From: Computational Models of Scientific Discovery and Theory Formation Edited by Jeff Shrager and Pat Langley. Morgan Kaufmann Publishers, Inc. 1990.

CHAPTER 6

A Unified Approach to Explanation and Theory Formation

BRIAN FALKENHAINER

1. Introduction

Deduction, abduction, and analogy are processes whose differences are normally reflected by distinct computational mechanisms. Furthermore, AI researchers typically decouple explanation and diagnosis from theory formation and discovery. Yet these tasks are intimately related and blend imperceptibly. Their integration into a unified view of explanation offers the potential for graceful degradation in the presence of an imperfect domain theory; in this approach, one provides a deductive explanation, if possible, and extends or revises the underlying theory when necessary to make explanation possible.

In this chapter, I suggest that procedural separations between deduction, abduction, and analogy are superfluous for the purpose of constructing plausible explanations of a given phenomenon. A single mechanism that proposes explanations of phenomena by their similarity to understood phenomena is sufficient, providing smoother adaptability to unanticipated or underspecified events and enabling transfer of knowledge from one domain to another. This *similarity-driven* view of explanation also lets one extend or revise imperfect theories when they fail to produce an explanation. Rather than being produced by separate processes, distinctions between the different explanation types result from the preferential ordering imposed when competing hypotheses are evaluated.

The plausibility of this conjecture is demonstrated by PHINEAS, a program that uses a single similarity-driven explanation mechanism to focus its search for explanations using its existing knowledge and to develop novel theories when its existing knowledge is insufficient. For example, when given only knowledge of liquid flow, the system is able to interpret the three situations shown in Figure 1:

a. A beaker contains more water than a vial to which it is connected by an unknown object. Why does the water level in the beaker decrease and the water level in the vial increase?

b. Two containers sharing a common wall of unknown substance each hold some solution. Why does one solution's level decrease and concentration increase while the other solution's level increases and concentration decreases?

c. What causes a hot brick and cold water to change to the same median temperature when the brick is immersed in the water?

In each case, PHINEAS bases its explanation on the case's similarity to liquid flow. In the first, the phenomenon most similar to an observation of liquid flow is liquid flow itself, thus suggesting that the unknown object may be a fluid path. In this work, identicality is viewed as an extreme form of similarity. The second behavior, called *osmosis*, represents a close generalization of liquid flow when viewed as flow of solute under osmotic pressure through a selective kind of fluid path. In the final "heat flow" observation, PHINEAS draws an across-domain analogy to liquid flow phenomena and conjectures the existence of a new type of fluid that affects an object's temperature. All three interpretations are produced by a single mechanism that forms its explanations from theories about phenomena most similar to the current situation.

This chapter begins with a discussion of the relationship between explanation and analogy and suggests that they share a common core, the search for explanatory similarity. It then describes PHINEAS, along with a detailed example of its operation. The system's behavior on a variety of examples is then discussed, which indicates success in achieving adaptability and provides impetus for a number of future research themes.

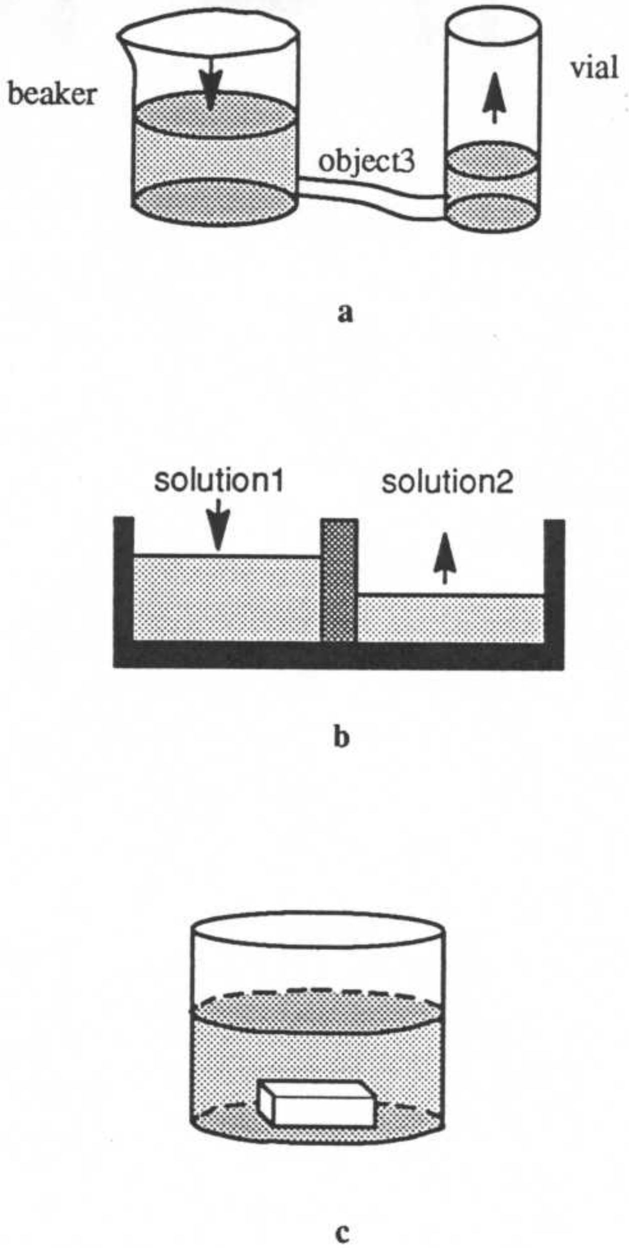


Figure 1. Three phenomena that PHINEAS explains by their similarity to liquid flow: (a) liquid flow, (b) osmosis, (c) heat flow.

2. Abduction as Similarity-Driven Explanation

Theory formation, explanation, and diagnosis all follow a pattern of reasoning called *abduction*, which can be defined as *inference to the best explanation*. Josephson, Chandrasekaran, Smith, and Tanner (1987) suggest that abduction is of the form:

$\mathcal{D}$ is a collection of data (facts, observations, givens);
$\mathcal{H}$ explains $\mathcal{D}$ ($\mathcal{H}$ would, if true, imply $\mathcal{D}$);
No other hypothesis explains $\mathcal{D}$ as well as $\mathcal{H}$;

Therefore, $\mathcal{H}$ is correct.

That is, if the hypothesis were true, it would explain the phenomenon.[1] There are two key phrases here. "If it were true" indicates that not all of the relevant knowledge may be available and that assumptions may be required to fill in the gaps. The process of finding the candidate hypotheses and of making assumptions along the way will be called the *interpretation-construction* task. The phrase "it would explain the phenomenon" indicates that the hypothesis would explain the phenomenon, not that it is the correct explanation. There may be other hypotheses that can also explain it. The process of deciding which hypothesis is the best explanation will be called the *interpretation-selection* task.

Abduction is traditionally characterized as using a fixed set of background theories. Assumptions needed to fill gaps due to incomplete knowledge of the situation are limited to ground atomic sentences (no new or revised rules are considered), as in

given $\text{CAUSE}(\mathcal{A}, \mathcal{C}), \mathcal{C}$ *infer* $\mathcal{A}$

These systems suffer from the *adaptability problem* (Falkenhainer, 1988): They are unable to revise or extend an imperfect domain theory to make conjectures about unanticipated events, and unable to apply knowledge of one domain to the understanding of another.

[1] There is a distinction between the abduction process, which is normally associated with backward chaining on a set of rules, and its ultimate product, a deductive proof tree typically having at least one assumption as a leaf. Throughout this chapter, I am primarily interested in the abstract product, independent of the chaining process, in which assuming some unknown antecedent facts is required to complete the explanation.

On the other hand, theory formation typically involves making assumptions about both the situation and the incompleteness or incorrectness of current theories. It includes inferences of the form

$$\textit{given}\ \text{CAUSE}(\mathcal{A}, \mathcal{C}) \wedge \mathcal{A} \Rightarrow \mathcal{C}, \mathcal{A}, \mathcal{C} \quad \textit{infer}\ \text{CAUSE}(\mathcal{A}, \mathcal{C})$$

Theory formation must face the problem of generating theory-revising hypotheses and establishing a preference among a possibly infinite set of hypotheses.

These problems can be resolved by noting the strong commonalities between traditional abduction and analogy and developing a model that encompasses both. For abduction, this unified model provides the power to extend the underlying domain theory when needed. For theory formation, it enables existing knowledge, possibly of other domains, to influence hypothesis generation and evaluation, thus taking into account knowledge of the way things normally behave in the world and the way theories about those behaviors are normally expressed. This view of explanation is based on the conjecture that search for similarity between the situation being explained and some understood phenomenon suffices as the central process model for explanation tasks. Two arguments support this view.

First, consider the traditional abduction task. Simple backward-chaining models work well for explaining atomic occurrences, such as `Wet(grass)`. However, as the complexity of the phenomenon being explained increases, the ability to backward chain to a small set of plausible candidates diminishes. One must consider the entirety of the situation and take into account all the interrelations between aspects. Hence, most abduction systems directed at complex phenomena are based on some form of macro-matching, typically in terms of schemas or frames, that seeks minimal hypothesis sets maximally fitting the data. This is true of script or schema-based models of story understanding (Charniak, 1972; DeJong, 1982; Mooney, 1987), process models for interpreting the behavior of a physical system (Forbus, 1986), and composite matching models of abduction and diagnosis (Josephson, Chandrasekaran, Smith, & Tanner, 1987; Reggia, 1983). The desire for a minimal, best match is also implicitly reflected in the Occam's razor heuristic found in simpler systems, which backward chain on one datum at a time (e.g., Pople, 1973). In other words, interpretation and explanation are a form of best match process, with the goal of matching the current situation to hypotheses that can explain it.

