ELSEVIER

Artificial Intelligence 72 (1995) 1–52

Artificial Intelligence

Computational research on interaction and agency

Philip E. Agre *

Department of Communication, University of California, San Diego, La Jolla, CA 92093-0503, USA

Abstract

Recent research in artificial intelligence has developed computational theories of agents' involvements in their environments. Although inspired by a great diversity of formalisms and architectures, these research projects are unified by a common concern: using principled characterizations of agents' interactions with their environments to guide analysis of living agents and design of artificial ones. This article offers a conceptual framework for such theories, surveys several other fields of research that hold the potential for dialogue with these new computational projects, and summarizes the principal contributions of the articles in this special double volume. It also briefly describes a case study in these ideas—a computer program called Toast that acts as a short-order breakfast cook. Because its designers have discovered useful structures in the world it inhabits, Toast can employ an extremely simple mechanism to decide what to do next.

1. Introduction

The papers in this special double volume illustrate an emerging way of doing research in artificial intelligence, which might be stated compactly as follows:

> Using principled characterizations of interactions between agents and their environments to guide explanation and design.

The purpose of this introduction is to explain this emerging style of research and to explore its relationship to other research in AI and elsewhere.

Let us begin with a familiar example. Consider a device (a "controller") that must direct the operations of an oil refinery. So far as control theory is concerned, an oil refinery is an enormous machine (the "plant") with a number of "control variables" that can be adjusted from the outside (the settings of various valves and burners) and a number of "output variables" whose values at any given moment can be determined from

* E-mail: pagre@ucsd.edu. Telephone: (619) 534-6328. Fax: (619) 534-7315.

SSDI 0004-3702(94)00054-9

the outside (the readings on various sensors and gauges). The task of the "controller", let us say, is to stabilize some of the output variables around certain values while maintaining other variables within certain fixed ranges. In concrete terms, the controller must adjust the valves and burners to sustain a fixed flow of oil without the plant blowing up.

Given a proposed design for this controller, how do we know whether it will work? It is impossible to answer this question simply by analyzing the controller itself. Nor, obviously, does it suffice to analyze the plant in isolation. Instead, it is crucial to analyze how the controller will interact with the plant. Given any particular set of initial values, and supposing for simplicity that the interaction is not stochastic, the combined system of plant plus controller will follow a determinate trajectory. The designer's goal is to ensure that the entire family of these interaction trajectories has certain properties. One way to characterize this family of trajectories is in terms of a differential equation that relates changes in the control variables to the current values (and perhaps the ongoing rates of change, or past values, or both) of the output variables.

The controller in this example might be regarded as an agent interacting with its environment, namely the plant, and differential equations provide one way of characterizing such interactions. Control theory, of course, provides only one way of thinking about interactions. It is tied to a particular model of interaction (through output and control variables), its historical development has been profoundly influenced by the need for safety and conservatism in relatively well-behaved systems, and it is thoroughly mathematical. A principled characterization of interaction, though, need not have any of these qualities to provide a useful guide to the design of artificial agents and the explanation of natural ones. Indeed, we have deliberately chosen the vague word "principled" (as opposed to, say, "formal") in order to include an unforeseeable range of possible types of theories of interaction. The important thing is that our characterization of interaction should allow us to address questions like these:

- What will our agent do in a given environment?
- Under what conditions will it achieve its goals or maintain desired relationships with other things?
- In what kinds of environments will it work?
- How do particular aspects of an environment, such as topography or mutability or the workings of artifacts, affect particular types of agents' abilities to engage in interactions that have particular properties?
- What forms of interaction require an agent to employ particular elements of internal architecture, such as memory?
- What forms of interaction permit an agent to learn particular knowledge or skills?

To ask these questions, we do not need to make any *a priori* assumptions about the architecture of our agents. To the contrary, the point is to understand, in as general a way as possible, the relationships among the properties of agents, environments, and forms of interaction between them. Of course, it is doubtful that any single theory can give a complete account of this vast topic. The papers in this special double volume, though, each provide detailed examples of the analysis of interactions within some particular domain of architectures and environments. This special double volume is thus explicitly ecumenical in approach, advocating no single architecture and no single formalism. Through the shared themes that arise within the principled characterization of interactions, we hope that each project can benefit from the others, and that readers can benefit from the three-dimensional picture of research in this area that this approach can offer.

This introduction cannot attempt a complete synthesis of research in this area, nor is it a manifesto representing a definite group or movement. Instead, it offers one perspective on how the research reported in this special double volume is situated in the larger intellectual and technical world. It is organized as follows. Section 2 outlines a series of themes that arise when doing computational research on interaction between agents and their environments, together with examples and conceptual discussion. In so doing, it also specifies more precisely the territory of research covered by this special double volume. Section 3 describes the conceptual connections between the research reported here and research in other fields. These connections may provide inspiration for further computational research on interaction. Section 4 summarizes the individual papers in this special double volume, offering comments on their distinctive contributions and their relationships to one another. Section 5 presents a case study in the ideas of the special double volume. Section 6 concludes with a prophesy and plea for interdisciplinary research.

2. Studying interaction

It is far too early to assemble a rulebook for research into computational theories of interaction and agency. It is possible, though, to convey some of the intuitions that have been developing through the progress of research in this area, both through the computational research reported here and through the existing traditions of research upon which it builds. Putting words to these intuitions is a hazardous matter, and the words offered here should be understood as heuristic devices, as first passes, and as invitations to formulate things in different ways, through different metaphors.

2.1. Mapping the territory

First it is necessary to define carefully the scope of research reported here. Let us imagine research on agents interacting with the world to be arrayed in a two-dimensional field, with one axis corresponding to the number of agents involved in the interaction and the other axis corresponding to the degree of realism with which the world is modeled. (See Fig. 1.) A single agent interacting with a very simple world would lie toward the origin of this diagram. Any project to model human life, or the lives of most animals, in a realistic way would lie in the upper-right corner of the diagram, and that is surely the future ideal of much of the field. As it is, most current research clusters in three areas:

(1) Research that explores single agents interacting with relatively simple environments, where particular aspects of the environment are analyzed in enough detail to bring out larger points.

(2) Research on relatively complex forms of interaction among several agents in extremely simple environments, where the interaction is largely symbolic and

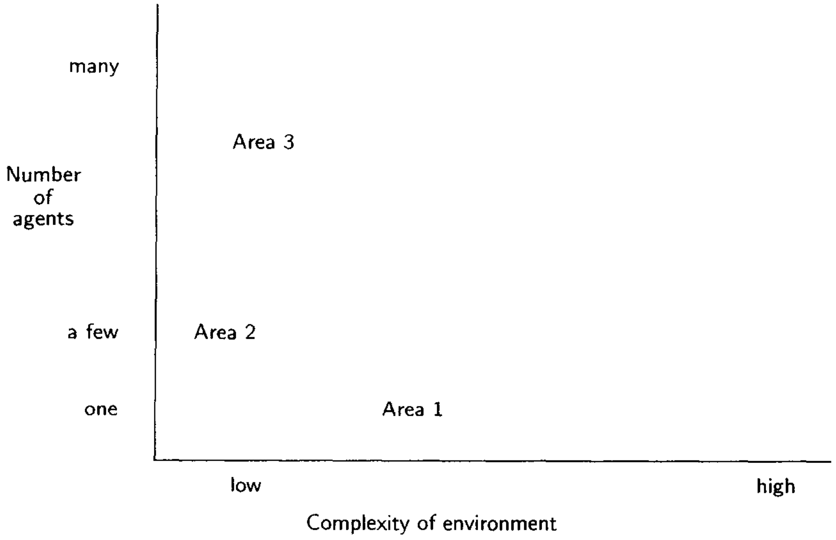

Fig. 1. Current AI research falls mostly into three clusters, which can be contrasted according to the degree of complexity of the environments they deal with and the number of interacting agents they employ.

depends little on the agents' bodies. The emphasis is on the logical structure of the interaction.

(3) Research on relatively simple interaction among numerous agents in slightly more complex environments, where the interaction does depend in some way on the agents being embodied. The agents may be physical robots or simulations, and the emphasis is on the emergence of order from simple forms of interaction.

The papers in this special double volume lie exclusively in the first of these three clusters, with the exception of the paper by Shoham and Tennenholtz, which lies in the third. As a result, numerous important issues go unaddressed here, including symbolic forms of interaction among agents. Integration of the three approaches, leading to an exploration of the middle regions of the diagram, is obviously an important goal for the future, and we hope that the analyses developed here will contribute their part to that project.

2.2. *Planning and reaction*

Computational research on interaction between agents and their environments has historically been structured by two sets of ideas: a dominant tradition focused on "planning" and a subordinate tradition focused on "reaction". Careful consideration of these ideas will make the distinctive position of the research reported here much clearer.

Although the term "planning" is not always used with great precision, let it refer here to the notion of organizing action through the construction and execution of computer-program-like symbolic structures called plans (cf. [7]). This idea can be traced backwards through the history of AI to a number of sources. Perhaps the most important of these is Karl Lashley's 1951 lecture "The problem of serial order in behavior" [49]. As a neurophysiologist vitally concerned with the workings of human brains, Lashley urged an understanding of cognitive processes whose prototype was the phonetic structure of language. Utterances of language have a formal structure of great intricacy whose basic elements, the phonemes, follow upon one another so rapidly that the structure simply could not emerge through the chaining together of behavioristic stimuli and responses. It follows, Lashley argued, that the brain must be capable of generating these structures on its own internal resources.

Moreover, Lashley proposed understanding all human action on the model of language. The job of the brain was to string together the "expressive elements" (by analogy to words or lexical units) by means of "the syntax of the act" (the grammar of action) in accord with the "determining tendency" that the action is intended to express. Although the theoretical vocabulary has changed, the general shape of this proposal was enormously influential [3]. In a more familiar AI vocabulary, Lashley is suggesting that the brain generates sequences of primitive actions by applying stored habitual schemata. The vague idea of "determining tendency", which Lashley abstracted by analogy to the semantic content being expressed by a linguistic utterance, has been replaced by the simpler notion of the goal to be achieved at the end of an action sequence.

Another influential early proposal, better known in the AI world because it was accompanied by computer models, was Allen Newell and Herbert Simon's computational model of problem solving based on search [59]. Although linguistic metaphors were not central to their exposition, their proposal was similar to Lashley's. Thought was held to consist in a process of search through a space of possible sequences of "operators", some of which correspond to desirable situations which might be understood as problem solutions or goals. The term "planning" entered the AI lexicon as one of the heuristic devices that could abbreviate these searches. According to Newell and Simon's conception, planning takes place when a coarser search space is used to guide the exploration of a finer (and thus combinatorially much larger) search space. This notion of nested search spaces aligned neatly with the formal concept of the hierarchical decomposition of action that was already found in research on linguistics. Each utterance has a grammatical structure that can be drawn as a hierarchical parse tree, with each lexical item itself having a hierarchical structure of syllables and phonemes. To researchers such as Newell and Simon, hierarchical decomposition held the promise of a universal structuring principle for human cognition.

The ideas proposed by Lashley and by Newell and Simon were combined in the first synthesis of the computational theory of planning, *Plans and the Structure of Behavior* by George Miller, Eugene Galanter, and Karl Pribram [57]. There one encounters the first recognizable definition of "Plans":

> A Plan is any hierarchical process in the organism that can control the order in which a sequence of operations is to be performed. [57, p. 16]

Note that a Plan here is not necessarily a symbolic mental structure. It is less specific than that: a "hierarchical process" specified in terms of its ability to structure (in Lashley's terms) the serial order of the organism's behavior. Miller, Galanter, and Pribram's conception of a Plan shaped later AI research in numerous ways. But the most important of these for present purposes is a persistent ambiguity throughout the whole of their book between two conceptions of Plans and their use:

(1) A notion of "Plans", a relatively fixed repertoire of commonly employed structures of action. In more recent AI work, this would be called a "plan library". Miller, Galanter, and Pribram give no account, however, of where these Plans come from. The Plans are hierarchical in their structure, and they can be assembled into larger structures by treating them as elements in a larger hierarchy.

(2) A notion of "the Plan", a hierarchical structure or process which provides a sort of running transcription (in linguistic terms, a parse tree) of the organism's behavior. No commitment is made here to the mechanisms by which this Plan arises, and it could perfectly well be improvised from moment to moment, with the sole constraint that it be possible to the process in retrospect as having been hierarchical in nature.

These two concepts correspond conceptually to two strands of research in AI, which are commonly known as "planning" and "reaction". In Miller, Galanter, and Pribram's book, though, they are conflated in a wide variety of ways. Brief reflection on them makes it clear why. A computational theory of action has at least two central goals:

- to explain how action has the structure it does, and
- to explain how actions are chosen that are appropriate to the circumstances in which they are taken.

The notion of "Plans" addresses the question of structure: action has the structure it does, says this theory, because it arises through the execution of things called Plans which have that same structure. Yet this theory does not provide a convincing account of how these actions are adapted to their circumstances. Of course, if the organism is wholly in control of the circumstances then rational decisions about action can be made *a priori*, before the execution of one of these Plans. And this may indeed be true for short stretches, as when uttering a single word or phrase. But Miller, Galanter, and Pribram wished to explain the whole structure of everyday life, in which a wide variety of contingencies arise.

The notion of "the Plan" addresses this need. It allows for a greater degree of improvisation, since elements can be added to the hierarchical structure of the Plan at any time, including at the very moment when those Plan elements are about to be executed. But it offers no account of the reason why action has the structure it does. Action is still hierarchical in nature, but the particular shape of the hierarchy is wholly unspecified. Miller, Galanter, and Pribram do not seem aware of the problem, most likely because they do not clearly distinguish between their two proposals, shifting frequently back and forth between them as the details of their argument demand.

This ambiguity in Miller, Galanter, and Pribram's book foreshadowed the outlines of three subsequent decades of research. Starting with Fikes and Nilsson's STRIPS program [22], a long tradition of research focused its attention on the first of Miller, Galanter, and Pribram's concepts, that of Plans which are constructed and executed as packages, and which might be stored in Plan libraries to provide an organism or robot with a repertoire of habitual patterns of action for future occasions. STRIPS did address the question of improvisation in a simple way through certain flexibilities in the execution process [23,24]. But for the next decade or so, research generally focused upon the plan-construction process, assuming plan execution to be a relatively simple matter. This line of research into plan-construction shifted to a new phase in the late 1980s as researchers began to cast the classical problems of plan-construction in much more formal terms, and to explore the mathematical questions to which these formalized problems gave rise [14,31,42,55]. Yet all along, half of Miller, Galanter, and Pribram's original story was missing.

This situation was remedied in the mid 1980s with the rise of what has come to be called "situated action" or (somewhat unfortunately, in my own view) "reactive planning" [4,26,28,68,70]. These "reactive" systems should be understood not as a radical departure, but as filling in a hole in the existing system of ideas around planning—as reinventing the other half of Miller, Galanter, and Pribram's theory. Here the emphasis was on interaction with the environment and on the role of tightly coupled perception-action loops in organizing activity. Just as planning offered no robust account of moment-to-moment interaction with the world, reaction offered no robust account of how the organism or robot's actions could be guaranteed to "work", understood as rational, and so on. Conflict between the two schools of research has often been heated, as each school has been able to point at substantial weaknesses of the other's mechanisms without always possessing the necessary concepts to appreciate the weaknesses of its own.

Observing this impasse, a substantial literature immediately grew up attempting to synthesize the planning and reaction theories through "hybrid architectures" (e.g., [25,62]. Just as Miller, Galanter, and Pribram attempted (probably without realizing it) to reconcile planning and reaction through rhetorical ambiguity and logical improvisation, the designers of hybrid architectures recapitulate in computational terms (again, probably without realizing it) this same attempt at reconciliation. In each case, the implicit project is to fashion a whole theory out of two half-theories that presuppose incompatible views of action. Although it is conceivable that the resulting theory might work out, and probable that the resulting architectures may have some practical applications, such research will most likely be frustrated in its forward progress by its lack of a consistent conceptual framework. These things are easy to see with the benefit of hindsight, of course, but it is important to recognize them nonetheless because of their substantial implications for computational theories of action.

It is the central purpose of this special double volume to overcome the conceptual impasse between planning and reaction. The point is not necessarily to offer a better and newer architecture, though research informed by new ideas will presumably lead in that direction, but to identify some concepts and methods of research that will reconcile the unhappy split between planning and reaction by providing interesting accounts of both the structure of behavior and the dynamics of an agent's interactions with its environment. Symbolic plans might play a role in this story or they might not, but they are not ruled out *a priori*.

