Benchmarks, Test Beds, Controlled Experimentation, and the Design of Agent Architectures

Steve Hanks, Martha E. Pollack, and Paul R. Cohen

■ The methodological underpinnings of AI are slowly changing. Benchmarks, test beds, and controlled experimentation are becoming more common. Although we are optimistic that this change can solidify the science of AI, we also recognize a set of difficult issues concerning the appropriate use of this methodology. We discuss these issues as they relate to research on agent design. We survey existing test beds for agents and argue for appropriate caution in their use. We end with a debate on the proper role of experimental methodology in the design and validation of planning agents.

In recent years, increasing numbers of AI research projects have involved *controlled experimentation*, in which a researcher varies the features of a system or the environment in which it is embedded and measures the effects of these variations on aspects of system performance. At the same time, two research tools have gained currency: *benchmarks*, precisely defined, standardized tasks, and *test beds*, challenging environments in which AI programs can be studied. In our view, the move toward more principled experimental methods is uncontroversially a good thing; indeed, we are optimistic that it will solidify the science of AI. However, we also recognize some issues concerning the appropriate use of these methods. First, benchmarks and test beds no more guarantee important results than, say, microscopes and Bunsen burners. They are simply part of the apparatus of empirical AI. It is up to the researcher to discriminate between uninteresting and important phenomena and to follow up reports of experiments with thorough explanations of their results. Second, there is little agreement about what a representative benchmark or test-bed problem is. A third and related concern is that results obtained with benchmarks and test beds are often not general. Fourth, because benchmarks and test beds are attractive to program managers and others who provide funding, there is a real danger that researchers will aim for the prescribed benchmark target when funding is perceived to be the reward. In sum, we are concerned that benchmarks and test beds, if not carefully used, will provide only a comfortable illusion of scientific progress—controlled experimentation with reproducible problems and environments and objective performance measures—but no generalizable, significant results.

Benchmarks and test beds serve at least two different purposes. One is to provide metrics for comparing competing systems. Comparison metrics are valuable for some purposes, but performance comparisons do not constitute scientific progress unless they suggest or provide evidence for explanatory theories of performance differences. The scientific value of well-crafted benchmarks and test beds is their power to highlight interesting aspects of system performance, but this value is realized only if the researcher can adequately explain why his or her system behaves the way it does.

A benchmark is illuminating to the degree that it tells us something we want to know about the behavior of a program.

The experimental control that can be achieved with test beds can help us explain why systems behave as they do. AI systems are intended to be deployed in large, extremely complex environments, and test beds serve as simplified, simulated versions of these environments, in which the experimenter has access to particular aspects of the environment, and other aspects are allowed to vary randomly. The experimental process consists in the researcher varying the features of the test-bed environment, the benchmark task, or the embedded system and measuring the resulting effects on system performance. A fundamental question exists, however, about the viability of this approach. The concern grows out of the tension between realism and the possibility of experimental control. On the one hand, controlled experiments seem, at least currently, to be feasible only for simplified systems operating in highly idealized environments. On the other hand, our ultimate interest is not simplified systems and environments but, rather, real-world systems deployed in complex environments. It is not always obvious whether the lessons learned from the simplified systems are generally applicable, but neither is it obvious how to perform systematic experiments without the simplifications.

Researchers disagree about how best to proceed in light of this tension. One approach is to maintain systematicity in experiments and look for ways to translate the results of the experiments into general principles that apply to more complex systems and environments. The alternative is to focus on more realistic systems and environments and to try to conduct systematic experiments on them directly. Much of this article focuses on a comparison of these approaches.

Although benchmarks, test beds, and controlled experimentation are increasingly important in a number of subareas of AI, including natural language understanding and machine learning, we focus our discussion on its role in agent design. We begin, in Benchmarks and Test Beds, by describing some of the criteria for good benchmarks and test beds and discussing some of the potential difficulties encountered in their design. In Current Issues in Agent Design, we discuss the range of features that a test bed for agent design might have. In Test-Bed Implementations, we survey existing test beds for agent design with these features in mind. Finally, in Discussion, we return to the general issue of experimental methodology in agent design and discuss some unresolved questions concerning its use. Our points will become increasingly more controversial as the article proceeds, and indeed, by the end of the article, we will no longer speak with one voice.

Benchmarks and Test Beds

Benchmarks are a common tool in computer science. In the design of central processing units (CPUs), for example, matrix multiplication is a good benchmark task because it is representative of an important class of numeric processing problems, which, in turn, is representative of a wider class of computational problems—those that do not involve significant amounts of input-output. The matrix multiplication problem can be described precisely and rigorously. Moreover, matrix multiplication is illuminating: It tells the CPU designer something interesting about CPU, namely, its processing speed. In other words, if we are interested in processing speed as a measure of performance, then matrix multiplication is a good benchmark: Good performance on matrix multiplication problems predicts good performance on the large class of numeric tasks for which the processor is being designed.

An early benchmark task for AI planning programs was the Sussman anomaly (the three-block problem) (Sussman 1975). The Sussman anomaly helped many researchers elucidate how their planners worked. It was popular because like matrix multiplication, it was representative of an important class of problems, those involving interactions among conjunctive subgoals, and it was easy to describe.

A benchmark is illuminating to the degree that it tells us something we want to know about the behavior of a program. Our goals as scientists, engineers, and consumers dictate what we want to know. Sometimes we are most interested in the system's raw performance. In buying a workstation, we might be impressed with the rate at which a particular machine performs matrix multiplication. Likewise, as the potential user of an AI search algorithm, we might be impressed with the performance of the min-conflicts heuristic algorithm on the million-queens problem (Minton et al. 1990). As scientists and engineers, however, our interests are different. In these roles, we want to understand why a system behaves the way it does. What is it about the Cray architecture that allows high-performance matrix multiplication? Why does the min-conflicts heuristic algorithm solve increasingly difficult *n*-queens problems in roughly constant time?

Understanding a system's behavior on a benchmark task requires a model of the task, so our goals as scientists and engineers will often be served only by benchmark tasks that we understand well enough to model precisely, especially for cases in which we expect a program to pass the benchmark test. Without a model of the task, it is difficult to see what has been accomplished: We risk finding ourselves in the position of knowing simply that our system produced the successful behavior—passing the benchmark.

Models are also important when we design benchmarks to be failed, but in this case, we need a model of the factors that make the benchmark difficult. For example, we learn more about articulation by asking a human to say "black back brake block" repeatedly than we do from having the person say the equally unpronounceable sentence "alckb bcak raebk lbcko." Both sentences are extremely difficult, but the former is more illuminating because we have models of phonetics that explain why it is difficult. Experiments can tell us which design choices lead to good performance on benchmark tasks, but we need good models of these tasks to explain why it is so. However, building a good model tends to require a simple problem, and there is always the danger that a simple problem will not be especially illuminating.

Benchmarks ideally are problems that are both amenable to precise analysis and representative of a more complex and sophisticated reality. Unfortunately, the current state of the field often elevates these problems to a new status: They become interesting for their own sake rather than for their help in understanding a system's behavior on larger, more interesting tasks. Cohen's (1991) survey of papers from the 1990 National Conference on Artificial Intelligence found that 63 percent of the papers focused on benchmark problems such as *n* queens, the Yale shooting problem, and Sussman's anomaly. However, few of these papers made explicit the connection between the benchmark problems and any other task. Without this additional analysis, it is difficult to say whether these problems are representative of others we presumably care about and, therefore, exactly why the reported solutions are themselves interesting.

As AI begins to focus less on component technologies and more on complete, integrated systems, these traditional benchmarks might reveal their limitations. For example, although we might use *n* queens to test the capability and speed of a constraint-satisfaction algorithm embedded in, say, a factory scheduler, this benchmark will not tell us whether the quality of a schedule is appropriate given time constraints and other goals of the program. However, it is far from obvious that any benchmark can be devised for such a case. Benchmarks are problems that everyone can try to solve with his/her own system, so the definition of a benchmark cannot depend on any system-specific details, nor can the scoring criteria. What a researcher learns about a system from performance on a benchmark is liable to be inversely proportional to the size, complexity, and specificity of the system.

Thus, the conscientious researcher, intent on evaluating a system, faces an uncomfortable choice. The behaviors of the system's components can be evaluated individually on benchmark tasks, or the system's behaviors—not necessarily those of individual components—can be evaluated by task-specific criteria. On the one hand, the researcher learns, say, that the embedded constraint-satisfaction algorithm is extremely slow and won't scale up; on the other, he/she learns that the system nonetheless produces robust, timely schedules for the particular job shop modeled. Neither result is likely to evoke interest outside the researcher's own laboratory. Why should the rest of us care that an inefficient algorithm suffices to solve an applied problem that doesn't concern us? The difficulty is that as our attention turns to integrated programs, benchmark scores for component processes might be at variance with or poorly predict task-specific measures.

Benchmarks ideally are problems that are both amenable to precise analysis and representative of a more complex and sophisticated reality.

The potential mismatch between benchmark scores and performance on real tasks is also a concern for researchers who are developing test beds. Although some test beds are no more than an interface to specify parameters of a benchmark problem and instrumentation to measure performance, those described in this article provide rich environments that present a wide range of challenges to planners and related AI programs. You can design a lot of tasks for your planning system in TILEWORLD, PHOENIX, and the other test beds discussed here. You can study a lot of phenomena—real-time satisficing, graceful degradation under resource restrictions, path planning and navigation, sensor fusion, various kinds of learning, and so on. However, each of these general behaviors will be implemented in a particular way depending on the specific test bed and system being developed. Graceful degradation in a simplified TILEWORLD agent might have little

Benchmarks and test beds do not currently bridge the gap between general and specific problems and solutions.

in common with what we call graceful degradation in a complex system deployed to perform a real task, just as aggressive behavior in seagulls has little in common with aggressive behavior in teenage boys. McDermott's (1981) wishful mnemonic problem has not gone away: Two test-bed researchers might each claim to have achieved graceful degradation under resource restrictions, but it is more accurate to say that each has achieved something that he or she calls graceful degradation. Test beds make it easier to build programs that exhibit diverse behaviors, but researchers have to face the problem of understanding what like-named behaviors have in common.

Benchmarks and test beds do not currently bridge the gap between general and specific problems and solutions. A gap exists between the benchmark *n*-queens problem and another, domain-specific problem that you care about. A gap exists between the test-bed problem of having too few bulldozers to fight fires in the PHOENIX simulation and a general resource-limited planning problem. Those of us who build and work with test beds appreciate the opportunities they provide to study many phenomena, but we also recognize the difficulties involved in finding test-bed–specific problems that satisfy the criteria of benchmarks: They are simultaneously representative of larger, more interesting problems; easy to describe; and illuminating.

Current Issues in Agent Design

Despite the difficulties in designing test beds and perhaps because of the promise associated with test-bed–based experimentation, a number of test-bed systems for studying agent design have been developed to date. In Test-Bed Implementations, we survey some of them. This section motivates the survey by describing some significant research issues in agent design and noting corresponding features that test beds should exhibit. Much current research in agent design builds on the classical planning paradigm that characterized the field for several years, so our section begins with a short explanation of this paradigm.

The *classical planning paradigm* assumes an environment that is both controlled and simple. The planning agent is generally assumed to have complete control over the environment, which means that its intended actions are the only events that can change the world's state and, furthermore, that the effects of its actions are fully known, both to the agent and to the system designer. The agent is usually assumed to possess complete and error-free information about the state of the world when it begins planning. Because it knows what the initial state of the world is, what actions it intends to carry out, and what the effects of those actions will be, it can, at least in principle, predict exactly what the state of the world will be when it finishes acting. In other words, it knows ahead of time whether a particular plan will or will not achieve its goal.

Classical planners embody strong simplifying assumptions both in the sense that their capabilities (the class of problems they can solve) tend to be limited and in the sense that the worlds in which they operate tend to be small, exhibiting few features and a limited physics. Planners are generally tested in domains with few planning operators, on goals with few conjuncts, and on models of the world in which few features are explicitly modeled. Performance tends to degrade when the number of operators, goal conjuncts, or environmental features increases. Just as *control* means that the planner can, in principle, prove that its plan will work, the *simplifying assumptions* mean that the planner can as a practical matter generate the proof. Control and simplifying assumptions, therefore, allow the planner the luxury of generating provably correct plans prior to execution time.

