Where does the sun go when it's hiding?
A computational model of human explanation

Dissertation Prospectus

Derek Devnich, M.A.

Committee:

John E. Hummel, Chair
William F. Brewer
Gerald DeJong
Gary S. Dell
Brian H. Ross

I. Introduction and road map

Humans use explanation in a wide array of situations. We invoke them in circumstances ranging from everyday problem solving (e.g., "Maybe the car won't start because it's out of gas.") to predicting the behavior of complex systems (e.g., "The price of gold will rise because the Federal Reserve has lowered interest rates.") to abstract philosophy (e.g., "Because the innate tendency of humanity is one of violent aggression, life in the state of nature is nasty, brutish, and short.") As these examples illustrate, the purpose of an explanation is to provide a reason *why* something is the case by building a useful causal theory.

The empirical literature on explanation is large and diverse. One class of studies takes a natural-history approach to documenting the content of various explanations (e.g., children's mental models of the day-night cycle, Vosniadou & Brewer, 1994). A second class investigates the beneficial effects of explanation for other tasks, such as transfer in problem solving (e.g., Chi et al., 1989) or evaluating one's domain knowledge (e.g., Rozenblit & Keil, 2002). A third class does not require subjects to generate explanations at all, but rather asks them to evaluate explanations constructed by an experimenter; most of the work on teleological explanations has this flavor (e.g., Lombrozo & Carey, 2006; Kelemen, 1999).

However, this literature is incomplete: These studies require people to generate and evaluate explanations, but they do not provide much evidence about how the process of generating an explanation works. The closest approximation comes from Vosniadou and Brewer (1994), whose data suggest that people explain complex phenomena by analogy to their pre-existing world knowledge. Although the authors' point is persuasive, the details of the process remain unspecified. One possible reason for this lack of empirical data is that there is no theoretical framework to scaffold such research. The goal of the present research is to fill this theoretical gap by using a computational model of human explanation generation.

My starting assumption is that the purpose of explanation is to infer the causes of a state of the world in order to make those causes available for the performance of other tasks such as problem solving and prediction. The model I have developed, Persephone[1], explicitly represents the contents of world states, as well as causal relations between those states. It infers causal models for to-be-explained states of interest by analogy to pre-existing knowledge retrieved from its long-term memory.

I will begin by reviewing the existing literature on explanation, which provides some plausible constraints on the explanation process. This review includes a discussion of the domain to be modeled, children's explanations of the day-night cycle. I will then describe the model's architecture and provide an example of its operation.

II. How people explain: A review of the literature

1. Explanation incorporates prior knowledge into the current situation

Explanation is involved in many cognitive processes, including problem solving, prediction, knowledge assessment, and mental simulation. In all of these applications, the process of explanation *per se* exhibits the same requirement: In order to successfully explain, the explainer needs to be able to (1) make inferences about a new situation based on prior

[1] Persephone was a Greek goddess; the ancient Greeks believed that her comings and goings from Hades caused the changing of the seasons.

information, and (2) make inferences about cause and effect relations between aspects of that situation (Keil, 2006; Lombrozo, 2006). Both types of inferences require the explainer to draw on prior knowledge in order to expand their understanding of the current situation.

Predicating unstated information for problem solving

Many real-world problems are actually symptoms of more fundamental underlying causes. In other words, if the current situation is a symptom of an underlying cause, the reasoner needs to change the cause, not the symptom. Explaining the underlying causes of symptoms as a prelude to problem solving occurs in fields ranging from medicine (a doctor wants to treat the disease, not the symptom) to software engineering (the programmer wants to fix the code that causes a crash, not the crash itself). The first step to solving these problems is reasoning backwards from the manifest symptoms in order to discover the fundamental cause. In order to do this, the reasoner may need to make additional supporting inferences. Thus, one of the key ingredients of explanation is predication: Calling into mind and integrating prior knowledge that can help us bridge the gap between the current situation and its ultimate cause.

Evidence for the role of predication in explanation comes from the showing that explaining improves problem solving performance by prompting problem solvers to predicate properties of the problem and the content domain that they might not predicate otherwise (Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Chi, de Leeuw, Chiu, & LaVancher, 1994). Chi et al. (1989) taught their subjects about force in Newtonian mechanics. Subjects were then given problems with detailed solutions and asked to explain these solutions. After explaining the solutions, the subjects attempted to solve novel transfer problems. The quality of subjects' explanations for the worked problems predicted their success in subsequent problem solving. Good explainers made more statements relating problem components to one another and to principles in the instruction materials. They couched their explanations in terms of goals and mechanisms, frequently mentioning preconditions for actions, the consequences of those actions, and the ultimate goal of the actions. They also made more self-monitoring statements of the form "I understand [something]" or "I don't understand [something]". In short, the good explainers were aware that they had relevant knowledge and tried to make unstated aspects of the problem explicit by incorporating that knowledge.