To reconcile planning and reaction, the important thing is to focus upon the structures of interaction between agents and their environments. Every agent that undertakes actions in some world has a structure of interaction with its environment, whether it is symbolic or connectionist, whether it has internal state or not, and so forth. To focus on interactions is not to legislate these things ahead of time. Nonetheless, a focus on interactions does impose a stiff constraint on the research process. Given an agent interacting with an environment, one must ask this question: "why do we think it should work?". Of course, the notion of "working" has no single definition, and different research programs can pursue a wide range of notions of "working" with equal legitimacy. The proposal of this special double volume is that approaches to this question will require researchers to formulate principled characterizations of the agent's interactions with its environment. As noted at the outset, the phrase "principled characterizations" is designed to cast a wide net, including both formal and informal theories, symbolic and quantitative theories, explanatory and prescriptive theories, biological and social theories, and so forth.

Whatever its faults, research on planning does at least offer a clear account of why the agent's actions ought to work. The environment is usually assumed to be basically stable, in the sense that the agent is the only significant source of disruption within it, and interaction proceeds in ways that can be anticipated in advance through some kind of search process. The agent itself formulates all of the characterizations of its interaction with its environment that it needs, and if its search for an adequate plan halts then the designer can be assured that that plan will actually work—the search process is effectively proving the theorem that some such plan will work. (In the case of probabilistic planning [45], the proposition being "proved" will be probabilistic in nature.) The problem, of course, is that the design of such provably correct plan-construction systems requires that highly restrictive conditions be imposed upon the world—roughly, that the world be representable using a formalism within which a proof of correctness can be performed as a practical matter. Of course, historically most such "proofs" have been informal. The point, though, has been to construct systems that produce correct plans if they halt with any plan at all.

The accomplishments of this research should certainly not be underestimated. It is not a simple matter to obtain any kind of correctness proof in domains as formally complex as those that AI research has investigated. Research on principled characterization of agent-environment interaction will surely build upon this existing work in a wide variety of ways. At the same time, it will also incorporate a wide variety of other influences. The remainder of this section sketches the outlines of the approach to AI research that results from this still emerging synthesis.

2.3. Correspondence and convergence

How does one argue that a particular agent-environment interaction will work? By far the most common approach in AI research has been to formulate arguments in terms of correspondence between internal representations and the outside world. In a simple form, such arguments work by induction: if it is assumed that the agent has correct knowledge of the world at some initial time, and if we can demonstrate that correctness of knowledge is preserved from one unitary action to the next, then it follows that the agent's knowledge of the world will remain correct for as long as it takes actions. If the correctness of an agent's actions is guaranteed by the correctness of its reasoning, the agent can then be shown to "work". This section contrasts this *correspondence method* of argument with a broader *convergence method* that is employed by many of the papers in this special double volume.

The correspondence method may seem unfamiliar when stated in the abstract form provided in the previous paragraph. Nonetheless, it is precisely what is at stake in attempts to solve the frame problem. Understood in its broadest terms, the frame problem is a lemma that must be proven in the midst of any attempt to design a plan-construction program. It asks, given that the agent correctly anticipates what the world will be like up to a certain point, how can it infer what the world will be like after a particular action is taken? In particular, which of the agent's beliefs can be assumed to stay in correspondence with the world after the action is taken? Answering these questions is a difficult matter, since it can take real work to infer all of the consequences of a given action. These consequences might be hard to catalog, yet we would not wish an agent to become disabled worrying that opening a door might have consequences far beyond its reasonable surmises, for example causing the sun to fall from the sky. Technical and philosophical research into the frame problem has determined that it can be usefully decomposed into a variety of separate problems [65], but this decomposition does not matter for the purposes of the present argument. The point is simply that the frame problem arises as part of any attempt to argue for an agent's correctness in a given environment by means of the correspondence method.

Some authors have made strong claims about the theoretical implications of the frame problem. Toth [73], for example, argues that the frame problem is fatal for a certain conception of AI research, whose unit of analysis is the individual's cognitive process. Likewise, the difficulty of bounding the necessary inferences from a given action is reminiscent of Dreyfus' [20] argument that real-life reasoning takes place against a large enough unarticulated background that attempts at logical formalization necessarily encounter an infinite regress of rules-about-how-to-apply-rules. These arguments should not be interpreted as grounds for the categorical rejection of a tradition of research, but rather as roughly indicating the contours of a complex phenomenon. Indeed, the paradoxes of the frame problem may simply be, at least in part, an inherent condition of life that is "solved" piecemeal by real agents through learning in particular cases. Although it is not possible to resolve the question here, we can explore how the question arises through the concepts that have historically guided AI research.

The correspondence method makes fairly specific assumptions about the process through which agents choose actions. These assumptions are not necessarily architectural in nature: the reasoning processes that encounter a version of the frame problem might operate on symbolic structures in an agent's memory, but they might also potentially be encoded in hardwired circuitry, simulated through neural networks, or subserved implicitly in the operation of other types of machinery. The point, though, is that the designer is approaching the design process in a certain way, maintaining a sense of the representational content of various machine states and making sure at all points that the correspondence method of argument constrains the design process.

This procedure might be contrasted with the *convergence method*. Here the design process is also constrained by an argument about correctness, whether formal or informal. The difference is that the method of argument focuses upon the agent's behavior and not on its internal states. Put another way, the method focuses upon particular relationships between the agent and its environment, characterizing these relationships in principled ways and making arguments about their invariants and their evolution. This approach is not wholly distinct from the correspondence method; it is a larger category that includes the correspondence method as a particular case—the relationship in that case being one of semantic correspondence. The term "convergence" is a little misleading in suggesting that the agent is necessarily evaluated by its eventual arrival at some kind of goal, but many other kinds of evaluation are possible as well.

The point is that principled arguments about correctness can be formulated in a variety of ways other than in terms of correspondence. Some of the most important invariants identified by these arguments might be located in the physical world, without any regard for the agent's internal states. To take a trivial example, an agent that performs an exhaustive search of a finite physical territory, placing breadcrumbs on each spot already searched and continually homing in on spots without breadcrumbs, can be easily demonstrated to find what it is looking for within a certain amount of time. Early AI theorists referred to this sort of thing as "external memory" and did not regard it as significantly different in its implications for cognitive architecture than internal memory [1,2]. Although this view is surely too simple once we take account of the geography of the physical world and the capacities of our organisms' physical bodies, note that the proof of correctness for this simple agent is entirely familiar from proofs of program correctness in computer science. The proof involves an invariant (the total of spots searched and unsearched), a progress function (the number of spots still unsearched), and a convergence condition (no more spots unsearched). Other arguments for correctness might employ considerably different methods. In each case, though, the argument will depend on some kind of principled characterization of the interaction—not necessarily of every detail of the interaction, just enough of its properties to allow an adequate argument to be formulated.

Another example may help to illustrate one of the correspondence method's inherent limitations for research on agents in environments with much qualitative structure. Consider an agent with a traditional set of symbolic "beliefs" that employs logical reasoning, based on these beliefs, to decide what actions seem indicated in particular situations. Since the agent is a finite being in a complicated world, it will probably have mistaken beliefs occasionally. Yet it may still be possible to demonstrate that the agent will necessarily achieve its goals anyway. Such a demonstration might proceed in several different ways, but in each case it will take into account specific properties of the relationship between the agent and its environment. For example, different paths may be easily distinguishable so long as the agent is registering certain properties of its environment, thus guarding against ending up on unintended routes. Likewise, the environment may be provided with signs that disambiguate all of the ambiguous situations that the agent might encounter. As with any other theory of erroneous beliefs (for example in computer perception research), demonstrating such things in a principled way requires an error model—a theory of the circumstances through which errors in belief might take place (cf. [77]). One might be able to demonstrate that mistaken beliefs will necessarily get corrected, that mistaken actions will necessarily provoke safe indications of the difficulty, or that uncorrected mistaken beliefs will lead at worst to alternative solution paths that are quantifiably less desirable than the optimum. In each case, the argument proceeds on the convergence model, even though symbolic beliefs are present and other, interrelated arguments might rest on the correspondence model.

A simple example might be provided by the control-theoretic notion of robust control: uncertainty about the plant is characterized by assuming that plant parameters are within certain known bounds. A controller that is robust under these conditions produces the desired behavior for all the possible plants consistent with these constraints.

These examples are obviously simple and abstract, meant for illustration. The rest of this introductory article will explain some of the concepts that have led the authors in this special double volume to develop more sophisticated forms of argument about interactions between agents and their environments.

2.4. Aerial and ground views

Another helpful distinction in research on interaction and agency is that between aerial and ground views of an agent's activities. When designing agents that operate in abstract territories such as search spaces, and that do not have bodies (simulated or not) in any real sense, it can be easy to lose the distinction between what the agent knows about a situation and what the designer knows about that situation. This distinction is not crucial when the design process is being constrained by the correspondence method, since in that case it is important for the agent to maintain enough knowledge about its environment to permit a proof to be constructed that the agent will do the things it is supposed to do. The agent need not be capable of actually performing that proof, since it suffices for the designer to have conducted the proof in a generalized way off-line, but the whole point of the correspondence method is that the agent knows those facts which permit it to get along successfully.

This is not true with the convergence method. An agent designed using the convergence method might be spoken of as having knowledge (or it might not), but this knowledge need not necessarily support a proof of convergence. The designer might be able to demonstrate that a large number of conditions about the agent's environment, together with the agent's internal states, afford a proof that the agent will be able to achieve its goals. For example, an agent that relies upon posted signs to find its way around will have great trouble in a world where such signs are sparse, but if the designer knows one particular world to have been adequately posted then it will be possible to prove that the agent will get where it is going, regardless of whether the agent itself can be sure of this.

This too is obviously a trivial example, and it is also a conceptually straightforward example in the sense that the agent is spoken of as having knowledge—the only question concerns the relationship between the agent's knowledge-set and the designer's. Things become more interesting when the agent is understood to have different kinds of knowledge from the designer—for example, indexical knowledge of its relationship to its surroundings—or when the agent is spoken of in wholly different terms, without reference to notions of knowledge. In such cases, it becomes particularly important to maintain a rigorous conceptual separation between the agent and the designer, so that the particularities of the agent's relationship to its environment can come out in full relief—and so that new theories about knowledge can arise that are rooted in the agent's having a body, being located in a physical environment, interacting with artifacts, and so forth.

The general point is that agents who are interacting with physical worlds have bodies, and embodiment has pervasive consequences for computational theories of action, knowledge, perception, and learning. In large part this is due to locality: agents with bodies, embedded in physical environments, only have direct access to limited regions of the world. These limited regions are not accidental or arbitrary in their shape, but have a structure that is made from the geometry of the space, the shapes of physical objects like hills and roads and walls and tools, and the causal interconnection of things. The simple partialness of the agent's access to these things already has significant consequences for theories of action that require agents to have substantially complete world models.

But more subtly, the local structure of an agent's involvements with the world brings to those involvements a pervasive indexicality: the agent is involved with *this* place, faces in *this* direction, interacts *now* with *these* artifacts. The agent does not necessarily know where it is, nor what time it is, nor what its heading is, nor which particular stone or can-opener or McDonald's it might be dealing with at any given time. Given this fact, the correspondence method, at least in its traditional forms, would suggest ensuring that the agent always know the answers to all these questions. The agent might possess a compass and a clock and a map, objects might be labeled with their identities, and so on. Another approach, compatible with the looser demands of the convergence method, is to explore the relationships between indexical knowledge ("this bike here now") and the more objective kinds of knowledge ("Karen's bike in Miami on Christmas"). Perception and action, after all, are inherently indexical in character: your retinas do not register "red" at a specific latitude and longitude, but rather "red here". Likewise, your hand does not close the door to room 317, but rather "this door". To interact with the world is to do things with your body, and your body is a physical thing that participates in the same locality and the same concrete particularity as any other.

It is this materiality of embodied action that makes the distinction between the aerial view and the ground view so compelling. The designer, taking a metaphorical position above the territory, can know a wide variety of things that an agent resting in a particular spot on the ground, with a particular heading, may not know. The agent's knowledge and ignorance are structured phenomena, and it is the designer's job to understand those structures. An agent might be going around in circles without ever knowing it, but the designer might be in a position to characterize the conditions under which the agent can and cannot avoid such a fate. An agent might be at risk of running out of stove burners without realizing it until it is too late, but the designer might be able to demonstrate that stove burners will necessarily be plentiful unless certain supplies run low. An agent may continually lose track of its tools, but the designer might be able to demonstrate that the tools will remain accessible so long as the agent makes a habit of putting them back where they belong.

2.5. *Structure in the agent and the world*

Having established the outlines of this emerging style of research, what kinds of things can be learned from it? Perhaps the most important lessons concern the ways in which agents are adapted to their environments. Although notions of adaptation are perhaps most familiar from biology, the most important ideas about adaptation in the history of AI are actually sociological. In his pre-AI book *Administrative Behavior* [71], Simon outlined many ways in which social organizations compensate for the "limited rationality" of their members. The orchestration of numerous workers within a larger organization, Simon argued, compensates for the individual's limited capacity for work. Likewise, the division of labor and the assignment of specialized tasks to individuals compensates for their limited abilities to learn new tasks. The flow of structured information through the organization compensates for their limited knowledge, and the precise formats of that organization, together with the precise definition of individual tasks, compensate for individuals' limited abilities to absorb information and apply it usefully in making decisions. Finally, Simon believed that the hierarchical structure of bureaucracies compensates for individuals' limited abilities to adopt their own values and goals.

Whatever the value system implicit in this analysis, the general form of argument has had an important influence on AI. This influence, though, has been indirect. When Simon moved from studies of organizations to studies of individual cognition in the 1950's, most of these ideas about the individual's cognitive environment did not survive the transition. The individual imagined in AI research has generally been isolated and self-reliant, except in the matter of goals, which in practice have almost invariably been assigned from the outside by the system's designer. Yet almost through the inherent logic of the enterprise, AI researchers have rediscovered the general form of argument outlined in *Administrative Behavior*: structure in the world compensates for the weaknesses of cognitive architectures. Some of these weaknesses might be imposed by the designers, for example when the goal is to explain human processing limitations, or they might derive from the weaknesses of all known architectures, or they might be inherent computational limitations deriving from undecidable problems and the like.

Despite this insight, perhaps the most significant shortcoming of research on "reactive" systems and "hybrid" planning-reaction architectures is a relative lack of concepts for discussing the useful structures of the world. When breaking free from the safe and constrained microworlds of classical planning research, these new schools emphasized environments which are "uncertain", "unpredictable", "complex", "changing", and the like (e.g., [33,40]). Unfortunately, it is next to impossible to say anything very general about environments that are characterized in these negative terms, as not-this and not-that. It is quite plausible that, by some measure, most environments are in fact wholly untenable, in the sense that they are so uncertain, unpredictable, and so forth that no organism above a certain primitive level could possibly survive in them. (Of course, these negative characteristics may be viewed as constitutive of warfare and other profoundly adversarial activities, and they are certainly part of American military discourse.) But given that the classical problems of plan-construction rapidly become intractable or undecidable as the qualitative complexity of the agent's environment increases (or, more precisely, as the planning formalism registers more and more of that complexity), it becomes imperative to discover the features of a given world—and of the agent's interaction with the world—that make life in that world tenable.

To be sure, AI research has employed a variety of concepts of "structure". Marr [52], for example, outlined a method of vision research that posits a set of modules, each

[21] T. Hogg and B.A. Huberman, Artificial intelligence and large scale computation: A physics perspective, *Phys. Reports* **156** (1987) 227-310.

[22] T. Hogg and J.O. Kephart, Phase transitions in high-dimensional pattern classification, *Comput. Syst. Sci. Eng.* **5** (4) (1990) 223-232.

[23] T. Hogg and C.P. Williams, Solving the really hard problems with cooperative search, in: *Proceedings AAAI-93*, Washington, DC (1993) 231-236.

[24] T. Hogg and C.P. Williams, Expected gains from parallelizing constraint solving for hard problems, in: *Proceedings AAAI-94*, Seattle, WA (1994) 331-336.

[25] T. Hogg and C.P. Williams, The hardest constraint problems: a double phase transition, *Artif. Intell.* **69** (1994) 359-377.

[26] J.H. Holland, *Adaptation in Natural and Artificial Systems* (University of Michigan Press, Ann Arbor, MI, 1975).

[27] B.A. Huberman and T. Hogg, Phase transitions in artificial intelligence systems, *Artif. Intell.* **33** (1987) 155-171.

[28] M.T. Jones and P.E. Plassmann, A parallel graph coloring heuristic, *SIAM J. Sci. Comput.* **14** (3) (1993) 654-669.