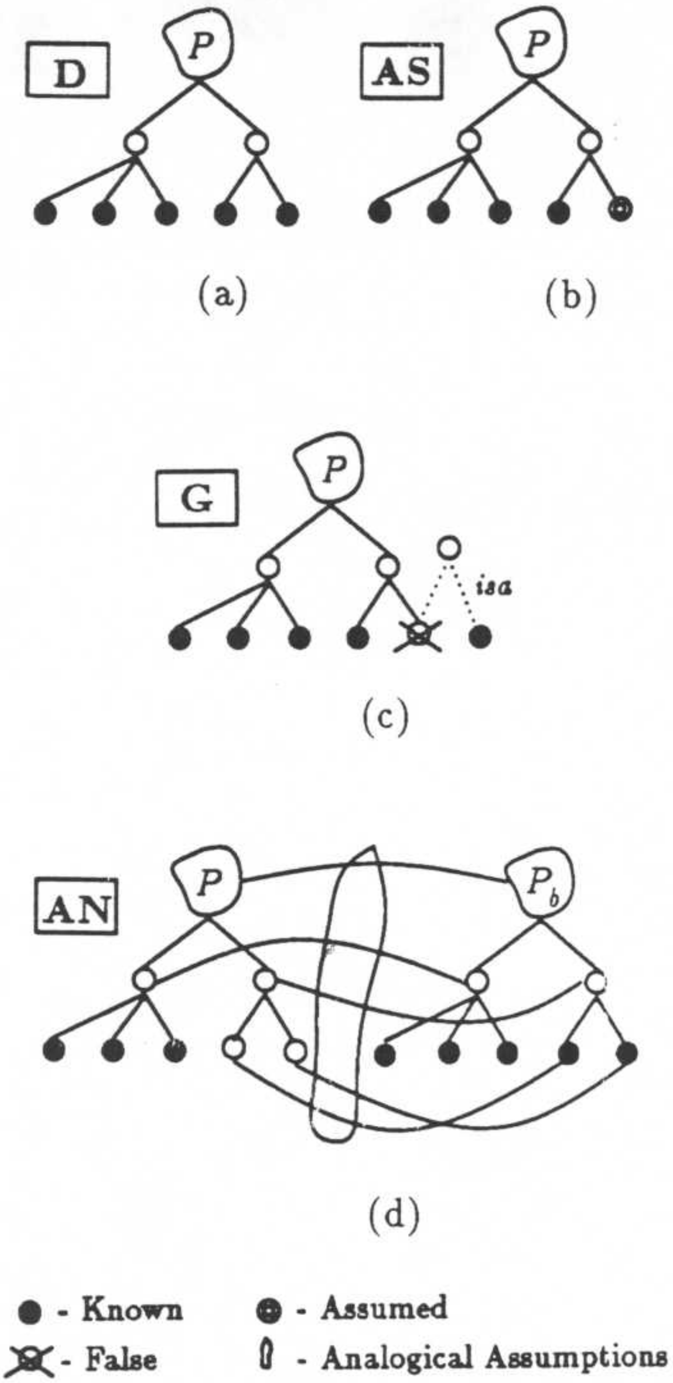
A second argument involves the explanation scenarios shown in Figure 2, which are summarized below:

Deduction scenario. Given phenomenon $\mathcal{P}$, where $\mathcal{P}$ represents a set of observables, a complete explanation of $\mathcal{P}$ deductively follows from existing knowledge. The only open question is whether it is *the* explanation, as there may be others. For example, suppose fluid flow is observed, and all the preconditions for fluid flow are known to hold (the source pressure is greater than the destination pressure, the fluid path is open, and so forth). Then a fluid flow explanation directly follows. Given the observed behavior and the existing preconditions, we could say that the situation is *literally similar* (Gentner, 1983) to liquid flow.

Assumption scenario. Phenomenon $\mathcal{P}$ is given, where $\mathcal{P}$ represents a set of observables. No explanation can be found using current knowledge because the status of some requisite facts is unknown. However, a complete explanation follows from the union of existing knowledge and a consistent set of assumptions about the missing facts. For example, if one observes liquid flow but does not know if the fluid path valve is open or closed, one can assume that the valve is open if there is no evidence to the contrary.

Generalization scenario. Phenomenon $\mathcal{P}$ is given, where $\mathcal{P}$ represents a set of observables. Existing knowledge indicates that candidate explanation $\mathcal{E}$ cannot apply because condition C_1 is known to be false in the current situation. However, $\mathcal{E}$ does follow if condition C_1 is replaced by the next most general relation since C_1's sibling is true in the current situation. This is a standard knowledge-base refinement scenario (e.g., Winston, Mitchell, & Buchanan, 1985) and is closely related to approaches that generalize from a set of examples (Hayes-Roth & McDermott, 1978; Winston, 1975).

Analogy scenario. Phenomenon $\mathcal{P}$ is given, where $\mathcal{P}$ represents a set of observables. No candidate explanation $\mathcal{E}$ is available directly, but explanation $\mathcal{E}_b$ is available if a series of analogical assumptions are made, that is, if the situation explained by $\mathcal{E}_b$ is assumed to be analogous to the current situation. For example, if heat flow is observed but little is known about heat phenomena, then an explanation may be constructed by analogy to liquid flow.



(a)

(b)

(c)

(d)

Figure 2. Four alternative explanation scenarios: (a) deduction scenario, (b) assumption scenario, (c) generalization scenario, (d) analogy scenario. In each, $\mathcal{P}$ is the phenomenon being explained and implications flow from the antecedents below to the explained consequents above.

Each scenario requires the interpretation-construction task: retrieve from memory explanatory hypotheses that match the current situation. Each also requires the interpretation-selection task: select from a set of candidate hypotheses the one that is most probable, plausible, or coherent. Importantly, each scenario represents the same process when viewed as different forms of similarity to an existing theory:

- *Deduction scenario*: complete match of identical features
- *Assumption scenario*: partial match of identical features
- *Generalization scenario*: matches between features having a close generalization
- *Analogy scenario*: a range of matches between different features and relations

A system based on this view would offer the best explanation available, ranging from application of an existing theory to distant analogy. It relies on the following conjecture:

> *Similarity conjecture*: All interpretation-construction tasks may be characterized as the search for maximal explanatory similarity between the situation being explained and some previously explained scenario. The previous situation may be drawn from an actual experience, a prototypical experience, or an imagined scenario derivable from general knowledge.

This conjecture suggests that there is no need for a strong distinction between deductive explanation processes and analogical explanation processes. The same basic process may be used in each explanation scenario, with distinctions between them emerging from how well existing knowledge supports the explanation. Deductive operations correspond to the high confidence derived from identicality matches. A corollary to the similarity conjecture is that the same basic processes are at work in both scientific theory formation and in everyday interpretation and hypothesis formation, as suggested by Leatherdale (1974).

The benefits of this view are that it suggests using a single computational architecture for explanation processes. Distinctions between explanation types influence only the weighing of evidence and the decision as to whether a new conjecture represents a revision of existing knowledge or a new separate body of knowledge. This chapter seeks to demonstrate the feasibility of this view.

3. The PHINEAS System

The similarity-driven model of explanation discussed in the previous section is illustrated by PHINEAS, a program that offers qualitative explanations of time-varying physical behaviors. The system uses remindings of similar experiences to suggest plausible hypotheses and uses qualitative simulation as a form of *gedanken* experiment to analyze the consistency and adequacy of these hypotheses. This section begins with a discussion of the representations used in PHINEAS to describe observations and to reason about their underlying causes. It then presents an overview of the PHINEAS system and the preference criteria that gives rise to its intuitively appealing, flexible behavior.

3.1 Representation

PHINEAS' theories about the physical world and its methods for using these to generate predictions are based on research in qualitative physics (Bobrow, 1985). Given a qualitative model of a particular physical configuration, a *qualitative simulator* produces a description of the possible behaviors for the given situation, called an *envisionment*. An envisionment describes physical states and the possible transitions between them. Each state represents an interval of time during which the qualitative description of behavior does not change. A specific behavior of the system through time, either observed or predicted, may then be represented as a single path through the envisionment. I will refer to such a path as a *history*, after Hayes (1979). For example, Figure 3(a) shows a beaker connected to a vial and an observed qualitative history for this configuration.

The present work uses Forbus' (1984) *qualitative process theory* as the primary formalism to represent and reason about physical change. In QP theory, a situation is represented as a collection of objects (e.g., contained liquid), a set of relationships between them (e.g., connected), and a set of process schemas that account for all changes in the world (e.g., liquid flow). Each object has a set of continuous *quantities*, such as `Temperature` and `Pressure`. Each quantity has an amount, expressed as `A[Temperature(brick)]`, and a derivative, expressed as `D[Temperature(brick)]`.

Process definitions have five components: *individuals*, *preconditions*, *quantity conditions*, *relations*, and *influences*. The individuals specify

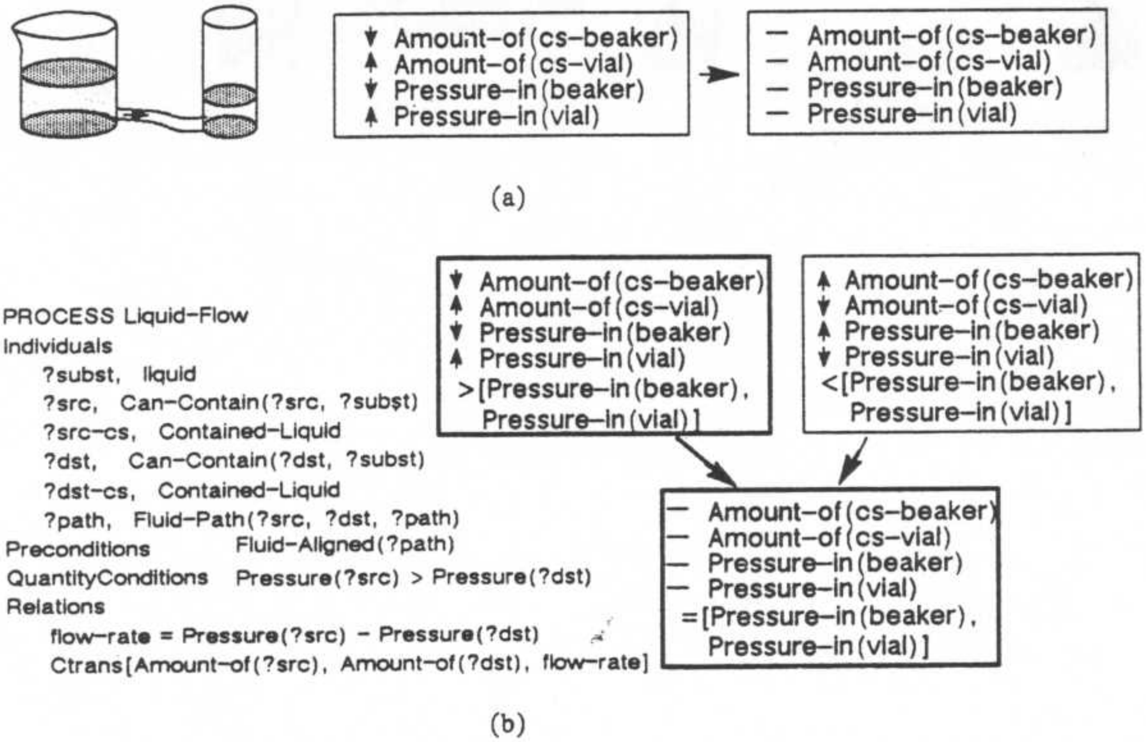


Figure 3. Qualitative physics representations: (a) qualitative observation of liquid flow from a beaker to a vial; (b) liquid flow process model and corresponding envisionment.

the objects involved in the process when it is active, the preconditions and quantity conditions indicate when the process will be active, and the relations and influences specify what relations will hold while the process is active. Figure 3(b) shows a typical QP theory definition for the liquid flow process and the envisionment it produces for the beaker-vial configuration.[2]

The explanatory consistency of a proposed model is established if there is a path through the envisionment derived from the model that corresponds to the measurements (Forbus, 1986). For example, the darkened two-state path of Figure 3(b) corresponds to the observation in Figure 3(a).

