Chapter 9

Scientific Discovery Processes in Children, Adults, and Machines

David Klahr
Carnegie Mellon University

Allen Newell's primary research goals were so fundamental, and his accomplishments so prodigious, that two fields—cognitive psychology and artificial intelligence (AI)—trace their ancestry to his pioneering work with Herbert Simon in the late 1950s. Newell and Simon started investigating the nature of intelligence by using very simple domains: closed-form games and puzzles. Forty years later, their successes are indicated by the fact that we are able to use the current versions of the methodologies and theories that they invented to investigate the cognitive processes that support scientific discovery: a domain that represents one of the pinnacles of human intelligence. The early tensions and mutual interactions between psychological approaches and artificial intelligence approaches remain in the studies of scientific discovery: In psychology, the research goal is to determine just how people manage to do science, whereas in AI the goal is to build systems that can make discoveries. This work has produced an accumulating body of evidence that there can really be a "science of science." As a result, the old view of scientific discovery—that it is mystical, ineffable, transcendent, unknowable—is giving way to both a descriptive and a synthetic science of discovery. The descriptive side is mainly from cognitive psychology, and the synthetic side is mainly from machine learning. Early interest in the psychology of science can be traced to Bruner, Goodnow, and Austin (1956), Wason (1960), and Simon (1966, 1973), among others. The state of the art as of a dozen years ago is summarized in Tweney, Doherty, and Mynatt (1981). The more recent resurgence of interest in the "cognitive science of science" can be attributed to Simon and his colleagues (Cheng & Simon, 1992; Kulkarni & Simon, 1988; Langley, Simon, Bradshaw, & Zytkow, 1987; Qin & Simon, 1990; Valdez-Perez, Simon, & Murphy, 1992). But psychologists were not the first, nor the only, scientists to argue for the ultimate knowability of the process of scientific discovery More than 50 years ago, Einstein wrote: "The whole of science is nothing more than a refinement of every day thinking" (Physics & reality, 1936, reprinted in Einstein, 1950, p. 59). He also wrote, "The scientific way of forming concepts differs from that which we use in our daily life, not basically, but merely in the more precise definition of concepts and conclusions; more painstaking and systematic choice of experimental material, and greater logical economy ("The common language of science," 1941, reprinted in Einstein, 1950, p. 98).

ED-?
economy"

So the basic premise—that scientific thinking involves some of the same processes used by ordinary folks—is not new. What *is* new is what we have learned in recent years about the psychological process underlying scientific discovery: the "precise definitions," "systematic choices," and "logical economy" of which Einstein speaks. These are the processes that empower scientific discovery, and that is what I address in this chapter.

There are five parts to this chapter: In the first part, I describe a framework for characterizing the discovery process. Next, I describe the psychological processes used by adults and children when they are engaged in scientific discovery. I summarize the results of empirical studies in my lab, as well as a few studies by others who have also been looking at developmental differences in scientific discovery processes. In the third part of the chapter, I say a bit about machine discovery systems. These systems continue the two-faceted approach that manifested itself in the earliest days of artificial intelligence. Some are computational models of human discovery processes, whereas others are a species of machine learning systems designed to support scientific discovery by machines. In the fourth section I attempt to characterize different approaches within cognitive science to understanding discovery, and finally, in the fifth part, I talk about the frontiers of this research.

SCIENTIFIC DISCOVERY AS DUAL SEARCH

Our research is based on the idea that scientific discovery is a type of problem solving in which there are two problem spaces: a space of hypotheses and a space of experiments. Both of these problem spaces require heuristics for constraining search. This dual search notion is an extension of Simon and Lea's (1974) generalized rule inducer. In our model, hypotheses correspond to GRI's rules, and experiments correspond to instances. In order further specify this very general characterization, Kevin Dunbar and I proposed a framework that we called SDDS, for "scientific discovery as dual search." The

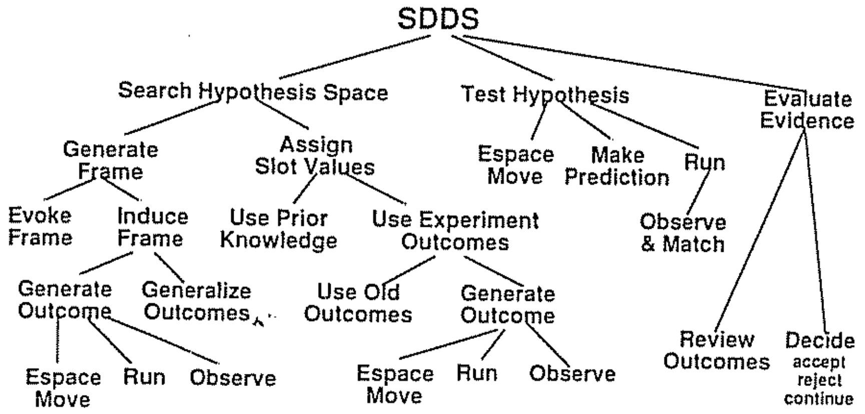


FIG. 9.1 SDDS framework

SDDS framework (Fig. 9.1) depicts the relationship among the component processes that coordinate the dual search (Klahr & Dunbar, 1988).

The three top-level components control the entire process: First, you have to Search the Hypothesis Space, then you have to Test that Hypothesis, and then you have to Evaluate the Evidence in order to decide whether the cumulative evidence—as well as other considerations—warrants acceptance, rejection, or continued consideration of the current hypothesis. That is a pretty conventional view of the discovery process. Now let's descend a level, and look more carefully at how hypotheses are generated.

There are two subcomponents for Search Hypothesis Space. One component generates the broad scope for the hypotheses, and the second component refines it and further specifies it. Because we use Minsky's "frame" notion, for representing hypotheses, we show this as first Generating a Frame and then Assigning Slot Values.

Where do these initial frames and their associated slot values come from? We propose two different types of sources for new hypotheses. One source is prior knowledge stored in memory, and the other source is the external world. The two different sources are evoked in both Generate Frame and in Assign Slot Values.

Generate Frame has two subcomponents corresponding to the two ways that a frame may be generated.

Evoke Frame is a search of memory for information that could be used to construct a frame. Prior knowledge plays an important role here. In cognitive psychology, several mechanisms have been proposed to account for the way in which initial hypotheses are generated. These include memory search, analogical mapping, remindings, and view instantiation (Dunbar & Schunn, 1990; Gentner, 1983; Gick & Holyoak, 1983; Klahr & Dunbar, 1988; Ross, 1984; Shrager, 1987). Each of these mechanisms emphasizes a different aspect of the way in which search in the hypothesis space is initiated. Although the SDDS framework doesn't have anything to add to these views, there is an important distinction between this form of frame generation and the other process under the Generate Frame node: Induce Frame.

Induce Frame generates a new frame by induction over a series of outcomes (Holland, Holyoak, Nisbett, & Thagard, 1986). It includes two subprocesses: The first Generates an Outcome, and the second Generalizes over the results of that (and other) outcomes to produce a frame.

This first process is of particular interest, for it calls for an experiment to be run (via E-Space Move). But this is an odd sort of experiment: It is not testing any hypothesis, because we don't have one yet. We are still in the part of the model that is searching for a hypothesis.

By including E-Space Move in this portion of the framework, we acknowledge the importance of running so-called "experiments" in the absence of a clear theory. This corresponds to pretheoretical observations and measurements of how one thing affects another with no clear-cut theory. This is not the conventionally assigned role for experimentation, but we all know how important it is.

Notice also that the E-Space Move occurs in two additional parts of the framework: not only in the service of inducing a frame, but also under Assign Slot Values, when the theory has been partially specified, and one is seeking a bit more constraint on the theory. E-Space Move also occurs under Test Hypothesis in its "traditional role" in the evaluation of fully specified hypotheses. By calling this a "move" in a "space" we emphasize the fact that the decision about what kind of data to collect, or what kind of observation to make in a problem-solving task that requires constrained search in a very large space.

If we move back up to the distinction between Generate Frame and Assign Slot Values, we can see that the two processes correspond to major and minor moves in the hypothesis space. Generate Frame involves the creation of a new hypothesis that may involve entirely new structural relations among its elements, whereas Assign Slot Values takes the structure—the frame, that is—as given, and refines some of its unresolved elements.