[29] A. Kamath, R. Motwani, K. Palem and P. Spirakis, Tail bounds for occupancy and the satisfiability threshold conjecture, in: S. Goldwasser, ed., *Proceedings 35th Symposium on Foundations of Computer Science* (1994) 592-603.

[30] P. Kanerva, Self-propagating search: A unified theory of memory, Technical Report CSLI-84-7, Stanford University, CA (1984).

[31] R.M. Karp and J. Pearl, Searching for an optimal path in a tree with random costs, *Artif. Intell.* **21** (1983) 99-116.

[32] S. Kirkpatrick, C.D. Gelatt and M.P. Vecchi, Optimization by simulated annealing, *Science* **220** (1983) 671-680.

[33] S. Kirkpatrick and B. Selman, Critical behavior in the satisfiability of random boolean expressions, *Science* **264** (1994) 1297-1301.

[34] J.R. Koza, *Genetic Programming: On Programming Computers by Means of Natural Selection and Genetics* (MIT Press, Cambridge, MA, 1992).

[35] T. Larrabee and Y. Tsuji, Evidence for a satisfiability threshold for random 3CNF formulas, in: H. Hirsh et al., eds., *AAAI Spring Symposium on AI and NP-Hard Problems* (AAAI, Menlo Park, CA, 1993).

[36] M. Lau and T. Okagaki, Applications of the phase transition theory in visual recognition and classifications, *J. Visual Commun. Image Representation* **5** (1) (1994) 88-94.

[37] G. Lewandowski and A. Condon, Experiments with parallel graph coloring heuristics, in: *Proceedings 2nd DIMACS Challenge* (1993).

[38] A.K. Mackworth, Constraint satisfaction, in: S. Shapiro and D. Eckroth, eds., *Encyclopedia of AI* (Wiley, New York, 1987) 205-211.

[39] C.J.H. McDiarmid and G.M.A. Provan, An expected-cost analysis of backtracking and non-backtracking algorithms, in: J. Mylopoulos and R. Reiter, eds., *Proceedings IJCAI-91*, Sydney, Australia (1991) 172-177.

[40] S. Minton, M.D. Johnston, A.B. Philips and P. Laird, Minimizing conflicts: a heuristic repair method for constraint satisfaction and scheduling problems, *Artif. Intell.* **58** (1992) 161-205.

[41] S. Minton and I. Underwood, Small is beautiful: a brute-force approach to learning first-order formulas, in: *Proceedings AAAI-94*, Seattle, WA (1994) 168-174.

[42] D. Mitchell, B. Selman and H. Levesque, Hard and easy distributions of SAT problems, in: *Proceedings AAAI-92*, San Jose, CA (1992) 459-465.

[43] A. Nijenhuis and H.S. Wilf, *Combinatorial Algorithms for Computers and Calculators* (Academic Press, New York, 2nd ed., 1978).

[44] J. Pearl, *Heuristics: Intelligent Search Strategies for Computer Problem Solving* (Addison-Wesley, Reading, MA, 1984).

[45] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (Morgan Kaufmann, San Mateo, CA, 1988).

[46] P. Prosser, An empirical study of phase transitions in binary constraint satisfaction problems, Technical Report AISL-49-93, Department of Computer Science, University of Strathclyde, Glasgow, Scotland (1993); also: *Artif. Intell.* **81** (1996) 81-109 (this volume).

include what AI has historically called "the domain"? The various papers in this special double volume, which are summarized in a later section of this introduction, make numerous suggestions. By way of general orientation, though, here are some general categories that might aid the search for structure:

- *Artifacts.* How do the properties of tools simplify the reasoning that decisions that agents must undertake in choosing actions? How about buildings and streets? How about clothing and furniture?
- *Signs.* Where are signs placed in the particular world being studied? What do they say, and what assumptions about knowledge do they make? What other sorts of symbolic labels are placed on things? What kinds of instructions are provided? Does the language in those instructions have any reliable properties?
- *Physical dynamics.* What rhythms are established in a particular category of physical interactions with an environment? What properties of those interactions are conserved or remain invariant? Under what conditions do they converge to attractors or remain bounded by certain envelopes? Why?
- *Customs.* What conventions do the agents in this world maintain? If the agents can rely upon one another to maintain these customs, how can this simplify their reasoning? What invariants do these customs maintain in the physical world?
- *Practical constraints.* What orderings upon actions are dictated by plain physical practicality? A cupboard must normally be open before objects can be retrieved from it. Although often possible, it is usually impractical to put on your pants after putting on your shoes. You cannot normally pick something up without being near it. You cannot bake bread every day without periodically obtaining fresh supplies of flour. The sheer mass of such constraints will tend to channel activity in particular ways.
- *Learning situations.* In what situations are agents called upon to do something new? Do those situations have any reliable properties? Does anyone or anything ensure that agents need only perform reasoning that is incrementally more complex than they have performed in the past? When and how can the agent get help?
- *Mutual adaptation.* Does some pressure operate to incrementally adapt various entities to one another? Examples might include biological coevolution, accumulation of shared knowledge in joint activity, and moving parts wearing together. Each of these cases obviously has its own particular logic and its own way of conceiving adaptation.
- *Inertia.* Are there limits to the possible rates of change of important things in the world? Does this inertia provide agents in that world with a margin of safety in uncertain situations? Does it guarantee that dangerous situations will be detectable before they cause permanent damage?
- *Locality.* Are the effects of actions confined to relatively limited parts of the world? Such locality effects can arise either through physical distance or more subtle routes of causal connection. Do these effects simplify reasoning or perception? In particular, do they guarantee that particular important circumstances will be perceptible when the agent needs them to be? Do they provide provable bounds on the possible harm that mistakes can cause?
- *Stabilization.* (The term derives from the article by Hammond, Converse, and

Grass.) What actions do agents in this world take to ensure that the world maintains its computationally useful properties?

- *Geometry.* What properties of the physical environment bound the complexity of the reasoning required to act or learn in it? Are there useful notions of "diameter", "bottleneck", "critical path", "hillclimbing", and so forth?

These are elementary examples of the kinds of structure that one might seek in the environment. In hinting at their computational ramifications here, it has been necessary to employ terms such as "reasoning" and "perception" that might tend to presuppose particular architectures or philosophies of activity. But such commitments are not necessary *a priori*. As the articles in this special double volume will illustrate, analysis of the relationship between agent architectures and structures of the world can be conducted on a wide range of agents in a wide range of worlds. In each case, since the point of structure is to simplify computations, the questions one might address to the world will be shaped by the needs of the architecture. What is the architecture good and bad at? When has it been fragile or incapable of scaling up? When have the necessary computations been impossible or intractable? And so forth.

In particular, one should distinguish between two uses that a designer might make of these structures in the world, corresponding to the aerial view or the ground view. As part of the aerial view of the agent's activities, knowledge about structures of the world might enable the designer to prove that the agent's activities will necessarily have certain properties. For example, these structures might inspire the development of formalisms and methods of argument that enable the designer to demonstrate convergence to set goals. As a separate matter, knowledge about structures in the world might also be part of the *agent's* ground view of its situation. The agent might engage in explicit symbolic reasoning about these structures, it might communicate about these structures with other agents, or it might even use its ideas about the structures to set about creating, maintaining, or restoring them as the need arises. The agent's understandings of its environment need not correspond to the designer's understandings. The agent might have a subset of the designer's understanding of the world, or it might have a simplified or comparatively shallow version of the designer's understanding, or it might slowly discover that understanding for itself by increments and approximations. Clearly distinguishing between the aerial and ground views of the world will help designers keep track of the wide range of design options that are compatible with any particular designer's understanding of the agent's world.

A focus on structures in the world and upon principled characterizations of interactions has a further benefit. When research is focused upon architectures and mechanisms, little intellectual room exists for interchange between researchers pursuing different lines of research. Discoveries about structures in the world and properties of interactions, on the other hand, might be useful to researchers employing radically different architectures. Different research projects may employ incommensurable vocabularies, but each project can learn valuable lessons from the ways that the others have moved back and forth between the design of agent architectures and the exploration of structures of the world and properties of interactions.

2.6. Units of analysis

All of this attention to activities in the world can be misleading when it is viewed from within the context of the history of AI. Founders of AI such as Newell and Simon were engaged in a fight against behaviorism, as were other authors such as Lashley whose ideas were influential in the development of AI. As such, embedded in the concepts and values of AI is a powerful allergy to behaviorism born of the field's founding battles. One factor contributing to this allergy is a powerful distinction made within AI between "cognition" occurring inside agents and "the world" located outside of them. The earliest texts of AI were mostly framed with terms like "thinking" that pointed to internal cognition, perhaps with occasional perceptions and actions but with no strong sense of an embodied agent's continual and structured involvement in an outside world.

Perhaps as a result, research on agents' interactions with their environments can sound like covert advocacy of behaviorism. Indeed, it is easy enough for research that would rebel against AI's conventional ideas to slip into a reliance on behaviorist-type ideas such as the stimulus-response chains that Lashley argued against in his article on the structure of serially ordered behavior. The central conceptual challenge for computational research on interaction and agency is to formulate AI problems and methods in a way which does not fall into either extreme.

Borrowing a term from sociology, the conceptual issue here concerns the "units of analysis" within which research is conducted. There can be little doubt that human beings and other creatures have skins and skulls which provide a certain degree of causal isolation between the things that occur within them and the things that occur outside of them. To the extent that interactions between agents and their environments provide a useful focus of research, though, it will be necessary to define concepts that cross the boundaries between inside and outside. In other words, AI research will have to develop units of analysis that refer to interactions and not simply to an agent plus a world considered as two separate entities. This proposition can sound forbidding to people trained in computer science, inasmuch as an interaction is not a "thing" that can be spoken of as causing effects to happen, or else as being the object of effects caused elsewhere. Yet to speak of interactions as units of analysis is stronger than simply studying interactions: it requires that at least some of one's fundamental concepts be defined in terms of interactions and their properties. What does this mean?

To reconstruct in computational terms the idea of the interaction as a unit of analysis, let us return to the control-theory example with which this introduction began. A controller attached to a refinery (or, for that matter, to a walking robot) will receive a long series of inputs through its sensors, and it will produce a long series of outputs as well. Long-term observation of these numbers may reveal that they converge to certain values, or that they enter into an oscillation with a certain amplitude and frequency. Does the controller cause this behavior? Does the plant cause it? Of course, the behavior results from the interaction of the two, and responsibility for it cannot be pinned down any more precisely. That does not make the behavior mysterious; it only means that the structure of the behavior is (to employ one more term from sociology) "located" in the interaction between the controller and plant, and not in either of them separately.

The notion of units of analysis becomes more important when the agent and its environment are continually influencing each other, so that each one changes through the course of the interaction. If we watch the interaction proceed for a moderately long period of time, so that both organism and environment have had a chance to change in large and complex ways from their original states, then it can be a challenge to specify what the agent and its surroundings even *are*. Of course, one might make a list of every molecule or variable setting or memory address or synaptic weight, but such an enumeration would probably not be a useful or parsimonious description. In such a case, the very identity of the agent, as well as the identities of the various things in its environment, can only be conceptualized in terms of the interaction through which they arrived at their current states. Again, nothing is mysterious about this. The challenge for research is to develop principled ways of talking about it that allow useful arguments to be made about the properties of the interaction, and thence about the rationale behind the agent's architecture and design.

These ideas allow us to reformulate in a more sophisticated way the insights about "structure in the world" described in the previous subsection. The point is not exactly that the world has structure all by itself, but rather that the world has the kind of structure that makes a difference to the workings of that particular agent. This is a property of the relationship between the agent and the world, not of the world alone. For example, we might discover that certain tools have come to be designed in such a way that human hands are minimally likely to slip when using them as they are customarily used. Such tools are well adapted to their customary use, but they might be poorly adapted to use by other species or for other purposes. It is only in a very narrow sense, then, that the tool's adaptation is a property simply of the tool. It is better to conceptualize it as a property of the relationship among a number of entities (tool, hand, materials being worked on, etc.), and specifically as something that only makes sense in the context of a particular form of interaction among those entities. The unit of analysis in this case, then, is not the tool but rather the customary way of using the tool to interact with the world.

This is progress, but much remains to be done. The account of "interaction" in this special double volume is almost wholly individualistic in nature. Its units of analysis, likewise, frame research issues in terms of a single agent's interactions with a structured environment. To make full sense of these interactions from the designer's aerial view, though, it will be important for research on embodied interaction to merge with computational research on social interactions. Tools and the customary ways of using them, for example, are generally not properties of an individual's activities but of a culture's. Cultures provide forms of embodied interaction that offer us considerable guidance in adapting ourselves to a complex world, and computational research holds as much promise for analysis of these settings as it does for the more particular types of interaction treated here.

2.7. *Representation*

The revised theoretical orientation suggested here clears some new space for computational research on representation. So long as research is guided by the correspondence method and the maintenance of objective world models, representations have very specific jobs to do and in consequence are highly constrained in their forms and uses. Within the broader perspective suggested here, though, new possibilities open up. Some of these have already been sketched. Perhaps most fundamentally, designers need to understand the respective roles of indexical and objective forms of representations ("a couple of feet straight ahead" versus "latitude 41, longitude 13"). Indexical representations are more closely tied, in causal and epistemological terms, to the agent's immediate circumstances, but they are not as well suited for other purposes, such as distributing knowledge about spaces and times to agents at distant or unknown locations.

Reacting against conventional theories that have seemed to import an encumbering system of philosophical and architectural assumptions, computational research on situated action has been deeply ambivalent about the concept of representation. Authors such as Brooks [10] and Beer [9] have been willing to say that their agents employ no representations at all. Since representations obviously exist (inner monologues, visual imagery, tactile maps, etc.), this raises the question of what purposes representations actually serve. Brooks and Beer concentrate their attention on insects, and it is common to suppose that representations are late evolutionary developments [43]. Agre and Chapman [4] take another approach, describing a notion of "indexical-functional representation" in which the representational elements are not internal symbolic structures but stable interactional relationships between agents and the objects that serve particular functional roles in the agents' activities. Whether these things really deserve to be called "representations" is a valid question. The important thing, though, is not to provoke a binary argument framed in terms of "representations versus no representations", but to explore interactional concepts which might do similar theoretical work while providing alternatives to the correspondence model.

As the papers in this special double volume demonstrate, it is possible to take a variety of approaches to representation. The perspective advocated here does not dictate any single approach to the question, and the reality of the matter might be complex and heterogeneous. For present purposes, it will be valuable to review some of the history of the AI notion of representation. Most of the earliest explicit theorizing about representation in AI was tied to architectural assumptions and processing mechanisms. Quillian [66], for example, explored mechanisms for automated reasoning in network-like structures that resembled the structure of the brain, at least in the sense of consisting of a small set of basic units joined by connections that can transmit simple signals. Faced with the difficulty of building representations of complex things within these semantic network structures, AI researchers invested great effort into making their semantics clear (e.g., Woods [79]), with the result that the structures were eventually understood by most of the AI community as merely notational schemes for modified first-order logic [38]. Subsequent AI research on representation has primarily been concerned with logical semantics, which has widely been viewed as providing foundations for the whole of AI work [27]. Concern with the physical realization of logical reasoning has lately taken the form of complexity-theoretic analyses of the problems of making crucial types of inferences within logics with particular sets of expressive features [12]. Complexity analysis, though, does not yield detailed information about the consequences of particular architecture choices. Connectionist research, due to its strong focus on the possibilities of a particular class of architectures, has resumed the type of close analysis of distributed inference mechanisms that Quillian began [41].

The story of representational research in AI, then, has had two interacting aspects, semantics and physical realization, whose constraints upon one another have been explored in fits and starts. Computational research on agency and interaction will surely have these two aspects as well, but now each aspect will be placed in the context of agents' involvements in their environments. A basic observation in this regard, already remarked upon, is the inherent efficiency virtues of indexical representations that are tied to direct sensorimotor interaction with an environment. The research in this special double volume is deeply concerned with the physical realization of agents' reasoning (if, indeed, the vocabulary of "reasoning" is employed), but it has no fully developed interactional account of the meaning of representations, and in particular the relationship between what AI has historically understood as "internal" and "external" representations.

As already remarked, external representational materials are likely to provide a substantial amount of useful structure in the everyday world, including things like signs and instructions. What do people do with representational materials, how do these activities complement internal uses of representations, what role do representation-mediated interactions play in the rise of internal representations, and what properties do internal representational reasoning have as a result? These topics have been thoroughly investigated in a variety of other fields [16,17,29,37,50,80], so rather than speculating on the directions that future computational research in this area could take, let us simply survey some possible connections to these other fields that future research could develop more concretely.