For example, when talking through a problem in which weights are suspended from a pulley, subjects encountered the equation: $\mathbf{T - m_1 g = m_1 a}$.[2] When asked to explain this equation, a poor explainer said: "Okay, cause the acceleration is due to gravity." This is just a restatement of the problem. In contrast, one of the good explainers said:

> "Okay, so it's basically a way of adding them together and seeing if there is anything left over. And if there is anything left over, it equals the force: mass times acceleration."

This explainer articulated the goal that the equation serves (to see whether the forces balance), and integrated more general domain knowledge by recognizing the equivalence of force and the terms given in the equation. All of this articulated information is unstated in the original problem. In a similar vein, when explaining the forces acting on a block resting on an inclined plane, subjects confronted this statement:

[2] $\mathbf{m_1}$ is the mass of a weight suspended from a pulley, **a** is the acceleration of that mass, **g** is acceleration due to gravity, and **T** is the tension on the line suspending the mass from the pulley. The equation states that the force experienced by an object suspended from a pulley is the difference between the upward pull of the line and the downward pull of gravity.

It is convenient to choose the x-axis of our reference frame to be along the incline and the y-axis to be normal to the incline.

One good explainer said:

> "…and it is very, umm, wise to choose a reference frame that's parallel to the incline, parallel and normal to the incline, because that way, you'll only have to split up *mg*, the other forces are already, component vectors for you."

As in the previous example, this explainer has integrated more general schematic knowledge about how this class of problems is solved (breaking forces into their component vectors) into the current problem, and used it to articulate a general principle.

In a later study, Chi et al. (1994) again found that explanation aided in later problem solving. Subjects studied a brief synopsis on the human circulatory system. The subjects were divided into two groups. The experimental group explained each point of the synopsis, while the control group read the synopsis without explaining. After the study phase, all of the subjects attempted to answer questions about additional properties of the circulatory system that were not described in the original text. The results showed that explainers were most concerned with incorporating knowledge about cause-effect relations in the problem domain. For instance, in response to the statement, **During strenuous exercise, tissues need more oxygen,** an explainer said:

> "During exercise, the tissues, um, are used more, and since they are used more, they need more oxygen and nutrients. And um the blood, blood's transporting it [O2 and nutrients] to them."

Here the explainer articulates a causal model in which the circulatory system works harder when tissues demand more oxygen. Although straightforward, this model does not appear anywhere in the text; the reader has to infer it using the other information in the text and their own background knowledge.

Two general observations are in order. First, all of the good explainers asked – at least implicitly - *Why* is this statement here? What purpose does it serve? They assumed that the written statements fit into some kind of causal structure, and that their task was to discover that structure. This is clear in the circulatory system example above, in which the explainer takes a set of facts and reformats them so that one subset of the facts causes another subset. This preference for cause and effect also drove the kinds of explanations that good explainers produced in Chi et al. (1989). The good explainers were interested in solving physics problems, which required them to apply rules that they had learned for the domain (Newell & Simon, 1972; Anderson & Lebiere, 1998). Each rule is, in effect, a small causal model of the form, *If X is the case, do Y in order to produce Z.* In order to successfully solve a problem, one needs both the right set of rules and a problem representation that is compatible with those rules. This is what the good explainers are producing: A representation of the problem that is more "fundamental" by virtue of the fact that it matches the causal structure of their rule set for that domain.

The second general observation is that Chi's subjects did not produce their explanations as a single, complete thought. Their instructions were to read each statement in the worked

problem, explain it, and then move on to the next statement. They had no difficulty constructing their explanations under these constraints. This suggests that explanation is a naturally constructive process, in which a person can repeatedly retrieve additional information from memory with which to augment their growing explanation.

These results impose several constraints on a model of explanation. First, explanation requires integrating domain knowledge with the current situation or problem. In this way, it is reminiscent of the task that models of analogy typically solve: Analogical inference is the process of using familiar ("source") knowledge to reason about a novel ("target") domain (e.g., Gentner, 1983; Holyoak & Thagard, 1995). Second, explainers have a preference for integrating causal information about the relevant domain into their explanations, so a model of explanation must have an adequate representation of causality. Third, explainers have the ability to integrate new information into their evolving explanations iteratively, so a model must demonstrate the same ability. Although extant models of analogy are well suited to address the first of these constraints, the last two constraints are not addressed by any extant model of analogy.