Once again, the location of E-Space Move reflects the fact that much of the experimentation that takes place within a paradigm is not of the grand hypothesis testing type, but rather the more data-driven attempt to induce a new theory, or to refine an existing theory by discovering a better set of slot values.

Let me summarize the main points of our theoretical orientation. Scientific discovery is comprised of three main components:

1. *Searching the hypothesis space.* The process of generating new hypotheses is a type of problem solving in which the initial state consists of some

knowledge about a domain, and the goal state is a hypothesis that can account for some or all of that knowledge in a more concise or universal form. Once generated, hypotheses are evaluated for their initial plausibility. Expertise plays a role here, as subjects' familiarity with a domain tends to give them strong biases about what is plausible in the domain. Plausibility, in turn, affects the order in which hypotheses are evaluated: Highly likely hypotheses tend to be tested before unlikely hypotheses (Klayman & Ha, 1987; Wason, 1968). Furthermore, subjects may adopt different experimental strategies for evaluating plausible and implausible hypotheses.

2. *Searching the experiment space.* One of the most important constraints on this search is the need to produce experiments that will yield interpretable outcomes. For human discovery systems, this requires domain-general knowledge about one's own information-processing limitations, as well as domain-specific knowledge about the pragmatic constraints of the particular discovery context. As we will see, there are important developmental differences in people's ability to constrain search in the experiment space.

3. *Evaluating evidence.* In contrast to the binary feedback provided to subjects in the typical psychology experiment, real-world evidence evaluation is not very straightforward. Relevant features must first be extracted, potential noise must be suppressed or corrected, and the resulting internal representation must be compared with earlier predictions. When people are reasoning about real world context, their prior knowledge imposes strong theoretical biases. These biases influence not only the initial strength with which hypotheses are held—and hence the amount of disconfirming evidence necessary to refute them—but also the features in the evidence that will be attended to and encoded.

Each of these three components is a potential source of developmental change, and most psychologists have studied them in isolation. But such decomposition begs the very question of interest: the coordination of search in two spaces. We wanted to try a different approach. We wanted to study discovery behavior in situations that required coordinated search in both the experiment space and the hypothesis space. In order to do this, we set up laboratory situations that were designed to place subjects in various parts of this framework and then looked at how they managed the dual search process.

LABORATORY INVESTIGATIONS OF SCIENTIFIC REASONING

Given this goal of studying scientific reasoning in the psychology lab, and given our additional goal of addressing some developmental questions, and inspired by earlier work with Jeff Shrager (Shrager & Klahr, 1986) we decided to study scientific discovery by using the device shown in Fig. 9.2a.

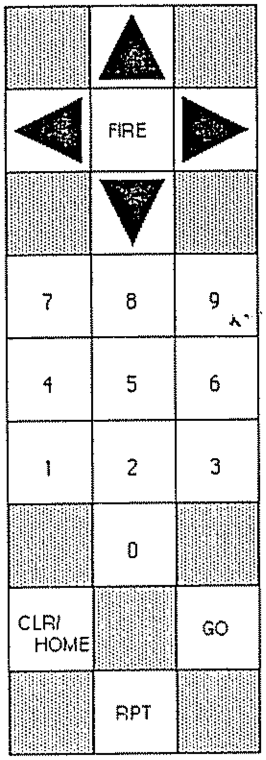


FIG 9 2 BigTrak keypad.

We used a computer-controlled toy robot tank called BigTrak. It is a battery-operated programmable self-contained vehicle, about 2 ft long. The functions on the BigTrak keypad interface are depicted in Fig. 9.2. The basic execution cycle involves first clearing the memory with the CLR key and then entering a series of up to 16 instructions, each consisting of a function key (the command) and a one- or two-digit number (the argument). The five command keys are ↑, move forward; ↓, move backward; ←, rotate left; →, rotate right; and FIRE. When the GO key is pressed, BigTrak executes the program. For example, suppose you pressed the following series of keys:

CLR ↑5 ← 7 ↑ 3 → 15 FIRE 2 ↓ 8 GO

When the GO key was pressed, BigTrak would move forward 5 ft, rotate counterclockwise 42 degrees (corresponding to 7 minutes on an ordinary clock face), move forward 3 ft, rotate clockwise 90 degrees, fire (its "laser cannon") twice, and back up 8 ft.

ED: 1 LINE SHORT

Procedure

Our procedure had three phases. In the first, subjects were introduced to BigTrak and instructed on the use of each basic command. Subjects were instructed in how to generate verbal protocols. During this phase, the RPT key was not visible. Subjects were trained to criterion on how to write a series of commands to accomplish a specified maneuver. The end of this phase corresponded to a scientist having a basic amount of knowledge about a domain.

In the second phase, subjects were shown the RPT key. They were told that it required a numeric parameter (N), and that there could be only one RPT N in a program. They were told that their task was to find out how RPT worked by writing programs and observing the results. This corresponded to a new problem in the domain: an unresolved question in an otherwise familiar context.

Finally, in the third phase, subjects could formulate hypotheses about RPT and run experiments to test those hypotheses. This required decisions about hypotheses and decisions about experiments. Subjects were never told whether or not they had discovered how RPT worked. They had to decide when to terminate search.

The task has several properties that make it appropriate for studying scientific discovery in the laboratory:

1. Prior knowledge can influence initial hypotheses as well as the strength with which subjects hold them.
2. Subjects have to design and evaluate their own experiments.
3. The mapping between experimental outcomes and hypotheses is non trivial.
4. We do not tell subjects whether or not they have in discovered a true hypothesis That is for them to decide.
5. The task is interesting and challenging for a wide range of ages.

ED - ?
non trivial.

Hypothetical Behavior

What would you do, if faced with this problem? What kind of scientific reasoning would *you* use, if asked to figure out how the RPT key worked? Figure 9.3 shows a hypothetical sequence of hypotheses, predictions, and experimental outcomes.

Hypothesis x (Hx) says that when you put in a RPT and a number, the whole program repeats that many times. That turns out to be a very popular hypothesis. Many subjects start out with this one, or something very similar to it.

ED: 1 LINE SHORT

Your first BigTrak Experiments

Hx: RPT N repeats the entire program N times.
Hy: RPT N repeats the Nth step once.

E1: ↑1 RPT 1	
Hx prediction:	↑2
Hy prediction:	↑2
BT's behavior:	↑2

E2: ↑1 FIRE 2 ↓ 1 RPT 2	
Hx prediction:	↑1 FIRE 2 ↓1 ↑1 FIRE 2 ↓ 1
Hy prediction:	↑1 FIRE 2 ↓1 FIRE 2
BT's behavior:	↑1 FIRE 2 ↓1 FIRE 2 ↓ 1

FIG 9.3 Hypothetical behavior in BigTrak task

Hypothesis y (Hy) is a little odd: it says that RPT N takes the Nth step in the program and repeats it one more time. That is not a very popular hypothesis. Very few subjects start out with it.

What about experiments? What kind of program would you write in order to test your hypotheses? Suppose you want to start simple, just to see what might happen. So you write Experiment 1: (↑1 RPT 1). Although this is a simple and easy to observe experiment, it is not very informative, because if Hx is right, BigTrak will go forward two times, but it will do the same thing if Hy is right. So this experiment can't discriminate between the two hypotheses. (This experiment might not be a total loss if BigTrak did something inconsistent with both hypotheses, but it doesn't.)

How about Experiment 2: ↑ 1 FIRE 2 ↓ 1 RPT 2.

Now the two hypotheses make distinctive predictions:

Hx predicts ↑ 1 FIRE 2 ↓ 1 ↑ 1 FIRE 2 ↓ 1.
Hy predicts ↑ 1 FIRE 2 ↓ 1 FIRE 2.

So Experiment 2 is critical with respect to the two hypotheses. It also has some nice properties; is pretty short, so you can keep track of what is going on, and it has easily distinguishable components, so each piece of behavior is highly informative.

So you enter the program shown in E2, and you run it. BigTrak goes like this:

↑ 1 FIRE 2 ↓ 1 FIRE 2 ↓ 1

Which is not what either theory predicted.