3. Connecting to other fields

As the articles in this special double volume demonstrate, computational research on interaction and agency can benefit from contact with a wide variety of other fields of research. The common denominator of these contacts is the abstract notion of interaction, although an extraordinary number of other strands run through the various fields as well. This section offers a very brief outline of some of these fields and their potential connections to AI research. Control theory has been discussed briefly above, and several other fields might be mentioned as well, including philosophical logic, sociolinguistics, decision theory, and several varieties of psychology.

3.1. Dynamical systems theory

In recent years, researchers in numerous fields have developed mathematical models of dynamical systems, which are defined generally as any systems that can be described in terms of the changes over time of a set of interacting variables. Many such "systems", of course, have no very useful properties. And others are simple linear systems whose properties can be analyzed with traditional mathematical tools. Yet others fall within categories of differential equations for which robust solution or analysis methods are known. Dynamical systems theory extends the categories of systems for which useful analyses can be made. The most important cases are those in which the development over time of a system is driven by the repeated application of the same principle, for example the laws of mechanics or natural selection or economic choice. Since AI pursues computational understandings of organisms at all levels, from the neurological and mechanical to the social, the full range of these cases should ultimately be relevant to AI research.

A challenge for the relationship between AI and dynamical systems theory is to reconcile the quantitative and qualitative aspects of the various systems that AI research seeks to understand. Differential equations describe systems that can be characterized in terms of numerical variables, but symbolic systems require other types of analysis. In its broadest definition, systems theory is general enough to provide definitions of even very complex symbolic systems. But it does not follow that general results exist that cast useful light upon those systems. As the various fields develop, they will most likely begin to overlap in their approaches, particularly as the conceptions of structure in the world that inform the research in this special double volume also continue to influence research in other fields.

3.2. *Robotics and vision*

Inasmuch as roboticists construct actual embodied agents, research on robotics has led to several forms of principled analysis of interactions between agents and environments. The most innovative of these have been on the lower levels where the main analytical tools are built upon the theories of kinematics and dynamics in physics. Raibert [67], for example, analyzes the various types of symmetry found in animals' gaits and demonstrates how these might be represented mathematically and used to simplify analysis and synthesis of walking and running machines. The designer of a running machine cannot impose any arbitrary pattern of leg movements and foot landings that might come to mind. Only certain cycles of movement are physically possible, and principled analysis allows this space of possibilities to be characterized.

Or consider the theory of force control [78]. Whereas a position control system directs the movements of a robot effector by specifying a sequence of physical locations that it should occupy, force control is defined in terms of dynamic relationships between the robot and its physical environment, such as a specific force vector that should describe the robot's pressure upon a picked-up part regardless of any changes or variations in the parts' shapes. Because force relationships are indexical or relational in nature, it can be easier to build a sensor for them than for objective quantities like absolute position (unless, of course, the robot and the workpiece are made of rigid elements and fixed to the floor). The point is not that position control is useless, but that the space of possible designs is structured in large part by the kinds of epistemological considerations discussed in this special double volume.

Robot design must also be informed by dynamical analysis of the interactions among objects in the world that the robot's actions will set in motion. Mason [53], for example, presents a mathematical analysis of the interactions that arise when a robot must push an object across a surface. A part might move in a variety of ways due to the vagaries of friction, and anticipation of the space of possible trajectories allows motor plans to be fashioned that move the part into a desired configuration without wasted motion and even without sensors. In situations with greater uncertainty, such a system might visually observe the part's progress and update the dynamic analysis and motor strategy as new information becomes available.

Because computer vision programs can be presented with digital snapshots taken at distant places and times, research in vision has not always been forced to confront the embodied nature of visual activity. The theories of Marr, for example, do not envision an agent that is interacting in any complex way with its environment, assuming instead that the purpose of vision is to construct a three-dimensional model of the world with little reference to its purpose. But recent research on visual systems for robots has begun to demonstrate the depth of rethinking that the construction of embodied agents demands. Ballard [8], for example, presents a series of experiments in "animate vision", in which the architecture of visual processing is interconnected in tight and principled ways with the architecture of motor control. A paradigm of this kind of interconnection might be vergence control, in which the physical configuration of the visual system (eye orientation, for example) is dynamically adjusted to permit stereo focus upon a particular object at some determinate distance. And the architectural boundaries among reasoning, learning, and perception start to disappear altogether once one starts modeling the active choices agents make about what to perceive based on what information they need (cf. [15]).

3.3. Biology

Ecology and evolutionary biology offer several powerful concepts for thinking about the relationship between organisms and their environments. Perhaps the central such concept is adaptation. For an organism to be adapted to an environment is not a simple thing, since neither the organism nor the environment are likely to be simple themselves. Some aspects of adaptation, of course, can be explained in relative isolation from this full complexity, for example in terms of the mechanics of flying or swimming. The challenge for a full explanation of adaptation is that every organism's life has many aspects—locomotion, respiration, avoiding predation, finding and eating food and water, regulating body temperature, social interaction, and so forth—each of which brings its own adaptational demands.

Biologists sum up these demands by speaking of a particular species as filling a "niche" in its local ecosystem. Each component of the ecosystem provides part of the adaptational context for the others, and the result is a tendency for all of the elements to coevolve and to become adapted to one another in intricate ways. As the ecosystem changes through exogenous influences and the internal interactions of its various components, the adaptational demands of the niche will change as well. This dynamic notion of an agent's environment is much more complex and subtle than the conceptions historically employed by computational research.

At the same time, biology has historically had fairly simple concepts to describe organisms' activities in their environments and their interactions with one another, and it is here that computational ideas may make significant contributions. Research on "artificial life" [46,74] has commenced precisely this project. Building and analyzing artificial and simulated creatures may help clarify many biological concepts by forcing unarticulated assumptions and unasked questions to the surface.

3.4. Activity theory

Activity theory is a school of sociologically oriented psychology and education research that developed from the writings of the Russian psychologist Lev Vygotsky [76]. Vygotsky believed that human cognition is profoundly shaped by culture, and in particular that cognitive processes arise through the internalization of patterns of social interaction among people. Vygotsky believed that the process of learning has a great deal of structure. Specifically, he believed that most learning takes place in what he calls the "zone of proximal development". Watching caretakers and children interact in the context of shared activities like games and chores, he observed that the caretakers endeavored to dynamically shift the division of labor between the two, with the aim of ensuring that the child's portion of the activity lay near the outer edge of the child's current capabilities. Thus spared from overly simple and overly difficult tasks, the child could focus on incremental learning. Vygotsky argued that these complex structures of learning are reflected in the child's developing cognition, so that the child's processes of thinking could be viewed as internalizing the patterns of social interaction that gave rise to them.

A strong believer in the cultural dimensions of cognition, Vygotsky also emphasized the role of cultural artifacts such as tools in shaping cognition. The invention or refinement of a tool is an important event, inasmuch as it effectively encodes in a physical material the result of a beneficial process of thinking and experimenting. In learning to use the tool according to its affordances and the customs surrounding its use, future generations will be spared the tedious and haphazard burden of reinventing it. Moreover, in order to use a tool it is usually not necessary to fully understand the reasoning behind it, much less the alternative designs for the tool that had been tried out and discarded. The environment of daily life includes a rich collection of cultural artifacts—the tools and other artifacts to which Vygotsky's arguments apply—that provide a tremendous amount of support to individuals and groups who are organizing their activities. These artifacts include kitchen utensils, buildings and streets, machines like cars and computers, clothing, and much else.

Subsequent research has developed Vygotsky's ideas in numerous directions. The term "activity theory" was coined by Vygotsky's follower Leontiev, who proposed a conceptual framework for analyzing larger "activity systems" beyond the simple parent-child dyad. Leontiev shares with much AI research an interest in the process through which activities become habitual or automatic, no longer requiring conscious structuring and guidance. Activity theory has been brought to the English-speaking world by a number of psychologists and educationalists who are looking for ways to place children's cognition and learning in larger social contexts [60]. Engeström [21] has considerably broadened the activity theory framework to provide a principled means of intervening in complex organizational settings to bring about changes in the local activity system through the development among its members of an "expanded" awareness of its actual dynamics.

3.5. Genetic epistemology

Another relevant school of developmental psychology is that founded by Jean Piaget [63], whose work on "genetic epistemology" traces the ways in which the child grasps the nature of reality through its interactions with its environment. In contrast to Vygotsky, Piaget focuses upon the child as an individual figuring things out through a process that has been likened to scientific experimentation. He argued that the child's relationship to its environment proceeds through a series of discrete and identifiable "stages", each of which is defined by a different form of epistemology. A child in a very early sensorimotor stage, for example, might have difficulty connecting the object that it encounters on one occasion to the same object a few moments later, after it has been momentarily obscured. Later, though, the child will come to understand the "permanence" of objects across time, and this is the beginning of the child's understanding of a world that exists independently of thought.

Drescher [19] has conducted an extensive program of computational research based on Piaget's theories. Building on some suggestions of Piaget's, he has implemented a computer system that employs "sensorimotor schemata" to learn and represent knowledge. A schema, in Drescher's usage, does not represent the world through correspondence or mirroring. Instead, it states (roughly) "when the perceivable world is like *this*, and you take *this* action, then the world is likely to turn out like *this*". Such schemata can be learned through a relatively simple process of induction by simply trying numerous simple actions in numerous situations. More complex cognitive structures can then be built up by "chaining" these schemata, a process similar in form to the assembly of new plans through the stringing together of existing plans and primitive actions in traditional AI planning research. Another, more advanced mechanism is the creation of "synthetic items", which function similarly to "items" of sensory input except that they represent the much more abstract proposition that a certain schema is likely to be applicable in the ongoing situation. Drescher presents some extremely detailed scenarios that describe how the creation of these complex cognitive structures through the simpler sensorimotor schemata explains various features of the developmental process that Piaget traced in his work, including the detailed sequence of substages through which the child passes during the period leading up to the full establishment of object permanence. These scenarios allow Drescher to construct a theory of cognitive architecture that is consistent with Piaget's theories, as well as with recent empirical claims that Piaget underestimated the amount of innate cognitive structure in the infant. Specifically, he argues that the infant, in constructing its own cognitive apparatus in the process of development, effectively *re*constructs the functionality of many innate peripheral faculties, thereby allowing them to be integrated with one another more effectively than is possible on a simple modular design.

For present purposes, the great strength of Drescher's work is that its scenarios include numerous arguments that depend upon on the structure of the environment in which the human infant lives. Within Piagetian theory, the most fundamental fact about this environment is precisely the existence of permanent objects. When you rest something on a table and cover it with a cloth, it stays there until you take the cloth back off. When you put something in the refrigerator, it stays there until someone takes it back out again. As you move around a stationary object, the views of that object that become available to your eyes possess some stable, reliable, and predictable relationships to one another. And so forth. This observation about the environment of human activity gives substance to the scenarios, which follow the child through the discovery of a wide variety of simple but fundamental interactional regularities.

3.6. European phenomenology

Phenomenology is a branch of philosophy whose goal is to develop good vocabularies for describing the experience of ordinary activities. Put in plain language, phenomenology provides words for answering the question, "what is it like?". Although Merleau-Ponty's [56] phenomenological analysis of human embodiment has a straightforward relevance to the analyses in this special double volume, phenomenology is chiefly known in AI through the influence of Martin Heidegger, whose book *Being and Time* [39] provides a phenomenology of ordinary routine activities such as carpentry. Heidegger's work is notorious for its obscurity, and those who have been inspired by his writing to criticize various central tenets of AI routinely find themselves in the impossible situation of translating between intellectual languages and communities that could hardly be more different. In particular, attempts to read Heidegger as directly specifying alternative algorithms or architectures of cognition are doomed to especially intractable confusion, inasmuch as it was very much Heidegger's goal to avoid expressing himself in such terms.

Nonetheless, Heidegger's writing can, if handled with care, provide useful guidance for the development of computational theories of interaction. Heidegger places great emphasis on the customary forms of activity that structure much of everyday life, and in particular upon the customary uses of tools that give a conventional structure to actions, space, and materials. This "structure", for Heidegger, is not a cognitive symbolic structure, but a structure of experience within which things take on particular interrelated meanings. He emphasizes, for example, that we do not normally relate to a pencil as *this* particular pencil, but rather as *a* pencil that is used in a certain habitual way. We can choose to withdraw the pencil from this ordinary, routine kind of relationship to our activities, staring at it as an object of curiosity or levity or scientific inquiry, but that is a very different experience from simply using it to write. This idea suggests, in a loose sort of way, investigating how a computational theory of interaction might give a different status to routine interactions with generic things (writing with a pencil) than to exceptional interactions with specific things (examining or measuring this particular pencil).

But research like Heidegger's can have its most productive influence upon AI when AI itself recovers a sense of its own historical development. Despite the efforts of Dreyfus, Heidegger's work is not directly addressed to AI as it exists today, but rather to a larger tradition of which AI is one part. If Heidegger's analysis of the history of philosophical ideas can be viewed as indicating paths not taken—and, therefore, alternative ways in which AI research might be conceptualized—then it need not be taken as posing an all-or-nothing challenge to AI's foundations, but rather one critical perspective to assist in the field's self-examination and evolution.

3.7. Buddhist phenomenology

Whereas the European phenomenological literature is relatively recent (while, of course, building upon much older philosophical tradition), Buddhism has an ancient and continuously developed phenomenological system. The whole point of Buddhism is to seek enlightenment by using systematic meditation to pursue mindful awareness of one's own cognition, and Buddhist scholars have developed extensive descriptive accounts of cognitive processes that provide guidance for this process. These systems of description have evolved historically as communities of meditators have found previous formulations inadequate to describe their own experiences. Central to this evolving intellectual system is the idea of illusion, which holds that any particular conceptualization of reality must be understood as an imposition, perhaps possessing heuristic value but not providing any definitive or exhaustive representation. Prior to the cultivation of mindful awareness, cognition proceeds in a ceaseless cycle of imposing specific, prestructured interpretations upon the world and contracting desires and drives based on the unrecognized illusions that result. Mindfulness does not eliminate thought or paralyze action, but it does liberate the individual from the illusion that thought transparently grasps reality or inevitably compels action.

These ideas may seem distant from the concerns of computational research. Yet Varela, Thompson, and Rosch [75] have argued that the connections are actually numerous and deep. Both traditions of inquiry, after all, are concerned with the mind and its relationships with reality. The pivot through which these authors develop the connections between the two traditions is Maturana and Varela's notion of "structural coupling" [54]. Structural coupling is a biological notion rooted in the theory of evolution. Evolved species are adapted to their environments, and this adaptation ought to be conceptualized in interactional terms:

- the organism interacts in complex ways with its environment;
- this interaction both sustains the organism's internal functioning and has some range of effects upon the environment as well;
- the organism's internal structures and the structures of its external environment have both changed over historical time through mutual adaptation of species and ecosystem; and
- the changes in these structures have accumulated to such a degree that it is difficult if not impossible to understand them except in the context of their interaction.

The structures of the organism and its environment are, in this sense, "coupled" to one another. Varela, Thompson, and Rosch point out that this coupling is analogous in certain ways to the Buddhist notion of "codependent arising", which describes the experience of cognition: one does not experience cognition as rising up, searching for something in reality, and then settling upon it; nor does one experience reality as invading oneself and bringing a previously dormant cognition back to life. Instead, the processes of cognition and the structures of reality arise together, each proposing the other as its own illusory validation. Although further research will need to flesh out this analogy in more detail, it is certainly stimulating, and leads to innovative investigations of the processes through which perception and action guide one another in embodied activity.

This type of research has many skeptics among technical people. And indeed, phenomenology and computational research are not straightforwardly commensurable. Allowing each to influence the other requires drawing out the most promising analogies between them and pursuing the suggestions for research that these analogies might generate. Unfortunately, the phenomenological method strikes many people in the cognitive science tradition as akin to introspection, which was once pursued systematically by Russian and German psychologists but ultimately ran afoul of its lack of conceptual precision and empirical reproducibility. Phenomenology, though, makes no claim to identify internal mental mechanisms but only to provide compelling and detailed descriptions of experience. More importantly, phenomenological methodology, particularly in the Buddhist version, is simply far more rigorous than introspectionism, having developed over a long period in extensive communities of investigators.

3.8. Sociology

Sociology has numerous schools and subdisciplines, many of which have potential connections to research on interaction in AI. Perhaps the central question of all sociological research, though, is the question of social order: in virtue of what does society seem to have a relatively stable structure? It is impossible to survey the many formulations of this question in a short space, much less the available answers to it. Nonetheless, research within sociological schools such as ethnomethodology and symbolic interactionism has been distinguished by a commitment to detailed empirical investigation of how, in fact, people actually enact the structures of society in their dealings with one another.