Most current work on agent architectures aims toward relaxing these assumptions. Reactive systems, for example, deal with the problem that the world can change unpredictably between plan time and execution time by deciding what to do at execution time instead of generating a plan prior to execution. Case-based planners confront the simplicity problem by storing only the essential details of a solution, allowing the planner to concentrate on the relevant features of a new problem.

Next we describe some specific issues that have recently attracted the attention of planning researchers and, therefore, guide decisions about what features a planning test bed might exhibit.

Exogenous events: Perhaps the most limiting assumption of the classical planning worlds (most notably, the blocks world) is that no exogenous, or unplanned, events can occur. Relaxing this assumption makes the process of predicting the effects of plans more difficult (Hanks 1990b) and also introduces the need to react to unplanned events as they occur at execution time (Agre and Chapman 1987; Firby 1989). The time cost of planning becomes important in a world that allows

unplanned changes: The longer the agent takes to plan, the more likely it is that the world has changed significantly between the time the plan was generated and the time it is executed (Bratman, Israel, and Pollack 1988; Russell and Wefald 1991; Dean and Boddy 1988).

Complexity of the world: A realistic world has many features. Even a simple block has color, mass, texture, smell, and so on, although many of these features will be irrelevant to many tasks. A realistic world also has a complex causal structure: Changes in one aspect of the world can change many other aspects, even though most of those changes might again be irrelevant to any particular problem. Reasoning about more realistic models of the world requires the ability to represent and make predictions about complex mechanisms (Weld and deKleer 1989) as well as the ability to recognize and focus attention on those aspects of the world relevant to the problem at hand (Hanks 1990a). A test bed for exploring realistically complex planning problems should itself provide a complexity and diversity of features.

Quality and cost of sensing and effecting: Sensing and effecting, generally ignored by the classical planners, are neither perfect nor cost free. An agent must therefore incorporate incorrect and noisy sensor reports into its predictive model of the world (Hanks and McDermott 1994) and must plan sensing actions to improve its state of information, taking into account both the benefit of the information and the cost of acquiring it (Chrisman and Simmons 1991). Thus, a test bed for studying agent design might be populated with agents having imperfect sensors and effectors. The test bed needs to make a clean distinction between the agent and the simulated world, the agent's sensing and effecting capabilities defining the interface.

Measures of plan quality: Classical planners are provided with a goal state to achieve, and they stop when their plans can achieve this state. However, simple achievement of a goal state is an inadequate measure of success; it does not take into account the cost of achieving the goal, and it also does not admit the possibility of partial goal satisfaction. Haddawy and Hanks (1993) and Wellman and Doyle (1991) explore the relationship between goal expressions and utility functions. A test bed for exploring richer notions of success and failure should allow the designer to pose problems involving partial satisfaction of desired states, forcing the planner to trade the benefits of achieving the goal against the cost of achieving it. The problem of balancing cost against solution quality becomes more difficult when the agent is actually planning for a sequence of problems over time, some of which might not even have been made explicit when it begins to plan.

Multiple agents: Allowing multiple agents to act in the world introduces new problems: how behaviors are coordinated, how the agents should communicate, how the effects of simultaneous actions differ from the effects of those actions performed serially. Multiple-agent planning is an active research area (Bond and Gasser 1988), and a test bed for exploring these research issues must allow coordinated behavior and communication among the agents that inhabit it.

Allowing multiple agents to act in the world introduces new problems.

In addition to the functions required to make the test bed challenging, we also identify some design issues that tend to make a test bed more useful to prospective users:

A clean interface: It is important to maintain a clear distinction between the agent and the world in which the agent is operating. The natural separation is through the agent's sensors and effectors, so the interface should be clean, well defined, and well documented. A designer must be able to determine easily what actions are available to the agent, how the actions are executed by the test bed, and how information about the world is communicated back to the agent.

A well-defined model of time: Test beds must present a reasonable model of passing time to simulate exogenous events and simultaneous action and to define clearly the time cost of reasoning and acting. (This problem is a general one in simulation and modeling. See Law and Kelton [1981], for example.) However, the test bed must somehow be able to communicate how much simulated time has elapsed. Making sense of experimental results requires a way to reconcile the test bed's measure of time with that used by the agent.

Supporting experimentation: Testing an agent architecture amounts to assessing its performance over a variety of sample problems and conditions. Controlled experiments require that problems and environmental conditions be varied in a controlled fashion. A test bed should therefore provide a convenient way for the experimenter to vary the behavior of the worlds in which the agent is to be tested. The experimenter must also be able to monitor the agent's behavior in the test-bed world (Langley and Drummond 1990). Although it is far from clear at this

point what statistics should be used in such an assessment, the test bed must allow performance statistics to be gathered. It is also useful for the data to be formatted automatically for analysis using statistical software packages.

Test-Bed Implementations

Previous sections provided the motivations for simulated test-bed worlds and discussed some of the problems that might be explored in them. This section surveys several of the simulated worlds available to the community. Our survey is not exhaustive, nor is our selection of test beds meant to imply that they are the best available. For each test bed, we describe the sort of world the test bed is supposed to simulate and the research problems it was designed to test, we discuss the interface between the agent and the world and that between the researcher and the system (agent plus world), and we summarize the main methodological commitments associated with the test bed.

Grid Worlds

Several test-bed worlds have been organized around the theme that the agent is situated in a rectangular two-dimensional grid, and its main task is to push tiles around the grid. We first discuss the TILEWORLD of Pollack and Ringuette (1990), then the independently developed NASA (National Aeronautics and Space Administration) TILEWORLD (NTW) (Philips and Bresina 1991) and the MICE simulator (Montgomery et al. 1992).

Pollack and Ringuette (1990) report on the TILEWORLD test bed, a system designed to support controlled experiments with agent architectures situated in dynamic and unpredictable environments. The world consists of a rectangular grid on which can be placed the agent, some tiles, some obstacles, and some holes. Each object occupies one cell of the grid. The agent can move up, down, left, and right unless doing so would cause it to run into the world's boundaries or an obstacle. When a tile is in a cell adjacent to the agent, the agent can push the tile by moving in its direction. The agent's goal is to fill holes with tiles. Each hole has a capacity *C* and a score *S*. When the agent pushes *C* tiles into a hole, the hole disappears, and the trial's score increases by *S*. Each trial has a time limit, and the agent's performance is measured by the trial's score at its completion.[1]

The TILEWORLD environment includes exogenous events: Objects in the world can appear and disappear during a simulation. The experimenter can control the rate at which these objects appear and disappear as well as certain characteristics (capacity and score) of the newly created objects. The ability to control these parameters is an important feature of TILEWORLD because it allows systematic exploration of worlds with various characteristics (for example, worlds that change relatively quickly or slowly). The goal of such exploration is to find systematic relationships between world characteristics and corresponding characteristics of the embedded agent. The TILEWORLD system is distributed with a basic agent design, which is also parameterized to allow manipulation by the experimenter (see the following discussion).

The interface between the agent and the world allows the agent to take one of four primitive actions at any time: move left, move right, move up, and move down. Some or all of the primitive actions might be infeasible at a given time, for example, if an obstacle is blocking the way. The effects of each action are predetermined and deterministic: The agent always moves to the appropriate adjacent cell if it chooses to do so and if the move is feasible. It never ends up in a different cell by accident. Tiles and obstacles are characterized by their types and their location on the grid. Each takes up exactly one cell. Holes, which can occupy one or more cells, are characterized by location, capacity, and score.

Holes, obstacles, and tiles appear and disappear probabilistically, according to parameter settings established by the researcher prior to any trial. The probabilities are independent of one another; a single probability governs the appearance of tiles, and it is the same regardless of the time, the location, or any other parameter in the game.

TILEWORLD has no explicit sensing operators. The agent is provided with a data structure that describes the world's state in complete detail and with complete accuracy. The use of this information is left to the designer of the embedded agent; for example, he or she can design mechanisms that distort the information to introduce inaccuracy.

The researcher describes a world by specifying the size of the grid; the duration of the game; and the probability parameters governing the appearance and disappearance rates of tiles, obstacles, and holes and the distribution of hole scores and capacities. The experimenter can control additional environmental characteristics; for example, the experimenter can decide whether hole scores remain constant until the hole disappears or whether the

score decreases over time. To facilitate experimentation, the system provides mechanisms for specifying suites of experiments, which can then be run without intervention, and recording performance data.

Three related qualities characterize TILEWORLD: its abstract nature, its simplicity, and its parameterized nature. TILEWORLD is not an attempt to model any particular planning domain; instead, the world might be used to pose paradigmatic planning problems in the abstract. It is a simple world that presents the agent with only a few possibilities for action; objects have few attributes, and the occurrence and effects of exogenous events are not complex. The world's simplicity means that a few parameters define a world instance completely, and these parameters can be varied as experiments are performed.

TILEWORLD was originally developed to investigate a particular agent architecture, IRMA (intelligent resource-limited machine architecture) (Bratman, Israel, and Pollack 1988), and, in fact, is distributed to the research community with an embedded IRMA agent. IRMA actually specifies a space of agent architectures; in other words, there is a range of agent architectures within the IRMA framework. The embedded TILEWORLD agent is parameterized to allow exploration of the design choices consistent with the IRMA specifications.

The interface between a TILEWORLD agent and its environment works as follows: When the agent wants to perform some action, it calls the simulator as a subroutine, specifying the action it wants to perform along with an indication of the amount of time that has elapsed since its last call (representing the amount of time it spent reasoning about what to do). The simulator then updates the world, both to reflect exogenous events that took place during that period and to reflect the agent's new actions. The resulting world is then passed back to the agent (in a data structure called the world).

This approach to agent-environment interface places the responsibility for specifying sensing and effecting conditions on the agent designer. If the agent uses the world data structure directly, it will always have a complete and correct model. Incomplete or noisy sensing can be achieved by manipulating this data structure before the agent is allowed to use it. Similarly, imprecision in effecting change has to be specified within the agent itself.

NTW (Philips and Bresina 1991; Philips et al. 1991) is an independently developed test bed that is also organized around the theme of a two-dimensional grid with tiles. Exogenous events in NTW consist of winds that can blow tiles across the grid. NTW has no obstacles or holes.

Two features distinguish the two simulators. First, the NTW simulator has no built-in measure of success that is analogous to the notion of a score. What the agent is supposed to do and what constitutes success is left entirely to the experimenter. The second is the nature of the interface between the agent and its environment. The TILEWORLD agent calls the simulator as a subroutine and passes information back and forth using a shared data structure. The NTW agent and the world simulator run asynchronously: The agent posts commands to the world, which are put in a queue and eventually executed. Operators can be programmed to fail probabilistically: A grasp operation might not result in the agent holding the tile, and a move might result in the agent being displaced to an adjacent location other than the one intended. The agent is given no indication of whether an operator has succeeded or failed and must explicitly sense the world to ascertain the effects of its actions.

MICE (Montgomery and Durfee 1990; Montgomery et al. 1992) is another grid-oriented simulator, designed to support research into coordinating the problem-solving behavior of multiple autonomous agents. The basic layout of MICE consists only of a grid and various agents, although agents can be used to simulate objects, such as tiles and forest fires.

The basic MICE operator is the move command, moving the agent from one grid cell to an adjacent cell. The link command is an abstract version of a grasp operator; the agent uses it to pick up objects. The world is populated only with agents, but they can be diverse. MICE has no explicit provision for exogenous events, although they can be simulated to some extent by implementing agents that have the desired effects on the world (making a grid cell wet and slippery to simulate rain, for example).

The main difference between the MICE simulator and the NTW and TILEWORLD simulators is that MICE makes even less of a commitment to a world physics; the experimenter defines an agent's sensing and effecting capabilities and also the effect of actions taken simultaneously by the agents. MICE might be viewed more as a framework for building test beds rather than a simulator in and of itself. (The MICE designers have built versions of TILEWORLD and PHOENIX

using this platform. See Montgomery and Durfee [1990], for example.)