Now you have to look carefully at the behavior, and, if you are very discerning, you notice that it repeated the last two steps. You also notice that you used a 2 as the value of N. If you are really on the ball here, you hypothesize that RPT N repeats the last N instructions one time. And that's the way the original BigTrak really worked.

So now you have discovered how RPT works: It repeats the last N instructions one time. And you did it with only three hypotheses and two experiments.

Performance: Adults Versus Children

How did our subjects do? In one of our studies (Dunbar & Klahr, 1989) we used two groups of subjects: Carnegie Mellon University (CMU) undergraduates, and children between the ages of 8 and 11 years. Table 9.1 shows the overall results. Recall that the RPT key takes the N instructions preceding the RPT instruction and it repeats that sequence one more time. It's a pretty nonintuitive function, and it was not easy to discover.

Children's success rate was very low. Only 2 of 22 children were successful, although 12 of the unsuccessful children were sure they had discovered the correct rule, and they terminated their experiments quite satisfied with their discovery. In contrast, nearly all of the adults discovered the correct rule. But it was not a trivial task for them. In fact, with respect to average time, number of hypotheses, and number of experiments, the adults were not very different from the children. The explanation for these vastly different success rates must lie at a deeper level. We need to look more closely at the nature of the hypothesis space and the experiment space.

Table 9.2 lists the more common hypotheses that subjects proposed in order of decreasing popularity or plausibility. Recall that the correct rule is number 5: Repeat the last N steps once. On the right side of the table

TABLE 9.1
Overall Performance of Children and Adults on BigTrak Task

	Adults	*Children*
Solvers	19 of 20	2 of 22
Mean time	20 min	20 min
Number of hypotheses	4.6	3.3
Number of programs	18	13

TABLE 9.2
Common Hypotheses (in Decreasing Order of Popularity or Plausibility)

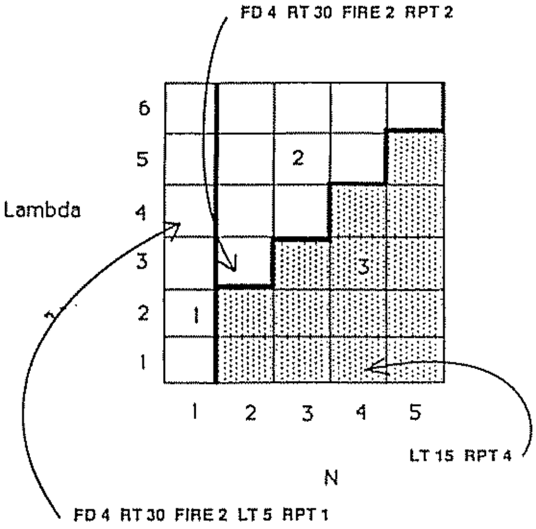
RPT N tells BigTrak to:	*Role of N*
Repeat the entire program N times	Counter
Repeat the last step N times	Counter
Repeat the subsequent steps N times	Counter
Repeat the entire program once	nil
Repeat the last N steps once	Selector
Repeat the Nth step once	Selector
Repeat the first N steps once.	Selector
Repeat the entire program *f*(N) times.	Counter

hypotheses are classified according to the role that they assign to the parameter that goes with the RPT command, shown here as "the role of N." In hypotheses 1, 2, 3, and 8, N counts the number of repetitions. We call these Counter hypotheses. In hypotheses 5, 6, and 7, N determines which segment of the program will be selected to be repeated again. We call these Selector hypotheses. This distinction, between Counters and Selectors, turns out to be a very useful distinction in our subsequent experiments. Search in the BigTrak hypothesis space can involve local search among Counters or among Selectors, or it can involve more far-ranging search between counter frames and selector frames.

What about the BigTrak experiment space? How can we characterize it? By one reckoning, it is enormous: For example, there are over 5^{15} distinct programs that subjects could write. However we have found that we can adequately characterize the experiment space in terms of just two parameters. The first is λ—the length of the program preceding the RPT. The second is the value of N—the argument that RPT takes. Because both parameters must be less than or equal to 15, there are 225 "cells" in the λ-N space. The regions and their general properties are depicted in Fig. 9.4.

We have divided the E-space into three regions, according to their general informativeness with respect to alternative hypotheses. Rather than go into details, I'll just remind you about the earlier example in which some experiments were very poor at distinguishing among competing theories, whereas others were very effective. The most important thing to note is that region 2 is particularly informative. This is where the program length is greater than the value of N.

This analysis of the H-space and the E-space revealed a couple of interesting things about how subjects went about this task. We found that there were two distinct types of subjects, with fundamentally different strategies. We distinguished between the two groups on the basis of how much information they had when they changed from a counter frame to a selector frame. If they made the switch without having seen the result of a region



Region 1: Poor discriminating power.

Region 2: Maximally discriminates among all of the common hypotheses. Can distinguish selectors from counters, and *which* selector or counter.

Region 3: Confusing for selector rules, because N gets truncated to λ and a sequence of experiments that varies N has no effect.

FIG. 9.4. BigTrak experiment space

2 experiment, then we called them "Theorists," because they could not have based their decision on conclusive experimental evidence. On the other hand, if they made the switch from Counters to Selectors only after running region 2 experiments, then we called them "Experimenters." (By the way, this analysis only makes sense for the adults, because so few children proposed selectors.)

The two kinds of strategies were accompanied by other differences (shown in Table 9.3). Experimenters took twice as long to discover how RPT worked; they explored much more of the experiment space, and they conducted many more experiments without any active hypothesis. That is, they spent a lot of time down in the lower left-hand region of the SDDS framework, as they ran experiments in order to generate a data pattern over which they could induce a frame.

TABLE 9.3
Performance Differences Between Theorists and Experimenters on BigTrak Task

	Theorists	*Experimenters*
Defining property	State selector frame without sufficient evidence	State selector frame only after sufficient evidence
Time (min)	11	25
Total experiments	9.3	18.4
Experiments without hypotheses	0.8	6.1
Comments about experiment space	5.9	0.9
E-space cells used	5.7	9.9

This tendency to suspend the hypotheses testing mode while attempting to discover some kind of regularity in the data suggested to us that we needed to find out a lot more about how subjects searched the experiment space, and about how different goals might influence that search. We also began to look at developmental differences in some of the key components of the SDDS model. As a developmentalist, I was particularly interested in addressing two long-standing disputes over the developmental course of scientific reasoning skills:

1. The *domain-specific* or *domain-general* debate asks whether there are any general, domain-independent rules used in scientific reasoning, or whether all developmental improvements can be attributed to domain-specific acquisitions. Of course, the question is not limited to scientific reasoning skills: it pervades all of cognitive development. It is analogous to the distinction in AI between *weak methods* and *knowledge-rich* approaches. Like most of the dichotomies in psychology, this one should not be overemphasized, because it is not a clear-cut distinction. But developmentalists devote a lot of energy to arguing about it, so I wanted to address it in my work.

2. The *child-as-scientist* debate asks whether or not it makes sense to describe the young child as a scientist. Some folks say, "yes, of course," and others say, "obviously not." Unfortunately, one can find empirical support for each position. On the one hand, results of formal studies, as well as abundant everyday experience, provide evidence that trained scientists, and even untrained adults, commonly outperform children on a variety scientific reasoning tasks (Kuhn, 1989). On the other hand, the empirical literature on scientific reasoning shows that adults demonstrate systematic and serious flaws in their reasoning, whereas young children are capable of surprisingly competent reasoning about hypotheses testing and experimentation (Brewer & Samarapungavan, 1991; Schauble, 1990; Vosniadou & Brewer, 1992).

It is clear that a one-bit answer to either of these questions will be inadequate. The questions have to be addressed in more depth We decided to recast the questions in terms of the components of the SDDS framework. In particular, we decided to use the BigTrak paradigm in such a way that we could focus on developmental differences in the heuristics used to constrain search in the experiment space.