The sociological research program that has had the greatest influence upon thinking in AI is arguably that of Lucy Suchman in her critique of AI planning research in *Plans and the Structure of Behavior* [72]. Suchman observed some people attempting to use a photocopier that had been equipped with a device that, based on AI planning theories, attempted to guide its users through complex copying operations by constructing a plan and then presenting the successive steps of the plan to the users. The users experienced a wide range of difficulties using these instructions, and they interpreted them in complex ways based on the situation as it presented itself in the moment. In particular, far from executing the instructions in the manner of a computer program, the people employed the instructions as resources—and as one set of resources among many—in figuring out what actions to take next. Both through Suchman's influence and other developments, a number of AI research projects have investigated the complex and varied uses that might actually be made of plans [5,30,64]. The deeper point, though, concerns the many subtle and improvised ways in which people structure their actions in accord with the demands of moment-to-moment meaningful interaction. These phenomena may lead future computational research to rethink its basic concepts in ways that can do justice to the improvisatory nature of human action.

3.9. Anthropology

Historically, anthropology differs from sociology in that it studies "them" rather than "us". Remarkably, even as this distinction has become untenable, anthropology has retained its distinctive character through its focus upon culture—and specifically upon the notion that cultures differ from one another in profound ways. Exploring the question of social order in a large variety of settings has led anthropologists to investigate numerous features of life that are normally too familiar to attract much attention. Among these is the role of habitual activities and customary artifacts in defining and maintaining a social order. Pierre Bourdieu [11], for example, disagreeing with a long tradition that locates social order in loud and visible things such as laws and ceremonies and conflicts, suggested to the contrary that the social order can best be found in the most ordinary details of everyday activities, and particularly in the habitual structuring of everyday uses of artifacts like houses, hearths, tools, clothing, pathways, and so forth. Anthropologists refer to this kind of theory as "practice anthropology" [61] because of its emphasis on the hidden order to be found in everyday cultural practices. While perhaps exaggerated in its emphasis on structured habit as opposed to conscious choice, this style of research has had a massive and generally salutary influence as anthropologists have chosen to view ever more ordinary and quotidian aspects of life as important and meaningful, and as legitimate topics of research.

In connecting anthropology with computational research on interaction and agency, a significant obstacle is the differing scales of research. Computational researchers must get things working, and that requires analyzing very small and specific actions. Whereas several schools of sociology, including those previously mentioned, have engaged in microscopic studies of human interaction, anthropology has mostly been concerned with larger things. Even when an author such as Bourdieu speaks of the fine details of habitual activities, it is rare for anything like a worked-out grammar of those activities to be provided. The focus, instead, is on articulating a set of analytical categories that allow apt descriptions to be given—descriptions that allow things on very different scales to be fitted together, so that economic structures, for example, can be related to the ways in which people teach and learn skills.

In this regard, a particularly promising analytical framework can be found in the work of Jean Lave. In her book *Cognition in Practice* [51], she provides a set of categories for analyzing people's interactions with their worlds on several different scales (more accurately, "levels"), from the moment-to-moment interleaving of different tasks to the historical structuring of an arena of activity such as a kitchen or supermarket. In contrast to much research in cognitive science and AI, she rejects the notion that people decide what to do by solving "problems" that can be abstracted from the complex and interconnected details of moment-to-moment activity. The people she observes do not so much solve problems as work through complicated dilemmas, resolving things just enough to keep moving. Her resolute focus upon units of analysis defined in terms of interactions leads her to theories that are hard to reconcile with computational research as it has historically been practiced. And indeed, numerous details will have to be worked out and rethought in order to strike up a productive relationship between conceptual systems such as Lave's and the concepts that guide computational research on interaction. An emphasis upon formulating computational ideas in terms of interactions, though, will ensure that the units of analysis in the two research projects are at least commensurable.

4. Papers in this double volume

The papers collected here are a diverse group, deriving from a remarkable variety of disciplinary backgrounds and technical literatures. Although they are described here within the agenda and vocabulary of this introductory article, it bears repeating that they each represent a distinctive approach to the issues. They should be understood as voices in a conversation, with numerous and subtle points of interconnection among them. Despite the temptation to impose an artificial structure upon them by sorting them into topical groups, they are arranged in alphabetical order by the first author.

4.1. Arbib and Liaw

Arbib and Liaw present an evolutionary scenario for explaining the complex functionality of the nervous system. Taking as their model the visual system of the frog, they summarize the evidence that motivates their model. Rather than directly specifying the operation of neurons, they frame their theory at an abstract level, in terms of the interacting "schemata" that give rise to the observed patterns of behavior. Schemata are abstract units of computational functionality that can be implemented on a variety of hardware substrates. In particular, schemata provide a level of abstraction that allows brains and computer hardware to be discussed in a common vocabulary. Arbib and Liaw and their colleagues have developed general formal models of schemata that allow precise accounts of particular systems to be formulated and reasoned about in principled ways.

Beginning with the life and ways of the frog, Arbib and Liaw develop an approach to the study of the visual system that places it squarely within the context of an embodied agent's interactions with its environment. Having done so, they discuss the issues that arise in making the difficult transition from sensorimotor behavior to symbolic reasoning. In particular, they sketch some processes through which novel schemata might arise in response to the demands of novel situations. They emphasize, however, that cognition within the schema model is not controlled by a centralized device but is a matter of cooperation and competition among a distributed set of schemata. Furthermore, these schemata do not employ any single representation scheme, but rather a patchwork of partial representations, each of which captures a particular aspect of the agent's interactions within a particular mode of processing.

Arbib and Liaw's argument illustrates an inversion of priorities that is common in computational research on interaction and agency. AI has traditionally been concerned with "higher" cognitive functions such as the construction of innovative plans to solve arbitrary goals, with less attention to evolutionarily prior phenomena. This is natural enough if one believes that, in fact, "thought" is a phenomenon that can be defined and studied by itself, without reference to the whole background of "low-level" processing and routine activity against which thought takes place. Rejecting this point of view leads to a considerably different approach: an emphasis on routine activities, on sensorimotor interactions with the world, and upon the ways in which "low-level" functions provide the functional, developmental, and evolutionary basis for the higher functions. In practice this means that the higher functions generally suffer the same methodological postponement that older AI research had visited upon the lower ones. The ultimate challenge, of course, is not to declare one set of functions to be more important than the other, but rather to provide substantive accounts of their interrelationship. Arbib and Liaw suggest some ways of seeing the higher functions as continuous with the lower ones. Although the higher and lower functions are made of the same stuff, so to speak, they do differ in the sense that the higher functions require schemata to be replicated and synthesized in a way that the lower functions do not. As Arbib and Liaw point out, this evolutionary shift is congruent with the theoretical movement that Newell referred to as the "Great Move" [58]—from a focus upon hardware, with its static interconnections, to a focus upon the interconnectable symbolic structures of higher thought.

4.2. Barto, Bradtke, and Singh

Barto, Bradtke, and Singh review and synthesize a great deal of research on the use of dynamic programming, applying their unified understanding of this class of algorithms to real-time control. Each of these algorithms enables an agent to learn how to improve its efficiency in achieving goals when interacting with dynamic, and possibly stochastic, systems. Through repeated trials of actual or simulated control of a given system, the agent draws on its accumulating experiences to produce improved control strategies. These methods differ from AI's heuristic state-space search techniques in that they must repetitively visit a large number of states, as opposed to threading their way through an explosive number of states. By improving their heuristic evaluation function using principles from dynamic programming, they improve their search strategy. Unlike classical dynamic programming systems, the algorithms described by Barto, Bradtke, and Singh do not need to visit all of the possible states. As the algorithm converges, effort is focused increasingly upon those states which must actually be visited by an optimal controller. It is possible to prove fairly strong results about the conditions under which these learning methods will converge to optimal controllers.

This research unifies results from a number of fields. In particular, by pointing out the relevance of asynchronous dynamic programming methods to research on learning in stochastic environments, it greatly strengthens the connections between AI research on search and learning and control theory research on adaptive control methods. The result is a nearly exhaustive investigation of a set of weak methods that can be applied to a wide range of problems. On the other hand, as with any weak method of any generality, the results guarantee convergence without making any strong promises about how long convergence will take. A project for future research will be to understand how the learning methods might be specialized to take advantage of particular kinds of structure in the environment.

4.3. Basye, Dean, and Kaelbling

Basye, Dean, and Kaelbling develop a series of algorithms for probabilistically solving the problem of "system identification". On an abstract level, system identification is the problem of reconstructing the structure of a state-transition graph by sampling its input–output behavior. That is, the algorithm is presented with a series of discrete options (such as turning left or right) and, upon choosing one of those options, is told what information can now be "seen" (red or green; hot, warm, or cold; etc.). Although system identification problems have been investigated in a wide variety of settings, their relevance here is to the problem of discovering the structure of an environment by traveling around in it according to some strategy. And since Basye, Dean, and Kaelbling have real robotic applications in mind, they have extended the problem to assume that the information about the environment that the agent receives is only probabilistically correct, perhaps because of noise in the operation of its sensors. As a result, their algorithms do not guarantee perfect correctness but a certain specifiable likelihood of correctness.

The problem cannot be solved in its most general form, since insufficient information may be available to sort through the fog and actually pin down which states are which. Intuitively, the difficulty is that the agent never knows where it is, has no guarantees that it can return to where it came from, and has no perfectly reliable way of knowing if its present location is the same as its location at any previous time. Its exploration, moreover, must begin wherever it happens to be located; it cannot, in other words, jump to an arbitrary location. Therefore, the authors explore the structures in the environment which can be exploited by particular search strategies to provably reduce uncertainty. It transpires that the problem can be solved in probabilistic polynomial time, provided that the world has certain properties such as reliable landmarks or tightly constrained structures which can be mapped with greater certainty than wide-open fields of densely interconnected vertices.

Although clearly simplified in relation to many real environments, Basye, Dean, and Kaelbling's paper is a sophisticated study in the interaction between learning, partial knowledge, action strategies, and environment structures. Their agent is not omniscient, does not reliably know where it is, may have wildly mistaken ideas about the structure of the environment, and follows trajectories that the designer can only characterize in abstract terms. Despite this, it is possible to characterize the agent's interactions with the environment in sufficient detail to demonstrate that the resulting models of the environment will converge to accuracy. Although it is convenient to explain the algorithm using the spatial metaphors of travel through a graph-structured space, the results will apply to environments in which the state-transitions represent other kinds of changes, such as the workings of artifacts. An important project for future research will be to understand what structures of particular categories of environments, especially these not-literally-spatial ones, correspond to the formal properties of graphs that permit Basye, Dean, and Kaelbling to prove their results.

4.4. Beer

Beer applies the mathematical machinery of dynamical systems theory to the formalization of agent–environment interactions. Specifically, he proposes viewing agent and environment as two coupled dynamical systems, so that the interaction between them can be viewed as the trajectory of one large system whose variables are simply the variables of both agent and environment together. This proposal provides a straightforward reading of the general notion of making interaction, not the internal cognitive processing, the unit of analysis for AI research. Having defined things in this way, dynamical systems theory provides an extensive vocabulary for discussing the space of possibilities through which a given agent-environment system travels. A given region of the space, for example, might form a basin within which all possible initial configurations eventually settle into a stable, periodic "limit set". The interaction might be defined as adaptive in relation to an arbitrary condition upon its trajectory.

One benefit of this general approach is in analysis. Research on interaction and agency will only progress if it becomes possible to inspect particular performances, and to characterize general categories of them, so as to understand what the agent is really doing and why. Since any given interaction can be understood as a trajectory through a space, this trajectory can be submitted to analysis using a variety of tools. Particular trajectories can be visualized by being plotted, though presumably in a reduced subset of the dimensions of the full coupled dynamical system. Beer provides several examples of this kind of analysis, and of the conclusions that can be drawn from it.

Beer's particular domain is a walking robotic insect whose leg parameters are driven by a simple neural network. The weights of this network, in turn, are set by a genetic algorithm that simulates many different settings of the weights and homes in through incremental, evolutionary refinements on a set of weights that maximizes certain measurements of the simulated insect's performances. Analysis of these performances demonstrates that the neural network has settled upon patterns of interaction with the environment (the insect walks on a horizontal floor) that correspond to the gaits used by insects. Moreover, when the sensors measuring the positions of the insect's legs are unreliable, the genetic algorithm settles upon a set of weights that permit the insect to switch among different dynamics for generating gaits as the situation demands.

Beer emphasizes, though, that his principal commitment is not to this particular architecture but to the dynamical-systems framework. He presents these discoveries as prototypes of an emerging style of AI research in which an agent's embodiment is accorded a central role. Once this is done, he argues, all of the traditional categories of AI research must be rethought. His robotic insects, for example, have internal states but do not have anything resembling traditional symbolic notions of representations. Of course, one might vindicate the notion of representation by defining it widely enough to include all possible uses of internal state. But so long as the notion retains any real content, Beer argues, his insects fall outside of it. Instead, the internal states in the robotic insects are grounded in, and take their functional "meanings" in relation to, the agent's interactions with its environment.

Dynamical systems theory provides a highly general framework for formalizing agent-environment interactions. Perhaps the principal challenge for research within this framework will be to formulate systems-theoretic definitions that capture the particular kinds of structure encountered in more general categories of agent-environment interaction. The structures of tool use, for example, presumably correspond to particular properties of enormous dynamical systems. But can these constraints be captured by relatively compact and comprehensible formulas for characterizing those properties? This question will presumably not have a single, simple answer. As Beer's analysis shows, development of this theory will require the elucidation of new, more appropriate conceptions of categories as basic as "representation". AI having been decades in the making, its reconstruction in terms of interactions will take unpredictable forms as well.

4.5. Donald

Donald presents a formalism for reasoning about the computational properties of distributed sensor systems. Given two sensor systems arranged in the world, one would like to ask a series of questions modeled on the theory of computational complexity: Can one sensor system detect every condition in the world that the other can? Are the two sensor systems equal in their sensory powers? If not, can we define precisely what would need to be added to the "weaker" sensor system to make it equivalent in power to the "stronger" one? And most generally, does there exist a formal sense in which information is "conserved" in the movement from one design to the other (for example, a design that involves two relatively simple mobile agents communicating via flashing lights versus a single relatively complex agent performing all of the necessary computations on its own)?

Formalizing these questions brings forth a large number of points that usually remain in the background of a design process. For example, a great deal of information is encoded into the calibration of sensors, and the formalism makes it possible to explain precisely what this information amounts to, and how it compares to the addition of another sensor or the addition of extra capabilities to an existing one. To take another example, it may turn out that one sensor system cannot be transformed into another, seemingly similar one, without the expenditure of considerable computational effort because that transformation would require a particular computation to be inverted; inverse problems, of course, are frequently much harder to solve than the problems they invert.

The form of analysis made concrete in Donald's paper would be extremely useful if extended to other aspects of the design of autonomous agents. An informal model in this regard might be found in Braitenberg's [13] speculations about the capacities of various kinds of agent machinery. Intuitively speaking, as an agent's machinery grows more sophisticated, it ought to be able to participate in a growing range of interactions with a given environment. Different forms of interaction would thus fall into a hierarchy, according to which categories of agent machinery are capable of participating in them. Of course, this hierarchy depends on the particular environment being considered, and the rearrangements of the hierarchy in different environments would provide a valuable indication of the degrees and kinds of adaptation that different varieties of agent machinery possess to environments with particular properties. These kinds of understandings would qualitatively improve our abilities to design novel situated agents and to understand and explain the ones that already exist.

4.6. Hammond, Converse, and Grass

Hammond, Converse, and Grass wish to develop computational models of what they call "long-term activity". Whereas classical planning was defined in terms of an agent pursuing a goal (and then, presumably, going to sleep or asking for a new goal), Hammond et al. wish to understand the strategies by which an agent can engage in productive activity in a given environment over long periods. In their contribution to this special double volume, they explore a category of action policies that contribute to orderly action over long periods. They refer to these policies as "stabilization"—actively changing the environment so as to maintain in effect the properties that the agent's actions rely upon. A simple example would be putting away your tools when you are finished with them. Hammond et al. provide a helpful taxonomy of the types of stabilization, with many examples, and they embody some of these examples in a simple demonstration program.

The underlying argument in Hammond et al. is important for the general project of making computational theories of interaction and agency. An omnipotent and omniscient agent would not need to put its tools away, since it would have no trouble finding them the next time it needs them. Agents with more realistic capacities, by contrast, need the world to have relatively stable properties. This observation takes on a specific form in the context of the case-based architectures that Hammond et al. employ. Their emphasis is not upon unique, complex, creative forms of reasoning but upon the stockpiling of "cases" that permit newly arising situations to be usefully assimilated by precedents from situations that have gone before. An agent with a large collection of cases will be able to act in a sophisticated fashion without necessarily engaging in sophisticated computation. But a collection of cases is only useful if those cases actually arise in the future. If the tools are always left in different places then new cases might be required much more frequently than if they are always left in the same place. It therefore makes sense, other things being equal, to actively manipulate the world so that the same cases tend to arise over and over.