The PHOENIX Test Bed

PHOENIX (Hart and Cohen 1990; Greenberg and Westbrook 1990) is a framework for implementing and testing multiple autonomous agents in a complex environment. The scenario is fire fighting; the world consists of a map with varying terrain, elevations, and weather. Fires can start at any location and spread depending on the surrounding terrain. Agents are fire-fighting units (commonly bulldozers) that change the terrain to control the fires.

It is helpful to distinguish the PHOENIX simulator from the PHOENIX environment and PHOENIX agents. The simulator has three main functions: (1) to maintain and update the map; (2) to synchronize the activities of the environment and the agents, which are implemented as independent tasks; and (3) to gather data. The PHOENIX environment includes a representation of Yellowstone National Park (from Defense Mapping Agency data) and the tasks that implement fires. PHOENIX agents generate tasks that simulate a fire boss, several bulldozers, watchtowers, helicopters, fuel tankers, and so on. Agent tasks include moving across the map, cutting a fire line, predicting the course of fires, planning the attack on the fire by several bulldozers, monitoring progress and detecting failures in expectations, and recovering from failure. Tasks insert themselves (by sending messages) onto a timeline maintained by the PHOENIX simulation. Tasks run intermittently and sometimes periodically.

PHOENIX agents sense and change the PHOENIX environment by sending messages to the object managing the map, but the simulator makes no attempt to control the form of the messages. Thus, PHOENIX agents have no predetermined set of operators. The PHOENIX environment contains only two kinds of objects: agents and fires. However, each cell of the map of the environment contains information that agents and fires use to determine their behavior. For example, bulldozers travel quicker on cells that are designated *blacktop road*, and fires burn faster in the direction designated *uphill*. Exogenous events are also implemented as tasks and influence other tasks indirectly. For example, wind causes fires to burn faster.

Tasks make their effects known by sending messages to the simulator. The form of these messages is not restricted; any task can, in principle, find out anything about the world and effect any change. The simulator enforces no model of sensing. It provides information about the world (the characteristics of a cell in the map, for example) by responding to messages but does not restrict its answers. However, the PHOENIX agents have limited sensory and physical abilities; for example, bulldozers have a 200-meter radius of view (although the view is not affected by elevation), and they move and cut fire lines at rates codified by the U.S. Forestry Service.

Defining an environment consists of defining a map—the topographic features for a land area, including ground cover, elevation, roads, rivers, and buildings—and processes within the environment, such as fires and wind. Defining an agent is generally more complicated because it involves designing sensors, effectors, a planner, a reactive component, internal maps of the environment, and so on.

PHOENIX includes an experiment-running facility that includes a language for specifying scripts for changes in weather, fires starting, and other events. It also allows for agents' behavior to be monitored, producing data files that can be read by data-manipulation and statistical packages. The design of the PHOENIX system is modular, and other test beds have been developed rapidly by swapping out the Yellowstone map and the PHOENIX agent definitions and swapping in, for example, a world of shipping lanes, ports, docks, ships, and roads.

PHOENIX differs from the previous simulators in that it tries to provide a realistic simulation of a single domain rather than implement an abstract domain-independent task environment. Apart from this difference, however, it is similar to the MICE simulator in that it enforces few constraints on how agents and exogenous events can sense or change the world. The simulator maintains the map and schedules activities, but, like MICE, much of the domain's physics lies in definitions of the individual tasks.

TRUCKWORLD

TRUCKWORLD (Firby and Hanks 1987; Nguyen, Hanks, and Thomas 1993) is a multiagent test bed designed to test theories of reactive execution (Firby 1989) and provide motivating examples for a theory of reasoning about dynamic and uncertain worlds (Hanks 1993; Hanks and McDermott 1994). The main commitment is to provide a realistic world for its agents but without physical sensors or effectors.[2]

An agent is a truck consisting of two arms; two cargo bays; several sensors; and various other components, such as a fuel tank, a set of tires, and direction and speed controllers. It operates in a world consisting of roads and locations. Roads connect the locations, which are populated with objects. The simulator itself places few restrictions on the behavior of objects, which can be complex. TRUCKWORLD can model objects such as fuel drums, which the truck can use to increase its fuel level; tire chains, which help it drive safely down slippery roads; vending machines, which require money and produce a product; and bombs, which tend to break unprotected objects in their immediate vicinity.

Exogenous events such as rainstorms occur periodically in the world. A rainstorm makes all roads in its vicinity wet, and dirt roads become muddy for a while. The truck runs the risk of getting stuck in the mud if it travels on a muddy road without proper tires. Objects in the vicinity of a rainstorm get wet, too, which might affect their behavior (a match might not ignite anymore, a plant might start growing). The occurrence of events can depend both on random chance and on characteristics of the world (rainstorms might be more likely at certain locations or at certain times of day).

TRUCKWORLD provides a wide variety of (simulated) sensors: Cameras report visual features of objects, sonars report whether there is an object at a location, scales report an object's weight, and X-ray machines report on objects within a closed container. Sensors typically have noise parameters: A camera sometimes reports an incorrect but close color for an object, and such a report is more likely at night than during the day. A scale reports the object's true weight distorted according to a user-supplied noise distribution; a sonar occasionally incorrectly reports that an object is present.

A variety of communication devices are available: Radios allow connection among agents; loudspeakers produce sounds that can be detected by microphones in the vicinity. Motion detectors notice when objects appear or disappear from their immediate vicinity. Tape recorders are activated when a sound is produced, and an agent can retrieve the recorded message later.

Communication between an agent and the simulator is tightly controlled: Each agent and the simulator itself run as separate processes, communicating over two channels. The agent performs actions and gets sensor reports over the command channel and uses the control channel to manipulate the simulator's internal state (for example, to connect or disconnect from the simulator, to advance the simulator's clock, or to collect statistics about the world). Multiple agents communicate using only the communication devices the world provides for them.

There were two main goals in designing TRUCKWORLD: (1) to provide a test bed that generates interesting problems both in deliberative and in reactive reasoning without committing to a particular problem domain and (2) to provide significant constraints on the agent's effecting and sensing capabilities and on the causal structure of the world but still allow the system to be extended to meet the designer's needs.

TRUCKWORLD occupies a position between simple abstract simulators such as TILEWORLD and NTW, a domain-specific simulator such as PHOENIX, and a test-bed–building platform such as MICE. TRUCKWORLD implements a specific set of operators for the agent (unlike MICE) but provides fewer constraints than do TILEWORLD or PHOENIX on the nature of the other objects in the world and on the interaction between the agent and these objects.

Summary

We looked at five systems for implementing planning test beds: the parameterizable TILEWORLD and NTW, the multiagent MICE platform, the PHOENIX fire-fighting simulation, and the TRUCKWORLD simulator. Although there are many differences in what features each system offers and what design decisions each makes, we can identify three main areas in which the systems differ:

Domain dependence: PHOENIX strives for a realistic depiction of a single domain, and TILEWORLD, NTW, and MICE try to describe abstract worlds and operators that affect the world. There is an obvious trade-off in this decision: A researcher using a domain-dependent simulator might be able to demonstrate that a program is an effective problem solver in the domain but might have difficulty going on to conclude that the architecture is effective for dealing with other domains. A researcher using an abstract simulator might be able to build a system based on what he or she judges to be general problem-solving principles, but then the difficulty is in establishing that these principles apply to any realistic domain.

Definition of sensors and effectors: The question arises about whether or to what extent the simulator should define the agent's sensing and effecting capabilities. At

one extreme, we have the PHOENIX simulation, which does not itself impose any constraints on environment dynamics or the information agents can find out about their environment. All such constraints are specified in the agent definitions and are merely enforced by the simulator. MICE and NTW represent the other extreme: The simulator defines an agent, as well as a world physics, supplying a set of sensing and effecting operations as part of the world. TRUCKWORLD partially defines the truck's effecting capabilities; it defines a set of primitive commands, but the exact effect of these commands depends on the objects being manipulated. Objects and their interactions are defined by the experimenter. TRUCKWORLD does not, however, define a set of sensing operations. Sensors are objects defined by the experimenter that happen to send sensory information back over the command channel.

Parameterizability: TILEWORLD and NTW have a built-in set of parameters that characterize the behavior of a world. These parameters facilitate experimentation; by varying the world's parameters systematically and matching them against various agent designs, one might be able to come up with agent types that perform well for particular world types. The price one pays for this ability to perform experiments is in simplicity and control. A world that is fully characterized by a small number of parameters must be simple, and furthermore, the parameters must characterize completely the nature of the agent's behavior in this world. PHOENIX allows the experimenter to specify values for parameters such as wind speed and also to write scripts for how parameters change over time during an experiment. PHOENIX also provides a mechanism called *alligator clips* for recording the values of parameters during experiments.

We are again faced with a trade-off: In TILEWORLD, NTW, and PHOENIX, one might be able to demonstrate a systematic relationship between a world's characteristics and an agent's performance. Such demonstrations, however, must be supplemented with convincing arguments that these relationships will be mirrored in a more realistic world, and it is far from easy to make such arguments. In TRUCKWORLD, one can demonstrate that the agent performs well on more complex problems, but it might be difficult to demonstrate precisely the reasons for this success and to apply these reasons to other domains.

Discussion

The discussion to this point has touched on mainly uncontroversial points: the need for introducing more rigorous empirical methods into planning research and the roles that test-bed environments and benchmark tasks might play. The question of what the ultimate goal of these research efforts is, as well as how the goal might best be pursued, is the subject of some disagreement among the authors. The following three subsections reflect this disagreement and represent the authors' personal opinions. In the first subsection, Hanks argues against a program of controlled experimentation in small, artificially simple worlds. Pollack defends such a program in the second subsection. In the third, Cohen addresses the problem of generalizing results from test-bed experiments.

The Danger of Experimentation in the Small (Steve Hanks)

The planning community has been pushed (or has pushed itself) in two directions recently, and these directions seem at odds. We see pressure to apply our representations and algorithms to more realistic domains, and at the same time, we feel the need to evaluate our systems more rigorously than by announcing a system's ability to solve a few small, carefully chosen problems. The problem is that programs that operate in more realistic domains tend to be bigger and more complicated, and big complicated programs are more difficult to understand and evaluate.

In writing this article, we agreed on the following two objectives as researchers in planning: (1) to build systems that extend the functions of existing systems—that solve larger or more complicated problems or solve existing problems better—and (2) to understand how and why these systems work. Further, running experiments is a good way (although not the only way) to accomplish the goal of understanding the systems we build. We tended to disagree, however, on the best way to achieve these objectives, in particular on the issues of what form an experimental methodology should take and what role it should play in the system-building process.

Here I discuss a particular methodological approach, which I call *experimentation in the small*. Langley and Drummond (1990) advocate this position in the abstract. Pollack and Ringuette (1990) and Kinny and Georgeff (1991) explore it concretely using an implemented test bed and a suite of experiments. I take the methodological commitments of this approach to be the following:

First, the researcher conducts experiments

in a test-bed world that is significantly simpler than the world in which the agent is ultimately to be deployed. In particular, the world is supposed to exhibit particular interesting characteristics but will be artificially simple in other aspects.

Second, the test bed provides a set of parameters that govern the world's behavior. Experimentation is a process of matching characteristics of the agent's problem-solving methods with the world's parameter values; the goal of experimentation is to discover relationships between these two sets of characteristics that predict good (or bad) performance.

The main point of my discussion is that experimentation in small, controlled worlds is not, in and of itself, an effective way to establish meaningful relationships between agents and their environments. I show that the nature of the relationships established by these experiments is inherently connected with the implementation details both of the agent and of the test-bed worlds. The hard part remains: generalizing beyond the particulars of the world or even arguing that a particular test-bed world is appropriate for studying a particular agent architecture. The experiments themselves do not provide guidance in this task and might even tend to hinder it.

I use the TILEWORLD test bed and experiments from Pollack and Ringuette (1990) and Kinny and Georgeff (1991) to make these points. My goal in doing so is not to single out this particular work for criticism. I do so, first, because it's important to discuss the concrete results that can be expected from these experimental endeavors, and second, these two pieces of work are rare examples of systematic experimentation with agent architectures in small, controlled worlds.