Predicating unstated information for prediction

Chi and colleagues have shown that explanation incorporates causal information into the explainer's understanding and representation of the current situation. In addition to allowing one to infer the causes of the current situation, a causal model allows prediction of additional consequences of those causes. As such, explanation and prediction are intimately related. Ahn, Brewer, & Mooney (1992) point out that when making predictions, simply having a large amount of arbitrary domain information is not enough; the information must be embedded in a causal model in order to be useful.

Ahn et al. (1992) presented subjects with novel situations that could invoke pre-existing knowledge or not. An example of a novel situation that invokes pre-existing knowledge (at least in University of Illinois undergraduates) is the Korean custom of Kyeah, which is a communal lending arrangement. In this arrangement, each member of the group contributes a small amount of money to a communal fund each month. Every month, a different member of the group receives the entire communal fund and uses it to make a large purchase. The lending continues until every member of the group has had the opportunity to make a large purchase. Although the exact arrangement is unfamiliar to an American audience, the arrangement presupposes schematic knowledge about reciprocity that is familiar. An example of a novel situation that does not invoke pre-existing knowledge (once again, for University of Illinois undergraduates) is the Pacific Northwest Indian custom of Potlatch, which is a form of aggressive gift-giving. In this ceremony, a chief gives large amounts of wealth to a rival chief and other guests in order to prove his superiority to the rival.

The subjects were given descriptions of each situation and answered questions that assessed their understanding of the constraints for that situation. For example, the Kyeah arrangement involves pooling resources and making loans to each person in turn. Questions that evaluated understanding the constraints of the situation were items such as "Will the arrangement work if the first person is known to be unreliable?" (no), and "Does the arrangement require a particular number of people?" (no). Here, the subjects successfully infer the situational constraints because they understand the base cause of those constraints, reciprocity. By predicating the base cause, the subjects are able to predict the consequences of it (or consequences of its absence). In contrast, subjects who are asked to make predictions about the

Potlatch examples do not have any causal knowledge of the domain. Without understanding of the base cause, the details of the manifest situation appear arbitrary, and the subjects find it impossible to make predictions. (The details of Potlatch ceremony become non-arbitrary once you understand it as analogous to warfare (Goldman, 1981).)

2. Explanation and the evaluation of understanding

Generating explanations is easy: Subjects find it trivial to produce some kind of explanation when asked (Chi et al., 1989; Chi et al., 1994; Vosniadou & Brewer, 1994; Taylor, Landy, & Ross, unpublished data). In contrast, the process of evaluating explanations – determining whether the explanation you have constructed is correct - appears to be quite laborious (Clement, 2003; 2006). John Clement's (2003, in press) verbal protocols of experts conducting thought experiments demonstrate the difficulty of evaluating explanations. For example, consider the following problem:

> *Two springs are suspended from a beam. They are identical in every way (same material, same length), except that one spring is twice the diameter of the other. If an identical weight is attached to each of the springs, which one will stretch more?*

The problem requires that one make a prediction based on one's understanding of how springs work; in other words, build a causal model or explanation for springs, and then use it to derive a prediction. Most of the subjects found this problem to be at the limits of their abilities. They took a trial-and-error approach, proposing various transformations of the problem (e.g., "What if the springs were square?") and then examining their implications. Although the subjects found it almost trivially easy to propose new transformations, they found it extraordinarily difficult to evaluate the correctness of those transformations.

This provides evidence that checking an explanation for internal consistency occurs after, and separately from, the postulation of that explanation. This means that it is possible (and even desirable) to model the process of explanation generation separately from the process of evaluating those explanations. However, this assertion is tempered by the second constraint: Much of what we see in the mental simulation protocols is people checking their explanations and then retreating and revising. A complete theory of the relationship between explanation and other cognitive processes (such as problem solving and mental simulation) will require us to model the evaluation process as well. Although I do not intend to model the evaluation process in this work, it remains as an important theoretical gap.

3. Explanation uses analogy to incorporate prior knowledge

We have seen many examples of subjects generating explanations that incorporate their prior knowledge. However, none of the studies described thus far tell us *how* explainers select the appropriate knowledge and incorporate it into their explanations. In their studies of children's mental models of the day-night cycle, Vosniadou and Brewer provide persuasive evidence that children construct mental models of the day-night cycle on the fly (i.e., explain the day-night cycle) by making analogies to pre-existing world knowledge (Vosniadou & Brewer, 1994; Brewer, in press). Using a structured interview, they asked elementary school-age children (first, third, and fifth grade) to describe how day and night come about. The interview

asked questions about the sun (for example: Where is the sun at night? How does this happen? Does the sun move? Does the earth move?), questions asking for an explanation of the day-night process (Here is a person on the Earth; can you make it day for this person? Can you make it night for this person? Tell me again how this happens), questions about the movement of the moon, and questions about the disappearance of stars during the day.