Note the form of this argument: faced with a seeming lack of generality in their architecture, Hammond et al. did not immediately decide to make the architecture more general. Instead, they sought structures in the world—and, more specifically, in agents' interactions with the world—that, once properly articulated, actually revealed an adaptive "fit" between the architecture and its environment. This is similar to Simon's approach in *Administrative Behavior* [71], where he explained the functioning of organizations largely in terms of compensating for the limited rationality of individual employees. Hammond et al. are, of course, dealing with individual agents in a wider variety of environments, but the similarities remain. It could have transpired that no viable compensatory structures were found, in which case suspicion might have been transferred back to the architecture. But generalization of the architecture should be the second line of defense, not the first.

The approach of Hammond et al. ought to find application in a wider variety of settings. Social and organizational activities have their own forms of stabilization, and techniques of stabilization are supported by cultures in many ways, from artifacts such as toolboxes to teaching methods to linguistic phrases such as "this goes here". Analysis of the limits of stabilization, moreover, might lead to the discovery of new (seeming, apparent) weaknesses in the case-based architecture, which might in turn provoke a search for further types of structure in agents' interactions with the world.

4.7. Hayes-Roth

Hayes-Roth introduces the concept of the "niches" that can be occupied by particular categories of agent architectures. While inspired by the biological concept of a niche, Hayes-Roth defines a niche according to several dimensions, each calling for a particular architectural approach: perceptual strategies, control mode, reasoning choices, reasoning methods, and meta-control strategies. For example, some environments, perhaps due to their high reliability and their high demands for efficiency, call for control modes based on strict linear sequencing of actions; other environments, by contrast, may call for actions to be improvised based on relatively complex moment-to-moment adaptation to evolving circumstances.

Rather than looking for a single super-architecture that is equally responsive to the entire territory of niches, Hayes-Roth has developed an architecture that is capable of dynamically adapting itself to changing conditions, synthesizing control policies that select and combine certain elements of the system's architectural repertoire according to its analysis of the demands of the situation. This kind of dynamic adaptation is necessary in Hayes-Roth's target domain of intensive-care monitoring, an extraordinarily complex environment whose demands can qualitatively shift among extreme positions. When a patient has a sudden medical crisis, for example, long-term tracking and reasoning must give way to a much more urgent form of processing that is capable of rapid responses to shifting states. Likewise, the patient's response to treatment may drift into an unfamiliar pattern, requiring the agent to change its processing mode into a much more active policy of probing and diagnosing to determine what might be going on. This reasoning, in turn, might shift between more qualitative forms based on past precedents and more quantitative forms based on simulation, depending on what kinds of information and symbolic knowledge might be available.

Hayes-Roth's system is still evolving. In evaluating it, Hayes-Roth insists upon the accumulation of empirical experience in complex real-life domains such as intensive care monitoring. As a strategic matter, the architecture can be deemed promising if it is able to shift gracefully among the various modes of operation that changing conditions require of it. Detailed analysis of whether the system behaves optimally within each of its many modes will be required later on, of course, once each facet of the system is equipped with the mass of detailed knowledge that it will require. But qualitatively accurate responsiveness within a relatively parsimonious architectural framework will provide promising signs for future development.

4.8. Horswill

Horswill presents a methodology for the construction of specialized agent architectures. Observing that AI has long pursued the goal of wholly general architectures that can be adapted to arbitrary circumstances, Horswill considers the contrary project, a search for architectures that are maximally adapted to particular environments. He suggests a process of incremental refinement in which structures of the environment are, so to speak, "folded in" to the agent's computations as assumptions, yielding simpler versions of the architecture that require simpler forms of computation and perhaps, in extreme cases, no computation at all. Experience with this method ought to lead designers to fill out a space of possible designs, a kind of lattice structure within which a designer can move downward as new environmental regularities are discovered and upward as those regularities prove false or unstable.

His examples are chosen from the construction of an autonomous robot designed to provide tours of an office space. Suitable constraints are discovered in the level floor, reliable visual properties, and independence of variables in search spaces afforded by this environment, leading to a particularly simple agent design. The same design process in a different environment, of course, might lead to the discovery of different regularities and the making of different simplifications to the agent architecture. Horswill emphasizes that his analysis of his robot's architecture is largely retrospective. The point of the design methodology is not to provide a simple algorithm from which optimal designs can be cranked out, but to provide a framework for thinking within which the generalization and specialization of designs can be undertaken in a conscious and deliberate way.

Horswill's paper expresses in a particularly clear way a theme that runs throughout these papers: the desirability of parsimony in architectures. When the units of analysis for design and analysis are defined in terms of interactions, the mutual fit between an agent and its environment becomes the most important source of guidance for the design process. A highly general architecture may be able to function well in a wide variety of circumstances, assuming that its computations are not impossibly cumbersome, but this very generality will produce a great deal of "slack" in the architecture's relationship to the environment. By aiming for simple machinery, and by shifting the primary explanatory burden to interactions and not to the architecture, designers such as Horswill are forced to pay ever more detailed attention to the environment and the agent's place within it.

4.9. *Kirsh*

Kirsh explores the wide variety of ways in which people employ the space around them to complement their cognition. If we watch people as they work we note that they constantly manage the resources around them, not just to get things done, but for cognitive ends—to highlight opportunities, to encode useful information, and to keep the task-relevant complexity of the world to a manageable level. These cognitively oriented manipulations of the environment happen on all scales, from a slight repositioning of a single workpiece to a long-term structuring of a whole workplace. Considering a striking range of cases, he distinguishes among three phenomena: spatial arrangements of tools and materials that simplify an agent's choices among alternatives, spatial arrangements that simplify the gathering of information through perception, and spatial arrangements that permit calculations to be formulated in a way that fits better with the capacities of internal cognition. As we observe an individual interacting with a complicated array of physical things in an environment—particularly when participating in a familiar activity in a familiar setting—it can become difficult to draw lines between the "internal" and "external" aspects of cognition. Of course, it is simple enough to make one list of the causal events going on with hands and artifacts and a second list of the causal events going on within brains, but the fact is that these two categories of events are continually triggering one another, so that it is difficult to make sense of them except as a members of a closely coupled system.

It is here that the case for an interactional unit of analysis in computational research on situated agency starts to become compelling. This is not to say that analyses based upon traditional theories of cognition must be abandoned. To the contrary, Kirsh uses theories of cognition-as-search to provide an intuitive explanation of why certain spatial arrangements of things lessen the burden upon internal cognition. The resulting picture of interaction, though, takes those traditional concepts in new directions, placing them in the larger context of an agent's involvement in a highly structured environment.

Kirsh's analysis brings out some of the enormous complexity of the phenomenon of "adaptation". The metaphors used to explain adaptation are frequently structural: the agent is spoken of as "well-fitted" to its surroundings. Yet if we ask whether a particular cognitive architecture is well-adapted to a given environment, the question only makes sense in the context of a potentially elaborate set of practices by which the agent actively manipulates its surroundings from moment to moment to achieve that "fit". Many of these practices are cultural in nature, must be learned by the agent, are supported by artifacts, and so forth. Furthermore, the means by which agents actively manage their workplaces are inseparable from the means by which they actually get useful work done in those settings. With this realization, the boundary between "perception" and "action" becomes complicated, and it becomes necessary to take care about what these terms—so sharply distinguished by many conventional AI architectures—are to mean.

4.10. Lespérance and Levesque

Lespérance and Levesque adapt methods from philosophical logic to give an account of the distinction between objective knowledge and indexical knowledge. Their point of departure is the observation that agents routinely know things in indexical terms that they do not know in objective terms. For example, it is common to know things like "something red just went by here" without having any objective name for "here" (such as a conventional place name or a latitude and longitude) or any objective knowledge of the current or recent time (such as a clock reading). Of course, the designer or another outside observer might have this knowledge from an aerial point of view. But down on the ground, the world is immediately tangible in indexical forms. Requiring an agent to represent the world in objective terms, then, would impose a wholly unnecessary epistemological burden, as well as requiring that knowledge that is actually independent of objective information (like the right way to core an apple, which operates regardless of what county one inhabits or what month it is) must be formulated in unnecessarily cumbersome ways, quantifying over the possible places and times rather than in indexical terms.

Formalizing indexical knowledge accurately, though, presents significant challenges. Many of these pertain to time. Events can have a range of complex relationships to "now" and to various "thens", and Lespérance and Levesque develop a fairly sophisticated logic of time that permits a wide variety of types of partial knowledge to be expressed accurately. They are also able to express a wide variety of "knowledge preconditions" for action. A simple example is that you cannot call me on the phone without knowing my number. A more complex example is that you cannot reliably place a letter in my mailbox if you are only aware of being "here", as opposed to being on my front step.

The logical formalism that Lespérance and Levesque have developed is meant, as they explain, solely as an account of the "knowledge level" of indexical and objective reasoning. That is, they do not provide any account of how these forms of reasoning might be realized in hardware. It would be a mistake to assume that an agent would have to manipulate a mass of symbolic formulae corresponding to those in Lespérance and Levesque's paper. Instead, it is possible that their formalism is best employed by the designer as a tool for analyzing (and, of course, designing) an agent's patterns of reasoning. Before this possibility can be realized, though, it will be necessary to explore the computational properties of the formalism and the ways that it can be fitted to particular classes of machinery. Simple and straightforward realizations of their theory will of course be possible through the use of general-purpose logical theorem-proving programs. This approach is most likely impractical, though, and more sophisticated kinds of physical realization will probably require the logic to be adjusted in various ways. Research in this area is bound to produce an expanded understanding of the computational properties of various forms of situated reasoning.

4.11. Lyons and Hendriks

Lyons and Hendriks present an architecture for the automatic incremental synthesis of agents that participate in complex, structured interactions with their environments. Their research is founded upon a formal framework for the characterization and analysis of agent–environment interactions. The basic idea is to model the agent and environment as interacting mathematical automata. Each automaton is assembled from a vocabulary of basic computing elements, and the behavior and interaction of agent and environment can be modeled in terms of the trajectories followed by these automata as they evolve according to a fixed set of formal rules. This approach allows one to make precise a long list of important questions about interaction, most particularly whether the interaction will eventually converge to a specific desired state. Although the method is only as powerful as the proof techniques for demonstrating such conclusions within it, it stands as one of the most thoroughly worked out frameworks for analyzing qualitatively complex interactions.

In their paper, Lyons and Hendriks employ their automata-theoretic formalism to motivate the design of a system for controlling industrial robots as they engage in complex assembly tasks. Their architecture has two components, a "reactor" that employs a fixed circuit-structure to control the robot's moment-to-moment interaction with its environment, and a "planner" that is capable of incrementally adding to the reactor's structure so as to extend its behavioral repertoire. The automata-theoretic formalism provides Lyons and Hendriks with a principled basis for designing a language that a programmer can use to represent "dynamics" of interaction. A robot can "participate" in one of these dynamics just in case it can sense particular kinds of situations, and take particular actions in them, that will guarantee that the joint agent–environment system will evolve in a particular way.

This is a different and more complex concept than the traditional notion of "executing a plan to achieve a goal". First of all, Lyons and Hendriks take for granted that only a certain proportion of the action in the world will be controllable by the agent (for example, through the movement of its limbs). Secondly, the "planner" does not envision a definite sequence of actions and world-states through which the "execution" will travel. Instead, it specifies a potentially large and complex space of possible trajectories whose destinations can be sufficiently influenced through the adoption of particular action policies that can be physically realized by the reactor, through the particular kind of machinery of which the reactor is made.

4.12. Rosenschein and Kaelbling

Rosenschein and Kaelbling present a view of representation and control based on the theory of situated automata. They observe that AI ideas about representation have frequently been based on mathematical logic, or upon notations that can be formalized in logical terms. Unfortunately, these ideas have traditionally been accompanied by specific architectural commitments, according to which knowledge is formulated through structures modeled on the techniques of symbolic programming. Thought, in this view, is a matter of the explicit computational manipulation of these symbolic structures by mechanisms such as theorem-proving programs. The extreme inefficiency of most such schemes has cast shadows on formal logic as a research tool in AI. Rosenschein and Kaelbling point out, however, that the basic point of logic is not architectural but semantic: it is a formal means of sorting out the meanings of representational elements, with no inherent commitments about the manner in which these elements are physically realized.

Pursuing this observation, Rosenschein and Kaelbling present an agent synthesis methodology in which the machinery being generated is unusually simple and straightforward. They present a logical formalism that allows them to represent the workings of a specific, wholly traditional class of digital machinery. The representational elements here are not symbolic structures but values in registers and on wires. Logical formalization permits the designer to give a precise account of the meanings of individual elements in terms of their correspondence to the world, and logical notation provides the basis for a set of languages for specifying the machinery for newly designed agents. The resulting circuitry need not be specified in complete detail. To the contrary, the compilers for these languages can perform a wide variety of manipulations on the logical forms and circuitry representation, and these manipulations can be proven to preserve the intended meanings of the computations because of the clear formal semantics of the underlying logical formalism.

The devices that are synthesized through Rosenschein and Kaelbling's methods are embodied agents whose activities take place across time. The authors point out that this is quite a different picture from the traditional notion of "solving a problem" by mapping a single, isolated input onto a single, isolated output. Instead, the picture is more like that of control theory, with a continual stream of inputs and a continual stream of outputs—in this case, tied to a discrete digital clock. The logic includes operators that can represent the relationships between values on adjacent ticks of this clock, thereby making it possible to reason in a principled way about the meanings of computational processes that unfold over a series of time units. A further valuable step would be to employ these methods to formalize the time-structures of activity in particular kinds of environments, in which strong guarantees might become possible regarding the correspondences between time-extended computations inside the agent and time-extended processes occurring in its surroundings.

4.13. Schoppers

Schoppers presents an architecture that combines a modal logic of time and belief with a control-theoretic philosophy of an agent's relationship to its environment. Rather than engaging in complex symbolic reasoning on-line, Schoppers' program compiles a sensorimotor decision tree that interacts with a variety of asynchronously operating subsystem controllers within a robot. One of these subsystems monitors the information available from the various sensors and maintains a consistent set of beliefs. This approach permits the agent to take advantage of complex dynamics within its relationship to its environment, intervening with specific corrective actions only when these dynamics are not headed for desired states. It also affords a high degree of parallelism in the agent's execution, as well as considerable resilience in the face of unexpected perturbations.

A reformulation of traditional AI ideas within a control-theoretic vocabulary leads Schoppers to fresh perspectives on a variety of AI issues. His point of departure is the observation that it is impossible to guarantee any sort of iron-clad coupling between the agent's internal states and the world outside. Instead, the agent can rely upon a variety of factors to ensure that it remains adequately coupled to the world. These include physical inertia, which ensures that incorrect actions undertaken based on transiently mistaken perceptions or deductions about the world cannot do too much harm before they are corrected. They also include the structure of the space around the agent, with its strong locality effects, so that the agent will necessarily get a better look at any object that it is in a position to affect. The result is a distinctive system modularity that focuses on managing the agent's relationship to its environment rather than upon dictating a predetermined sequence of actions.

Schoppers applies his architecture to the control of a rescue robot operating in space. It would be valuable to apply Schoppers' framework to environments with more and different types of interactional regularities, such as those involving interaction with artifacts and real-time cooperative interaction with other agents. Future research could characterize in more detail the loose coupling between robot and environment that is recognized by Schoppers' approach.

4.14. Shoham and Tennenholtz

Shoham and Tennenholtz explore in mathematical terms the conditions under which large numbers of simple agents can be programmed to avoid colliding with one another. They observe that strategies for programming such agents can be arrayed along a continuum, from one extreme at which the programmers specify detailed paths for each individual agent, to another extreme at which the agents engage in negotiations of unbounded complexity. In the middle region between these extremes are a wide variety of possible "social laws" that might guide agents' actions. While the agents themselves might develop these social laws through systematic reasoning or incremental evolution, Shoham and Tennenholtz focus on the problem of off-line methods for designing these laws.

Their paper develops in two stages. In the first stage, they consider at length a particular case study, in which the agents attempt to avoid colliding while traveling in a grid. The challenge is to define a social law that permits the designers to prove mathematically that the agents will reach their goals without colliding. This is difficult when the designers have limited knowledge of the precise arrangement of the agents upon the grid. In the second part of their paper, Shoham and Tennenholtz sketch a general formalism for proving things about social laws. In particular, they explore the computational complexity of the automatic synthesis of provably correct social laws for large numbers of agents. Although this problem is unsurprisingly intractable in the general case, they specific various conditions under which it can be made tractable.