The Original TILEWORLD Experiments

The planning agent studied in Pollack and Ringuette (1990) is an implementation of the IRMA architecture (Bratman, Israel, and Pollack 1988). One of the key ideas advanced in their paper is that one way for agents to cope with a changing environment is to filter out (avoid considering) options that conflict with their current intentions (the filtering mechanism) unless the option is especially promising (the filter-override mechanism). IRMA also suggests that the agent separate what to do (deliberation) from how to do it (planning). The TILEWORLD agent thus chooses its actions in two phases: The deliberation phase chooses a hole to fill (we call it the *current hole*), then the planning phase plans a sequence of moves that can fill the current hole. The agent's filter-override mechanism governs when the agent abandons the current hole in favor of a new alternative.

The TILEWORLD agent has three components:

First is the *filter-override mechanism*, a test applied to a newly appeared hole that determines whether the task of filling the current hole should be reconsidered in light of the new option(s).[3] Only one filter-override mechanism was implemented: a threshold v such that a new hole would be considered as an alternative to the current hole just in case its score exceeded the score of the current hole by at least v points.

Second is the *deliberator*, a procedure that chooses the next hole to work on. Two alternatives were implemented. The simpler (highest score, or HS) deliberator always chooses the hole with the highest score. The more complicated (likely value, or LV) deliberator divides the hole's score by an estimate of the cost of filling it: the sum of the distances of the n closest tiles, where n is the hole's capacity.

Third is the *path planner*. Given a hole to fill, the path planner uses breadth-first search to generate the optimal sequence of moves to fill it with tiles. The choice of a path planner was not among the agent parameters varied experimentally; only the optimal path planner was implemented.

The experiments show the following results: (1) An agent that acts in parallel with reasoning performs slightly better than an agent that acts and reasons serially. (2) The more sophisticated LV deliberator performs somewhat better than the simpler HS deliberator. (3) The filter-override mechanism at best has no effect on the agent's performance and, in some cases, makes it perform worse.

Hanks and Badr (1991) analyze these experiments in detail. Here, I want to discuss some issues relevant to the question of what this experimental paradigm can be expected to accomplish. In particular, I want to stress the need for caution in interpreting these results. There is a large gap between the effort's larger goal of establishing general relationships between agent designs and environmental conditions and the information that is actually presented in the paper. I don't see this gap as a fault of the paper—which presents preliminary work—but it is important to keep the results in perspective.

The connection between a general architecture for problem solving (in this case, IRMA) and the particular results reported must be interpreted, taking into account many design and implementation decisions: (1) the way in

which the IRMA architecture was realized in the TILEWORLD agent (for example, equating deliberation with choosing which hole to fill and planning with generating a sequence of moves to fill the hole), (2) the implementation of these modules in the TILEWORLD agent (for example, what the specific algorithms for deliberation and path planning are and how they interact), and (3) the implementation of the TILEWORLD simulator (for example, the choice of what environmental parameters can be varied, the interaction among the different parameters and between the agent and the simulator, and the simplifying assumptions built into the world itself).

Consider the first result, for example, and the broader conclusions we might be able to draw from it. TILEWORLD uses a simulated notion of serial and parallel reasoning. In fact, the act cycle and reasoning cycle run sequentially, but they are constrained to take the same amount of time. Is this implementation detail important to assess the benefit of acting in parallel with reasoning? I'm not sure. In the current implementation, the agent cannot be interrupted during its reasoning cycle by changes to the world that occur during the concurrent act cycle. This deviation from truly parallel reasoning and acting strikes me as significant. In any event, the speedup result must be interpreted with an understanding of the particular implementation and cannot be interpreted more broadly without further analysis.

The second result, suggesting that the LV deliberator performs better than the HS deliberator, must also be interpreted in the context of the particular implementation. Hanks and Badr (1991) note that one part of the TILEWORLD agent is the path-planning algorithm, which (1) solves the problem optimally; (2) is not subject to experimental variation; and (3) is written in C, presumably for efficiency reasons. To what extent do the experimental results depend on the ability to solve the path-planning subproblem quickly and optimally? Hanks and Badr (1991) show that the fast path planner has a greater effect on the system's performance than does variation in the deliberator (which was one of the parameters varied experimentally). Given this fact, we should be cautious about interpreting the experimental result too broadly. Would an agent actually benefit from a more sophisticated deliberator if it were unable to solve the path-planning subproblem quickly and optimally? This question would have to be answered to apply the result beyond the specific implementation and experimental setting examined.

The final result—that the filter-override mechanism does not generally improve the agent's performance—strikes me as the one most closely related to the specific agent and environment implementations. Hanks and Badr recognize the problem that the environment did not challenge the deliberator, thus rendering a fast preliminary filtering mechanism unnecessary. They propose making the environment more challenging, specifically by making the world change more quickly (that is, by changing the parameters that govern the world's behavior).

Another interpretation of the same result is that TILEWORLD is inherently not a good test of an IRMA-like filtering mechanism. The justification for a filtering and override mechanism is that the filter override benefits the problem solver when deliberation is complex and difficult but, at the same time, when deliberation at least potentially benefits the planner significantly.

Put another way, deliberation is really a matter of predicting the future state of the world and choosing one's actions to maximize utility given the predicted future. The problem with TILEWORLD is that there is little to predict. Tiles appear and disappear at random and with no pattern. The effects of the agent's actions are localized. On balance, there is little to be gained from thinking hard about the world, which Hanks and Badr (1991) show by demonstrating that there is little benefit to be had even by implementing a deliberator that computes the agent's optimal course of action given current information. If deliberation is either easy to do or doesn't benefit the agent significantly, then there is no need for a surrogate for deliberation such as the filter override. Hanks and Badr mention the possibility of making the deliberation process more expensive but not the possibility of changing the world (for example, giving it more causal structure or making the agent's reward structure more complex) to give more potential payoff to the deliberation process.

The point of this discussion is to demonstrate the difficulty of interpreting experimental results such as those reported in Pollack and Ringuette (1990) or, more specifically, the difficulty associated with applying the results to any circumstances other than those under which the experiments were conducted. In the next subsec-

tion, I discuss the implications of the general paradigm of experimentation in the small, but first, I want to discuss some follow-up experiments in the TILEWORLD environment.

Subsequent TILEWORLD Experiments The experiments in Pollack and Ringuette (1990) tried to establish a relationship between the agent's commitment to its current plan—its willingness to abandon its current goal to consider a new option, or *boldness* as it was called—and the rate at which the world changes. Kinny and Georgeff (1991) try to make this relationship precise and provide additional empirical support. They begin their experimental inquiry by further simplifying the test-bed world: "[The TILEWORLD] was considered too rich for the investigative experiments we had planned. Therefore, to reduce the complexity of the object-level reasoning required of our agent, we employed a simplified TILEWORLD with no tiles" (Kinny and Georgeff [1991], p. 83). The agent's task in this simplified TILEWORLD is to move itself to a hole on the board, at which point it is awarded the hole's score. The agent is provided with perfect, immediate, and cost-free information about the world's current state.

Once again, the planning agent is configured around the tasks of deciding which hole to pursue, deciding which path to take to the chosen path, and deciding whether to pursue a new hole that appears during execution.

The agent always chooses the hole with the highest ratio of score to distance. It adopts a new hole according to its filter-override policy, also called its degree of commitment or degree of boldness. Degree of boldness is a number *b*—the agent automatically reconsiders its choice of hole after executing *b* steps of its path toward the hole it is currently pursuing. A *bold agent,* therefore, tends to make a choice and stick with it. A *cautious agent* tends to reconsider more often and is more likely to abandon its old choice of hole in favor of a new one.

Another agent parameter is its planning time, a number *p* set by the experimenter. The path planner produces an optimal path to the current hole, and the planning-time parameter dictates that it took *p* time units to do so. It is important to point out two things. First, the number *p* bears no necessary relationship to the amount of time that it actually takes to generate the plan. Planning time is a constant set by the experimenter and does not depend on the time it takes to build a plan for the current path. Second, increasing or decreasing *p* has no effect on solution quality. Kinny and Georgeff are not exploring the trade-off between planning time and plan quality. The path planner always returns an optimal path; the planning-time parameter makes it seem like it took *p* time units to do so.

A single parameter causes variation in the world: γ, which is the ratio of the agent's clock rate to the rate at which the world changes. Large values of γ indicate that the world changes frequently relative to the amount of time it takes the agent to act. The agent's effectiveness is measured by dividing the number of points the agent actually scores by the sum of the scores for all the holes that appear during the game.

… deliberation is really a matter of predicting the future state of the world.

The experiments showed various relationships between effectiveness, rate of world change, commitment, and planning time: (1) Effectiveness decreases as the rate of world change (γ) increases. (2) As planning time approaches 0, an agent that reconsidered its options (choice of hole) after every step performs better than an agent that never reconsidered. (3) As γ increases, an agent that reconsidered often tends to perform better than an agent that reconsiders infrequently, planning time held constant. (4) When the cost of planning is high, an agent that reconsiders infrequently tends to perform better than one that did so frequently, rate of world change held constant.

These experiments used an agent that reconsidered its current hole after a fixed number of steps *b*. If the agent instead reconsidered its choice of path either after *b* steps or at the time the target hole disappeared, then the bold agent outperformed the cautious agent, regardless of the values of *p* and γ. Performance was improved further by reconsidering the choice of target when a hole appeared closer to the agent than the current target.[4]

Once again, I want to point out the difficulty in applying these results to situations

other than the specific experimental environment. Doing so requires evaluating what the simplifications are to TILEWORLD and how they affect the complexity of the deliberation task, evaluating how well the definitions of *boldness* and *planning time* apply to different domains, and so on. To what extent does the last result, for example, depend on the fact that the agent was provided with complete, instantaneous, correct, and cost-free information about changes to the world?

Analysis How do these experiments advance the cause of building intelligent agents? I think it's clear that the agents presented in these papers do not, in and of themselves, constitute significant progress. Both operate in extremely simple domains, and the actual planning algorithm consists of using a shallow estimate of a hole's value to focus the agent's attention, then applying an optimal algorithm to plan a path to the chosen hole. This strategy is feasible only because the test bed is so simple: The agent has, at most, four possible primitive actions; it doesn't have to reason about the indirect effects of its actions; it has complete, perfect, and cost-free information about the world; its goals are all of the same form and do not interact strongly; and so on.

The argument must therefore be advanced that these experimental results will somehow inform or constrain the design of a more interesting agent. Such an argument ultimately requires translating these results into general relationships that apply to significantly different domains and agents, and I pointed out how tricky it will be to establish any applicability beyond the experimental test bed itself. A crucial part of this extensibility argument will be that certain aspects of the world—those that the test bed was designed to simulate more or less realistically—can be considered in isolation; that is, studying certain aspects of the world in isolation can lead to constraints and principles that still apply when the architecture is deployed in a world in which the test bed's simplifying assumptions are relaxed.

Finding a general and useful interpretation for experimental results is a crucial part of the process of controlled experimentation. One immediately faces the trade-off between stating the relationships in such a way that they are not so general as to be uninformative and stating them so that they are not so specific that they don't generalize outside the particular agent and world in which the experiments were conducted.

Both TILEWORLD papers discuss the difference between a bold and a cautious agent, for example. These general terms are supposed to suggest an agent's willingness to reassess its plan commitments as it executes its plans: A bold agent rarely reconsiders its plans; a cautious agent does so frequently.[5]

The two main results from Kinny and Georgeff (1991) can be stated as follows: First, it's a good policy for an agent to be more cautious as the world changes more rapidly. In other words, planning ahead doesn't do much good when the world changes a lot before the plan is executed or while the plan is being executed. Second, it's a good policy for an agent to rethink its commitment to a goal when the goal disappears or when a goal appears that superficially looks more promising than its current goal. Both results turn out to be robust, holding as various other parameters both of the agent and of the world are varied. Stated this way, the relationships seem pretty straightforward; I would be surprised to hear about an agent that did not adopt these policies either explicitly or implicitly. The question is, therefore, whether the relationships stated in these general terms provide significant guidance to those that build other agents.