Vosniadou and Brewer hypothesized that children would explain the day-night cycle using their knowledge of the everyday physical world. In this case, the relevant knowledge was the children's experience with the behavior of everyday light sources such as light bulbs. In the children's experience, ways of reducing light from a light source (such as a light bulb) include moving the light further away, moving an object in front of the light, moving the light behind an object, switching the light off, and turning away from the light. In addition, the children have made observations about the day-night cycle that they will incorporate into their explanations; for example, the sun is in the sky during the day but not at night, whereas the moon is in the sky at night but not during the day (although this is not actually true, most children believe it to be true). Applied to the day-night cycle, this suggests explanations such as the sun moves further away at night, clouds block the sun at night, the sun goes under the earth (it's hiding), the sun turns off at night, or the earth rotates away from the sun. This kind of mapping and inference is the sine qua non of analogy. The children have asserted the correspondence of two objects, the sun and a light bulb. Once they have established this correspondence, their desire for analogical coherence (corresponding objects fill corresponding roles) makes them willing to map objects that fill corresponding roles in the sun/light bulb scenarios, even if those objects are highly dissimilar (e.g., the earth rotating away from the sun fills the same role as a person who turns away from a lamp). Once the poorly understood target situation (in this case, the solar system) is completely mapped, the children can generate additional inferences about it from the well understood source (in this case, the light bulb).

To evaluate this hypothesis, Vosniadou and Brewer classified the children's responses to the questions as using one of sixteen mental models, or explanations, for the day-night cycle (Vosniadou & Brewer, 1994, pp. 162-163):

1. Sun is hidden by clouds or darkness
2. Sun and moon move up and down to the ground
3. Sun and moon move up and down to the other side of the earth
4. Sun and moon move up and down, unspecified
5. Sun moves out into space
6. Sun and moon revolve around the earth every day
7. Earth and moon revolve around the sun every day
8. Earth rotates up and down, with the sun and moon fixed at opposite sides
9. Earth rotates up and down, with the sun fixed but the moon moving
10. Earth rotates around its axis, with the sun and moon fixed at opposite sides
11. Earth rotates around its axis, with the sun fixed but the earth moving
12. Earth rotates in an unspecified direction
13. Mixed: Earth rotates and the sun moves up and down [children were assigned to one of the mixed classifications if they used elements from multiple explanations that were not consistent with one another]
14. Mixed: Earth rotates and revolves
15. Mixed: General

16. Undetermined [children were assigned to the undetermined classification if they did not give comprehensible responses or there response was some variant of "just because" such as "God made it that way"]

These data closely adhere to Vosniadou and Brewer's predictions: The explanations incorporate physical mechanisms with which the children are already familiar from every day experience, such as rotation, height, occlusion, and distance. The children attempted to use analogy to incorporate pre-existing knowledge of physical mechanisms into a causal model that correctly predicts the observed behavior of the system. This enabled the majority of children produce sensible and coherent explanations for this novel domain. Although the explanations are wrong by the standards of astronomical science, they correctly integrate the children's pre-existing causal knowledge about the world with the novel domain to the best of their understanding (the study thus provides a window into the explanations of true novices). However, this explanation process is error-prone: Some children do produce mixed or inconsistent explanations. Many of these explanations were classified as mixed because children produced different explanations at different points during the interview. For example:

E: Where is the sun at night?
C: Behind the moon.
E: How does this happen?
C: The earth is rotating. The earth is going around in circles on its axis. And it's making the sun move toward the moon. It looks like it because the earth is rotating.
E: Does the earth move?
C: Yes.
E: Does the sun move?
C: Some, maybe, yes. *(p. 144)*

Vosniadou and Brewer attribute these "wandering" explanations to the children's attempts to reconcile what they had heard (e.g., the earth rotates) with what is observable in the world (the sun is hidden at night). As with Clement's (2003, in press) work, this result highlights the difficulty of evaluating an explanation by showing that children do not necessarily notice when their explanations are inconsistent.

III. A brief description of the model

With our description of the empirical literature on explanation generation in hand, we can turn to the characteristics that a computational model of explanation generation would need to exhibit in order to capture the essential features of the phenomenon. Briefly, a model must be able to:

1. Generalize its knowledge to new situations
2. Represent causality
3. Iteratively extend its explanations
4. The model must remember its explanations for future use

In the following sections, I will describe possible implementations for each of these essential features.

1. The model must generalize its knowledge to new situations

The flexibility with which we generate explanations depends upon two kinds of flexibility in the representations and processes underlying those explanations. The first is *relational* flexibility: The ability to represent roles independently of their arguments. This enables the reasoner to recognize that a particular element is the *same* element across multiple situations. In other words, the reasoner can recognize that the sun is the *same* sun, even when it is filling different roles (revolving, radiating, bending the fabric of space-time, etc). This means that the reasoner can learn about the sun independently of the roles in which it appears, allowing the reasoner to generalize across situations.