Shoham and Tennenholtz's paper occupies a distinctive place among the papers in this special double volume. It is the only paper to deal with large numbers of agents, and with the use of customs to provide reliable structure in agents' interactions with the world. Nonetheless, their paper fits comfortably with the others in the sense that their agents are embodied. Their bodies are surely primitive, but it does matter to the definition of the problem, and to the proofs of correctness, that the agents have locations, occupy space, and have limited perceptual and motor capabilities. The social laws that Shoham and Tennenholtz specify for the agents traveling on the grid require the agents to use the space in specific ways by moving about in relatively conventional patterns. In particular, their proofs require them to characterize these emergent patterns of movement in enough detail to demonstrate that they converge. Their paper is thus a simple example of the ways in which interactional customs can provide reliable structure. The agents need not be able to prove that their social laws are adequate; they need only follow those laws.

Future research along these lines might explore the ways in which agents can improvise their interactions with one another. It would probably be impractical to posit agents which invent completely innovative ways of interacting every time they encounter one another; customs, after all, have the important computational benefit of making these kinds of impossibly open-ended reasoning processes unnecessary. Yet customs do evolve with time, and agents do improvise their interactions in a variety of ways, from incremental optimizations to private deals (both formal and informal) among small numbers of agents who deal with one another regularly. Although human beings clearly engage in a great deal of this sort of thing, computational research should probably begin with simple cases and work upward.

4.15. Webber et al.

Webber and her colleagues describe a project to build an system that can animate the movements of a human figure as it follows instructions written in English. Instructions, Webber et al. point out, differ in numerous ways from computer programs, as well as from the symbolic structures that AI has long referred to as "plans". The interpretation of instructions appears to be conditioned by the situation in which the instructions are given. This context dependence of instructions is reflected as well in the linguistic forms commonly found in instructions, for instance in users' manuals for machines, and Webber et al. adduce numerous examples from the naturally occurring instructions that they have studied.

These insights have numerous implications for research on computational theories of interaction and agency. They illustrate one sense in which the "higher-level" functions of language use and symbolic reasoning must interact with the "lower-level" functions of sensorimotor interaction. In particular, they suggest that the conventional modularity that separates the interpretation of linguistic meaning from motor skills might have to be rethought. They also suggest the significant role of pragmatics—features of language that relate to the situation of language-use—in the situated interpretation of instructions. Finally, they force clear thinking about notions such as intentions and expectations that are central to cognitive theories of action.

As this ambitious project develops, it will no doubt encounter other features of language and thought that relate to agents' interactions with their environments. The expectations upon which people rely in interpreting instructions are cultural, in the sense that different cultures organize their concepts about action and interaction in different ways. Interactions between people can presuppose a wide and very subtle range of shared background understandings, for example when the participants in an interaction are members of the same profession, and thus possess a shared vocabulary and a shared experience of training, or members of the same family or circle of friends, and thus possess a shared background of references to things that have happened in the past. A difficult challenge is to understand the senses in which these phenomena are grounded in embodied activities.

4.16. Whitehead and Lin

Whitehead and Lin explore a number of algorithms for learning to engage in serially ordered behavior within the technical framework of reinforcement learning. Historically, of course, reinforcement learning has a close association with the behaviorist school of psychology that the founders of artificial intelligence sought to overcome. Whitehead and Lin are not nearly behaviorists, but their work is clearly part of an alternative tradition within AI. Whereas the main stream of AI research focused on complex cognitive processes internal to agents, other work retained a focus upon agents' interactions with their environments. Behaviorists took this focus to extremes, arguing that it was pointless or meaningless to posit internal cognitive processing. But their search for means of explaining behavior based on sequences of stimuli and responses, guided through the learning process by positive and negative reinforcement, counterbalances explanatory principles based wholly on internal processing.

The architectures that Whitehead and Lin explore do not maintain complete world models. To the contrary, they maintain very simple representations of the world that are grounded in sensorimotor experience. As a result, it becomes necessary for their architectures to actively interpret their available sensory input in functional terms. The agent must actively decide which available stimuli to pay attention to, and it must learn which of these stimuli is likely to permit the agent to accurately predict the degree of "payoff" which its actions will receive. A stimulus that has low predictive value in a particular situation is most likely capable of being generated by things in the world with differing functional significances, with the result that it does not provide information that allows the agent to choose correctly among possible actions. A stimulus that has high predictive value, on the other hand, is probably generated by those things in the world whose states are relevant to the agent's decisions. This deep idea connects the indexicality and active nature of perception with the practicalities of learning from limited information.

The most immediate difficulty with this proposal is that it requires the agent's actions to be functions of its immediately available inputs. Whitehead and Lin therefore extend their analysis to architectures that can maintain limited types of internal state. In contrast once again to architectures that assume that a complex internal model is kept up to date, the authors explore much simpler schemes in which the architecture itself synthesizes state elements which assist it in predicting payoffs. The result is a notion of internal representation tied to the meaningful aspects of the agent's interactions with the world.

This model obviously requires much further development before it can undertake more complex tasks. One aspect of this development might be a more extensive analysis of the structures in the world that permit the algorithms to work well or poorly. The synthesis of internal states tied to functionally significant properties of the agent's sensorimotor interactions is a powerful idea, and it might work best when dealing with artifacts whose functional states are meant to be readily distinguishable, or in environments which have been heavily marked with indications of their normal roles in customary forms of activity. In such settings, the synthesis of internal states might be channeled in comprehensible ways, corresponding not simply to the objective structure of the environment but to the structure of the agent's involvements in it.

5. Case study

An informal review of research that I conducted with Ian Horswill [6] will provide an instructive case study in the themes of this special double volume. This research explores one of the ways in which cultural artifacts support activity by simplifying computational tasks that would otherwise be extremely complex. Its ideas are embodied in a computer program called Toast that acts as a short-order cook, cooking a continual stream of breakfast dishes by interleaving the various actions. It does so without having to construct any symbolic plans, perform any search, or engage in any explicit reasoning about the future. It can do so because certain properties of the artifacts of cooking tend to reduce the computational complexity of decisions about what to do next—or at least to permit simple strategies such as "find something that needs doing and do it" to provably converge to certain kinds of goals. The point is not that all activity is like this, either inside or outside the kitchen, but to indicate some of the ways in which structures in the world can simplify computational problems.

5.1. Model of action

One place to begin the story is with the assumptions of the classical planning literature. This literature takes a definite stand on the nature of action. Although some authors have explored the consequences of relaxing or complicating one or more of these assumptions by certain increments (for example, by introducing probabilities or concurrency), this underlying model of action continues to anchor the literature by providing a set of default assumptions for new projects. The model begins with the idea of "actions" and "situations" as discrete entities, so that the effects of an action can be represented in terms of the transition from one clearly defined situation to another. The result, of course, is that the agent's actual and potential activities can be represented in terms of the possible routes through a directed graph whose vertices correspond to situations and whose arcs are labeled with the actions which can lead from one situation to the next.

A great deal of planning research is concerned, implicitly or explicitly, with the structure of this graph. This structure is affected by many things, most prominently the agent's repertoire of actions, the representation scheme employed to identify the possible actions and dissect the possible situations in the world, and the structure of the world itself. If an agent is going to take actions in the world by executing a plan, that plan must be guaranteed (at least probabilistically) to trace a path through the graph that arrives at a desired end-point from a given beginning-point—or, more precisely, from any beginning-point that is consistent with whatever knowledge the agent has about the beginning-point. Whether, and how efficiently, it should be possible to discover such a plan will depend on the structure of the state graph (large or small, high or low branching factor, clear landmarks, etc.) and on the ways in which the structure of the graph can be exploited in designing algorithms to search it.

Investigation of the computational properties of the state-space graph structure, though, is conceptually independent of the idea of a plan or the idea of activity as plan-execution. The upshot of our research is that the world includes structures that permit a great deal of action to be conducted through simple forms of improvisation without the necessity of explicit plan-construction. It is sometimes necessary to engage in symbolic reasoning about the future, of course, and to make representations of action to help guide future activities. But we would like to suggest that these more complex forms of reasoning about action are delimited and controlled to a substantial extent by the structures in the world that support simpler forms of moment-to-moment action choice.

Our domain, once again, is that of cooking breakfast in a short-order restaurant, and I wish to make clear our intentions in choosing this domain. Cooking is an attractive domain (cf. [32,47]) because it is fairly complicated but still routine, has fairly well-defined properties but regularly admits of uncertainty and surprise, and has plainly been organized in customary ways to allow action to be driven in large part by vision (cf. [4,8,48]). We do not claim to analyze all of the complexities of actual breakfast-cooking, of course (cf. [34]). Rather, we formalize the activities of cooking breakfast for purposes of our analysis using the formal methods of the classical planning literature. In employing these methods our purpose is not to endorse the assumptions that underlie them, but rather to demonstrate how the research process points beyond them. Finally, our goal is not to invent a sophisticated new architecture for making breakfast, but rather to discover structures within the domain that make the invention of sophisticated architectures unnecessary. The real work, in other words, is taking place at the designer's level, in the "aerial view", discovering regularities that can permit an agent operating at the "ground level" to get along with relatively simple policies.

To explore the structure of cooking world, we elaborate the traditional formal framework by using an object-centered representation of action. The objects in question are those found in cooking tasks, such as pots and pans, tools and utensils, and materials such as food ingredients. The agent's actions all pertain in some way to these objects: moving them, transforming them, mixing them, cleaning them, and so forth. The state of the world can be decomposed into the states of these objects and a small number of possible relationships among them. The states of an egg, for example, can include being intact, being broken, being beaten, and being cooked. A bowl can be filled, empty-and-dirty, and empty-and-clean.

These descriptions of states obviously fail to capture all of the properties that the objects could possibly have. The formalization of these actions is analogous to the model of actions employed by a classical planning program: each possible action has a set of preconditions and a set of effects. The difference is that these preconditions and effects must be expressed in terms of the properties and relationships of objects. The action of cleaning a given spoon, for example, has no preconditions at all, since it makes sense to clean a spoon regardless of what state it is in; the effect of this action is to move the spoon into the "clean" state. The action of beating an egg with a fork in a bowl has the preconditions that the fork be clean (if the set of states is more elaborate, of course, the fork can be in the state of being dirty-with-beaten-egg, so that the egg-stirring fork need not be cleaned after each episode of stirring), that the egg be broken, and that the broken egg be located in the bowl; its effects are that the egg moves into the "beaten" state, the fork moves into the "dirty" state, and the beaten egg remains in the bowl.

Much of the formalism, then, concerns the states of objects. In particular, the state of the world at any given moment will consist in large part of the states of all objects. As with any conventional formalism, it would be possible to generate a graph structure that contains all of the possible world-states and the actions that can be taken to move from one world-state to the next. If the kitchen contains a large number of objects, of course, this graph will be enormous because of the large number of actions that can be taken at any moment and the huge number of possible combinations of individual object-states.

The enormity of this graph obviously conceals a great deal of structure within it. This becomes evident if we represent the state-space graph in another, object-centered way. If we neglect for the moment the relationships among objects, we can view each object as having its own state graph. The structure of this graph will depend on what type of object it is, so that the graph for eggs has one structure, which might include states corresponding to "intact", "broken", "beaten", and "cooked"; and the graph for forks will have another structure, which might have the states "clean" and "dirty". Given such graphs for each type of object, the state space of the whole world can be understood as the cross-product of the state-space graphs for each individual object. In fact, the whole world's state-space graph is a subset of this much larger cross-product graph, since it only includes actions that can actually be taken with the objects that are present. A world without forks, for example, will include no state-transitions in which eggs are beaten.

This idea of decomposing state graphs by interpreting them as the products of graphs for individual objects has already been introduced by Harel [34,35], who refers to his notation for these graphs as statecharts. Simply representing planning problems within such a notation, of course, does not change their inherent complexity. If the problem of identifying a correct plan within a given graph is unsolvable or intractable in the search space corresponding to the product graph then it is equally unsolvable or intractable when the graph is drawn in a different way. The purpose of the object-centered state-graph formalism, then, is not simply to reveal the implicit structure of domains like making breakfast as they are already defined, but also to provide a language within which to express additional structures that might be discovered within them. Such additional structure might transform cooking breakfast from a computationally difficult domain into a much more straightforward one.

Additional structure can indeed be found in the domain by categorizing the state-graphs for the types of objects actually found in kitchens. Let us consider two major categories, which might be called tools and materials. Informally speaking, materials include items of food like eggs, cups of water, and pats of butter. Tools include things like forks and spatulas which are primarily used to do things to materials. Every tool has a distinguished state in which it is clean, dry, and ready to use. Materials tend to have original, raw states, and they tend to pass through a series of further states as things are done to them with tools. Tools, furthermore, can be cleaned at any time, regardless of what state they are in, without the necessity of invoking other objects that might be in inconvenient states themselves. If a sponge or brush is used to clean a tool, then it will always be available and in a suitable state. These two categories, tools and materials, cover a large proportion of the objects found in kitchens, and their properties are much more specific than the worst, most complex state graphs and actions that might be imagined in the abstract.

5.2. *Formalism*

Given these intuitions, let us outline a simple formalism for domains that involve objects and actions. Such a domain will have a set of *object types*. (The term "object" can be used instead when the context makes clear that one is speaking of an object type and not a particular concrete object.) Each object type has an associated state graph, which is a finite directed graph whose vertices are called *states* and whose arcs are called *operations*. Note that the "operation" is the arc itself, not a label on the arc. Each operation is thus unique and is not shared by different object types. The domain will also have a set of *action types*, each of which has an associated set of operations drawn from the graphs associated with the domain's object types. For example, the action type of beating an egg might have two operations, corresponding to the egg's transition from "broken" to "beaten" and the fork's transition from "clean" to "dirty".

Let us say that an action is *focused* if it consists of a single operation (that is, if it involves a single object). A state in a given object type's state graph is *free* if it can be reached from any other state in that graph using only focused operations. A *tool*, then, is an object with at least one free state in its state graph. Each tool will have a distinguished free state, its *normal* state. An example of a normal state is "clean".

Given a set of tool types, it becomes possible to define a *material.* The basic idea is that one uses clean tools to do things to materials. A *tool action* is an action involving some finite number of tools (most commonly one tool), and possibly also one object class which is not a tool. A *normal tool action* is a tool action in which the actions involving tools require that those tools originate in their normal states. A *material* is an object with an acyclic state graph which includes a particular, distinguished state, the *raw* state, from which any other state in the graph can be reached purely by means of normal tool actions. The material might have other operations in its state graph besides the ones included in normal tool actions.

A *cooking task* is a task which has these four properties:

- all of the objects are tools and materials,
- enough tools exist to perform each of the actions required by each type of material,
- every instance of material starts out in its raw state, and
- the goal is to move some of the materials, all of which are instances of different material types, into other particular states.

Informally, it is possible to solve a cooking problem by repeatedly applying a simple policy:

- Choose a material that has a goal state but is not yet in it.
- Determine its current state, look up in a table which state it must pass through next in order to reach its goal state, and then look up in another table a normal tool action that is capable of affecting this necessary state change.
- Inspect the list of tool types required by this action. If the world contains a tool in its normal state for every one of these tool types then employ these tools to execute the action, thus causing the material and all of the tools to potentially change their states.
- If there exists a tool type in the required action that does not correspond to any tool in the world which is currently in its normal state, then choose one of these problematic tool types.
- Choose a tool of this type. Determine its current state, look up in a table which state it must pass through next in order to reach its goal state, and then look up in another table a focused action that can effect this necessary state change. Then take that action.

It is easy to see why this simple policy works. Each action either moves a material toward its goal state or moves a tool toward its normal state. When every tool is in its normal state (if not before), it becomes possible to move a material toward its goal state. Since every material type's state graph is finite, it is possible to calculate the total number of state transitions that the materials mentioned in the goal must go through. Likewise, since every tool type's state graph is finite as well, it is possible to place an upper bound on the number of state transitions that the tools in the world must go through in order for an action upon a material to become possible. Since every action reduces one of these quantities, since the total distance of the materials from their goals necessarily decreases whenever all of the tools are in their normal states, and since no action ever increases the total distance of the materials from their goals, it follows that the materials with goal states will eventually reach them.