Of course, the first relationship can be restated much more specifically, mentioning the agent's goals (to move to holes), its problem-solving strategy (to choose a hole using a heuristic, then plan a path optimally to the hole, using exactly p units of time to do so), its definition of boldness (the number of operators it executes before replanning), and the nature of its world-change parameter (the rate at which holes randomly appear). Interpreted in this light, the result is much less obvious. It does provide significant guidance to somebody who wants to design an agent using the architecture so described, to act effectively in a world so described, given problems of the sort so described, but nobody really wants to do it. Thus, the problem is how to interpret this more specific relationship in a broader context. What if the agent doesn't have immediate, perfect, and cost-free information about the appearance of holes? What if the designer does not have an optimal and efficient path planner at his/her disposal? What if the appearance of holes is not truly random but operates according to some richer causal structure? Do the same relationships still hold? For that matter, are the same relationships even meaningful?

The main point here is that experimentation does not provide us automatically with meaningful relationships between agents and environments. Claiming that a specific experimental relationship establishes a connection between boldness and the rate of world change constitutes a form of wishful thinking.[6] It translates a specific relationship between a particular implemented agent and a particular simulated world into terms that are intuitive, broad, and imprecise. Giving intuitive names to these characteristics and to their relationship does not make them meaningful or broadly applicable. The real contribution of such an analysis would be to come up with the right way of characterizing the agent, the world, and their relationship in terms that are not so specific as to be applicable only to the experimental domain but not so vague as to be vacuously true. Thus far, the experimental work has not focused on this question; in fact, it's worth asking whether running experiments in artificially small, simple worlds is the right place to start looking for these relationships at all.

Examining Environmental Features in Isolation I turn now to the second assumption underlying experimentation in the small: A particular characteristic of a realistic world can be studied in isolation, and good solutions to the restricted problem lead to good solutions in the realistic world. TILEWORLD, for example, focuses on unplanned change in the form of the random appearance and disappearance of tiles and holes (or just holes in the case of the simplified TILEWORLD) but simplifies away other aspects of the world.

This scaling assumption is absolutely crucial to the whole experimental paradigm, and I have not seen it defended in the literature. In fact, the only explicit mention of the assumption I have found appears in Philips et al. (1991, p. 1):

> We are not suggesting that studies of these attributes in isolation are sufficient to guarantee the obvious goals of good methodology, brilliant architectures, or first-class results; however, we are suggesting that such isolation facilitates the achievement of such goals. Working on a real-world problem has obvious benefits, but to understand the systems that we build we must isolate attributes and carry out systematic experimentation.

My own work leads me to believe that it will be difficult to isolate particular aspects of a large planning problem. In Hanks (1990b), for example, I confront the problem of reasoning about plans in an uncertain world. Unplanned, random change—such as tiles and holes appearing and disappearing—is one source of uncertainty, but there are others: The agent can have incomplete or incorrect information about the world's initial state, have an incomplete model of its own actions, and might not have enough to time to consider explicitly every outcome of its plan. I see no way to separate one of these factors from the others in any principled way; therefore, I see no way that studying the simplified problem of a world in which all uncertainty is the result of unplanned, random change can shed light on the larger problem of reasoning about plans in an uncertain world.

It's not even clear whether the problem that the TILEWORLD papers claim to be investigating—the decision of when it is advantageous to act as opposed to deliberate—can be considered in a context in which all exogenous change is random. The decision about whether to plan or act depends both on the world and on the agent's ability to predict the world; the better it is at reasoning about the effects of its actions, the more benefit can be derived from thinking ahead.

TILEWORLD trivializes the prediction process by making the world essentially unpredictable: Tiles and holes appear and disappear at random. The agent, therefore, has no incentive to reason about what tiles might appear or disappear or where they might appear, which greatly simplifies the question of whether it should deliberate or act. Can we therefore apply the experimental results established in TILEWORLD to worlds in which prediction is a difficult problem?[7]

Experimentation in the small depends on the ability to study particular aspects of a realistic world in isolation and to apply solutions to the small problems to a more realistic world. I have seen no indication that such studies can in fact be performed; in fact, neither TILEWORLD paper argues that random, unplanned change is a reasonable feature for isolated study. An experimenter using these worlds, therefore, runs the risk of solving problems in a way that cannot be extended to more realistic worlds and, at the same time, of making his/her job artificially difficult for having studied the problem in isolation. Kinny and Georgeff (1991) state that "[simulated worlds] should ideally capture the essential features of real-world domains while permitting flexible, accurate, and reproducible control of the world's characteristics" (p. 82). Their proposition is appealing, but

the fact is we don't know what it means to "capture the essential features of real-world domains," much less whether it is possible to do so in a system that allows "reproducible control of the world's characteristics." Conducting experiments in small, controlled worlds carries with it the responsibility of considering the implications of the simplifications that were made to allow the experimentation in the first place.

However at this point, we must remind ourselves of our ultimate goals: to build systems that solve interesting problems and to understand why they do so. Research decisions must be oriented toward solving problems, not toward satisfying methodological goals. The ultimate danger of experimentation in the small is that it entices us into solving problems that we understand rather than problems that are interesting. At best, it gives the mistaken impression that we are making progress toward our real goal. At worst, over time it confounds us to the point that we believe that our real goal is the solution of the small, controlled problems.

Conclusion In no way should this section be taken as an argument against using experimental methods to validate theories or programs. In fact, I think the need for experimentation is manifest; we need to understand why and how well our ideas and our architectures work, and we will not always be able to do so using analytic methods. Neither am I opposed to conducting these experiments in controlled, overly simplified worlds. I can imagine, for example, a researcher implementing some idea in a system, then building a small world that isolates the essence of this idea, then using the small world to explore the idea further. I object, however, when attention turns to the experimentation process itself instead of the ideas that are to be tested and when the assumptions inherent in the small world are adopted without regard to the relationships the world is supposed to demonstrate.

The ultimate value—arguably the only value—of experimentation is to constrain or otherwise inform the designer of a system that solves interesting problems. To do so, the experimenter must demonstrate three things: (1) his/her results—the relationships he/she demonstrates between agent characteristics and world characteristics—extend beyond the particular agent, world, and problem specification studied; (2) the solution to the problem area studied in isolation will be applicable when the same problem area is encountered in a larger, more complex world; and (3) the relationship demonstrated experimentally actually constrains or somehow guides the design of a larger, more realistic agent. The experimental work I have seen has addressed none of these questions.

I originally stated that our two objectives as researchers are (1) building interesting systems and (2) understanding why they work. It seems to me that experimentation in the small adopts the position that these goals should be tackled in reverse order—that you can understand how an interesting system must be built without actually building one. I don't believe this case to be so; rather, we should be building systems and then applying analytic and experimental tools to understand why the systems did (or did not) work.

The Promise of Experimentation (Martha E. Pollack)

Steve Hanks believes that experimentation in the small is a dangerous enterprise. I believe that to the contrary, controlled experimentation—small, medium, or large—promises to help AI achieve the scientific maturity it has so long sought. In these comments, I try to defend this belief.

In his section entitled The Danger of Experimentation in the Small, Hanks begins by stating that the primary objectives of those studying agent design are "(1) to build systems that extend the functionality of existing systems . . . and (2) to understand how and why these systems work" (emphasis mine). I would put the second point somewhat differently and claim that we aim to understand "how and why such systems can work." This change is not a minor matter of wording but, rather, a fundamental disagreement about research methodology. Hanks believes that complex-system building must precede experimental analysis, but I believe that these two activities can and should proceed in parallel. Hanks does not object to all experimentation, only to experimentation in the small, that is, experimentation using simplified systems and environments. I claim not only that such experimentation can be informative, but that given our current state of knowledge about system design, controlled experimentation often requires such simplifications. Thus, in my view, Hanks's position is tantamount to an injunction against all experimentation in AI; in other words, it is a call for the maintenance of the status quo in AI methodology.

It is important to be clear about what con-

stitutes an understanding of how and why certain autonomous agents work. In my view, this understanding consists of a theory that explains how alternative design choices affect agent behavior in alternative environments; that is, it will largely consist of claims having the form, A system with some identifiable properties *S*, when situated in an environment with identifiable properties *E*, will exhibit behavior with identifiable properties *B*.[8]

The goal of experimentation in AI (and arguably, a primary goal of the science of AI taken as a whole) is to elucidate the relationships between sets of properties *S*, *E*, and *B*, as previously defined. I argue that for experimentation to succeed in meeting this goal, two types of simplification must be made. The first type is inherent in the notion of experimental design. Experimentation necessarily involves selective attention to, and manipulation of, certain characteristics of the phenomena being investigated. Such selectivity and control constitute a type of simplification of the phenomena. The second type of simplification that is currently needed arises from our existing abilities to build complex AI systems. Large, complex systems that tackle interesting problems are generally not principled enough to allow the experimenter to meaningfully probe the design choices underlying them. Moreover, they are designed for environments in which it might be difficult or impossible to isolate, manipulate, and measure particular characteristics. Finally, these systems do not generally include instrumentation to measure their performance, although it is conceivable that in many cases, this instrumentation could be added in a fairly straightforward way. Thus, these systems do not allow the experimenter sufficient access at least to *S* and *E* and, possibly, also to *B*; they are, in short, ill suited for controlled experimentation. In contrast, the kinds of simplified systems we described in Test-Bed Implementations, that is, test beds such as TILEWORLD, NTW, TRUCKWORLD, and PHOENIX and their embedded agents, are designed specifically to provide the control needed by the experimenter.

Hanks correctly notes that the simplifications required for experimentation introduce methodological challenges. In particular, he points out the issue of generalizability: How can a researcher guarantee that the simplifications made in the design of an experiment do not invalidate the generality of the results obtained? I believe that this issue is a serious one that poses a significant challenge to AI researchers. Moreover, I agree with Hanks that by and large, the controlled experimentation that has been performed to date in agent design—including my own work—has not adequately met this challenge. This inadequacy, however, is a result of the fact that so far painfully little controlled experimentation has been conducted in AI; as Hanks notes, the TILEWORLD experiments represent relatively "rare examples of systematic experimentation with agent architectures."[9] It is extremely difficult and often impossible to have confidence in the generality of the results obtained from a few experiments. The desire for robust, generalizable results should lead us to do more, not less, experimentation.

controlled experimentation —small, medium, or large— promises to help AI achieve the scientific maturity it has so long sought

The problem of generalizability is not unique to AI; it is inherent in the experimental methodology, a methodology that has been tremendously successful in, and indeed is the cornerstone of, many other sciences. I see nothing in AI's research agenda that would preclude its also benefiting from controlled experimentation. Of course, adopting the experimental method entails adapting it to the particulars of the AI research program. In my comments to follow, I give some necessarily sketchy suggestions about how we might adapt the methodology and, in particular, how the challenge of generalizability can be met in AI. Following Hanks, I also use TILEWORLD as an example.

Simplification in Experimentation

"Simplification, paring back the variables, far from invalidating results, is indeed required by the foundations of empirical design. The success of reductionism depends on measuring and reporting only that bit of cloth that can be understood and tested piecemeal" (Powers [1991], p. 355).

Experimentation mandates simplification. In investigating a complex phenomenon, the experimenter selectively attends to some aspects of it, namely, those that he/she

believes are relevant to his/her hypotheses. He/she exerts control over those aspects of the phenomenon, manipulating them as necessary to test his/her hypotheses. At the same time, he/she holds constant those influences that he/she believes are extraneous to his/her hypotheses and allows or even forces random variation in those influences that he/she believes are noise. This selective attention to, and intentional manipulation of, certain aspects of the phenomenon is the "paring back [of] the variables" noted in the previous quotation.

Does Hanks object to simplification as such; that is, does he believe that to be useful, a hypothesis about agent design cannot make reference only to some aspects of an agent's architecture or environment? Although he appears to be inclined toward this conclusion when he asserts his belief that "it will be quite difficult to isolate particular aspects of a large planning problem," this objection is not his primary one. Rather, what he views as dangerous is a particular way of achieving simplification in research on agent design, namely, by conducting experiments using highly simplified agents operating in highly simplified environments. This reliance on simplification defines what he terms "experimentation in the small." Hanks's introductory comments mention only objections to the use of simplified environments, but his criticisms of the TILEWORLD experiments show that he also objects to the use of highly simplified agents.