At the same time, explanation also requires *semantic* flexibility so that it can exploit partial but imperfect matches between the objects and relations in the current situation and the objects and relations encoded in potentially relevant schemas or examples in long-term memory. For example, imagine that someone knows that Susan owns a Civic, and this person wants know what else might be true given this fact. They might remember a prior case in which Bill owned a jeep, and they could use this prior example as a "source" analog (Gentner, 1983; Holyoak & Thagard, 1989) with which to reason about what Susan is likely to do with her Civic. However, they could only do so if their mental representations of the situations allowed them to tolerate the semantic differences between Susan and the Civic on one hand, and Bill and the jeep on the other (Hummel & Holyoak, 1997).

These two kinds of flexibility also characterize human reasoning using analogies, schemas and rules (Holyoak & Thagard, 1989, 1995; Hummel & Holyoak, 1997, 2003). Accordingly, the point of departure for my attempt to simulate explanation is a LISA, Hummel and Holyoak's (1997, 2003) model of analogy, relational reasoning, and schema induction.

Representation in LISA

LISA is a connectionist computing architecture whose representations and processes capture symbolic information by dynamically binding relational roles to their arguments. LISA represents propositions (such as *owns* (Susan, Civic)) using a hierarchy of distributed and progressively more localist (i.e. symbolic) nodes (Figure 1). At the bottom of the hierarchy, objects and relational roles are represented as patterns of activation distributed over units coding for their semantic features (the small circles in Figure 1). At the next level of the hierarchy, both objects and relational roles are represented by localist *object* and *role* units (large circles and triangles in Figure 1), which share bi-directional excitatory connections with the semantic units describing them. For example, the object unit *Susan* would share excitatory connections with semantics such as *human*, *female*, *blonde*, etc. Role-argument bindings are encoded by localist *sub-proposition* units (rectangles in Figure 1), which share bi-directional excitatory connections with the object and role units they bind together. At the top of the hierarchy, localist *proposition* units (ovals in Figure 1) bind individual sub-propositions together into complete propositions.

****
Figure 1 goes about here
****

LISA's knowledge representations are compartmentalized into "analogs," collections of units that represent the propositional content of individual events, stories, concepts, or schemas. Within an analog, a given object or role is represented by a single unit across all propositions in which it plays a role. For example, the object *jeep* would be represented by the same object unit in the proposition *drives* (Bill, jeep) and in *crashes* (Bill's brother, jeep). However, separate analogs do not share object, role, sub-proposition, or proposition units; the jeep is represented by one object unit in one analog and by a different object unit in another analog. In other words, object and role units do not represent objects or roles in the abstract; they represent specific instantiations or *tokens* of those objects or roles in specific analogs. (The same is true of sub-proposition and proposition units.) As such, I will collectively refer to object, role, sub-proposition, and proposition units as *token units*. In contrast to the token units, all analogs connect to the same pool of semantic units. The semantic units thus represent the abstract *types* to which the tokens refer.

Operations in LISA

The hierarchy depicted in Figure 1 represents the static structure of propositions within a single analog, both in LISA's long-term memory and (when a proposition becomes active) in its working memory. However, analogy requires discovering how roles, arguments, and their bindings correspond *across* analogs. In order to determine how the elements of one analog correspond to the elements of another, the first analog must communicate information about its contents and their bindings to the second analog. It does so by synchrony (i.e. timing) of firing: Relational roles fire in synchrony with the arguments to which they are bound, and separate role-argument bindings fire out of synchrony with one another. From the point of view of the second, recipient analog, the first analog creates a pattern of activation on the semantic units representing the role and argument that are currently firing. Roles, arguments, and their upstream binding units become active in response to the semantic units to the extent that the roles and arguments are connected to those units; in other words, the roles and arguments in the second analog will become active to the extent that they are similar to the currently-firing role-argument binding in the first. LISA assumes that units in the second analog correspond to units in the first when those units are active at the same time.

For the purposes of LISA's operation, analogs are divided into three sets. A *driver* and one or more *recipients* are assumed to reside in *active memory* (a primed subset of long-term memory that is larger than working memory; Cowan, 2001); the remainder are dormant in long-term memory. All of LISA's operations are controlled by the driver. One at a time, propositions in the driver become active and enter the *phase set*, the set of active role bindings. The phase set is LISA's working memory, and like human working memory (see Cowan, 2001), it is limited to holding at most 4-6 role bindings at a time. The patterns of activation that propositions in the phase set generate on the semantic units excite other propositions in LISA's long-term memory during memory retrieval, and in its active memory during mapping, analogical inference and schema induction.