[2] The predicate `Ctrans` refers to "continuous transfer" and is a macro for the standard QP theory pair `I-[Amount-of(?src),flow-rate]` and `I+[Amount-of(?dst),flow- rate]`. See Falkenhainer (1988) and Forbus (1984) for more details.

3.1.1 INITIAL KNOWLEDGE

PHINEAS uses three sources of knowledge during its reasoning process. These include:

1. *Initial domain theory.* Domain knowledge consists of a collection of qualitative theories about physical processes (e.g., liquid flow), entities (e.g., fluid paths), and general physical principles (e.g., mechanical coupling). This qualitative knowledge is represented using the language of Forbus' (1984) QP theory.
2. *Prior experiences.* When comparing a new observation with prior experience, PHINEAS consults a library of previously observed phenomena (structure and behavior descriptions). Focus on relevant attributes is ensured by the storage of only those aspects of a prior situation that participated in its explanation. Past reasoning traces are summarized by storing with each state in an observation the instantiated collection of theories (e.g., process definitions) that were used to explain it; for example, `Liquid-Flow(beaker1,vial8, pipe2)` might be stored with an observation of liquid flow.

   Behaviors are indexed in memory via *behavioral abstractions*, which record abstract characterizations and summaries of the phenomenon not captured by the standard QP theory representation. These correspond to graphic characterizations (e.g., `linear`, `cyclic`, `asymptotic`), movement continuity (e.g., `corpuscular`, as in a ball, or `continuous`, as in liquid flowing), and movement type (e.g., `phase-change-movement` or `invariant-form-movement`). These are arranged in generalization hierarchies, forming a forest of behavioral abstractions similar to the memory organization used by Kolodner (1984).
3. *Observation.* The final source of PHINEAS' information is the observation targeted for explanation. The system records three classes of information: the original scenario description (e.g., `Open(beaker)`), the behavior across time (e.g., `Decreasing[Amount-of(alcohol)]`), and behavioral abstractions that apply to the observation (e.g., `asymptotic`).

3.1.2 OUTPUTS

In response to a given observation, PHINEAS attempts to produce an explanatory "theory" and the envisioned behaviors it predicts. A theory consists of a set of process descriptions, entity descriptions, and atomic facts. The process and entity descriptions may be elements of the existing domain theory or new postulated theories. The system makes this distinction during hypothesis evaluation. The atomic facts are assumptions about the scenario that are required to complete the explanation.

3.2 Process Components

As depicted in Figure 4, PHINEAS operates in four stages: *access*, *mapping/transfer*, *qualitative simulation*, and *revision*. Falkenhainer (1988) provides the details of each stage, but they are briefly reviewed in this section. Throughout the discussion, the term "base" refers to a recalled analogue, and "target" refers to the current situation to be explained.

3.2.1 THE ACCESS STAGE

A new observation triggers a search in memory for understood phenomena that exhibit analogous behavior. This retrieval process involves two stages. First, behavioral abstractions of the observed situation are used to focus attention on a potentially relevant subset of memory. Second, each phenomenon in this subset is inspected more carefully by matching its detailed structural and behavioral description to the current situation. This comparison is performed by the *structure-mapping engine* (SME) (Falkenhainer, Forbus, & Gentner, 1986, 1989).[3] This partial mapping provides an indication of what objects and quantities correspond by virtue of their behavioral similarity. It serves as an important source of constraint during the mapping process.

The match also indicates where the phenomena correspond and thus what portion of the base analogue's behavior should be considered relevant. The problem of relevant theory selection is solved by retrieving only those domain theories that had been used to explain the matched portions of the base situation. Each behavioral state indicates what

[3] SME is a flexible analogical matching system motivated by Gentner's (1983) ***structure-mapping theory*** of analogy. It may be configured to model a number of different theories of analogical mapping and is discussed further in Subsection 3.2.2.

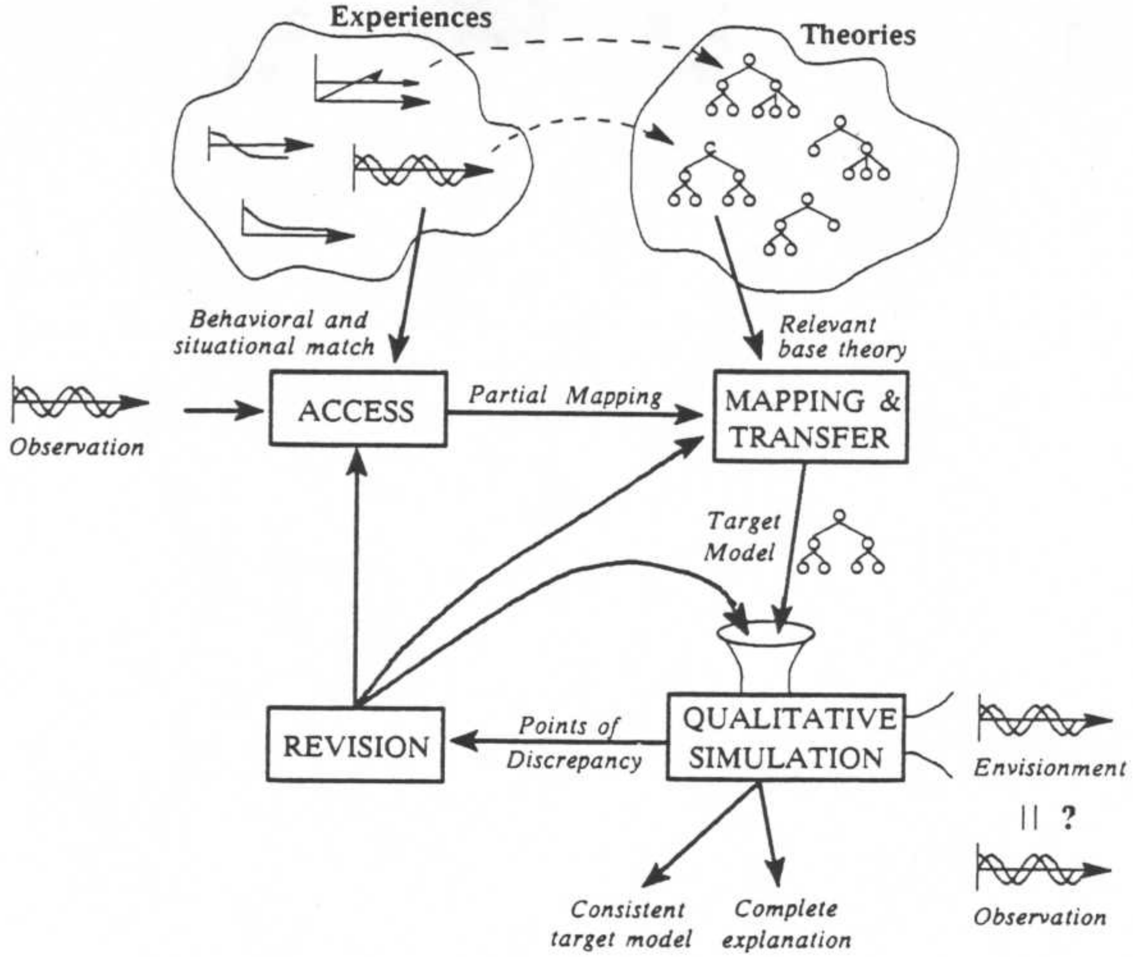
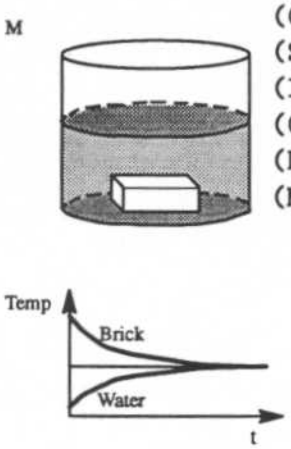


Figure 4. A functional decomposition of PHINEAS.

processes were active during that state. Thus, if the current observation matches only a subset of the states in the base observation, only those relevant process models are used.

The retrieved candidates are then ordered according to SME's evaluation score and are proposed one at a time as potential analogues on PHINEAS' global agenda.

The system's behavior can be clarified with an example. The caloric theory of heat, dominant during the eighteenth century, postulated a material heat substance called *caloric.* The temperature of an object was thought to be proportional to the amount of caloric present. Furthermore, caloric tended toward equilibrium, causing it to flow between bodies placed in contact until an equilibrium of their temperatures was



```
(Physob brick)
(Solid brick)
(Volume-solid brick)
(Liquid water1)
(Contained-liquid water1)
(Container-of water1 bucket)
(Substance-of water1 water)
(Immersed-in brick water1)
(Contained-in water1 bucket)
(Dual-approach-finish 2-obj-hf)
(Meets (Situation 2-obj-hf-sit0
          (Set (Decreasing (Temperature-in brick))
               (Increasing (Temperature-in water1))
               (Greater-Than (A (Temperature-in brick))
                             (A (Temperature-in water1)))))
       (Situation 2-obj-hf-sit1
          (Set (Constant (Temperature-in brick))
               (Constant (Temperature-in water1))
               (Equal-to (A (Temperature-in brick))
                         (A (Temperature-in water1))))))
```

Figure 5. An unexplained thermal situation. When a hot brick is immersed in cold water, the brick's temperature decreases and the water's temperature increases. This transitions to a state in which the temperatures are constant and equal.

achieved. Let us consider how PHINEAS achieves a naive level of the caloric view when it encounters thermal behavior for the first time.

The explanation task is illustrated in Figure 5. When a hot brick is immersed in cold water, their temperatures asymptotically approach each other until reaching equality. PHINEAS begins by searching memory for analogous behavior. First, the behavioral abstractions describing the observation are used to probe memory. In this case, **`dual-approach-finish`** applies, which characterizes two quantities asymptotically approaching each other and reaching equality. Only one candidate analogue demonstrates this abstract behavior—two-container liquid flow. This scenario describes liquid flowing from one container (`beaker3`) to another (`vial2`), through a pipe (`pipe1`) connecting them. Using SME to compare the current and recalled situations, PHINEAS determines that the roles of the beaker and vial in the liquid flow description correspond to the roles of the brick and water in the thermal situation, respectively. Additionally, it finds that **`pressure`** in the liquid flow situation corresponds to **`temperature`** in the thermal situation.