I alluded earlier to my belief that it is necessary to make significant simplifications in the agents and environments we use in conducting experimentation. Large, realistic systems have generally been built without the benefit of a principled understanding of agent design—precisely what experimentation (supplemented with theorizing) aims at. As a result, it is extraordinarily difficult to determine which mechanisms of these complex systems are responsible for which aspects of their behavior, in other words, to isolate the key properties of *S* and *B*. It is difficult to determine what in the system is essential to the observed behavior and what is instead an artifact of the way the system happened to be implemented. In addition, when these systems are deployed in real environments, there is no ready way to isolate and control key features of these environments, that is, to get a handle on *E*.

The test beds that we surveyed in Test-Bed Implementations are designed specifically to provide the researcher with the control needed to conduct experimentation—to enable him/her to control and monitor the conditions of the environment and to measure the behavior of a system embedded in the environment. In other words, a useful test bed will give the researcher a handle on *B* and on *E*.

To give the researcher a way to measure *B*, the test-bed designer specifies what counts as successful behavior in the test-bed environment and provides instrumentation that measures success. To give the researcher a way to control and monitor *E*, the test-bed designer selects some set of environmental features and provides instrumentation that allows the researcher to control these. One potential objection is that the test-bed designer thereby influences the experiments that can be conducted using the test bed; researchers might want to study other characteristics of *B* and *E* than those identified by the test-bed designer. However, this problem only exists if researchers are mandated to use particular test beds. The problem disappears if we leave the decision about which test bed to use to individual researchers. A test bed is just a tool, and it is up to the researcher to determine the best tool for his/her current task. Indeed, in some cases, researchers might need to build their own tools to pursue the questions of interest to them. It is worth noting, though, that some test beds might be more flexible than others, that is, might more readily suggest ways to model a variety of environmental features and/or behavioral aspects and, thus, be more amenable to modification by the test-bed users. Later, I suggest that flexibility is one of the strengths of the TILEWORLD system.

To this point, I have focused on how a test bed allows control of *B* and *E*. It is, of course, also necessary for the researcher to have control of the system features, *S*. One way to achieve this control is to use the same kind of parameterization in an agent embedded in a test-bed environment as is used in the environment itself. One of the more useful features of the TILEWORLD system is precisely that it provides the experimenter with control over the embedded system as well as over the environment.

Hanks does not dispute the claim that simplification of the kind provided by test-bed environments and agents provides experimental control. What worries him is that the price we might pay for this control is too high. His main argument is that the simplifications that provide the needed control also make it impossible to produce results that are in any sense real or generalizable, that is, can

be shown to be applicable to larger AI applications. Cohen, Hanks, and I all agree that this problem, often called *realism*, is the most difficult challenge facing researchers on agent design who adopt the experimental methodology we discuss in this article. However, we disagree about whether this difficulty in insurmountable.

Toward Realism The problem of realism is a challenge for experimentalists—for all experimentalists, not just those in AI. To achieve the experimental control they need, scientists in many disciplines have made use of simplified systems and have thus had to address the question of how the lessons they learn using these systems can be applied to more complex phenomena. However, the history of science is full of examples in which this challenge has been met successfully. For example, biologists have used the simple organisms *Drosophila* and *Escherichia coli* in numerous experiments aimed at understanding the fundamental mechanisms of genetics. The results of these experiments have had tremendous significance for the theory of inheritance in all organisms, including humans. Neurobiologists have used *aplysia*, animals with only a few neurons, to conduct experiments investigating neuroplasticity. Again, the results have been generalized to theories about the ways in which human brains function. As another example, engineers have built systems to simulate natural phenomena—wind tunnels and wave machines, for instance. These simulations abstract away from much of the complexity of real environments. Nonetheless, experiments conducted using them have provided many valuable lessons about the effects of the modeled phenomena on engineered artifacts such as airplanes.

Of course, merely pointing out that many other sciences have been able to meet the challenge of realism is not, in and of itself, enough to demonstrate that AI researchers concerned with agent design will be able to do so. What is needed is a closer look at how this challenge has been met. A widely used introductory textbook on statistics describes the process of achieving realism as follows:

> Most experimenters want to generalize their conclusions to some setting wider than that of the actual experiment. Statistical analysis of the original experiment cannot tell us how far the results will generalize. Rather the experimenter must argue based on an understanding of psychology or chemical engineering or education that the experimental results do describe the wider world. Other psychologists or engineers or educators may disagree. This is one reason why a single experiment is rarely completely convincing, despite the compelling logic of experimental design. The true scope of a new finding must usually be explored by a number of experiments in various settings.
>
> A convincing case that an experiment is sufficiently realistic to produce useful information is based not on statistics, but on the experimenter's knowledge of the subject-matter of the experiment (Moore and McCabe 1989, p. 270).

The key to achieving realism lies in the researcher's knowledge of the subject matter; the researcher must provide an argument, based on his/her understanding of the subject matter, that, in fact, the experimental results do describe the wider world. For such arguments to be satisfying, they must be informed by a rich theory of the phenomena in question. For the experimental program to succeed in AI, AI researchers will need to be more scrupulous about careful theory development; as I have claimed elsewhere (Pollack 1992), our field has not always valued theory development as an integral part of our work.

Research into agent design begins with a theory. Of course, the theory, in whole or in part, can be informed by the theorist's previous experiences with building large, interesting systems. An experimental research program on agent design includes the following components (see Cohen's [1991] MAD [modeling, analysis, and design] methodology): (1) a theory describing some aspect(s) of agent design—particularly, the agent's architecture, the environment, and the agent's behavior—and the purported effect of these design aspects on agent behavior in certain environments;[10] (2) an implemented test-bed environment and a description of the characteristics of the environment; (3) an implemented agent who will operate in the test bed; and (4) mappings describing the relationship between the real phenomena described by the theory and their intended analogs in the test-bed environment, the relationship between the agent architecture described in the theory and its realization in the implemented agent, and the relationship between the agent's design and its performance in the test-bed world.

A typical set of experiments will then evolve from some hypothesis, typically asserting that under the conditions of some given environment, some specified behavior will be observed in agents having some given architectural characteristics. Experiments can then be designed using the implemented (or operationalized) analog of this hypothesis, relating conditions in the test bed to observed behavior in the implemented agent. Such experiments can have several different types of result. They can confirm, deny, or suggest modifications to the hypotheses in the underlying theory. They can suggest needed changes to the test-bed system or to the mappings between the actual and the simulated environments. They can reveal flaws in the way the environment was modeled. They can suggest needed changes to the simplified agent or to the mappings between the actual and the simulated agents. They can reveal flaws in the way the agent was modeled or the way its behavior was measured. Perhaps most importantly, they can suggest additional experiments that should be performed, either using the same or another test bed and agent.

This last type of result is critical. Experimentation is an iterative process. Part of the experimental program is to refine the mapping between a theory and its realization in implemented systems. Part of the experimental program is to iteratively refine the experiments themselves. As Moore and McCabe (1989, p. 270) put it, "a single experiment is rarely completely convincing.... The true scope of a new finding must usually be explored by a number of experiments in various settings." To facilitate related experiments, great care must be given to the way in which theories are stated and to the way in which these theories are operationalized in experimental settings. Test beds and simplified agents make it possible to meet the latter requirement.

The TILEWORLD Experience

To make this discussion more concrete, I want to describe briefly some of the experiences we have had in conducting experiments using the TILEWORLD system. I focus on TILEWORLD because it is the experimental work with which I am most familiar and because Hanks addresses it in his comments. I do not mean to suggest that TILEWORLD is the ultimate test bed or one that all researchers should use in their work. On the contrary, for reasons I have already discussed, it is essential that AI researchers use a variety of test-bed systems in their experimentation. Moreover, TILEWORLD is an early, prototype test-bed system, and in using it, my research group and I have not only learned about agent design but also a great deal about the test-bed design. These lessons have led to a number of changes and extensions to the original system, which was reported on in Pollack and Ringuette (1990). Here I mention some of these changes; see also Pollack et al. (1993).

The initial goal in building TILEWORLD was to study a particular, well-developed theory of resource-limited reasoning, called IRMA, that we had previously developed (Bratman, Israel, and Pollack 1988; Bratman 1987; Pollack 1991). This theory was built on a detailed philosophical analysis of the role of intention in managing reasoning; the aim was to investigate certain underspecified aspects of this model. In particular, we began with a theoretically motivated strategy for coping with changing environments—the strategy of *commitment-based filtering.* Roughly speaking, this strategy involves committing to certain plans and tending to ignore options for action that are deemed incompatible with these plans. Filtering can be more or less strict, and we wanted to determine the environmental conditions under which stricter filtering was more advantageous. In addition, there are various ways to realize the notion of strictness, and we wanted to explore the effects of these alternatives on agent behavior in different environmental conditions. The environmental condition that we suspected to be most important was the average rate of change in the environment. Details are in Pollack and Ringuette (1990); this brief sketch is meant to highlight the fact that underlying our attempt to relate *S* (in this case, conditions on filtering), *E* (average rate of change), and *B* (the agent's overall performance) was a larger theory about the role of intentions in resource-limited reasoning.

The experiments that we conducted, as well as those performed by others using TILEWORLD (Kinny 1990; Kinny and Georgeff 1991; Kinny,. Georgeff, and Hendler 1992), led to each of the kind of results I described earlier:

First, they provided preliminary confirmation of some parts of the theory. Experimentation showed that strict filtering of incompatible options, coupled with an appropriate overriding mechanism, is viable at least under some circumstances (Kinny and Georgeff 1991; Kinny 1990). In other words, commitment to one's plans can be a valuable strategy for managing a changing environment. Experimentation also suggested needed

modifications to the theory. For example, one TILEWORLD user, John Oh, pointed out to us that the agent's performance is hindered by its inability to immediately adopt certain extremely promising options without deliberation. The original theory included a mechanism for short circuiting deliberation to eliminate a new option, but it lacked a mechanism for short circuiting deliberation to immediately adopt a new option. Thus, the theory needed to be modified to include a new mechanism of the latter type.

Second, the experiments suggested needed changes to the test-bed environment. As Hanks correctly points out, the original TILEWORLD test bed was extremely homogeneous—essentially, the world only presented one type of top-level goal (hole filling). This fact limited the range of experiments that could be conducted; there was no way to explore the behavior of agents who had to perform complex (and, thus, computationally costly) plan generation. Since the publication of Pollack and Ringuette (1990), researchers have increased the complexity of the TILEWORLD environment, so that they can study situations in which a wider range of options are presented to the agent (Pollack et al. 1993).

Third, the experiments also suggested needed changes to the agent embedded in the TILEWORLD environment. Early experiments showed that the simplifications researchers made in the deliberation and plan-generation component of the system were too extreme. Both processes were uniformly inexpensive, and we were thus unable adequately to explore the advantages of the filtering process, whose intent is to reduce the amount of deliberation and planning needed (Pollack and Ringuette 1990). This limitation subsequently led us to increase the complexity of the deliberation process. Note the interaction between this change and the previous one described; the added complexity in the agent depended on the added complexity in the environment.

Finally, the experiments suggested a large number of additional experiments that need to be conducted to expand and strengthen the original theory. Hanks, in fact, gives many examples of such experiments. He wonders about the significance of the agent's ability to perform some planning problems optimally. He suggests that the degree of (un)predictability in an environment might be an important influence on the value of committing to one's plans. He asks, "What if the agent doesn't have immediate, perfect, cost-free information about the appearance of holes? What if the designer does not have an optimal and efficient planner at his/her disposal?" Questions such as these are precisely what a theory of agent design should answer and directly suggest experiments that could be performed using TILEWORLD or other test-bed systems. We count as a success of our experience with TILEWORLD that it has led a number of researchers to ask just such questions. Moreover, TILEWORLD has proven to be flexible in the sense that it can readily be modified to support experiments investigating environmental and agent-design issues other than those for which it was originally designed.