LISA performs memory retrieval as a form of guided pattern recognition (Hummel & Holyoak, 1997). Patterns of activation generated on the semantic units by one proposition tend to activate other, similar, propositions in long-term memory, bringing them into active memory. For example, the patterns activated by the proposition *owns* (Susan, Civic) might activate the

proposition *owns* (Bill, jeep). Propositions in active memory are similar to those in long-term memory to the extent that they have similar roles and similar arguments with similar bindings.

****
Figure 2 goes about here
****

LISA discovers analogical mappings by learning which structures in the recipient tend to become active in response to structures in the driver. In this trivial analogy, Bill bound to *owner* activates Susan bound to *owner* in the target, and jeep bound to *owned* activates Civic bound to *owned*. LISA thus maps Susan to Bill and Civic to jeep. The same is true for corresponding roles of the *owning* relation, and the sub-proposition and proposition units binding those roles to their fillers.

LISA represents these correspondences as learned *mapping connections* between corresponding structures (e.g., between Susan and Bill). These connections serve not only to represent the learned mappings, but also to constrain future mappings: If LISA maps Bill to Susan in the context of *owning*, then the resulting mapping connection will cause Bill to directly activate (and therefore map to) Susan in subsequent propositions. The learned mapping connections play a central role in LISA's capacity for *self-supervised learning*, the core of its algorithm for analogical inference and schema induction. For example, once LISA maps Bill to Susan, jeep to Civic, *owner* to *owner*, and *owned* to *owned*, it can generate new units in the target corresponding to unmapped source units (such as the *wanting* relation). It can combine these newly-inferred units with pre-existing units (such as Susan) to form whole propositions (e.g., Susan wants to own a Civic; Holyoak & Thagard, 1989). Finally, augmented with a simple algorithm for intersection discovery, LISA's algorithm for analogical inference also provides a very natural account of the induction of abstract schemas from concrete examples (Hummel & Holyoak, 2003).

LISA's knowledge representations ("LISAese"), along with its algorithms for memory retrieval, mapping, inference and schema induction, provide a natural account of roughly 50 phenomena in the literature of analogical thinking, as well as 15 or more in cognitive development (see Doumas, et al., 2008; Hummel and Holyoak, 1997, 2003; Hummel & Ross, 2006; Morrison et al., 2004; Richland et al., 2006; Viskontas et al., 2004). These abilities derive from the fact that LISAese simultaneously enjoys the flexibility of distributed connectionist representations and the relational sophistication of symbolic approaches to knowledge representation. As such, they are an ideal platform on which to build a model of understanding and explanation.

2. The model must represent causality

Given that explanations invoke causal relations, how should they be represented? There are several options. At one extreme, once could represent causes as relational roles inside the pre-existing LISA architecture (Figure 3 is an example of this style of representation). Although tempting, this representational format is almost certainly wrong, at least as the exclusive basis for representing causal relations. Recall that working memory tasks such as analogy are capacity-limited (Cowan, 2001; Halford et al., 1998; Holyoak, 2005). LISA captures this limitation by using synchrony for dynamic role-filler binding: The model can only keep a limited number of

bindings simultaneously active and mutually *out* of synchrony with one another. If people represent causal relations exclusively as explicit propositions, then causal reasoning, like general relational reasoning, should be strongly tied to working memory capacity.[3]

****
Figure 3 goes about here
****

Evidence suggests that this is not the case: Causal reasoning and general relational reasoning appear to be separate capacities. Support for the difference between causal reasoning and general relational reasoning comes from studies of relational reasoning in children. Numerous studies have found children have a reduced ability to reason relationally, relative to adults (e.g., Halford, 2005; Richland, Morrison, and Holyoak, 2006; Sera and Smith, 1987; Smith, 1989). For example, Richland et al. gave elementary school-age children a picture analogy task in which they saw pairs of scenes, such as a picture of a dog chasing a cat chasing a mouse, paired with a picture depicting a child chasing a second child, who was in turn chasing a third child. The children had to tell the experimenter which object in one scene "goes with" a particular object in the other scene (e.g., which object in the second scene "goes with" the cat in the first). Richland et al. found that the children could identify the relationally similar object (e.g., the *chaser* in a scene in which one thing chases another) when the relation in question was a two-place relation (e.g., dog chases cat), but had difficulty when the relation was a three-place relation (or a pair of two-place relations, as in dog chases cat chases mouse). Adults find the same task trivially easy. In spite of these deficiencies in their capacity for generalized relational reasoning, the children in Vosniadou and Brewer's study show adult-like reasoning about cause and effect. Where their explanations lack sophistication, it is due to their lack of worldly knowledge. In fact, even young infants have rational expectations about cause-effect relations, and are troubled when those expectations are violated in domains they understand (such as physical reasoning; Baillergeon, 2004; Leslie, 1994).