The BT Microworld

For this study we moved from the original BigTrak toy to a computer microworld called BT. The toy tank became an animated "rocket ship" icon, and the BigTrak keypad became a screen display activated by pointing and clicking with a mouse (see Klahr, Fay, & Dunbar, 1993, for details). We explored the effect of domain-specific knowledge by manipulating the plausibility of hypotheses. Our goal was to investigate the extent to which prior knowledge—as manifested in hypothesis plausibility—influenced how people designed experiments and how they interpreted the results of those experiments.

Procedure. The study had three phases. The first and third phases were the same as in the previous study. Subjects learned about all the normal keys and were trained to criterion on getting BT to move around the screen. In the second phase, the RPT key was introduced as before. Subjects were told that their task was to find out how RPT worked by writing at least three programs and observing the results. But then we changed the procedure a bit, by suggesting one way that RPT might work. The experimenter said: "One way that RPT might work is": and then we stated one of four hypotheses listed next. Then we told subjects to write at least three programs to see if the repeat key really did work the way we had suggested, or some other way. The entire session lasted approximately 45 minutes.

ED - ?
work is ":

Throughout the study, we used only four rules for BT. The two popular, or plausible hypotheses were the two Counters:

A: Repeat the entire program *N* times.
B: Repeat the last step *N* times.

In contrast, there were two hypotheses that subjects were unlikely to propose. These are the two Selectors:

C: Repeat the *N*th step once.
D: Repeat the last *N* steps once.

TABLE 9.4
Design of BT Experiment:
Specific Hypotheses for Each Given–Actual Condition

Given Hypothesis	*Actual Rule*	
	Counter	*Selector*
Counter	B: Repeat last step N times ↓ A: Repeat entire program N times THEORY REFINEMENT	A: Repeat entire program N times. ↓ D: Repeat the last N steps once THEORY REPLACEMENT
Selector	D: Repeat the last step N steps once ↓ A: Repeat entire program N times. THEORY REPLACEMENT	C: Repeat step N once. ↓ D: Repeat the last N steps once THEORY REFINEMENT

Design. The design is shown in Table 9.4. We provided each subject with an initial hypothesis about how RPT might work. The Given hypothesis was always wrong. BT was always set to work according to some rule other than the Given rule. We called that the Actual rule. Both the Given and Actual could be either plausible (i.e., a Counter) or implausible (i.e., a Selector). In Counter → Counter and Selector → Selector conditions, the Given hypothesis was only "somewhat" wrong, in that it was from the same frame as the way that RPT actually worked. In Counter → Selector and Selector → Counter conditions, the Given was "very" wrong, in that it came from a different frame than the Actual rule. The subjects' task in the former situation corresponded to theory refinement, whereas in the latter situation it corresponded to theory replacement.

Subjects

We used four different groups of subjects, Carnegie Mellon (CM) undergraduates, Community College (CC) students, "sixth" graders (a mixed class of fifth to seventh graders, mean age 11 years), and third graders (mean age 9 years). CMs were mainly science or engineering majors, whereas the CCs had little training in mathematics or physical sciences. Children came primarily from academic and professional families. Most of the third graders had about 6 months of LOGO instruction. Note that CCs had less programming experience than the third graders.

Results. The proportion correct for each group in each condition is shown in Fig. 9.5. As we expected, domain-specific knowledge—manifested in expectations about what "repeat" might mean in this context—played an important role. Regardless of what the Given hypothesis was, subjects found it easier to discover Counters (81%) than Selectors (35%).

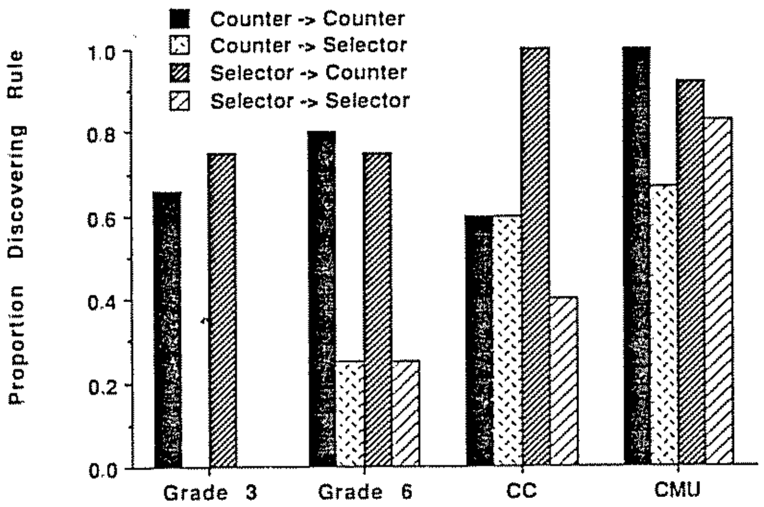


FIG. 9.5 Percentage correct.

There was also a main effect for group: The correct rule was discovered by 83% of the CMs, 65% of the CCs, 53% of the sixth graders, and 33% of the third graders. This group effect is attributable to the Actual = Selector conditions, in which 56% of the adults but only 13% of the children were successful. In fact, none of the third graders discovered Selectors. For Counters, adults and children were not as different in their success rates (88% vs. 75%).

What about subjects' reactions to the Given hypothesis? Recall that we presented subjects with either plausible or implausible hypotheses in order to determine the extent to which search in the hypothesis space was influenced by plausibility. This is one of the points at which domain-specific knowledge (which determines plausibility) might affect domain-general knowledge about experimental strategies.

Prior to running the first experiment, subjects were asked to predict what would happen. Their predictions indicated the extent to which they understood and accepted the Given hypotheses. Each subject's response to the Given hypothesis was assigned to one of three categories: I, Accept the Given hypotheses; II, accept the Given, but also propose an alternative; III, reject the Given, and propose an alternative. The number of subjects in each category is shown in Table 9.5 as a function of grade level and type of Given hypothesis. In both conditions, the adults always accepted the Given hypothesis, either on its own (category I), or in conjunction with an alternative that they proposed (category II). Adults never rejected the Given hypothesis. In contrast, no third grader and only two sixth graders ever proposed an alternative to compare to the Given (category II). Instead, children consid-

TABLE 9.5
Subjects' Responses to the Given Hypothesis

	Adults		*Children*	
Response to Given Hypothesis	*Counter*	*Selector*	*Counter*	*Selector*
Accept Given	.70	60	.71	.33
Accept Given and propose alternative	.30	.40	06	.06
Reject given; propose alternative	0	0	.23	.61

ED-?

one hypothesis

ered only one hypotheses at a time. When given Counters, they mainly accepted them, but when given Selectors, they mainly rejected them and proposed an alternative, which was usually a Counter of their own design.

This propensity to consider multiple versus single hypotheses affected the type of experimental goals set by the subjects. These goals, in turn, were used to impose constraints on search in the experiment space. We looked at these goals more closely by analyzing (a) what subjects said about experiments and (b) the features of the experiments that they actually wrote. Subjects' verbal protocols contain many statements indicating both explicit understanding of the experiment space dimensions, as well as what might be called a general notion of "good instrumentation": designing interpretable programs containing easily identifiable markers. Subjects made explicit statements about both kinds of knowledge. Here are some typical adult statements:

1. "I don't want to have two of the same move in there yet, I might not be able to tell if it was repeating the first one or if it was doing the next part of my sequence."
2. "I'm going to use a series of commands that will . . . that are easily distinguished from one another, and won't run it off the screen."
3. "So I'm going to pick two [commands] that are the direct opposite of each other, to see if they don't really have to be direct opposites but I'm just going to write a program that consists of two steps, that I could see easily."

Sixth graders were somewhat less articulate, but still showed a concern for both experiment space dimensions and program interpretability. In contrast, third graders rarely made such comments. The proportion of subjects making such comments is shown in the top row of Table 9.6.

At a finer level of detail, good instrumentation was assessed by the extent to which subjects observed three pragmatic constraints: (a) using standard units of rotation, such as 15 or 30 "minutes" (90 and 180 degrees), for rotate commands; (b) using small numeric arguments (values <5) on move commands, so that the actions of BT are not distorted by having it hit the

TABLE 9.6
Proportion of Self-Generated Constraints

	Student Group			
Constraint	*CMU*	*CC*	*Sixth*	*Third*
Explicit λ-N comments	83	.60	53	.20
Standard turn units	.92	.95	71	.53
Small arguments	.92	.85	65	.47
Proportion of programs in small[a] E-space region	50	63	.26	31

[a] 4 × 3 = 5% of 15 × 15

boundaries of the screen; and (c) using distinct commands in a program where possible. Programs constrained in these ways produce behavior that is easier to observe, encode, and remember. For both turns and moves, there was a strong effect of grade level.