One error that we made in the initial TILEWORLD experiments was a failure to be precise enough in the terminology we used to describe the theory and its realization in the test bed and simplified agent.[11] Instead of using qualitative terms, we should perhaps have developed quantitative analyses. For example, instead of describing environments as fast or slow relative to some arbitrary baseline, we might have defined the rate of environmental change as the ratio between the average period of time between changes in the environment and the average amount of time it takes an agent to form an arbitrary plan. Qualitative definitions such as this one would certainly have facilitated the specification of the mapping functions between real phenomena and the TILEWORLD operationalization of them.

It is clear that significant effort must be put into the development of vocabularies for describing agents and environments and their realizations in implemented systems. I agree completely with Hanks that the real contribution of this line of research will be "to come up with the right way of characterizing the agent, the world, and their relationship." This goal is the primary purpose of our ongoing work. However, I disagree strongly with Hanks when he goes on to claim that to date the terms used in the TILEWORLD studies (and in all other experimentation in the small) are "so specific as to be applicable only to the experimental domain [or] so vague as to be vacuously true."

Consider the TILEWORLD results that he describes as vacuously true. He states these in terms of the circumstances under which it is advantageous to reconsider the plans to which one has already committed (for example, be more inclined to reconsider when the world is changing more rapidly; reconsider when your goal becomes impossible). Howev-

er, what is most important about the early TILEWORLD results is that they support the idea that commitment is a good idea in the first place; the results, as described by Hanks, have to do with refinements to this basic idea. Kinny and Georgeff found that commitment led to the most effective behavior under all the conditions they studied, provided the agent was given a minimal override policy that allows for reconsideration of goals that have become unachievable.

The key idea of the IRMA theory is that it pays for an agent in a dynamic environment to commit to certain courses of action, even though the environment might change so that some of these courses of action cease to be optimal. *Local optimality*—always doing what is best at a given time—must be sacrificed in the interest of doing well enough overall; commitment to one's plans generally rules out local optimality but can help lead to overall satisficing, that is, good enough, behavior. Although I cannot restate the entire argument here (again, see Bratman, Israel, and Pollack [1988]; Bratman [1987]; Pollack [1991]), it should be said that this claim is far from being so obvious that all reasonable people would assent to it.[12] Hanks says that he would be "surprised to hear about an agent that did not adopt these policies," but in fact, the recent literature in agent design has been filled with examples of agents, specifically, the so-called reactive agents, that are notable precisely because they do not commit to any plans; instead, they decide at each point in time what action is appropriate (Agre and Chapman 1987; Brooks 1991; Schoppers 1987). A standard attempt to resolve the debate between those advocating reactiveness and those advocating deliberativeness has been to suggest a middle road: Rational agents sometimes should deliberate about, and commit to, plans, and other times, they should react more immediately to their environment. The TILEWORLD experiments conducted to date can be seen, at least in part, as an attempt to clarify the conditions under which each alternative is desirable.

The key idea of the IRMA theory is that it pays for an agent in a dynamic environment to commit to certain courses of action, even though the environment might change

Conclusion

In these comments, I distinguished between two kinds of simplification in experimentation: (1) investigating hypotheses that focus on particular characteristics of a system, its behavior, and its environment and (2) using simplified systems, operating in simplified environments, to conduct the experiments. I claimed that the former is essential to all experimentation, and that although in principle the latter is not necessary, de facto it is, given the current state of our science.

Although in his comments Hanks focuses on the difficulties involved in using test beds and simplified agents in experimentation, in his conclusion, he supports their use, provided that the hypotheses toward which they are directed were inspired by experiences with particular large-scale systems. Thus, he says that he is not "opposed to conducting... experiments in controlled, overly simplified worlds [and] can imagine, for example, a researcher implementing some idea in a system, then building a small world that isolates the essence of this idea, then using the small world to explore the idea further." Apparently, Hanks feels that the problem is not in the use of simplified systems and agents per se but, rather, in the fact that researchers who have to date used simplified systems and agents have been willing to investigate hypotheses that have been developed apart from the implementation of any particular system. Thus, it appears that the primary dispute between Hanks and myself has little to do with the use of test beds and simplified systems. We both agree that unprincipled fiddling with any systems (large or small) is just that. Experimentation must build on theorizing.[13] However, Hanks demands that any theory worth investigating must derive directly from a large, implemented system, but I see no need for this restriction. Sometimes, hypotheses about agent design can result from other avenues of inquiry—such as the philosophical theorizing that led to IRMA—and it might be more effective to explore these theories experimentally before investing in large, complex systems that embody them.

Generalization of Test-Bed Results (Paul R. Cohen)

Much of the preceding discussion touches on the problem of generalizing results from research with test beds. I do the reader no service by recounting my coauthors' arguments. Instead, I try to clarify what test beds are for, focusing on their role in the search for general rules of behavior.[14] I was struck by Steve Hanks's repeated assertion that results from the TILEWORLD studies are difficult to interpret, so this assertion serves as the launching point for my own comments. All empirical results are open to interpretation. Interpretation is our job. When we read the results of a study we have to ask ourselves, What do they mean? We can answer this question in several ways. First, we might say, Goodness gracious, this result deals a deadly blow to the prevail-

ing theory of, say, agent curiosity. Let's agree that this response is unlikely for two reasons: First, we don't have a theory of agent curiosity—or a theory of any other agent behavior—and death-dealing empirical results are, in any case, rare. Second, we might interpret a study as a chink in the armor of a prevailing theory; for example, results from astronomy sometimes are interpreted as troublesome for the big bang theory. This response, too, is unlikely because we don't have any theories that make predictions for results to contradict. Third, a study might be interpreted as supporting a prevailing theory, if we had any theories to support. Fourth, a result might suggest a theory or just a tentative explanation of an aspect of agent behavior. I interpret Kinny and Georgeff's paper in this way, as weak evidence for the theory that agents sometimes do better in unpredictable domains if they are bold. In addition, I have no sympathy for the complaint that the paper is difficult to interpret. Interpretation is our job, especially now when we have no theories to do the job for us. In short, we ought to ask what our few empirical results mean—what theories they suggest because we currently have no theories to provide interpretations—instead of assert strenuously that they mean nothing.

Let us recognize that empirical results are rarely general. Interpretations of results might be general, but results are invariably tied to an experimental setup. It is wrong to assert that because Kinny and Georgeff worked with a trivial test bed, their results have no general interpretation. I have already recounted one general interpretation: Bold agents sometimes do better in unpredictable domains. Moreover, every substantive word in this interpretation has a precise meaning in TILEWORLD. Thus, Kinny and Georgeff could say, Bold agents sometimes do better in an unpredictable environment, and here is what we mean by bold, agent, sometimes, better, and unpredictable. If you are interested in our theory, tell us what you mean by these terms, and let us see if the theory generalizes.

Nothing prevents us from inventing general theories as interpretations of results of test-bed studies, and nothing prevents us from designing additional studies to test predictions of these theories in several test beds. For example, two students in my research group explored whether bold PHOENIX agents do better as the PHOENIX environment becomes more unpredictable. The experiment proved technically difficult because PHOENIX agents rely heavily on failure-recovery strategies; so, it is difficult to get them to commit unswervingly to any plan for long. Their natural state is bold, and they fail catastrophically when we make them less so; so, the results were inconclusive. However, imagine the experiments had succeeded, and evidence was accrued that boldness really does help PHOENIX agents when the environment becomes more unpredictable. Then two research groups—mine and that of Kinny and Georgeff—would have demonstrated the same result, right? Whether you agree depends on what you mean by *the same result*. I mean the following: Kinny and Georgeff offered a mapping from terms in their theory (bold, agent, better, sometimes, unpredictable) to mechanisms in TILEWORLD, and I offered a mapping from the same terms to mechanisms in PHOENIX. We both found that a sentence composed from these terms—bold agents sometimes do better in an unpredictable environment—was empirically true. In reality, as I noted, we were unable to replicate Kinny and Georgeff's result. We failed for technical reasons; there was no easy way to create a PHOENIX agent that was not bold. Differences in experimental apparatus always make replication difficult. For example, TILEWORLD has just one agent and a limited provision for exogenous events; so, it would be difficult to use TILEWORLD to replicate results from PHOENIX. Still, these problems are only technical and do not provide a strong argument against the possibility of generalizing results from test-bed research.

Interpretation is our job. When we read the results of a study we have to ask ourselves, What do they mean?

Test beds have a role in three phases of research. In an *exploratory* phase, they provide the environments in which agents will behave in interesting ways. During exploration, we characterize these behaviors loosely; for example, we observe behaviors that appear bold or inquisitive. In exploratory research, the principal requirement of test beds is that they support the manifestation and observation of interesting behaviors, which is why I favor complex agents and test beds over simple ones. In a *confirmatory* phase, we tighten up the characterizations of behaviors and test specific hypotheses. In particular, we provide an operational definition of, say, boldness so that a data-collecting computer program can observe the agent's behavior and decide whether it is bold. We test hypotheses about the conditions in which boldness is a virtue, and when we are done, we have a set of results that describe precise, test-bed–specific conditions in which a precise, agent-specific behavior is good or bad. In confirmatory research, the primary

requirement of a test bed is that it provide experimental control and make running experiments and collecting data easy. For this reason, PHOENIX has a script mechanism for automatically running experiments and integrated data-collection, data-manipulation, and statistical packages. In the third phase, *generalization*, we attempt to replicate our results. As I described earlier, several research groups might attempt to replicate bold behavior under conditions comparable to those in the original experiment. Each group will have to design their own agent-specific, test-bed–specific definitions of *bold* and *comparable conditions*. For example, uncertainty about the environment might be induced in agents by rapidly changing wind speed in PHOENIX and erratically moving holes in TILEWORLD. To achieve this goal, test beds would have to be parameterizable, and researchers woule have to work closely during the generalization phase.

The boldness theory is general to the extent that boldness and unpredictability in TILEWORLD are phenomena similar to boldness and unpredictability in PHOENIX and other test beds. Similar agents in similar test beds are apt to manifest similar behaviors, but this similarity does not convince us that the behaviors are general. Generality is achieved when different agents in different test beds exhibit common behaviors in common conditions. The more the agents and test beds differ, the more difficult it is to show that behaviors and conditions are common. If we had theories of behavior, we could show how conditions and behaviors in different test beds are specializations of terms in our theories. However, we do not have theories; we must bootstrap theories from empirical studies. Our only hope is to rely on our imaginations and abilities to interpret behaviors and conditions in different test-bed studies as similar.

In conclusion, I believe results of test-bed research can be generalized. Some features of test beds will make it easier to observe, explain, and test hypotheses about agents' behaviors. Generalization is done by scientists, not apparatus, so I strongly disagree with any implication that particular kinds of test beds preclude generalization. Test beds offer researchers the opportunity to tell each other what they observed in particular conditions. When a researcher publishes an observation, other researchers are responsible for the hard work required to say, I observed the same thing!

Acknowledgments

Steve Hanks was supported in part by National Science Foundation (NSF) grants IRI-9008670 and IRI-9206733. Martha E. Pollack was supported by United States Air Force Office of Scientific Research contracts F49620-91-C-0005 and F49620-92-J-0422, Rome Laboratory, the Advanced Research Projects Agency (ARPA) contract F30602-93-C-0038, and NSF Young Investigator's Award IRI-9258392. Paul Cohen was supported in part by ARPA contract F30602-91-C-0076.

Notes

1. The scoring metric in TILEWORLD was later revised to make it easier to compare trials of varying length: Raw score was replaced with a normalized value called efficiency (Kinny and Georgeff 1991). A number of changes have been made to the TILEWORLD system since 1990, some of which are discussed in The Promise of Experimentation; see also Pollack et al. (1993). Code and documentation for TILEWORLD are available by sending mail to tileworld-request@cs.pitt.edu.

2. TRUCKWORLD code and documentation are available by sending mail to truckworld-users-request@cs.washington.edu.

3. The filtering mechanism itself in the original TILEWORLD agent is trivial: When the agent is working on filling a hole, the filter rejects all other holes; when the agent does not have a current hole, the filter accepts all holes.

4. In both experiments, the agent was automatically and immediately notified of the appearance and disappearance of holes.