A more dramatic demonstration of the dissociation between causal reasoning and general relational reasoning is the finding that rats correctly use causal models to determine what they can infer from observation and intervention. In work by Blaisdell, Sawa, Leising, and Waldmann (2006), rats learned one of two causal models. The first was a *common cause* model, in which the onset of a light predicted both the onset of a tone and the availability of food (light ➔ tone; light ➔ food). The second was a *causal chain* model, in which the onset of a tone predicted the onset of the light, which in turn predicted the availability of food (tone ➔ light ➔ food). An observer would conclude that the tone is associated with food in both conditions; it is not readily apparent by observation that the tone is the cause of food in the second case but not in the first. Indeed, rats trained on the common cause model or the causal chain model were equally likely to search for food when presented with the tone.

[3] For the sake of clarity, I will use the term "generalized relational reasoning" to refer to the ability to reason relationally about *arbitrary* domains (e.g., the ability to reason about the solar system based on one's understanding of lamps), an ability that is likely uniquely human (see Penn, Holyoak & Povenelli, in press). In contrast, it is probable that humans and other animals can engage in certain types of domain-specific relational reasoning using dedicated modules. As I will argue below, evidence for causal reasoning in rats and other non-human animals suggests that causal reasoning is *not* dependent on the capacity for generalized relational reasoning. Instead, it appears that the ability to learn and represent causal relations is more basic, more modular, phylogenetically older, and less dependent on working memory.

Blaisdell and colleagues point out that giving the rats the opportunity to intervene – in other words, giving them the opportunity to perform controlled experiments – could provide them with the information necessary to distinguish between the causal models (cf. Pearl, 2000). They trained the rats from the previous study to intervene by pressing a lever that produced the tone. Rats whose environment was dictated by the common cause model discovered that they could produce the tone independently of the light. These rats no longer looked for food when tested with the tone; they had learned that the tone and the food were causally independent (bar press ➔ tone; light ➔ food). In contrast, rats whose environment was dictated by the causal chain model discovered that they could produce the tone, which produced the light, which produced the food (bar press ➔ tone ➔ light ➔ food). These rats were just as likely to look for food when tested with the tone as they had been in the observation condition. Blaisdell and colleagues conclude that the rats were not just representing statistical associations between stimuli; they had inferred a causal model of their environment that was consistent with their experiments.

As neither rats nor infants are famous for their ability to perform generalized relational tasks such as analogy or problem solving, these findings indicate that causal reasoning cannot depend on general relational reasoning. This in turn suggests that causal relations do not depend on the same kinds of relational representations as general relational reasoning.

How, then, should causes be represented? The salient feature of causal models is that they represent contingencies between states of the world. Learning these contingencies is useful for any organism that wants to learn how its environment works, so it is unsurprising that many organisms represent causality in a similar (and normative; cf. Cheng, 1997) way. Therefore, a model that captures causality needs to represent states of the world and the contingent relationships between those states (in other words, which states are causes and which states are their effects). I will posit a new kind of representational unit - the *group* unit - to explicitly represent states and determine their place within a causal structure.

Representing causal relations with groups

People engage in general relational reasoning, and the LISA representational scheme does an excellent job of capturing this capacity. People also engage in causal reasoning – reasoning about contingencies between states - and we have good reason to believe that the current LISA representational scheme is not appropriate for capturing this capacity. Therefore, the guiding intuition for Persephone is that people represent *individual* states of the world as collections of LISA-like relational propositions, and represent contingencies *between* states of the world as relationships among group units that are specifically adapted to representing causality.

There are two types of group units, each performing a specific function. The first of these functions is to explicitly identify individual states of the world. A *state* group unit does this by forming a bilateral, excitatory connection with all of the propositions that collectively describe that state. The second function is to explicitly identify causal contingencies between states. A *link* group unit does this by forming bilateral, excitatory connections with a state group describing a cause and with a state group describing that cause's effect. Each group unit is connected to semantic units that specify whether it represents a cause state, an effect state, or a link between states.

A typical causal structure appears in Figure 4. Here, the causal model states that if the sun goes behind the hills, this causes it to disappear, which in turn causes night (NB: in this diagram, units below the level of sub-proposition have been removed for clarity).

****
Figure 4 goes about here
****

There are several advantages to this representation. The first is that causal contingencies are not represented as relational roles (i.e., Predicate units), and so do not occupy working memory slots. The second is that group units can take advantage of the same algorithms for mapping and inference that the model applies to its other units. This makes it possible for the model to infer causal information (i.e., groups and their connections) from prior examples; in short, it allows the model to make causal inferences by analogy. For example, Figure 5a shows how a child might infer the role of the moon in causing day by making an analogy to the model shown in Figure 4. Using LISA's preexisting algorithms for mapping and inference, the inference is straightforward (Figure 5b).