Another interesting difference between the children and the adults was the way in which adults limited their search to a small "corner" of the experiment space. We looked at the section of the E-space with λ between 1 and 4, and N between 1 and 3. This corresponds to only 5% of the full E-space. But we discovered that over half of the adults' experiments occurred within this small area. On the other hand, children's experiments were much more scattered throughout the space.

Overall, Table 9.6 shows us that both what subjects said and what they did produced different patterns for the different groups: Older subjects—even those with weak technical backgrounds—were better able than children to constrain their search in the experiment space and to design interpretable experiments.

What were subjects trying to do here? What were their experimental goals? How can we infer these goals from the kinds of experiments they ran? We reasoned as follows: If the experimental goal is to identify which of the program steps are repeated for Selector hypotheses, or to discriminate between Selectors and Counters, then subjects should write programs having more than N steps (i.e., with $\lambda > N$). (In programs where λ is several steps greater than N, it is easy to distinguish among repeats of all steps, first step, last step, and N steps.) On the other hand, if the goal is to demonstrate the effect of a Counter, then subjects should use larger values of N and (for pragmatic reasons) relatively short programs (i.e., programs with $\lambda \leq N$). This all works out to a prediction about the conditions under which λ should be greater than N. Figure 9.6 shows the proportion of subjects in each condition whose first programs had $\lambda > N$. Responses of both of the adult groups and the sixth graders were consistent with the normative account I just gave. Third graders showed the opposite pattern.

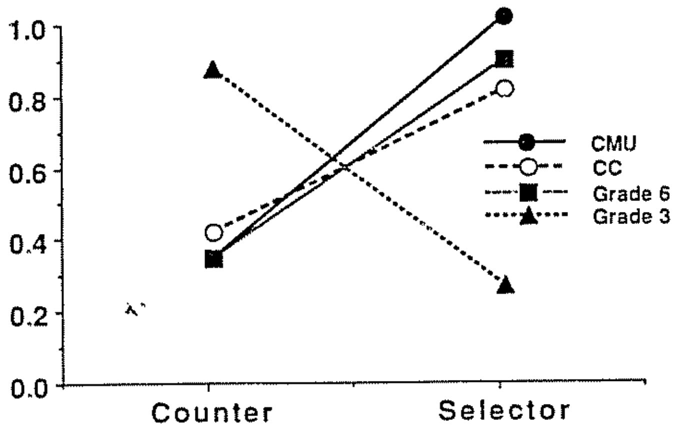


FIG. 9.6. λ-λ' on first experiment.

Heuristics for Constraining Search in the Experiment Space

These analyses reveal a distinctive pattern of results that differentiate the different groups. Our interpretation of these patterns is that they result from a set of domain-general heuristics that allow some subjects to constrain their search in the experiment space. These heuristics are differentially available to children and adults. Based on these and other analyses, we have proposed the following four heuristics:

E1. *Focus on one dimension of an experiment or hypothesis.* An incremental, conservative approach has been found to be effective in both concept attainment and hypothesis testing. This heuristic suggests that in moving from one experiment to the next or one hypothesis to the next one should decide upon the most important features of each and focus on just those features. Here, the CM adults stood apart from the other three groups. They were much more likely than any of the three other groups to make conservative moves—that is, to minimize differences in program content between one program and the next.

E2. *Use the plausibility of a hypothesis to choose experimental strategy.* In this study, we found that both children and adults varied their approach to confirmation and disconfirmation according to the plausibility of the currently held hypothesis. When hypotheses were plausible, subjects at all levels tended to set an experimental goal of demonstrating key features of the given hypothesis, rather than conducting experiments that could discriminate between rival hypotheses.

For implausible hypotheses, adults and young children used different strategies. Adults' response to implausibility was to propose hypotheses from frames other than the Given frame, and to conduct experiments that could discriminate between them. Our youngest children's response was to propose a hypothesis from a different, but plausible, frame and then to ignore the initial, and implausible, hypothesis while attempting to demonstrate the correctness of the plausible one. Third graders were particularly susceptible to this strategy.

E3. *Maintain observability.* As BT moves along the screen it leaves no permanent record of its behavior. Subjects must remember what BT actually did. Thus, one way to implement this heuristic is to write short programs. Adults almost always used it, whereas the youngest children often wrote programs that were very difficult to encode This heuristic depends on knowledge of one's own information-processing limitations as well a knowledge of the device Our finding that the third graders often failed to maintain observability may be a manifestation, in the realm of experimental design, of more general findings about the development of self-awareness of cognitive limitations.

E4. *Design experiments giving characteristic results.* This heuristic maximizes the interpretability of experimental outcomes. Physicians look for "markers" for diseases, and physicists design experiments in which suspected particles will leave "signatures." In the BT domain, this heuristic is instantiated as "use many distinct commands." On average, about half of all programs in each group did not contain any repeated commands. However, because third graders were more likely to use long programs, they were more likely to use repeated commands, which reduced the possibility of generating characteristic behavior.

Kulkarni and Simon (1988) proposed another heuristic called *Exploit surprising results.* They built it into in their computational model of Hans Krebs' discovery of how amino acids work in the kidney In the BT domain, it is manifested when subjects replace their current goal—such as trying to determine the number of times something gets repeated—with a new goal of determining why an unexpected program segment was repeated. This appears to be a very useful way to constrain search in the experiment space, but in our own studies the evidence for its use is not strong. The data supported it in one of our adult studies, but not in subsequent ones.

Although I have described five heuristics for constraining search in the experiment space, I have not said anything about how they get invoked, or how their inherent contradictions are handled. For example, E1 calls for conservative moves, whereas exploiting surprise calls for bold pursuit of a surprising result. Adults not only used these heuristics effectively, but also they were able to deal with these inherent contradictions. In contrast, children either failed to use some of these heuristics at all, or else they tended to let one of them dominate. We still have a lot to learn about how this heuristic conflict is resolved.

Heuristics for Modifying Hypotheses

These rules help to constrain search in the E-space. But experiments are supposed to be related to hypotheses, and subjects are supposed to use the results of experiments to modify their hypotheses. How do they do it? How should they do it? The overwhelming evidence, from our own research and from many other labs, is that people seem to follow the following heuristic for dealing with disconfirming evidence: "Don't give up the ship!" On the cover of the announcement of the *Mind Matters* symposium, the organizers included a remarkable quotation from Allen Newell:

> Working with theories is not like skeet shooting, where theories are lofted up and BANG, they are shot down with a falsification bullet, and that's the end of that theory Theories are more like graduate students, once admitted you try hard to avoid flunking them out, it being much better for them and for the world if they can become long-term contributors to society. (Newell, 1990)

The statement says as much about Newell the mentor and teacher as Newell the scientist, but I confine my remarks to the scientific claim in the statement. It bears on a part of the discovery process that remains quite undeveloped, in the SDDS framework as well as in machine discovery systems. Just how do people evaluate evidence that bears on their theory? One thing is clear from our studies and those in other labs: Newell's statement is correct with one modification Although most people like to treat other people's theories like clay pigeons, they do treat their own theories like their own graduate students. They nurture them, tolerate their failings, and strive mightily to improve them rather than abandon them. How do they manage do this while maintaining scientific respectability?

Some people have proposed a Bayesian framework for understanding the process of evidence interpretation and theory revision (Cheeseman, 1990), but such approaches do not attempt to get at the underlying psychological processes. The problem is that when you are faced with the canonical equation for revising your priors, you still have to estimate a series of conditional probabilities. How do you decide how likely the current evidence is, given that one or another hypothesis is true or false?