3.2.2 THE MAPPING AND TRANSFER STAGE

The objective of the second stage is to generate an initial hypothesis about the current observation. This stage has two components, *mapping* and *transfer*.

Given a candidate analogue, PHINEAS retrieves the models used to explain analogous aspects of the recalled experience. Mapping serves to complete the initial set of correspondences (*matching*) and to propose *candidate inferences* sanctioned by those correspondences (*carryover*). The model of mapping used in this work is called *contextual structure-mapping* (Falkenhainer, 1988, 1989), a knowledge-intensive adaptation of Gentner's (1983, 1988) structure-mapping theory of analogy. It uses knowledge of the various contextual factors affecting analogical interpretation, such as the role of each element in the two analogue descriptions, to analyze the similarity between their descriptions. The mapping is constructed by SME_{CSM}, that is, by SME configured to run the rules of contextual structure-mapping.

An important observation made in contextual structure-mapping is that the correspondences and inferences proposed by the mapping stage may be incomplete. The *transfer* stage analyzes the results of the mapping stage to elaborate its correspondences and minimize unnecessary conjectures. This centers around two issues. First, if the candidate inference references an object needed in the base scenario that has no apparent target correspondent, one must find a corresponding target object or conjecture its existence. These unknown objects are represented by `(:skolem` *base-object*`)` and are called *skolem objects*.[4] Second, candidate inferences represent relevant base expressions having no apparent correspondent in the target. However, they may not be applicable to the target. Thus, the domain theory is consulted, and more detail about each candidate inference is sought. Alternate, analogous target expressions may be found or a new vocabulary (predicates) may be created. If new information is found, mapping is repeated to see how it affects the overall mapping. A *map and analyze* cycle may ensue.

Let us return to the caloric heat flow example. Upon completion of access, PHINEAS attempts to map the relevant liquid flow domain theory into the current thermal situation. First, the domain theory

[4] The term *skolem object* derives from standard logical use of a *skolem constant* to denote the existence of an unknown object and enable removal of an existential quantifier.

Table 1. A mapping from the liquid flow process instance to the hot brick in cold water scenario, as generated by SME_{CSM}.

```
Match Hypotheses:  pressure-in(beaker3) ⟺ temperature-in(brick)
                     pressure-in(vial2) ⟺ temperature-in(water1)
                                beaker3 ⟺ brick
                                  vial2 ⟺ water1

Weight: 2.07

Candidate Inferences:

 Contained-Fluid((:skolem cs-water-beaker), (:skolem water), brick)
   EFFECTS  Quantity[amount-of(:skolem cs-water-beaker)]
            Quantity[temperature(:skolem cs-water-beaker)]
            Qprop[temperature(:skolem cs-water-beaker),
                  amount-of(:skolem cs-water-beaker)]
              •
              •
              •

 Contained-Fluid((:skolem cs-water-vial), (:skolem water), water1)
   EFFECTS  Quantity[amount-of(:skolem cs-water-vial)]
            Quantity[temperature(:skolem cs-water-vial)]
            Qprop[temperature(:skolem cs-water-vial),
                  amount-of(:skolem cs-water-vial)]
              •
              •
              •

 Liquid-Flow((:skolem cs-water-vial), (:skolem water), water1)
   PRECONDITIONS  Liquid(:skolem water)
                  Can-contain(brick, (:skolem water))
                  Fluid-path(:skolem pipe1)
                  temperature-in(:skolem cs-water-beaker)
                    > temperature-in(:skolem cs-water-vial)
   EFFECTS  Quantity[flow-rate]
            flow-rate = temperature-in(:skolem cs-water-beaker)
                      - temperature-in(:skolem cs-water-vial)
              •
              •
              •
```

used to explain the two-container liquid flow experience is retrieved. This consists of the liquid flow process and two instantiations of

```
(Contained-Fluid contained-fluid substance container),
```

one for the beaker water and one for the vial water. SME is then invoked with knowledge of the partial mapping established during access, giving the results shown in Table 1. Its candidate inferences propose a new contained-fluid relationship, in which the temperature of the container (`brick` and `water1`) is proportional to the amount of substance it contains. This substance is currently unknown but is analogous to the water in the liquid flow situation. Additionally, a new process is proposed: When two objects of differing temperature are connected by a physical path, the unknown substance continuously flows from the object of higher temperature to the one of lower temperature, at a rate equal to their difference in temperatures.

The candidate inferences are next passed to the transfer stage. It first determines that none of the proposed expressions is inconsistent in their current state. Next, these inferences are inspected for the presence of skolem objects, and four are found: `(:skolem cs-water-beaker)`, `(:skolem cs-water-vial)`, `(:skolem water)`, and `(:skolem pipe1)`. The first two are compound objects (objects defined solely by their constituents) and are therefore ignored. The unknown `(:skolem pipe1)` indicates that no correspondent for the pipe connecting the beaker and vial was found. However, when PHINEAS is given the task of locating an object satisfying the conjunction

```
(Physical-Path brick water1 ?pipe) ∧ (Fluid-Aligned ?pipe) ∧
   (Fluid-path ?pipe)
```

it finds that

```
(Physical-Path brick water1 (common-face brick water1))
```

and

```
(Fluid-Aligned (common-face brick water1))
```

are true in the current scenario and that the third conjunct can be assumed. Therefore, the system establishes `(common-face brick water1)` as the analogue for `pipe`. This demonstrates the utility of the transfer stage in filling out an incomplete analogical mapping. An analogy will often evoke additional information or perspectives about the two analogues made relevant by its consideration.

The remaining unknown, `(:skolem water)`, indicates that no correspondent for the water flowing from beaker to vial was found. Additionally, no correspondent is found when an object satisfying the relevant conditions is sought:

(Substance *?pipe*) ∧ (Liquid *?pipe*) ∧
(Can-Contain brick *?pipe*) ∧ (Can-Contain water1 *?pipe*)

However, when a new entity token is made for the missing water correspondent, a contradiction arises:

```
(Liquid sk-water-1) ∧ (Volume-Solid brick) ⇒
   ¬(Can-Contain brick sk-water-1)
```

As a result, `(Liquid sk-water-1)` is changed to `(Phase-1 sk- water-1)`, with `Phase-1` added as a new kind of `Phase`. This illustrates PHINEAS' ability to create new object tokens (`sk-water-1`) when it cannot resolve a skolem object produced by mapping. Further, it is able to distinguish between assuming the presence of an unobserved object and conjecturing a theoretically novel entity. This is important information that it can use in theory evaluation and selection.

At this point, the transfer task is completed, resulting in the model shown in Table 2. This model postulates that the brick and water each contain `sk-water1-1` and that their temperatures are proportional to the amount they contain. Additionally, it proposes the new `Process-1`, which might be called a *heat flow* process. This indicates that `sk-water1-1` will flow from the object of higher temperature to the object of lower temperature. PHINEAS does not generalize beyond replacing constants with variables, hence only the `brick` and `water1` are believed to contain `sk-water1-1`.[5]

3.2.3 THE VERIFICATION STAGE

Verification-based analogical learning (Falkenhainer, 1986, 1988) depicts analogical learning as an iterative process of hypothesis formation, verification, and revision, centered around the requirement to confirm ad-

[5] Research on explanation-based learning has shown that this is not sufficient to ensure proper generalization (DeJong & Mooney, 1986; Mitchell, Keller, & Kedar-Cabelli, 1986). Explanation-based generalization might be performed at this point, but it has not been necessary so far.

Table 2. A final PHINEAS hypothesis explaining the behavior of a hot brick in cold water. This hypothesis was derived from the system's theories about liquid flowing between two containers.

```
(DEFPROCESS  (PROCESS-1 ?SUBST ?SOURCE ?SRC-CS ?DESTINATION
              ?DST-CS ?PATH)
  INDIVIDUALS  ((?SUBST :CONDITIONS (SUBSTANCE ?SUBST)
                                     (PHASE-1 ?SUBST))
                (?SOURCE :CONDITIONS (CAN-CONTAIN ?SOURCE ?SUBST))
                (?SRC-CS :CONDITIONS (CONTAINED-FLUID-1 ?SRC-CS
                                      ?SUBST ?SOURCE))
                (?DESTINATION :CONDITIONS
                      (CAN-CONTAIN ?DESTINATION ?SUBST))
                (?DST-CS :CONDITIONS (CONTAINED-FLUID-1 ?DST-CS
                                      ?SUBST ?DESTINATION))
                (?PATH :CONDITIONS (FLUID-PATH ?PATH)
                                   (PHYSICAL-PATH ?SOURCE
                                    ?DESTINATION ?PATH)))
  PRECONDITIONS  ((FLUID-ALIGNED ?PATH))
  QUANTITYCONDITIONS  ((GREATER-THAN (A (TEMPERATURE-IN ?SOURCE))
                                     (A (TEMPERATURE-IN ?DESTINATION)))
                       (GREATER-THAN (A (AMOUNT-OF ?SRC-CS)) ZERO))
  RELATIONS  ((QUANTITY (FLOW-RATE ?SELF))
              (Q= (FLOW-RATE ?SELF)
                  (- (TEMPERATURE-IN ?SOURCE)
                     (TEMPERATURE-IN ?DESTINATION)))
              (GREATER-THAN (A (FLOW-RATE ?SELF)) ZERO))
  INFLUENCES ((CTRANS (AMOUNT-OF ?SRC-CS) (AMOUNT-OF ?DST-CS)
                      (A (FLOW-RATE ?SELF)))))

(DEFENTITY  (CONTAINED-FLUID-1 ?V-1 ?V-2 ?V-3)
    (CONTAINER-OF ?V-1 ?V-3)
    (SUBSTANCE-OF ?V-1 ?V-2)
    (QUANTITY (AMOUNT-OF ?V-1))
    (QUANTITY (TEMPERATURE-IN ?V-3))
    (QPROP (TEMPERATURE-IN ?V-3) (AMOUNT-OF ?V-1)))

(ASSUME (SUBSTANCE SK-WATER-1))
(ASSUME (PHASE-1 SK-WATER-1))
(ASSUME (CAN-CONTAIN BRICK SK-WATER-1))
(ASSUME (CONTAINED-FLUID-1 SK-CS-WATER-BEAKER-1 SK-WATER-1 BRICK))
(ASSUME (CAN-CONTAIN WATER1 SK-WATER-1))
(ASSUME (CONTAINED-FLUID-1 SK-CS-WATER-VIAL-1 SK-WATER-1 WATER1))
(ASSUME (FLUID-PATH (COMMON-FACE BRICK WATER1)))
```

equacy of use in explaining a given phenomenon. In PHINEAS, it sanctions the use of gedanken experiments in the form of qualitative simulations to analyze the adequacy of proposed models. Specifically, the predictions of a proposed model are compared against the observed behavior, enabling the system to test the validity of the analogy and sanction refinements where the analogy is incorrect. The system generates an envisionment of the scenario, which it then compares with the original observation. If the envisionment is consistent and complete with respect to the observation, then the explanation is considered successful. If it is inconsistent or fails to provide complete coverage, then revision is aimed at the points of discrepancy.