This argument obviously relies on a large number of simplifying assumptions. For example, relations among objects have not been taken into account and objects cannot be mixed together or split into pieces. As well, it has been assumed that no object can be committed to a purpose over a long period, thus temporarily making it incapable of being used for any other purpose. The major category of objects for which this latter condition holds are "containers" such as cups, plates, bowls, frying pans, and stove burners. This category also includes clamps and vises, though very few other kitchen implements. The major pitfall associated with containers is running out of them, and the key to avoiding this pitfall is simply to have enough of them at hand. If enough of them are not available then it will become necessary to engage in some type of scheduling. Potentially complex plan-construction thus has its place, but analysis of the world's structure can isolate this place to a relatively small corner of the total activity. (The intuition here is similar to that of the algorithms for efficient constraint satisfaction presented by Dechter and Pearl [18].) Other simplifications can likewise be remedied by a judicious combination of appeals to structure in the world and limited extentions to the architecture. My purpose here, though, is not to develop the formalism in enough detail to accommodate these possibilities—or even to thoroughly vindicate its usefulness. Instead, I wish to present them as an instance of the ideas in this special double volume. Referring back to the discussions in Section 2, let us consider these in turn.

- *Aerial and ground views.* The whole formalism of states, actions, tools, materials, and so forth is part of the designer's aerial view, not the agent's ground view. The agent can employ a simple policy that involves looking up certain information in tables, and the designer can prove that this policy will always lead to a correct outcome, if not necessarily an optimal one.
- *Structure in the world.* The domain of cooking breakfast was discovered to have some useful kinds of structure that could assist an agent in choosing actions in a simple way. This structure can be viewed as an abstraction hierarchy, with the actions on tools forming one layer of abstraction and the actions on materials forming another layer. The model can obviously be generalized to several layers of abstraction [44].
- *Located in the practices.* The "structure in the world" was not located in the objects (the tools and materials) all by themselves. Instead, it was located in the objects together with a customary set of practices for using them. It is conceivable that another culture might employ eggs and forks and spatulas in wholly different activities with different computational properties. The proofs here depended on these objects being used in the ways that are familiar from the simplest recipes in American kitchens.
- *Looking for structure.* The search for this structure was motivated by the great computational complexity of unconstrained plan-construction problems, and in particular by the enormous search spaces that planning methods face in most realistic domains. This structure compensates for the difficulty of searching huge spaces by ensuring that the necessary spaces are small, and indeed that subgoal interactions are so constrained that search becomes unnecessary.
- *Convergence.* The proof of correctness is precisely, in computer science terms, a

proof of convergence. It proceeds along the lines of classical program correctness proofs using progress functions that can be demonstrated to move continually toward the goal state of zero.

- *Cultural support.* The structure in the world is not a simple matter of physics but is located largely in artifacts such as tools. As Vygotsky suggested (see the account of Vygotsky's ideas above), the people who invented the artifacts of cooking effectively rendered concrete a type of knowledge for simplifying tasks without requiring everyone in future to understand this knowledge in any explicit way.

6. Conclusion

This introduction has sketched an emerging method of computational research on interaction and agency. It has placed this method in the context of a variety of other fields and it has illustrated them through summaries of the articles and a case study. The shape of future research in this area cannot be predicted in detail, this being the nature of research. The precedents offered by the papers in this double volume, though, do make clear that research on computational theories of interaction and agency provides a fertile territory for the cross-pollination of a wide variety of different fields, each with its own conception of interaction and its own models of agency. Changing the metaphor, perhaps the continuation of the trend will help to transform artificial intelligence from a self-contained discipline to a kind of interdisciplinary switchboard for the construction of principled characterizations of interaction between agents and their environments.

Acknowledgements

The electronic mail archive for this project contains over 4000 messages totaling nearly five megabytes (not including manuscripts). Clearly the editors ought to acknowledge the contributions of numerous individuals and organizations, and they are happy to do so. Danny Bobrow and Mike Brady supported the project over a long period. Approximately sixty referees wrote well over 300 pages of exceptionally useful comments on the manuscripts. The American Association for Artificial Intelligence and Philips Research Laboratories New York provided grants to support the Workshop on Computational Theories of Interaction and Agency at the University of Chicago in February 1993, in which the authors assembled to discuss early drafts of their papers. We wish to thank the University of Chicago Computer Science Department's AI Group for acting as our hosts for this workshop, and Tim Converse for coordinating the local arrangements. Finally, the authors themselves gracefully acceded to numerous requests for revisions and considerable delays in the mechanics of the editorial process. We hope that they and the reader will benefit from the result.

References

[1] P.E. Agre, The symbolic worldview: Reply to Vera and Simon, *Cogn. Sci.* **17** (1) (1993) 61-69.

[2] P.E. Agre, Interview with Allen Newell, *Artif. Intell.* **59** (1-2) (1993) 415-449.

[3] P.E. Agre, The soul gained and lost: Artificial intelligence as a philosophical project, *Stanford Humanities Review*, to appear.

[4] P.E. Agre and D. Chapman, Pengi: An implementation of a theory of activity, in: *Proceedings AAAI-87*, Seattle, WA (1987) 196-201.

[5] P.E. Agre and D. Chapman, What are plans for?, in: P. Maes, ed., *Designing Autonomous Agents: Theory and Practice from Biology to Engineering and Back* (MIT Press, Cambridge, MA, 1991).

[6] P.E. Agre and I. Horswill, Cultural support for improvisation, in: *Proceedings AAAI-92*, San Jose, CA (1992).

[7] J. Allen, J. Hendler and A. Tate, eds., *Readings in Planning* (Morgan Kaufmann, San Mateo, CA, 1990).

[8] D.H. Ballard, Animate vision, *Artif. Intell.* **48** (1) (1991) 57-86.

[9] R.D. Beer, *Intelligence as Adaptive Behavior: An Experiment in Computational Neuroethology* (Academic Press, Boston, MA, 1990).

[10] R.A. Brooks, Intelligence without representation, *Artif. Intell.* **47** (1-3) (1991) 139-160.

[11] P. Bourdieu, *Outline of a Theory of Practice*, translated by Richard Nice (Cambridge University Press, Cambridge, England, 1977). Originally published in French in 1972.

[12] R.J. Brachman and H.J. Levesque, The tractability of subsumption in frame-based description languages, in: *Proceedings AAAI-84*, Austin, TX (1984) 34-37.

[13] V. Braitenberg, *Vehicles: Experiments in Synthetic Psychology* (MIT Press, Cambridge, MA, 1984).

[14] D. Chapman, Planning for conjunctive goals, *Artif. Intell.* **32** (3) (1987) 333-377.

[15] D. Chapman, *Vision, Instruction, and Action* (MIT Press, Cambridge, MA, 1991).

[16] H.H. Clark and D. Wilkes-Gibbs, Referring as a collaborative process, *Cognition* **22** (1) (1986) 1-39.

[17] J.L. Comaroff and S. Roberts, *Rules and Processes: The Cultural Logic of Dispute in an African Context* (University of Chicago Press, Chicago, 1981).

[18] R. Dechter and J. Pearl, The anatomy of easy problems: a constraint-satisfaction formulation, in: *Proceedings IJCAI-85*, Los Angeles, CA (1985) 1066-1072.

[19] G.L. Drescher, *Made-Up Minds: A Constructivist Approach to Artificial Intelligence* (MIT Press, Cambridge, MA, 1991).

[20] H.L. Dreyfus, *What Computers Can't Do: A Critique of Artificial Reason* (Harper and Row, New York, 1972).

[21] Y. Engeström, *Learning by Expanding* (Orienta-Konsultit Oy, Helsinki, 1987).

[22] R.E. Fikes and N.J. Nilsson, STRIPS: a new approach to the application of theorem proving to problem solving, *Artif. Intell.* **2** (3) (1971) 189-208.

[23] R.E. Fikes, P.E. Hart and N.J. Nilsson, Learning and executing generalized robot plans, *Artif. Intell.* **3** (4) (1972) 251-288.

[24] R.E. Fikes, P.E. Hart and N.J. Nilsson, Some new directions in robot problem solving, in: B. Meltzer and D. Michie, eds., *Machine Intelligence* **7** (Wiley, New York, 1972).

[25] R.J. Firby, An investigation into reactive planning in complex domains, in: *Proceedings AAAI-87*, Seattle, WA (1987) 202-206.

[26] M.S. Fox and S. Smith, ISIS: A knowledge-based system for factory scheduling, *Expert Syst.* **1** (1) (1984) 25-49.

[27] M.R. Genesereth and N.J. Nilsson, *Logical Foundations of Artificial Intelligence* (Morgan Kaufmann, Los Altos, CA, 1987).

[28] M.P. Georgeff and A.L. Lansky, Reactive reasoning and planning, in: *Proceedings AAAI-87*, Seattle, WA (1987) 677-682.

[29] J. Goody, *The Logic of Writing and the Organization of Society* (Cambridge University Press, Cambridge, England, 1986).

[30] B.J. Grosz and C.L. Sidner, Plans for discourse, in: P.R. Cohen, J. Morgan and M.E. Pollack, *Intentions in Communication* (MIT Press, Cambridge, MA, 1988).

[31] N. Gupta and D. Nau, On the complexity of blocks-world planning, *Artif. Intell.* **56** (2) (1992) 223-254.

[32] K.J. Hammond, T. Converse and C. Martin, Integrating planning and acting in a case-based framework, in: *Proceedings AAAI-90*, Boston, MA (1990.

[33] S. Hanks and D. McDermott, Modeling a dynamic and uncertain world I: Symbolic and probabilistic reasoning about change, *Artif. Intell.* **66** (1) (1994) 1-55.

[34] C. Hardyment, *From Mangle to Microwave: The Mechanization of Household Work* (Polity Press, Oxford, England, 1988).

[35] D. Harel, Statecharts: A visual formalism for complex systems, *Sci. Comput. Program.* **8** (3) (1987) 231-274.

[36] D. Harel, On visual formalisms, *Commun. ACM* **31** (5) (1988) 514-530.

[37] H. Haste, Growing into rules, in: J. Bruner and H. Haste, eds., *Making Sense: The Child's Construction of the World* (Methuen, London, 1987).

[38] P.J. Hayes, In defense of logic, in: *Proceedings IJCAI-77*, Cambridge, MA (1977) 559-565.

[39] M. Heidegger, *Being and Time*, translated by J. Macquarrie and E. Robinson (Harper and Row, New York, 1961). Originally published in German in 1927.

[40] J. Hendler, ed., Planning in uncertain, unpredictable, or changing environments, proceedings of the AAAI symposium at Stanford, University of Maryland Systems Research Center Report SRC TR 90-45 (1990).

[41] G.E. Hinton and D.S. Touretzky, Symbols among the neurons: details of a connectionist inference architecture, in: *Proceedings IJCAI-85*, Los Angeles, CA (1985) 238-243.

[42] H.A. Kautz and E.P.D. Pednault, Planning and plan recognition, *AT & T Tech. J.* **67** (1) (1988) 25-41.

[43] D. Kirsh, Today the earwig, tomorrow man?, *Artif. Intell.* **47** (1-3) (1991) 161-184.

[44] C.A. Knoblock, Automatically generating abstractions for planning, *Artif. Intell.* **68** (2) (1994) 243-302.

[45] N. Kushmerick, S. Hanks and D.S. Weld, An algorithm for probabilistic least-commitment planning, in: *Proceedings AAAI-94*, Seattle, WA (1994).

[46] C.G. Langton, ed., *Artificial Life II: Proceedings of the Workshop on the Artificial Life*, Santa Fe, NM (1990).

[47] A.L. Lansky and D.S. Fogelsong, Localized representations and planning methods for parallel domains, in: *Proceedings AAAI-87*, Seattle, WA (1987) 240-245.

[48] J.H. Larkin, Display-based problem solving, in: D. Klahr and K. Kotovsky, eds., *Complex Information Processing: The Impact of Herbert A. Simon* (Erlbaum, Hillsdale, NJ, 1989).

[49] K.S. Lashley, The problem of serial order in behavior, in: L.A. Jeffress, ed., *Cerebral Mechanisms in Behavior: The Hixon Symposium* (Wiley, New York, 1951).

[50] B. Latour, Visualization and cognition: Thinking with eyes and hands, *Knowledge and Society: Studies in the Sociology of Culture Past and Present* **6** (1986) 1-40.

[51] J. Lave, *Cognition in Practice: Mind, Mathematics, and Culture in Everyday Life* (Cambridge University Press, Cambridge, England, 1988).

[52] D. Marr, *Vision* (Freeman, San Francisco, CA, 1982).

[53] M.T. Mason, Mechanics and planning of manipulator pushing operations, *Int. J. Rob. Res.* **5** (3) (1986) 53-71.

[54] H.R. Maturana and F.J. Varela, *The Tree of Knowledge: The Biological Roots of Human Understanding* (New Science Library, Boston, MA, 1987).

[55] D. McAllester and D. Rosenblitt, Systematic nonlinear planning, in: *Proceedings AAAI-91*, Anaheim, CA (1991) 634-639.

[56] M. Merleau-Ponty, *Phenomenology of Perception*, translated from the French by Colin Smith (Humanities Press, New York, 1962).

[57] G.A. Miller, E. Galanter and K.H. Pribram, *Plans and the Structure of Behavior* (Holt, New York, 1960).

[58] A. Newell, *Unified Theories of Cognition* (Harvard University Press, Cambridge, MA, 1990).

[59] A. Newell and H.A. Simon, GPS: A program that simulates human thought, in: E.A. Feigenbaum and J. Feldman, eds., *Computers and Thought* (McGraw-Hill, New York, 1963) 279-296.

[60] D. Newman, P. Griffin and M. Cole, *The Construction Zone: Working for Cognitive Change in School* (Cambridge University Press, Cambridge, England, 1989).

[61] S.B. Ortner, Theory in anthropology since the sixties, *Comparative Studies in Society and History* **26** (1) (1984) 126-166.

[62] D.W. Payton, J.K. Rosenblatt and D.M. Keirsey, Plan guided reaction, *IEEE Trans. Syst. Man Cybern.* **20** (6) (1990) 1370-1382.

[63] J. Piaget, *The Construction of Reality in the Child*, translated by Margaret Cook (Basic Books, New York, 1954).

[64] M.E. Pollack, The uses of plans, *Artif. Intell.* **57** (1) (1992) 43-68.

[65] Z.W. Pylyshyn, ed., *The Robot's Dilemma: The Frame Problem in Artificial Intelligence* (Ablex, Norwood, NJ, 1987).

[66] M.R. Quillian, Semantic memory, in: M. Minsky, ed., *Semantic Information Processing* (MIT Press, Cambridge, MA, 1968).

[67] M.H. Raibert, Running with symmetry, *Int. J. Rob. Res.* **5** (4) (1986) 3-19.

[68] S.J. Rosenschein and Leslie Pack Kaelbling, The synthesis of digital machines with provable epistemic properties, in: J. Halpern, ed., *Proceedings Conference on Theoretical Aspects of Reasoning About Knowledge*, Monterey, CA (1986).

[69] E.D. Sacerdoti, Planning in a hierarchy of abstraction spaces, *Artif. Intell.* **5** (2) (1974) 115-135.

[70] M. Schoppers, Universal plans for reactive robots in unpredictable environments, in: *Proceedings IJCAI-87*, Milan, Italy (1987) 1039-1046.

[71] H.A. Simon, *Administrative Behavior: A Study of Decision-Making Processes in Administrative Organization* (Macmillan, New York, 2nd ed., 1957).

[72] L.A. Suchman, *Plans and Situated Actions: The Problem of Human-Machine Communication* (Cambridge University Press, Cambridge, England, 1987).

[73] J.A. Toth, Review of Kenneth Ford and Patrick Hayes, eds., *Reasoning Agents in a Dynamic World: The Frame Problem Artif. Intell.* **73** (1995), to appear.

[74] F.J. Varela and P. Bourgine, eds., *Toward a Practice of Autonomous Systems: Proceedings of the First European Conference on Artificial Life* (MIT Press, Cambridge, MA, 1992).

[75] F.J. Varela, E. Thompson and E. Rosch, *The Embodied Mind: Cognitive Science and Human Experience* (MIT Press, Cambridge, MA, 1991).

[76] L.S. Vygotsky, *Mind in Society: The Development of Higher Psychological Processes*, M. Cole, V. John-Steiner, S. Scribner and E. Souberman, eds. (Harvard University Press, Cambridge, MA, 1978). Originally published in Russian in 1934.

[77] D.S. Weld, Reasoning about model accuracy, *Artif. Intell.* **56** (2) (1992) 255-300.

[78] D.E. Whitney, Historical perspective and state of the art in robot force control, *Int. J. Rob. Res.* **6** (1) (1987) 3-14.

[79] W.A. Woods, What's in a link?, in: D.G. Bobrow and A. Collins, eds., *Representation and Understanding: Studies in Cognitive Science* (New York, Academic Press, 1975).

[80] J. Yates, *Control through Communication: The Rise of System in American Management* (Johns Hopkins University Press, Baltimore, MD, 1989).