6402), 301.

Sutton, R. S. (1988). Learning to predict by the method of temporal differences. *Machine Learning*, *3*(1), 9-44.

Sutton, R. S. (1990a). Integrated architectures for learning, planning and reacting based on approximating dynamic programming. The Seventh International Conference on Machine Learning, 1990. San Mateo, CA: Morgan Kaufmann.

Sutton, R. S. (1990b). First results with Dyna: An integrated architecture for learning, planning and reacting. In W. T. Miller III, R. S. Sutton & P. J. Werbos (Eds.), *Neural networks for control.* Cambridge, MA: MIT Press.

Swaszek, P. F. (1985). *Quantization.* New York: van Nostrand Reinhold.

Tamminen, M. (1982). The extendible cell method for closest point problems. *BIT*, *22*, 27-41.

Taylor, J. R. (1989). *Linguistic categorization: Prototypes in linguistic theory*. Oxford: Clarendon Press.

Tinbergen, N. (1951). *The study of instinct*. Oxford: Clarendon Press.

Tinbergen, N. (1952). Derived activities: Their causation, biological significance, origin and emancipation during evolution. *Quarterly Review of Biology*, *27*, 1-32.

Tinbergen, N. (1953). *Social behaviour in animals: With special reference to vertebrates*. London: Methuen.

Tolman, E. C. (1932). Purposive Behavior in animals and men. New York: Century.

Touretzky, D. (1986). *The mathematics of inheritance systems*. San Mateo, CA: Morgan Kaufmann.

Trevarthen, C. (1977). Descriptive analyses of infant communicative behaviour. In H. R. Schaffer (Ed.), *Studies in mother-infant interaction*. London: Academic Press.

Trevarthen, C. (1979). Communication and co-operation in early infancy: A description of primary intersubjectivity. In M. Bullowa (Ed.), *Before Speech: The beginning of interpersonal communication*. Cambridge: Cambridge University Press.

Turing, A. (1950). Computing machinery and intelligence. *Mind*, *59*(236), 433-460.

Ullmann, S. (1951). *The principles of semantics*. Oxford: Blackwell.

Ullmann, S. (1962). *Semantics: An introduction to the science of meaning*. Oxford: Blackwell.

Üxküll, J. Baron von (1934). A stroll through the worlds of animals and men: A picture book of invisible worlds (trans.). *Semiotica*, *89*(4), 319-391.

Valiant, L. G. (1994). *Circuits of the mind*. Oxford: Oxford University Press.

van Essen, D. (1979). Hierarchical organization and functional streams in the visual cortex. *Annual Review of Neuroscience*, *2*, 227-263.

Vogel, T. (1991). *Learning in large state spaces with an application to bipedal robot walking* (doctoral dissertation). Technical Report 241. Computer Laboratory, Cambridge University.

Vygotsky, L. S. (1934/1986). *Thought and language* (trans.). Cambridge, MA: MIT Press.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes* (trans.). M. Cole, V. John-Steiner, S. Scribner & E. Souberman (Eds.). Cambridge, MA: Harvard University Press.

Waibel, A. (1988). *Prosody and speech recognition.* London: Pitman.

Warrington, E. (1975). The selective impairment of semantic memory. *Quarterly Journal of Experimental Psychology, 27,* 635-657.

Wason, P. C. (1981). Understanding the limits of formal thinking. In H. Parret and J. Bouveresse (Eds.), *Meaning and Understanding.* Berlin: Walther de Gruyter.

Watkins, C. J. C. H. (1989). *Learning from delayed rewards.* Unpublished doctoral dissertation. King's College, Cambridge University.

Watkins, C. J. C. H., & Dayan, P. (1992). Q-learning. *Machine Learning, 8*(3-4), 279-292.

Watt, R. (1987). Scanning from coarse to fine spatial scales in the human visual system after the onset of a stimulus. *Journal of Optical Society of America, A 4,* 2006-2021.

Weld, D. S. (1991). System dynamics and the qualification problem. In K. M. Ford & P. J. Hayes (Eds.), *Reasoning agents in a dynamic world: The frame problem.* Greenwich, CT: JAI Press.

Wertsch, J. V. (1985). *Vygotsky and the social formulation of mind.* Cambridge, MA: Harvard University Press.

Wertsch, J. V. (1991). *Voices of the mind.* London: Harvester.

Wierzbicka, A. (1990). 'Prototypes save': On the uses and abuses of the notion of 'prototype' in linguistics and related fields. In S. L. Tsohatzidis, *Meaning and prototypes: Studies in linguistic categorization.* London: Routledge.

Wisniewski, E. J. & Medin, D. L. (1994). On the interaction of theory and data in concept learning. *Cognitive Science, 18,* 221-281.

Witkin, A. (1986). Scale-space filtering. *Proceedings of the Ninth International Joint*

Conference on Artificial Intelligence, pp. 1019-1022. Los Altos, CA: Morgan Kauffman.

Wittgenstein, L. (1958). *Philosophical investigations* (2nd ed.). Oxford: Blackwell.

Wittgenstein, L. (1969). *On certainty*. Oxford: Blackwell.

Woods, W. A. (1975). What's in a link: Foundations for semantic networks. In D. G. Bobrow & A. Collins (Eds.), *Representation and understanding: Studies in cognitive science*. New York: Academic Press.

Zajonc, R. B. (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist, 35*, 151-175.