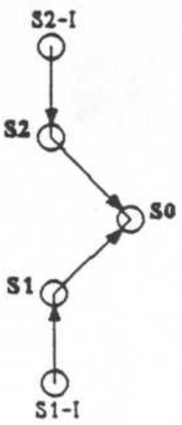
PHINEAS produces envisonments of predicted behavior using Forbus' (1988) *qualitative process engine* (QPE). The process of comparing and identifying points of discrepancy between the predicted and observed behaviors is performed by DeCoste's (1989) *dynamic across-time measurement interpretation* system, DATMI.

Only one test remains in the ongoing caloric example, to verify the adequacy of the model in explaining the original observation. As shown in Figure 6, the model produces a five-state envisionment, with state `S2` transitioning to state `S0`, demonstrating that the model is able to predict the observed temperature changes. In state `S2`, `Process-1` is active, the substance `sk-water-1` is flowing from the brick to the water, and the temperature of the brick is decreasing while the temperature of the water is increasing, each at a rate equal to the difference in their temperature. In state `S0`, the brick and water temperatures are equal, and all quantities are constant.

3.2.4 THE REVISION STAGE

If PHINEAS' initial hypothesis is inadequate, an attempt should be made to adapt it around points of inaccuracy. We advocate, but have not fully implemented, a model of revision that relies on past experiences to guide the formation and selection of revision hypotheses (Falkenhainer, 1988). It considers behavior analogous to the current anomaly and considers differences between the current anomalous situation and the prior situations that were consistently explained. This is the only component of PHINEAS that is not fully implemented.

Quantity	S2-I	S2	S0	S1-I	S1
Ds[FLOW-RATE(PI0)]	–	–	–	-1	-1
Ds[FLOW-RATE(PI1)]	-1	-1	–	–	–
Ds[AMOUNT-OF(SK-CS-WATER-BEAKER-1)]	-1	-1	0	1	1
Ds[AMOUNT-OF(SK-CS-WATER-VIAL-1)]	1	1	0	-1	-1
Ds[PRESSURE(SK-CS-WATER-BEAKER-1)]	-1	-1	0	1	1
Ds[PRESSURE(SK-CS-WATER-VIAL-1)]	1	1	0	-1	-1
Ds[TEMPERATURE-IN(BRICK)]	-1	-1	0	1	1
Ds[TEMPERATURE-IN(WATER1)]	1	1	0	-1	- 1
A[AMOUNT-OF(SK-CS-WATER-BEAKER-1)]	>0	>0	>0	>0	=0
A[AMOUNT-OF(SK-CS-WATER-VIAL-1)]	=0	>0	>0	>0	>0
A[TEMPERATURE-IN(BRICK)] A[TEMPERATURE-IN(WATER1)]	>	>	=	<	<
ACTIVE(PI0)	F	F	F	T	T
ACTIVE(PI1)	T	T	F	F	F



Processes:

```
  PI0:  PROCESS-1(SK-WATER-1 WATER1 SK-CS-WATER-VIAL-1 BRICK
SK-CS-WATER-BEAKER-1 (COMMON-FACE BRICK WATER1))
  PI1:  PROCESS-1(SK-WATER-1 BRICK SK-CS-WATER-BEAKER-1 WATER1
SK-CS-WATER-VIAL-1 (COMMON-FACE BRICK WATER1))
```

Figure 6. Envisionment produced by the hypothesized caloric model when applied to the brick immersed in water scenario. States are distinguished only by derivative and process values. They are split by QPE when this distinction produces a state lasting an interval of time (S2) and also lasting for an instant (S2-I).

3.3 Preference Criteria

PHINEAS is primarily concerned with *interpretation construction*—to find candidate explanations and the assumptions on which they rest. However, a system that exhaustively generated an unordered set of possible hypotheses would not be of much use. It should focus on the most promising explanations first and provide a preferential ordering on fully developed hypotheses. Correspondingly, PHINEAS incorporates two types of preference criteria; one influences the focus of problem-solving efforts and the other selects among competing, fully developed hypotheses.

The preceding sections presented PHINEAS as a sequential process concerned with the development of a single hypothesis. However, its operation is controlled by a task agenda that maintains multiple hypotheses in various stages of development. Eight task types are currently used, including `access`, `mapping`, `transfer`, and `simulate`. Repeatedly, the task–hypothesis pair at the front of the agenda is selected and executed, resulting in further development of its corresponding hypothesis. This task may in turn spawn other tasks, modify tasks waiting for execution, or signal the acceptance of a hypothesis, which halts the cycle.

The first type of preference criterion influences the ordering of tasks on PHINEAS's agenda and guides it toward developing the most promising hypotheses first. Each task is given a priority level, which induces a roughly depth-first behavior. In addition, the `mapping` and `transfer` tasks have an auxiliary score for sorting tasks within the same priority level. This auxiliary score is SME's evaluation metric for the match between the current observation and the task's associated base analogue. When determining which of two candidate analogues to consider next, PHINEAS selects the one with the higher similarity score. This metric supports the *similarity conjecture*—interpretation-construction tasks may be characterized as the search for maximal explanatory similarity between the situation being explained and some explainable scenario.

Once the system has formed a complete hypothesis (the output of the transfer task), it uses a second type of preference criterion. This criterion considers the characteristics of the hypothesis itself and enables selection among competing hypotheses. A complete account of theory selection requires consideration of many complex factors, such as a theory's plausibility, coherence, effect on prior beliefs, simplicity, and specificity in accounting for the phenomenon. Unfortunately, these are significant open research problems in their own right and are certainly beyond the scope of this chapter. However, a number of important, more specific preference criteria are readily available and have been found useful in PHINEAS for establishing preference between competing hypotheses. These are:

C_{CE} *Conjectured entities.* Does the hypothesis conjecture the existence of a novel kind of entity, and, if so, how many?

C_{VE} *Vocabulary extensions.* Does the hypothesis require the creation of new predicates, and, if so, how many?

C_{CA} *Composite assumptions.* Does the hypothesis conjecture the existence of new physical processes or new knowledge structures (e.g., schemas), and, if so, how many?

C_{AE} *Assumed entities.* Does the hypothesis assume the presence of a known type of entity not mentioned in the original scenario description, and, if so, how many?

C_{AA} *Atomic assumptions.* Does the hypothesis make additional assumptions about the properties and interrelationships of objects in the scenario, and, if so, how many?

The single-preference criterion used to evaluate a hypothesis or compare two competing hypotheses is a function of these five metrics. The method for combining them is adapted from Michalski (1983), who describes the use of a *lexicographic evaluation functional* (LEF) for evaluating alternate inductive concept descriptions. This approach specifies a list of elementary criterion–tolerance pairs, in which each elementary criterion is applied sequentially to prune the space of hypotheses. In PHINEAS, the elementary preference criteria are ordered according to an approximate measure of decreasing "cost":

$$\text{LEF} = (C_{CE},\ C_{VE},\ C_{CA},\ C_{AE},\ C_{AA}).$$

Thus, an explanation that postulates the existence of a novel kind of entity (C_{CE}) is at all times deemed inferior to one that does not. Each criterion returns a number ($N \geq 0$) as described above, where a value of zero indicates success and a value greater than zero indicates failure. The function is used to select the most preferable explanation(s) from a given set as follows: First, each proposed explanation is evaluated by criterion C_{CE}, and those that pass C_{CE} are retained. The process is repeated with the next criterion on the set of retained hypotheses until only a single hypothesis remains or the list of criteria is exhausted. If at any point all hypotheses evaluated by a particular criterion fail, the process stops, and the current set is returned in increasing order according to their score, N, for that criterion.

This evaluative function produces an interesting property when viewed from the perspective of the four explanation scenarios described in Section 2:

1. *Deductive scenario.* Given phenomenon $\mathcal{P}$, where $\mathcal{P}$ represents a set of observables, a complete explanation of $\mathcal{P}$ deductively follows from existing knowledge. This corresponds to explanations passing every criterion. It occurs when all the antecedent features of the base are present in the target.

2. *Assumption scenario.* No explanation can be grounded with current knowledge because not all the relevant facts are known. However, a complete explanation follows from the union of existing knowledge and a consistent set of assumptions about the missing facts. This

corresponds to explanations passing every criterion but one of the last two, C_{AE} and C_{AA}. It occurs when some of the antecedent features of the base have no correspondent in the target, but may be consistently assumed to hold in the target.

3. *Generalization scenario.* Existing knowledge indicates that candidate explanation $\mathcal{E}$ cannot apply because condition C_1 is known to be false in the current situation. However, $\mathcal{E}$ does follow if condition C_1 is replaced by the next most general relation since C_1's sibling is true in the current situation. This corresponds to explanations passing the first two criteria, C_{CE} and C_{VE}, but failing C_{CA}, in which a knowledge structure is viewed as "new" if it represents a modification of an existing knowledge structure.[6] It occurs when some of the antecedent or consequent features of the base match an analogous set of features in the target, thus mapping the base theory to a situation beyond its declared scope.
4. *Analogy scenario.* No candidate explanation $\mathcal{E}$ is available directly, but explanation $\mathcal{E}_b$ is available if a series of analogical assumptions are made, that is, if the situation explained by $\mathcal{E}_b$ is assumed analogous to the current situation. This corresponds to explanations failing one of the first three criteria, C_{CE}, C_{VE}, or C_{CA}. It occurs when some of the features of the base match an analogous set of features in the target or when new vocabulary must be created to complete the mapping.

All four scenarios arise as a result of the same basic mechanism. The evaluative function causes PHINEAS to propose standard, deductive explanations if any are found. In their absence, conventional abductive explanations will be preferred. If existing theories are insufficient to provide an explanation, explanations adapting knowledge of potentially analogous phenomena will be offered. By using similarity as the single source for explanation generation, PHINEAS is able to offer a "best guess" in the presence of an imperfect or incomplete domain theory.