Chinn and Brewer (1992) noted that most scientific discovery and theory revision systems assume that when empirical data conflict with the current theory, it is the theory that must be changed. In contrast, they argue, psychologists who study how people treat anomalous data have long recognized that people are pretty good at discounting some anomalies, some of the time. Chinn and Brewer go on to propose a taxonomy for the different ways that humans react to anomalous data. I have summarized their taxonomy in Table 9.7.

TABLE 9 7
Psychological Responses to Anomalous Data

Response to Anomalous Data About Theory A	*Accept Data as Valid?*	*Explain Why Data Accepted?*	*Change Theory in Any Way?*
Ignore data	No	—	—
Reject data	No	Yes	No
Exclude data from domain of theory	Maybe	Yes	No
Hold data in abeyance	Yes	No	No
Reinterpret data and retain A	Yes	Yes	No
Reinterpret data and tinker: A → A'	Yes	Yes	Minor
Accept data and change: A → B	Yes	Yes	Yes

Note From Chin and Brewer (1992) Adapted by permission

ED-7

From Chinn

The taxonomy includes seven kinds of responses to anomalies, which differ along three dimensions: (a) whether or not the anomalous data are accepted as valid, (b) whether or not the scientist explains how that judgment (i.e., to accept or reject the data) was made, and (c) whether or not the theory is changed as a result of the anomalous data. The normative, "skeet shooting" response is shown in category 7, but the table makes it clear that it is far from the only possible response. Category 2 is particularly interesting because it is frequently used when the scientist claims that an experimental error has occurred. Suppose you know that there is some probability of false positives or false negatives in your experimental outcomes. How does that affect the way you interpret the outcomes that do or don't agree with your predictions? We have been extending earlier work by Gorman (in press) on this question by presenting subjects with errorful feedback that includes both false positives and false negatives, in order to find out more about how subjects deal with such error (Penner, 1993).

Category 4 is also interesting: Here the scientist accepts the anomaly as valid, that is, not irrelevant to the domain, and not a fluke of experimental procedure. However, it is not yet clear what to do about it, other than to hope to come back to it at some point. This category has some of the properties of Kulkarni's and Simon's heuristic to exploit surprising results, for during that exploration, the scientist usually holds on to the current theory, but still peruses more information that may bear on it.

Category 4 also is characteristic of a strategy that another one of my graduate students, Chris Schunn (Schunn & Klahr, 1992), found in a more complex version of BT. He found that when subjects were attempting to discover a complex but decomposable rule, they would take an unexpected result and defer pursuing it for a while. He calls it the *put upon stack heuristic* (PUSH). Now PUSH and "Exploit surprises" are quite incompatible with one another, and at this point we do not know the conditions under which one or the other will dominate. Nor do we know which of these responses is favored by children. This remains a promising area for further research.

Young Children as Good Scientists

It is tempting to conclude—from the results of our studies with BT—that children simply don't understand the underlying logic of scientific reasoning. But consider the results of a recent study by Sodian, Zaitchick, and Carey (1991). They gave first and second graders a problem concerning a mouse in a house, and asked them to distinguish between conclusive and inconclusive experiments to find out something about the mouse. The story went something like this:

1. A mouse has been eating stuff in the kitchen at night.
2. It is either a big mouse or a little mouse.
3. We have a food box with a little hole, just wide enough for a little mouse, but too narrow for a big mouse.
4. We have another food box with a big hole, wide enough for either mouse.

Then the children were asked two questions (in counterbalanced order):

The Find Out question: Suppose we want to find out which mouse it is? Which box should we put out?

The Feed question: Suppose we want to be sure that the mouse gets the food. Which box should we put out?

For the Find Out question, the correct response is put out the box with the little hole: If the food is gone, then the mouse that took it must be small. If it is not gone, then the big mouse couldn't get to it. For the Feed question, the correct answer is to put out the box with the big door.

The majority of the first graders and most of the second graders gave the correct answer to both the Find Out question and the Feed question. Thus, the children demonstrated the ability to discriminate between testing a hypothesis and getting an effect. But notice: There were only two hypotheses, they were mutually exclusive and exhaustive, and the children did not have to search for them. Same thing for experiments. Under such conditions, even first-grade children show an ability to distinguish theory from evidence.

So: When do children think "scientifically"? I believe that our analysis, when combined with the related work from other laboratories, of the kind I just described, clarifies the conditions under which children's domain-general reasoning skills are adequate to successfully coordinate search for hypotheses and experiments Even first-grade children can exhibit an understanding of some basic components of the logic or scientific reasoning if at least these three conditions are satisfied:

1. Hypotheses must be easily accessible (such as the highly plausible Counters in our study) or few in number (as in the two-alternative situations used in Carey's lab).
2. The experimental alternatives must also be few in number so that E-space search demands are minimized (also as in the mouse experiments).
3. The domain must provide feedback relevant to discriminating among plausible hypotheses (as in region 2 experiments in BT studies).

In situations lacking any of these constraints, children will appear to be very poor scientists. Notice that this is not just another consequence of children's inadequate encoding or mnemonic skills. On the contrary. in our BT studies, when experimental outcomes were consistent with children's expectations, they were correctly encoded, even though they were much longer than those incorrectly encoded, but discrepant from children's expectations Instead, the adult superiority appears to derive from set of domain-general skills that go beyond the logic of confirmation and disconfirmation and deal with the coordination of search in two spaces.

MACHINES AS DISCOVERY SYSTEMS

So far I have talked about two of the three type of discovery systems mentioned in my title. Now I turn to the third type: machine discovery systems. As in many of the domains that have been approached by both the AI community and the cognitive psychology community, the strengths of one approach are the weaknesses of the other. The great advantage in studying humans is that they are obviously capable of making discoveries. All the great discoveries in the world are made by humans, and just about nothing of any importance has yet been discovered by a machine. In contrast, the advantage of focusing on machine discovery systems is that we know everything there is to know about how they work, because we built them, whereas we still have a lot to learn about human discovery systems (both old and young.)

ED - ?
and young(o ∋)

If you look at the existing machine discovery systems in terms of the SDDS model (Fig. 9.1), you find systems that address one or another parts of the overall process, but nothing that really takes on the entire framework (see Cheng, 1992). For example, Thagard (1989) constructed a system that models the way in which a body of evidence is evaluated in terms of currently competing theories. His system focused almost entirely on the Evaluate Evidence process.

But where do hypotheses come from? SDDS proposes two sources. One source, Search H-Space, relies heavily on analogy (see Shrager & Langley, 1990). For example, both Shrager's (1987) "view application" process and Falkenhainer's (1990) Phineas system attempt to reason about novel situations in terms of analogies and partial matches to prior knowledge structures. Both of these systems, then, correspond to the Evoke Frame node in SDDS. The other source of new hypotheses is via induction. In many cases they are induced from regularities in empirical data. This process, which corresponds to SDDS's Induce Frame node, is the domain of the original Bacon series, as well as more recent systems such as Nordhausen and Langley's (1990) IDA system. But it does not address problems of search in the experiment space. In contrast, both Kulkarni (1989) and Rajamoney (1990) have proposed systems that propose experiments to discriminate among candidate hypotheses. Each of these systems corresponds to the "conventional" use of experimentation that is represented in SDDS by the E-space search in the service of Test Hypothesis. And so on. I could continue this exercise, but the message is clear. There are machine discovery systems that focus on segments of the overall process shown in Fig. 9.1, but the UTD—the unified theory of discovery—is not with us yet.

DISCOVERING DISCOVERY PROCESS

Now I move up a level, from a description of how machines or humans do scientific discovery, to a characterization of the discovery processes used in the field. My analysis is based on two premises. The first premise is that the SDDS framework is applicable to any form of scientific discovery. The second premise is that people engaged in research on discovery processes are themselves engaged in scientific discovery: They are attempting to discover the discovery process. From these two premises, it follows that we can use the dual space concept to characterize our own endeavors.

I believe that most of the effort in the creation of computational models of discovery takes place in the Generate Frame part of Hypothesis Space Search. That is, the process of constructing such systems can be viewed as an attempt to evoke frames in the space of hypotheses stated as running programs. These hypotheses are instantiated as discovery systems, but they are only weakly constrained by empirical evidence from human performance. In general, this work is highly analytic: It is based on a normative analysis of what ought to be the case, with the assumptions derived from intuition or logic, rather than from induction over a rich database.