5. The terms can be defined precisely within the IRMA framework—they describe the sensitivity of the agent's filter-override mechanism—but presumably the terms and the associated relationships are intended to be applied to agents other than implementations of IRMA.

6. Compare McDermott (1981).

7. Chapman (1990) advances an even stronger view that randomness without structure actually makes planning more difficult.

8. For similar statements of this research paradigm, see Cohen, Howe, and Hart (1990); Rosenschein, Hayes-Roth, and Erman (1990); Pollack and Ringuette (1990); and Langley and Drummond (1990). Also see the paper by L. Chrisman, R. Caruana, and K. Carriker from the 1991 AAAI Fall Symposium Series on Sensory Aspects of Robotic Intelligence, "Intelligent Agent Design Issues: Internal Agent State and Incomplete Perception." Some researchers also split out the properties of the agent's task; in these comments, I consider the task specification to be part of the environment, but my argument does not depend on this consideration.

9. Although I believe the situation is changing; recent conference proceedings appear to include an increasing number of experimental papers on agent design, and in some other subfields of AI, notably machine learning and text understanding, there are many such papers.

10. A general question exists about the appropriate language for the researcher to use in articulating his/her theory. Sometimes, it will be the language of mathematics; other times a natural language, clearly used, can suffice.

11. Another error was the failure to provide a clean enough interface between the agent and the environment; it is more difficult than originally hoped to excise the IRMA-based embedded agent and replace it with an alternative. Also, as Hanks points out, we used an awkward mechanism, which has since been modified, for simulating concurrent acting and reasoning on a sequential machine.

12. If you don't believe me, I invite you to listen to the objections that are raised when I give talks describing IRMA.

13. An exploratory phase of experimentation can occur after initial attempts at verifying a particular theory and can sometimes look like fiddling, but this area is another matter.

14. Much of what I say arises from conversations with Bruce Porter of the University of Texas. Although I owe my current understanding of the issues to our discussions, I do not mean to imply that he agrees with everything here.

References

Agre, P., and Chapman, D. 1987. PENGI: An Implementation of a Theory of Activity. In Proceedings of the Sixth National Conference on Artificial Intelligence, 268–272. Menlo Park, Calif.: American Association for Artificial Intelligence.

Bond, A., and Gasser, L. 1988. *Readings in Distributed Artificial Intelligence.* Los Altos, Calif.: Morgan Kaufmann Publishers.

Bratman, M. 1987. *Intention, Plans, and Practical Reason*. Cambridge, Mass.: Harvard University Press.

Bratman, M.; Israel, J.; and Pollack, M. 1988. Plans and Resource-Bounded Practical Reasoning. *Computational Intelligence* 4:349–355.

Brooks, R. 1991. Intelligence without Reasoning. In Proceedings of the Twelfth International Joint Conference on Artificial Intelligence, 569–595. Menlo Park, Calif.: International Joint Conferences on Artificial Intelligence.

Chapman, D. 1990. On Choosing Domains for Agents. Position Paper presented at the NASA Ames Workshop on Benchmarks and Metrics, Moffett Field, California, June.

Chrisman, L., and Simmons, R. 1991. Senseful Planning: Focusing Perceptual Attention. In Proceedings of the Ninth National Conference on Artificial Intelligence, 756–761. Menlo Park, Calif.: American Association for Artificial Intelligence.

Cohen, P. 1991. A Survey of the Eighth National Conference on Artificial Intelligence: Pulling Together or Pulling Apart? *AI Magazine* 12:16–41.

Cohen, P.; Howe, A.; and Hart, D. 1990. Intelligent Real-Time Problem Solving: Issues and Examples. In *Intelligent Real-Time Problem Solving: Workshop Report,* ed L. Erman, IX-1–IX-33. Palo Alto, Calif.: Cimflex Teknowledge Corp.

Dean, T., and Boddy, M. 1988. An Analysis of Time-Dependent Planning. In Proceedings of the Seventh National Conference on Artificial Intelligence, 49–52. Menlo Park, Calif.: American Association for Artificial Intelligence.

Firby, R. J. 1989. Adaptive Execution in Complex Dynamic Worlds. Ph.D. diss., Dept. of Computer Science, Yale Univ.

Firby, R. J., and Hanks, S. 1987. A Simulator for Mobile Robot Planning. In Proceedings of the DARPA Knowledge-Based Planning Workshop, 23-1–23-7. Washington, D.C.: Defense Advanced Research Projects Agency.

Greenberg, M., and Westbrook, L. 1990. The PHOENIX Test Bed, Technical Report, COINS TR 90-19, Dept. of Computer and Information Science, Univ. of Massachusetts.

Haddawy, P., and Hanks, S. 1993. Utility Models for Goal-Directed Decision-Theoretic Planners, Technical Report, 93-06-04, Dept. of Computer Science and Engineering, Univ. of Washington.

Hanks, S. 1993. Modeling a Dynamic and Uncertain World II: Action Representation and Plan Evolution, Technical Report, 93-09-07, Dept. of Computer Science and Engineering, Univ. of Washington.

Hanks, S. 1990a. Practical Temporal Projection. In Proceedings of the Eighth National Conference on Artificial Intelligence, 158–163. Menlo Park, Calif.: American Association for Artificial Intelligence.

Hanks, S. 1990b. Projecting Plans for Uncertain Worlds. Ph.D. diss., Dept. of Computer Science, Yale Univ.

Hanks, S., and Badr, A. 1991. Critiquing the TILEWORLD: Agent Architectures, Planning Benchmarks, and Experimental Methodology, Technical Report, 91-10-31, Dept. of Computer Science and Engineering, Univ. of Washington.

Hanks, S., and McDermott, D. 1994. Modeling a Dynamic and Uncertain World I: Symbolic and Probabilistic Reasoning about Change. *Artificial Intelligence* 65(2). Forthcoming.

Hart, D., and Cohen, P. 1990. PHOENIX: A Test Bed for Shared Planning Research. In Proceedings of the NASA Ames Workshop on Benchmarks and Metrics, Moffett Field, California, June.

Kinny, D. 1990. Measuring the Effectiveness of Situated Agents, Technical Report 11, Australian AI Institute, Carlton, Australia.

Kinny, D., and Georgeff, M. 1991. Commitment and Effectiveness of Situated Agents. In Proceedings of the Twelfth International Joint Conference on Artificial Intelligence, 82–88. Menlo Park, Calif.: International Joint Conferences on Artificial Intelligence.

Kinny, D.; Georgeff, M.; and Hendler, J. 1992. Experiments in Optimal Sensing for Situated Agents. In Proceedings of the Second Pacific Rim International Conference on Artificial Intelligence, 1176–1182. Seoul, South Korea: Korean Information Science Society.

Langley, P., and Drummond, M. 1990. Toward an Experimental Science of Planning. In *Proceedings of*

the DARPA Workshop on Innovative Approaches to Planning, Scheduling, and Control, 109–114. San Mateo, Calif.: Morgan Kaufmann Publishers.

Law, A., and Kelton, W. 1981. *Simulation Modeling and Analysis.* New York: McGraw-Hill.

McDermott, D. 1981. Artificial Intelligence Meets Natural Stupidity. In *Mind Design: Essays in Philosophy, Psychology, and Artificial Intelligence*, ed. J. Haugland, 143–160. Cambridge, Mass.: The MIT Press.

Minton, S.; Johnston, M.; Philips, A.; and Laird, P. 1990. Solving Large-Scale Constraint Satisfaction and Scheduling Problems Using a Heuristic Repair Method. In Proceedings of the Ninth National Conference on Artificial Intelligence, 17–24. Menlo Park, Calif.: American Association for Artificial Intelligence.

Montgomery, T., and Durfee, E. 1990. Using MICE to Study Intelligent Dynamic Coordination. In Proceedings of the Second International Conference on Tools for Artificial Intelligence, 438–444. Washington, D.C.: Institute of Electrical and Electronics Engineers.

Montgomery, T.; Lee, J.; Musliner, D.; Durfee, E.; Darmouth, D.; and So, Y. 1992. MICE Users Guide, Technical Report, CSE-TR-64-90, Dept. of Electrical Engineering and Computer Science, Univ. of Michigan.

Moore, D., and McCabe, G. 1989. *Introduction to the Practice of Statistics.* New York: W. H. Freeman and Company.

Nguyen, D.; Hanks, S.; and Thomas, C. 1993. The TRUCKWORLD Manual, Technical Report, 93-09-08, Dept. of Computer Science and Engineering, Univ. of Washington.

Philips, A., and Bresina, J. 1991. NASA TILEWORLD Manual, Technical Report TR-FIA-91-04, NASA Ames Research Center, Mountain View, California.

Philips, A.; Swanson, K.; Drummond, M.; and Bresina, J. 1991. A Design Rationale for NASA TILEWORLD, Technical Report FIA-91-04, AI Research Branch, NASA Ames, Moffett Field, California.

Pollack, M. 1992. The Uses of Plans. *Artificial Intelligence* 57(1): 43–69.

Pollack, M. 1991. Overloading Intentions for Efficient Practical Reasoning. *Nous* 25(4): 513–536.

Pollack, M., and Ringuette, M. 1990. Introducing the TILEWORLD: Experimentally Evaluating Agent Architectures. In Proceedings of the Eighth National Conference on Artificial Intelligence, 183–189. Menlo Park, Calif.: American Association for Artificial Intelligence.

Pollack, M.; Joslin, D.; Nunes, A.; and Ur, S. 1993. Experimental Investigation of an Agent Design Strategy, Technical Report, Dept. of Computer Science, Univ. of Pittsburgh. Forthcoming.

Powers, R. 1991. *The Gold Bug Variations.* New York: William Morrow and Company.

Rosenschein, S.; Hayes-Roth, B.; and Erman, L. 1990. Notes on Methodologies for Evaluating IRTPS Systems. In *Intelligent Real-Time Problem Solving: Workshop Report*, ed. L. Erman, II-1–II-12. Palo Alto, Calif.: Cimflex Teknowledge Corp.

Russell, S., and Wefald, E. 1991. *Do the Right Thing: Studies in Limited Rationality.* Cambridge, Mass.: The MIT Press.

Schoppers, M. 1987. Universal Plans for Reactive Robots in Unpredictable Environments. Intelligence without Reasoning. In Proceedings of the Tenth International Joint Conference on Artificial Intelligence, 1039–1046. Menlo Park, Calif.: International Joint Conferences on Artificial Intelligence.

Sussman, G. 1975. *A Computer Model of Skill Acquisition.* New York: Elsevier Science Publishing Company.

Weld, D., and deKleer, J. 1989. *Readings in Qualitative Reasoning about Physical Systems.* Los Altos, Calif.: Morgan Kaufmann Publishers.

Wellman, M., and Doyle, J. 1991. Preferential Semantics for Goals. In Proceedings of the Ninth National Conference on Artificial Intelligence, 698–703. Menlo Park, Calif.: American Association for Artificial Intelligence.

Steve Hanks is an assistant professor of computer science and engineering at the University of Washington. He received a Ph.D. from Yale University in 1990 and an M.B.A. from the Wharton School in 1983. His research interests include automated planning and decision making, reasoning and decision making under uncertainty and with incomplete information, and decision support systems for complex and subjective domains.

Martha E. Pollack is associate professor of computer science and intelligent systems at the University of Pittsburgh. She received her Ph.D. from the University of Pennsylvania in 1986 and was employed at the AI Center, SRI International, from 1985 to 1991. A recipient of the Computers and Thought Award (1991) and a National Science Foundation Young Investigator's Award (1992), Pollack has research interests in resource-limited reasoning, plan generation and recognition, natural language processing, and AI methodology.

Paul R. Cohen is an associate professor of computer science at the University of Massachusetts at Amherst and director of the Experimental Knowledge Systems Laboratory. He received his Ph.D. from Stanford University in computer science and psychology in 1983. At Stanford, Cohen edited the *Handbook of Artificial Intelligence, Volume III*, with Edward Feigenbaum and Avron Barr and recently finished editing volume IV with them. He is currently completing a book entitled "Empirical Methods for Artificial Intelligence."