4. Examples of PHINEAS' Behavior

The previous section focused on PHINEAS' explanation of heat flow using a cross-domain analogy with liquid flow. This section describes

[6] The issue of whether to actually create a new knowledge structure or modify the existing one is an important but orthogonal issue. Here we are concerned with hypothesis evaluation rather than storage of an accepted hypothesis.

examples that begin to blur the distinction between analogous phenomena and identical phenomena. In each example, PHINEAS initially begins with knowledge of nine processes—liquid flow, liquid drain (to constantly empty an ideal sink), heat flow, boiling, heat-replenish (e.g., to constantly maintain the heat of a stove), dissolve, osmosis, linear motion, and spring-applied force. The section closes with a discussion of what PHINEAS' behavior indicates about the utility of the proposed analogical model of explanation.

4.1 Oscillation

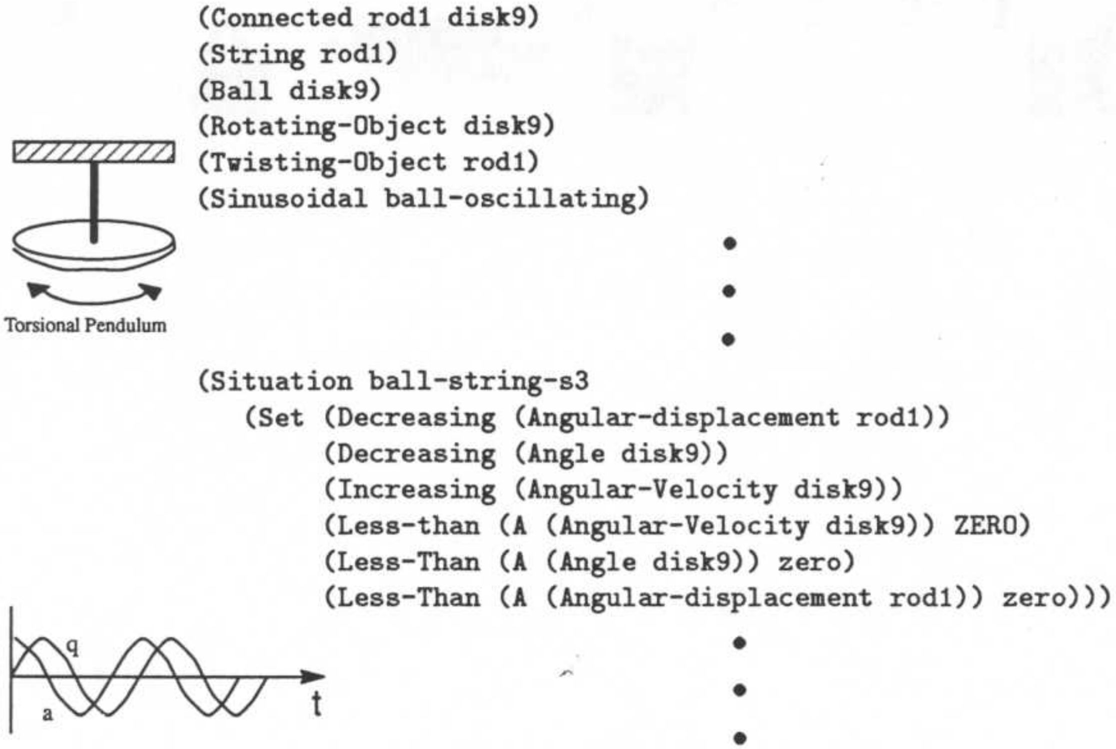
Oscillation is a common phenomenon in physical systems. PHINEAS' initial knowledge contains theories about a prototypical spring-mass system, in which a spring is anchored to a wall on one end and attached to a mobile mass on the other. If the block is pulled and then released, it will oscillate back and forth forever.[7] Drawing from this knowledge, PHINEAS is able to explain several examples of simple harmonic motion, such as an induction-capacitance (LC) circuit and a cantilever pendulum. Here we consider the behavior of a torsion oscillator.

PHINEAS is initially given a description of a disk rotating while suspended by a rubber rod, and the disk's sinusoidal behavior is represented as a cycle of eight qualitatively described temporal intervals (Figure 7). Each interval contains facts describing the derivatives and amounts of angle, and angular velocity. In addition, it is told that the disk is a rotating object and the rod is a twisting object.

When PHINEAS probes memory for prior experiences with sinusoidal oscillation, it finds the spring-mass system. A detailed comparison of this behavior and that of the rotating disk reveals a correspondence between the eight behavioral states of each system. This correspondence indicates that the compressing spring corresponds to the twisting rod and the translating block corresponds to the rotating disk. Additionally, position is mapped to angle, and velocity is mapped to angular velocity, due to their similar behavior.

With a behavioral correspondence established, PHINEAS fetches the domain theory used to explain the spring-mass system. This consists of a `Force` process that applies the spring's force to the attached block, a

[7] Modeling friction and resistance in oscillators is a difficult problem in QP theory. Ideal, frictionless oscillators are discussed throughout this section.



```
(Connected rod1 disk9)
(String rod1)
(Ball disk9)
(Rotating-Object disk9)
(Twisting-Object rod1)
(Sinusoidal ball-oscillating)
                                              •
                                              •
                                              •
(Situation ball-string-s3
   (Set (Decreasing (Angular-displacement rod1))
        (Decreasing (Angle disk9))
        (Increasing (Angular-Velocity disk9))
        (Less-than (A (Angular-Velocity disk9)) ZERO)
        (Less-Than (A (Angle disk9)) zero)
        (Less-Than (A (Angular-displacement rod1)) zero)))
                                              •
                                              •
                                              •
```

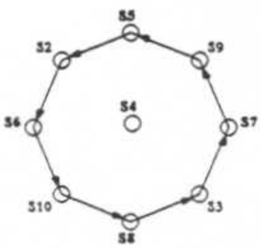
Figure 7. A torsional oscillator and its behavior when the disk is rotated and then released.

spring-mass-system object definition describing the system's total energy and the relationship between the block's position and the spring's displacement, and a **spring** object definition describing its restorative force as a function of displacement. When the spring-mass theory is mapped into the oscillating disk situation, transfer first examines each relation and finds no inconsistencies. The transfer phase next checks for skolem objects in the candidate inference and finds **(:skolem sm-sys)**. The symbol **sm-sys** is a token that represents the spring-mass system taken as a whole. This compound object token is replaced by **sk-sm-sys-23**, which represents the newly defined rod-disk system:

```
(spring-mass-system-22 sk-sm-sys-23 rod1 disk9)
```

The proposed model of the rotating disk scenario is now usable. When the model is applied to the disk–rod pair, it produces an envisionment containing an eight-state cycle, as shown in Figure 8. When PHINEAS examines the envisionment, it finds a perfect match between the observed and predicted behavior. Thus, the model is adequate and the explanation process is completed.

```
•
•
•
```



```
In behavioral segment 3
        (Angular-Displacement rod1) is Decreasing
        (Angle disk9) is Decreasing
        (Angular-Velocity disk9) is Increasing
        (A (angular-velocity disk9)) is Less Than zero
        (A (angle disk9)) is Less Than zero
        (A (angular-displacement rod1)) is Less Than zero

   Due to the following processes being active:
      FORCE-PROCESS(ROD1 RESTORATIVE-FORCE
                    DISK9 ANGULAR-VELOCITY)
```

```
•
•
•
```

Quantity	S5	S2	S6	S10	S8	S3	S7	S9	S4
Ds[ANGLE(DISK9)]	-1	-1	-1	0	1	1	1	0	0
Ds[ANGULAR-DISPLACEMENT(ROD1)]	-1	-1	-1	0	1	1	1	0	0
Ds[ANGULAR-VELOCITY(DISK9)]	-1	0	1	1	1	0	-1	-1	0
Ds[KINETIC-ENERGY(DISK9)]	1	0	-1	0	1	0	-1	0	0
Ds[POTENTIAL-ENERGY(ROD1)]	-1	0	1	0	-1	0	1	0	0
Ds[RESTORATIVE-FORCE(ROD1)]	1	1	1	0	-1	-1	-1	0	0
ACTIVE(PI0)	T	T	T	T	T	T	T	T	T
ACTIVE(PI1)	T	T	T	T	T	T	T	T	T
A[RESTORATIVE-FORCE(ROD1)]	<0	=0	>0	>0	>0	=0	<0	<0	=0
A[ANGULAR-VELOCITY(DISK9)]	<0	<0	<0	=0	>0	>0	>0	=0	=0

Processes:

```
   PI0: FORCE-PROCESS(ROD1,DISK9)
   PI1: DERIVATIVE-PROCESS(ANGLE(DISK9),ANGULAR-
VELOCITY(DISK9))
```

Figure 8. Complete envisionment produced by the hypothesized torsional oscillator model.

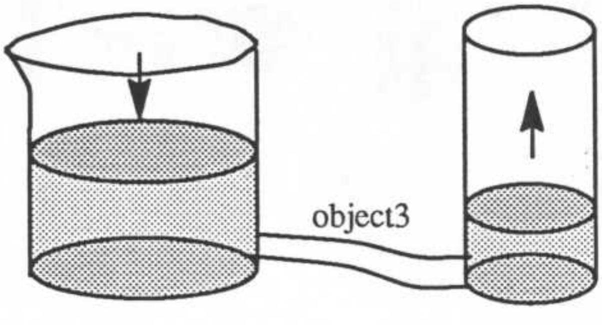


Figure 9. A beaker and a vial, each containing water, are connected by object3. What is causing the water in the beaker to decrease while the water in the vial is increasing?

4.2 Liquid Flow

Consider the scenario illustrated in Figure 9, in which a beaker and a vial, each containing water, are connected by object3. In this situation, the water in the beaker decreases while the water in the vial increases. To explain this phenomenon, PHINEAS begins by probing memory for the best set of candidate analogues. It finds four initial possibilities: **`two-container-liquid-flow`** (score = 28.07), **`leaky-container`** (score = 15.77), **`dissolving`** (score = 14.24), and **`boiling`** (score = 14.14).

Examining each of these possibilities, PHINEAS finds that not only is the **`two-container-liquid-flow`** scenario most similar to the current situation, it is the only candidate that produces a consistent set of predictions. Thus, the system concludes with the single assumption `(Fluid-Path object3)`, which is sufficient to completely explain the observed behavior. PHINEAS' explanation is shown in Table 3. Under this assumption, the situation is viewed as a normal instance of liquid flow, with `object3` serving as the fluid path. The conclusion that this is an instance of liquid flow arises because PHINEAS' model of liquid flow mapped to the current scenario without change, rather than because the liquid flow process was instantiated by deduction.