What about psychological studies of scientific reasoning of the type that I described today? I think that this work is mainly comprised of search in the space of experiments; moreover, this E-space search is not usually in the service of hypothesis testing, but rather it is mainly at the level of either evoking frames or filling slot values. Most of the effort in my own work has been focused on empirical studies about the nature of human thinking in situations that approximate "real" scientific discovery. So I would put most of the work from my lab, as well as many of the other psychological studies on discovery processes, in the regions where we have E-space search in the service of evoking or refining hypotheses.

Although these two approaches, the H-space and the E-space searches, start from quite different points, use different search processes, and use different criteria to evaluate their progress, I think that they are converging on the same general discoveries about the discoveryprocess. Indeed, that is what one would hope for, for our basic premise is that search in the two spaces should converge toward discovery. In this case, the entity doing the dual search is the field at large, rather than a single scientist, but I see convergence, nevertheless.

DISCOVERY FRONTIERS

One of the many remarkable things about Allen Newell that always impressed me was how cheerfully he could list all the current inadequacies and flaws in his current position on a topic. I think he could do that because he had the conviction that, for all its flaws, his game was the best game in town.

In Allen Newell's view, the current limitations of field simply presented yet another challenge, and he was always able to put a positive spin on them. In fact, he ended his book with a chapter that listed things that Soar had not done. But did he call the chapter "weaknesses and limitations"? Not at all. He called it "Along the Frontiers." I like that title: Not only does it imply discoveries yet to be made, but it also captures some of Allen Newell's enthusiastic optimism. How exciting to be on a frontier! So in this concluding section, I make a few comments about directions in which I believe that research on scientific reasoning should be extended.

More Space

The SDDS framework emphasis two primary spaces, but it is clear that scientific discovery takes place in several other spaces:

- The *instrumentation* space has been alluded to in my earlier discussions of how subjects decide about how to insert markers in their programs, but in the real scientific context, it is clearly a complex and fundamental space in its own right. From high-energy physics to cognitive neuroscience, advances in instrumentation are at the cutting edge of the science. Machine discovery systems do not worry about this much: They assume that the data are there, waiting to be analyzed by the discovery system, or else they postulate an idealized set of experiments to generate such data.

- The *representation* space has also received short shrift in my account, although its role is also crucial. Cheng and Simon (1992) argued that "law induction, and scientific discovery more generally, requires the right representation for success," and they compared the relative difficulty of mathematical and diagrammatic representations in Gallileo's research Finding the right representation is crucial, and finding it requires constrained search in a large space of possibilities.
- By *communication* space, I mean the set of choices that scientists must make about how to package, disseminate, promote, and defend their science, as well as what to read, whom to listen to, what meetings to attend. In many cases, these considerations, of audience, of intended impact, of how to tie ones work to the existing body of knowledge, have far-reaching impact on the kind of core science that one does (see Bazerman, 1988, for a discussion of how publication options impacted Newton's seminal work on light refraction).

These three do not exhaust the space of spaces Indeed, in a commentary on our SDDS model Newell (1989) proposed several others for the BigTrak world and he went on to link these multiple spaces to the "architectural features of the foothills (of rationality)."

> With Soar, we have finally found out how to have multiple problem spaces. Not one or two problems spaces, but problem spaces all the way down Furthermore, this is driven by the architecture—scratch an impasse, get another problem space. The proliferation of spaces may be modulatable ever so slightly by deliberation, but not much. Thus, the multiple problem-space character of a task is not a strategy choice for an intelligent agent or even a task characteristic. Multiple problem spaces are a feature of the foothills, created by the nature of the cognitive architecture. (p 432)

Complexity and Knowledge

Some machine discovery systems deal with enormously complex "real world" domains. However, much of the work on discovery—both the construction of machine discovery systems and the psychological studies of discovery—is in highly simplified domains. The BigTrak is a pale shadow of the complexity of domains in which real scientific discovery occurs. Therefore, the question remains about the extent to which our results would scale up when we move to domains in which either prior knowledge or inherent complexity of the domain is increased

We have been perusing this by looking at people's reasoning and experimental strategies in domains where they have strong intuitions and biases about how the physical world works. For example, David Penner and I (Penner & Klahr, 1993) have been studying how people generate experiments in order to determine the factors that affect the rate at which objects sink in water. In a related study in the domain of biology, Kevin Dunbar (1993a) created a computer-based microworld that captured several important features surrounding the discovery of the mechanisms of genetic control.

In addition to adding more knowledge, we need to add more complexity. At present, most discovery tasks studied in the psychology laboratory do not require subjects to decompose a complex phenomenon into its components in order to investigate them in isolation. However, as I mentioned earlier, in Chris Schunn's complex microworld we sometimes do see such a decomposition of a complex theory into its components, an independent investigation of each components, and then an assembly and integration of the components into a comprehensive model. This behavior does not reveal itself in simpler discovery experiments because it is not necessary.

But the most ambitious extension of the dual-search frame work is an ongoing study by Kevin Dunbar, who decided to move far beyond the microworld tasks into the real world of world-class scientists working in their labs (Dunbar, 1993b, 1995). Dunbar spent a year making daily observations in four different labs working in the area of molecular biology. He is using the SDDS framework to structure his observations and interpretations of what in happening in an ongoing, collaborative and cutting edge scientific endeavor. And, of course, he will use these observations to further elaborate the framework itself.

This is a very exciting undertaking, as it simultaneously moves along most of the fronts listed here. Not only does it involve multiple spaces, and more knowledge, but it will also address issues of social context and motivation, which you can see further down on the list of frontier issues.

Social Context

Clearly, the social context provides a rich source of knowledge and constraint in scientific discovery. Cooperation is important. So is competition. Sociologists and historians focus almost entirely on processes outside the individual that shape scientific discovery (Bijker, Hughes, & Pinch, 1987), but they are silent on the cognitive processes that are involved in this social exchange. Cognitive psychologists are just beginning to investigate the role of collaboration in scientific reasoning. But we have a long way to go.

Motivation

In our own work, we found a curious phenomenon: When we suggest hypotheses to subjects they are much less likely to believe them, and much more willing to reject them, than when they generated the very same hypotheses themselves. The motivation to prove the other person wrong and to prove oneself right is very strong.

Much of the best work in machine discovery is based on the historical record of the great scientists making the great discoveries. But we have been very selective in extracting information from those historical accounts. Such accounts are often filled with statements about excitement, astonishment, disappointment, envy, doubt, despair. Are these descriptions of emotional and motivation states irrelevant to understanding science? I doubt it.

Learning

Why does it take so long to train a scientist? Is it all due to the slow learning rate of humans and the huge amount of content knowledge and specific techniques necessary to work in a field? Why don't we start earlier then? Is it because we can't? That is, because the kind of domain-general search constraints are simply not available to young children? That is what the results reported here today imply, but we have a lot more work to do before we really understand the nature of these cognitive limitations.

Development

All of the discoveries I have been talking about so far are discoveries about things "out there": discoveries about devices, or about the planets, or about the kidney. What about discovery "in here"? When my colleague Bob Siegler (Siegler & Shipley, 1995) talks about discovery, he is talking about how children discover new strategies in doing arithmetic, or solving problems, or playing games. To what extent is what we have learned about discovery processes in the first sense relevant to discovery processes in the second sense? Is self-awareness of one's own discovery processes a useful skill for the scientist who is attempting to discover something about the world? Do the same heuristics and search constraints apply? It is clear that search in a large space faces those who would discover more about discovery, and our challenge is to see whether we can effectively constrain that search as we seek to discover discovery systems.

ACKNOWLEDGMENTS

Sections of this chapter are based on research done in collaboration with Kevin Dunbar and Anne L. Fay, and supported by grants from the National Institute of Child Health and Human Development (R01-HD25211) and the A. W. Mellon Foundation.

REFERENCES

Bazerman, C. (1988). *Shaping written knowledge: The genre and activity of the experimental article in science.* Madison: University of Wisconsin Press.