5. Discussion

This chapter has described a unified, similarity-driven method for explanation that seeks the best match between an observation to be explained and understood phenomena. This enables explanation from existing

Table 3. PHINEAS' analysis of the observation that the amount of water in the beaker is decreasing and the amount of water in the vial is increasing. This analysis requires only a single assumption: `object3` is a fluid path.

```
Hypotheses for theory OBJECT3-FLOW-THEORY-5 derived from
2-CONTAINER-LF:

       (ASSUME (FLUID-PATH OBJECT3))

Analysis of OBJECT3-LF according to theory OBJECT3-FLOW-THEORY-5

       In behavioral segment 1
           (PRESSURE-IN VIAL6) is Increasing
           (AMOUNT-OF CS-WATER-VIAL1) is Increasing
           (PRESSURE-IN BEAKER6) is Decreasing
           (AMOUNT-OF CS-WATER-BEAKER1) is Decreasing
           (A (PRESSURE-IN BEAKER6)) is Greater Than
              (A (PRESSURE-IN VIAL6))

           Due to the following processes being active:
               LIQUID-FLOW(WATER BEAKER6 CS-WATER-BEAKER1 VIAL6
                  CS-WATER-VIAL1 OBJECT3)

       In behavioral segment 2
           (PRESSURE-IN VIAL6) is Constant
           (AMOUNT-OF CS-WATER-VIAL1) is Constant
           (PRESSURE-IN BEAKER6) is Constant
           (AMOUNT-OF CS-WATER-BEAKER1) is Constant
           (A (PRESSURE-IN BEAKER6)) is Equal To
              (A (PRESSURE-IN VIAL6))

           There are no processes active.
```

theories if possible and theory formation or revision when necessary. Importantly, all explanations are formed with a single mechanism, with distinctions among deductive, abductive, and novel analogical explanations arising out of the evaluation process. Initial viability of the method has been demonstrated by PHINEAS on a variety of complex examples from several domains.

This work may be viewed as addressing problems in abduction. Traditional abduction systems reason from a fixed set of theories. However, most concepts, particularly when a theory is developing, do not lend themselves to precise, intensionally defined theories whose boundaries are perfectly specified by a set of necessary and sufficient conditions. Thus, in this work the underlying domain theory is assumed to be imperfect. Reasoning from similarities enables adaptation of the underlying domain theory when needed to explain observed phenomena.

This work may also be viewed as addressing problems in theory formation, in that it provides a way of constraining the set of possible revisions or extensions to existing theories. Reasoning from similarities suggests which theories are relevant and thus are candidates for revision. Also, it enables existing knowledge, possibly of other domains, to influence hypothesis generation and evaluation. It takes into account knowledge of the way things normally behave in the world and the ways theories about those behaviors are normally expressed.

This section briefly evaluates the viability of the approach, reviews related approaches to explanation, and closes the chapter with a discussion of plans for future research.

5.1 Viability of the Model

PHINEAS has been tested on over a dozen examples representing variations on a set of nine basic explanation tasks. In addition to those discussed previously, these examples include explanations of evaporation by analogy to boiling, liquid flow, and dissolving; osmosis by analogy to liquid flow; and floating of a balloon by analogy to an object floating in water.

5.1.1 SUCCESSES OF THE APPROACH

One of the goals of this work has been to show the feasibility of similarity as a single mechanism for both analogical and more traditional explanations involving deduction or abduction. In PHINEAS, the distinctions between deductive, abductive, and analogical explanations arise as an emergent result of the evaluation process. This offers an elegant, general alternative to special case solutions for theory inadequacy problems, such as generalizing a theory's preconditions or developing novel theories through cross-domain analogies. In the examples presented, PHINEAS was shown to explain a simple instance of liquid flow, apply linear oscillation concepts to an angular case, and develop a new "caloric" model of heat flow by analogy to liquid flow.

The primary region of flexibility and power corresponds to what I termed the "generalization scenario" (also classifiable as a form of within-domain analogy). PHINEAS adapts very well to situations close to, but not included in, the stated applicability boundaries of existing theories. An example of this is the mapping of liquid flow through a pipe to liquid flow through an open conduit. Informal experiments have shown that PHINEAS' behavior degrades smoothly with "analogical distance," the degree to which a distant, cross-domain analogy is required.

Due to attempts to find alternative justifications for unsupported dependencies, many of the within-domain analogies examined possess an another interesting characteristic. A new explanation may fall within the scope of existing knowledge, but the relevant explanation schema (i.e., QP theory process definition) is conservatively associated with a more restricted set of scenarios than may be applicable. This would correspond to the often observed comment "I never thought of it working for such a case, but I can see why it should." In other words, one often has the knowledge to safely extend a theory beyond its preconceived boundaries. When viewed from the perspective of explanation-based generalization (DeJong & Mooney, 1986; Mitchell, Keller, & Kedar-Cabelli, 1986), PHINEAS is able to apply a compiled explanation schema by adapting it to fit the current situation, rather than being forced to abandon the explanation schema and solve the current, similar explanation task from scratch. This view is similar in spirit to SWALE (Kass, 1986) and is examined further in Falkenhainer (1989).

5.1.2 LIMITATIONS OF THE APPROACH

PHINEAS falls short of a complete model of explanation in several ways. First, it will always produce a conjecture, no matter how weak, unless it cannot find a candidate analogue to initiate the explanation process or cannot form a hypothesis that is consistent with the observation. This is part of the intended design, and manifests itself in PHINEAS' ability and willingness to generalize a theory in response to an unanticipated observation. However, this is a two-edged sword; violations of existing knowledge are used to indicate new phenomena rather than to indicate false hypotheses. Although the preference criteria will ensure that such hypotheses are selected only if nothing better exists, an explanation system should also be able to know when it lacks knowledge. The framework needs an evaluative measure that takes into account the cost of overthrowing prior beliefs for the benefits of a more coherent belief state. Additional factors, such as plausibility and specificity in accounting for the phenomenon, are required as well.

PHINEAS also differs from more traditional abduction methods in its inability to recognize simultaneous instances of the same phenomenon. Extending the example of liquid flow between two containers in the previous section to "three-container liquid flow" demonstrates this limitation. Three containers connected in series (`can1` to `can2` by `pipe12` and `can2` to `can3` by `pipe23`) produce two simultaneous instantiations of the liquid flow process. When PHINEAS is given a description of the scenario, it finds two different analogies with the potential analogue, "two-container liquid flow." One analogy is with the `can2` to `can1` flow; the other is with the `can2` to `can3` flow. These are processed as two independent candidate explanations. Additionally, they are both adequate, since QPE fortuitously applies the proposed liquid flow model (intended for the `can2` to `can1` pair, for example) to both container pairs. Thus, the system concludes with two consistent, functionally equivalent explanations. It does not possess the knowledge that a single phenomenon was simply occurring twice. This is an important problem that research in analogy has yet to address.

Finally, PHINEAS lacks composability and flexibility to situations that do not fit prepackaged patterns. This is due in part to its inability to merge multiple analogies when forming candidate explanations, as described by Burstein (1983). It is also due to the granularity of theories considered. Schema-application approaches to explanation typically do not adapt as well as rule-chaining systems to novel configurations.

5.2 Related Approaches to Explanation

Although explanation systems differ along many dimensions, two aspects seem particularly relevant in forming class-wide comparisons with PHINEAS. First, knowledge content and use range from rule chaining and schema application to extensionally defined instance-based models. Second, explanation systems differ in how they treat lack of knowledge about the domain or lack of knowledge about the scenario to be explained. This subsection briefly reviews selected approaches to explanation along these two dimensions.

The traditional model of explanation in artificial intelligence depicts a knowledge-rich process that draws inferences from general, intensionally defined domain knowledge in the form of rules or schemas. This inference process further requires a match of antecedent information that is *complete* (all required features are present) and *exact* (each required feature is present in its prespecified form) to enable inference chaining. These characteristics lead to brittleness due to the lack of exact or complete matches to the real world and the need to anticipate all future scenarios. There have been attempts to remedy this situation. For example, Anderson's (1983) ACT system allows partial matching of antecedents using activation levels to control production-rule firing. Alternatively, probabilistic or default reasoning models enable inference in the presence of incomplete knowledge (e.g., Josephson, Chandrasekaran, Smith, & Tanner, 1987; Pearl, 1987). However, these systems still require exact matches of antecedent features and are thus insensitive to the presence of analogous but syntactically distinct features. These limitations are addressed by PHINEAS' analogy mechanism. It enables matching of analogous rather than identical features, reduces the need to have a precisely defined set of necessary and sufficient conditions for each theory, and enables knowledge of a familiar domain to aid reasoning about another domain.

This work shares much of the philosophy behind case-based reasoning, which uses similar past problem solving experiences to solve new cases (Hammond, 1989; Kolodner, Simpson, & Sycara-Cyranski, 1985). Systems in this paradigm have predominately been knowledge weak, with "match strength" used as a basis for believing that the current and retrieved cases share common principles. More recently, attempts have been made to include knowledge in the process, both in matching equivalent yet nonidentical features (Bareiss, Porter, & Wier, 1987) and in subsequent testing and transformation of the initial hypothesis through examination of deeper domain knowledge (Kass, Leake, & Owens, 1986; Simmons & Davis, 1987). Both processes are found in PHINEAS, in which the domain and the concern with across-domain analogies required more sophisticated representations and a more sophisticated notion of analogical similarity. Further, it required a deep causal analysis of the consistency of a hypothesis, both internally and with respect to the observation. Finally, neither the case-based or the traditional knowledge-intensive models tend to address problems in theory formation, such as anticipated yet unknown objects or the creation of new terms (e.g., postulating intrinsic properties of objects).[8]

Explanation systems also differ in their reaction to gaps in available knowledge. As in PHINEAS, most explanation systems can offer explanations in the presence of incomplete knowledge about the scenario to be explained. Probabilistic approaches (Buchanan & Shortliffe, 1984; Josephson, Chandrasekaran, Smith, & Tanner, 1987; Pearl, 1987) examine a priori probabilities assigned to antecedent information. When addressing open-ended, common-sense problems about the world, having such probabilities seems unrealistic. PHINEAS follows work in interpretation and story understanding (e.g., Charniak, 1988; DeJong, 1982), that tends to use schema-based models and identify the assumables as the unknown elements of a relevant and consistent schema. More work is needed to better understand what can be assumed and when.

Explanation systems rarely address a second type of knowledge gap—lack of applicable knowledge about the domain. However, there are a few exceptions in addition to PHINEAS. Pazzani's (1987) OCCAM can infer new causal rules by using knowledge of abstract patterns of causality (e.g., temporally and spatially connected events), and Rajamoney's (this volume) COAST revises existing theories primarily through experimentation. Falkenhainer and Rajamoney (1988) show how COAST and PHINEAS have been integrated, with similarity-driven explanation providing focus and experimentation providing empirical testing of hypotheses. O'Rorke et al. (this volume) introduce explicit metatheoret-

[8] See Karp (this volume), O'Rorke et al. (this volume), and Rajamoney (this volume) for alternative approaches to theory formation and explanation.

ical rules that elegantly give the effect of extending the basic notion of abduction to include assumption of new causal rules. However, that work is still in progress and has not yet addressed problems associated with focusing this process or relating elements of a developing theory to existing theories.

5.3 Directions for Future Research

The problem of retrieving a plausibly useful analogue from memory still stands as the least understood, most important unsolved problem in analogy. Some important progress has been made (Hammond, 1989; Kolodner, 1984), but models of access are still limited by simple representations and specialized forms of within-domain analogy. The two-stage mechanism described in this chapter (first use abstractions to focus on a candidate set, then use structural comparison to prune and order this set) sidesteps important issues. How are these abstractions formed for the stored situations? How are they recognized in the target situation? How are they organized so that an excessive number of analogues are not retrieved?

Composability is a fundamental requirement for any model of explanation. However, in its current form PHINEAS relies on a single analogous explanation structure to explain each new observation. Two capabilities are needed to address this limitation. First, the ability to draw from multiple sources of knowledge is required, as in Burstein's (1983) work on multiple analogies and their composition. Second, a theory revision ability is needed to let PHINEAS repair initial hypotheses that provide incomplete or inconsistent explanations. The two capabilities must interact, since an explanation's inadequacy may arise from an incomplete theory, which requires retrieval of additional knowledge, or a slightly incorrect theory, which requires modification of its components.

PHINEAS and most analogy systems built to date use analogy as their sole learning method. However, analogy, like any other single learning mechanism, is best viewed as a single component in a synergistic cooperation of learning methods. In addition to analogical inference, learning and explanation may be accomplished through sufficient knowledge of unexplained components (Hall, 1989) or abstract patterns of causality (Pazzani, 1987). In scientific investigation, an analogically derived hypothesis may suddenly "come to mind." However, this *flash of insight* may have been preceded by a tedious, incremental process in which data were collected and analyzed, patterns sought, and overall familiarity increased (Langley & Jones, 1988). In order to build a general investigative system, we must integrate analogy with directed experimentation, empirical learning, and analytic learning. Some work on developing a general protocol enabling such interaction has already begun (Falkenhainer & Rajamoney, 1988). However, the protocol leaves many questions unanswered, such as how to take advantage of prior problem solving and trend detection and how to integrate the results of analogy into memory.

Acknowledgements

This work has benefited from discussions with Ken Forbus, Dedre Gentner, John Collins, and Dennis DeCoste. Jeff Shrager and Pat Langley provided greatly appreciated writing assistance. Some of PHINEAS' components were written by Ken Forbus (QPE), Johan DeKleer (ATMS and portions of QPE's rule system), and Dennis DeCoste (DATMI).

This chapter describes work conducted as part of the author's doctoral disseration at the University of Illinois, Department of Computer Science. Support was provided through an IBM Graduate Fellowship and the Office of Naval Research, Contract No. N00014-85-K-0559.

References

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

Bareiss, E. R., Porter, B. W., & Wier, C. C. (1987). PROTOS: An exemplar-based learning apprentice. *Proceedings of the Fourth International Workshop on Machine Learning* (pp. 12–23). Irvine, CA: Morgan Kaufmann.

Bobrow, D. (Ed.). (1985). *Qualitative reasoning about physical systems.* Cambridge, MA: MIT Press.

Buchanan, B. G., & Shortliffe, E. H. (1984). *Rule-based expert systems: The MYCIN experiments of the Standord heuristic programming project.* Reading, MA: Addison-Wesley.

Burstein, M. (1983). Concept formation by incremental analogical reasoning and debugging. *Proceedings of the Second International Workshop on Machine Learning.* Monticello, IL. (Revised version appears in R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach* Vol. 2. San Mateo, CA: Morgan Kaufmann, 1986).

Charniak, E. (1972). *Towards a model of children's story comprehension.* Doctoral dissertation, Laboratory for Artificial Intelligence, Massachussetts Institutue of Technology, Cambridge, MA.

Charniak, E. (1988). Motivation analysis, abductive unification, and nonmonotonic equality. *Artificial Intelligence, 34*, 275–295.

DeCoste, D. (1989). *Dynamic across-time measurement interpretation: Maintaining qualitative understandings of physical system behavior.* Master's thesis, Department of Computer Science, University of Illinois at Urbana-Champaign.

DeJong, G. (1982). An overview of the FRUMP system. In W. Lehnert & M. Ringle (Eds.), *Strategies for natural language processing.* Hillsdale, NJ: Lawrence Erlbaum.

DeJong, G., & Mooney, R. (1986). Explanation-based learning: An alternative view. *Machine Learning, 1*, 145–176.

Falkenhainer, B. (1986). *An examination of the third stage in the analogy process: Verification-based analogical learning* (Technical Report UIUCDCS-R-86-1302). Urbana: University of Illinois, Department of Computer Science. *Proceedings of the Tenth International Joint Conference on Artificial Intelligence.*

Falkenhainer, B. (1988). *Learning from physical analogies: A study in analogy and the explanation process.* Doctoral dissertation, Department of Computer Science, University of Illinois at Urbana-Champaign.

Falkenhainer, B. (1989). *Contextual structure-mapping.* SSL Technical Report, Xerox PARC, 1989.

Falkenhainer, B., Forbus, K. D., & Gentner, D. (1986). The structure-mapping engine. *Proceedings of the Fifth National Conference on Artificial Intelligence* (pp. 272–277). Philadelphia: Morgan Kaufmann.

Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. *Artificial Intelligence, 41*, 1–63.

Falkenhainer, B., & Rajamoney, S. (1988). The interdependencies of theory formation, revision, and experimentation. *Proceedings of the Fifth International Conference on Machine Learning* (pp. 353–366). Ann Arbor, MI: Morgan Kaufmann.

Forbus, K. D. (1984). Qualitative process theory. *Artificial Intelligence, 24*, 85–168.

Forbus, K. D. (1986). Interpreting measurements of physical systems. *Proceedings of the Fith National Conference on Artificial Intelligence* (pp. 113–117). Philadelphia: Morgan Kaufman.

Forbus, K. D. (1988). The qualitative process engine, a study in assumption-based truth maintenance. *International Journal for Artificial Intelligence in Engineering, 3*, 200–215.

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

Gentner, D. (1988). Mechanisms of analogical learning. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning.* London: Cambridge University Press.

Hall, R. J. (1989). Learning by failing to explain: Using partial explanations to learn in incomplete or intractable domains. *Machine Learning, 3*, 45–77.

Hayes, P. J. (1979). The naive physics manifesto. In D. Michie (Ed.), *Expert systems in the micro-electronic age.* Edinburgh: Edinburgh University Press.

Hayes-Roth, F., & McDermott, J. (1978). An interference matching technique for inducing abstractions. *Communications of the ACM, 21*, 401–411.

Josephson, J. R., Chandrasekaran, B., Smith, J. W., & Tanner, M. C. (1987). A mechanism for forming composite explanatory hypotheses. *IEEE Transactions on Systems, Man, and Cybernetics, 17*, 445–454.

Jones, R., & Langley, P. (1988). A theory of scientific problem solving. *Proceedings of the Tenth Meeting of the Cognitive Science Society*, (pp. 244–250). Montreal: Lawrence Erlbaum.

Kass, A. (1986). Modifying explanations to understand stores. *Proceedings of the Eighth Meeting of the Cognitive Science Society* (pp. 691–696).

Kass, A., Leake, D., & Owens, C. (1986). SWALE, A program that explains. In R. Schank (Ed.), *Explanation patterns: Understanding mechanically and creatively.* Hillsdale, NJ: Lawrence Erlbaum.

Kolodner, J. (1984). *Retrieval and organizatinal strategies in conceptual meory: A computer model.* Hillsdale, NJ: Lawrence Erlbaum.

Kolodner, J., Simpson, R. L., & Sycara-Cyranski, K. (1985). A process model of case-based reasoning in problem solving. *Proceedings of the Ninth International Joint Conference on Artificial Intelligence.* (pp. 284–290). Los Angles, CA: Morgan Kaufmann.

Leatherdale, W. H. (1974). *The role of analogy, model and metaphor in science.* Amsterdam: North-Holland.

Michalski, R. S. (1983). A theory and methodology of inductive learning. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach.* San Mateo, CA: Morgan Kaufmann.

Mitchell, T. M., Keller, R. M., & Kedar-Cabelli, S. T. (1986). Explanation-based generalization: A unifying view. *Machine Learning, 1*, 47–80.

Mooney, R. (1987). *A general explanation-based learning mechanism and its application to narrative understanding.* Doctoral dissertation, Department of Computer Science, University of Illinois at Urbana-Champaign.

Pazzani, M. J. (1987). Inducing causal and social theories: A prerequisite for explanation-based learning. *Proceedings of the Fourth International Workshop on Machine Learning* (pp. 230–241). Irvine, CA: Morgan Kaufmann.

Pearl, J. (1987). Embracing causality in formal reasoning. *Proceedings of the Sixth National Conference on Artificial Intelligence* (pp. 369–373). Seattle, WA: Morgan Kaufmann.

Pople, H. (1973). On the mechanization of abductive logic. *Proceedings of the Third International Joint Conference on Artificial Intelligence* (pp. 147–152). Stanford, CA: Morgan Kaufmann.

Reggia, J. A. (1983). Diagnostic expert systems based on a set covering model. *International Journal of Man-Machine Studies, 19*, 437–460.

Simmons, R., & Davis, R. (1987). Generate, test and debug: Combining associational rules and causal models. *Proceedings of the Tenth International Joint Conference on Artificial Intelligence* (pp. 1071–1078). Milan, Italy: Morgan Kaufmann.

Smith, R., Winston, P., Mitchell, T. M., & Buchanan, B. G. (1985). Representation and use of explicit justification for knowledge base refinement. *Proceedings of the Ninth International Joint Conference on Artificial Intelligence* (pp. 673–680). Los Angeles, CA: Morgan Kaufmann.

Winston, P. (1975). Learning structural descriptions from examples. In P. Winston (Ed.), *The psychology of computer vision.* New York: McGraw-Hill.