Bijker, W. E., Hughes, T. P., & Pinch, T. (1987). *The social construction of technological systems: New directions in the sociology and history of technology.* Cambridge, MA: MIT Press.

Brewer, W. F., & Samarapungavan, A. (1991). Child theories versus scientific theories: Differences in reasoning or differences in knowledge? In R. R. Hoffman & D. S. Palermo (Eds.), *Cognition and the symbolic processes: Applied and ecological perspectives* (pp. 209–232). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking.* New York: Science Editions.

Cheeseman, P. (1990). On finding the most probable model. In J. Shrager & P. Langley (Eds.), *Computational models of scientific discovery and theory formation* (pp. 73–95). San Mateo, CA: Morgan Kaufman.

Cheng, P. C-H. (1992). Approaches, models and issues in computational scientific discovery. In M. T. Keane & K. Gilhooly (Eds.), *Advances in the psychology of thinking* (pp. 203–236). Hempstead, Herefordshire: Harvester-Wheatsheaf.

Cheng, P. C-H., & Simon, H. A. (1992). The right representation for discovery: Finding the conservation of momentum. *Machine Learning: Proceedings of the Ninth International Workshop* (ML92) (pp. 62–71). San Mateo, CA: Morgan Kaufmann.

Chinn, C. A., & Brewer, W. F. (1992). Psychological responses to anomalous data. *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society* (pp. 165–170). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dunbar, K. (1993a). Concept discovery in a scientific domain. *Cognitive Science, 17*(3), 397–434.

Dunbar, K. (1993b). In vivo cognition: Knowledge representation and change in real-world scientific laboratories. *Proceedings of the 60th Meeting of the Society for Research in Child Development.* New Orleans.

Dunbar, K. (1995). How scientists really reason: Scientific reasoning in real-world laboratories. In R. J. Sternberg & J. Davidson (Eds.), *Mechanisms of insight.* Cambridge, MA: MIT Press.

Dunbar, K., & Klahr, D. (1989). Developmental differences in scientific discovery. In D. Klahr & K. Kotovsky (Eds.), *Complex information processing: The impact of Herbert A. Simon* (pp. 109–143). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dunbar, K., & Schunn, C. D. (1990). The temporal nature of scientific discovery: The roles of priming and analogy. *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society* (pp. 93–100). Hillsdale, NJ: Lawrence Erlbaum Associates.

Einstein, A. (1950). *Out of my later years.* New York: Philosophical Library.

Falkenhainer, B. (1990). A unified approach to explanation and theory formation. In J. Shrager & P. Langley (Eds.), *Computational models of scientific discovery and theory formation* (pp. 157–196). San Mateo, CA: Morgan Kaufman.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science,* 7, 155–170.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogic transfer. *Cognitive Psychology,* 7, 1–38.

Gorman, M. E. (1992). Experimental simulations in falsification. In M. T. Keane & K. Gilhooly (Eds.), *Advances in the psychology of thinking* (pp. 147–176). Hempstead, Herefordshire: Harvester-Wheatsheaf.

Holland, J., Holyoak, K., Nisbett, R. E., & Thagard, P. (1986). *Induction: Processes of inference, learning, and discovery.* Cambridge, MA: MIT Press.

Kaplan, C. A., & Simon, H. A. (1990). In search of insight. *Cognitive Psychology,* 22, 374–419.

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Psychology, 12*, 1–55.

Klahr, D., Fay, A., & Dunbar, K. (1993). Heuristics for scientific experimentation: A developmental study. *Cognitive Psychology, 25*(1), 111–146.

Klayman, J., & Ha, Y. (1987). Confirmation, disconfirmation and information in hypothesis testing. *Psychological Review, 94*, 211–228.

Kuhn, D. (1989). Children and adults as intuitive scientists. *Psychological Review, 96*, 674–689.

Kulkarni, D. (1989). *The processes of scientific research: The strategy of experimentation*. Unpublished doctoral dissertation, School of Computer Science, Carnegie Mellon University, Pittsburgh.

Kulkarni, D., & Simon, H. A. (1988). The processes of scientific discovery: The strategy of experimentation. *Cognitive Science, 12*, 139–175.

Langley, P., Simon, H. A., Bradshaw, G. L., & Zytkow, J. M. (1987). *Scientific discovery: Computational explorations of the creative processes*. Cambridge, MA: MIT Press.

Newell, A. (1989). Putting it all together. In D. Klahr & K. Kotovsky (Eds.), *Complex information processing: The impact of Herbert A. Simon* (pp. 399–440). Hillsdale, NJ: Lawrence Erlbaum Associates.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

Nordhausen, B., & Langley, P. (1990). An integrated approach to empirical discovery. In J. Shrager & P. Langley (Eds.), *Computational models of scientific discovery and theory formation* (pp. 97–128). San Mateo, CA: Morgan Kaufman.

Penner, D. (1993). *Scientific reasoning in the presence of data errors*. Unpublished doctoral dissertation, Department of Psychology, Carnegie Mellon University, Pittsburgh.

Penner, D., & Klahr, D. (1993). *The interaction of domain-specific knowledge and domain-general discovery strategies: A study with sinking objects*. Working paper, Department of Psychology, Carnegie Mellon University, Pittsburgh.

Qin, Y., & Simon, H. A. (1990). Imagery and problem solving. *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society* (pp. 646–653). Hillsdale, NJ: Lawrence Erlbaum Associates.

Rajamoney, S. (1990). A computational approach to theory revision. In J. Shrager & P. Langley (Eds.), *Computational models of scientific discovery and theory formation* (pp. 225–253). San Mateo, CA: Morgan Kaufman.

Ross, B. H. (1984). Remindings and their effects in learning a cognitive skill. *Cognitive Psychology, 16*, 371–416.

Schauble, L. (1990). Belief revision in children: The role of prior knowledge and strategies for generating evidence. *Journal of Experimental Child Psychology, 49*, 31–57.

Schunn, C. D., & Klahr, D. (1992). Complexity management in a discovery task. *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society* (pp. 900–905). Cambridge, MA: MIT Press.

Shrager, J. (1987). Theory change via view application in instructionless learning. *Machine Learning, 2*, 247–276.

Shrager, J., & Klahr, D. (1986). Instructionless learning about a complex device. *International Journal of Man-Machine Studies, 25*, 153–189.

Shrager, J., & Langley, P. (Eds.). (1990). *Computational models of scientific discovery and theory formation*. San Mateo, CA: Morgan Kaufman.

Siegler, R. S., & Shipley, C. (1995). Variation, selection, and cognitive change. In T. Simon & G. Halford (Eds.), *Developing cognitive competence: New approaches to process modeling* (pp. 31–76). Hillsdale, NJ: Lawrence Erlbaum Associates.

Simon, H. A. (1966). Scientific Discovery and the Psychology of Problem Solving. In R. Colodny (Ed.), *Mind and cosmos* (pp. 22–39). Pittsburgh: University of Pittsburgh Press.

Simon, H. A. (1973). Does scientific discovery have a logic? *Philosophy of Science, 40*, 471–480.

ED-7
re: Simon (1966) —
Scientific Discovery
and the Psychology
of Problem Solving.

Simon, H. A., & Lea, G. (1974). Problem solving and rule induction: A unified view. In L. Gregg (Ed.), *Knowledge and cognition* (pp. 105–128). Hillsdale, NJ: Lawrence Erlbaum Associates.

Sodian, B., Zaitchik, D., & Carey, S. (1991). Young children's differentiation of hypothetical beliefs from evidence. *Child Development, 62*, 753–766.

Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences, 12*, 435–502.

Tweney, R. D., Doherty, M. E., & Mynatt, C. R. (Eds.) (1981). *On scientific thinking*. New York: Columbia University Press.

Valdes-Perez, R. E., Simon, H. A., & Murphy, R. F. (1992, July). Discovery of pathways in science. *Proceedings of the Machine Discovery Workshop, International Conference on Machine Learning* (pp. 51–57). Abendeen, Scotland.

Vosniadou, S., & Brewer, W. F. (1992). Mental models of the earth: A study of conceptual change in childhood. *Cognitive Psychology, 24*, 535–585.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology, 12*, 129–140.

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology, 20*, 273–281.