****
Figures 5a and 5b go about here
****

3. The model must iteratively elaborate its explanations

The example described in Figure 4 assumes that a complete causal model, relevant to the situation of interest, is already available for retrieval from LTM. This assumption ignores one of the primary benefits of explanation, which is the ability to construct more extensive causal models from smaller fragments. As noted previously (Chi et al., 1989; Chi et al., 1994; Vosniadou & Brewer, 1994), people frequently extend their explanations with newly-retrieved fragments, and a model of explanation generation should be able to demonstrate the same behavior.

Although it may not be immediately apparent, this requirement is problematic for a model of analogy. In analogy terms, we want the model to construct a large, elaborate target explanation using a series of small source analogs. As demonstrated below, the current LISA architecture cannot do this.

The problem of accommodation in analogy

The fundamental task of a model of analogy is to find correspondences between two different representations. In order to accomplish this task, the model must be willing to accommodate the differences between those two representations (for example, the model must be willing to assert that "sun" corresponds to "lamp," even though they are, in fact, different things). Put another way, the model's task is to find the best possible fit between the representations given their disparities.

To the extent that a model of analogy accommodates the differences between representations – that is, to the extent that it tries to find the best fit, even when all fits are poor –

it will be unwilling to reject a fit as being *too* poor. In the extreme, such a model could assert that "sun" corresponded to "aardvark" if that was the best match available. It could assert that "sun" corresponded to *both* "aardvark" and "soliloquy" if the second representation contained more items than the first. It could not guarantee that a role and its argument in one representation would correspond to a role and its argument in the other. In short, it would make poor analogies, at least as people understand them.

A successful model of analogy needs some constraints on its willingness to accommodate differences. For example, the LISA model imposes the one-to-one mapping constraint, forbidding any element from mapping to more than one element in the other representation (or "analog", in LISA parlance). It also imposes a unit type mapping constraint, so that only units of a given type can map to one another (for example, roles only map to roles, objects only map to objects, and so on). Together, these constraints mean that if two analogs have differing numbers of elements, some units in the larger analog will go unmapped. The model interprets this absence of mapping as a cue to infer into the smaller analog units corresponding to the unmapped units in the larger analogy. Thus, these constraints not only allow LISA to make sensible analogies, they allow it to make sensible inferences based on those analogies.

These constraints also mean that in order to make inferences about a target analog, the source analog used to drive inference *must have more elements than the target*. If the source has the same number of, or fewer elements than, the target, the model can find mappings for all them, leaving no unmapped elements to drive inference. This poses a challenge for iterative explanation, in which the reasoner retrieves a series of small source fragments in order to reason about an ever-expanding target explanation. If we want a newly-retrieved source fragment to drive inference of new elements in the target's causal chain, we must prevent that fragment from mapping to preexisting elements.

Iterative inference of a causal model

Persephone possesses two pieces of knowledge that allow it to successfully engage in iterative inference. The first is a memory for which target units mapped to some source unit during a previous iteration. The second is the direction of inference: Persephone knows whether it is explaining or predicting, and consequently whether it is trying to extend the causal chain in the target backwards or forwards. During an explanation (for example), Persephone posits that the earliest cause in the causal chain may itself be the effect of something else. It uses that cause to probe its long-term memory, searching for causal fragments with effects similar to the probing cause. Once it successfully retrieves a causal fragment, it will map the elements of the probing cause – the earliest cause in the target – onto the last effect in the retrieved source. The only elements in the target that Persephone allows to map are elements with no previous mappings and the elements of the propositions composing the earliest cause; all other mappings are prohibited. Thus, the number of source elements is greater than the *effective* number of target elements (i.e. the number of target elements eligible for mapping), and so the source can drive inference of new causal elements in the target.

****

Figures 6a, 6b, 6c, and 6d go about here

****

Figure 6 demonstrates a more realistic version of causal inference based on multiple retrieved fragments. In this example, a child knows that the sun is behind the earth at night and is trying to explain how this happens. Using the proposition *behind* (sun, earth), they probe long-term memory for similar effects and retrieve a familiar schema stating that when one object moves around another object, it ends up behind that second object (Figure 6a). Using this retrieved source, the child infers that the sun moves around the earth, causing it to end up behind the earth (Figure 6b). But what causes the sun to move around the earth? The child can posit that the newly-inferred cause is, in turn, the effect of something else. The child probes long-term memory using the newly-inferred cause and retrieves an example from previous experience. The example states that if you are facing a lamp that is stationary and you rotate away from it, the lamp appears to move around you (Figure 6c). Because all of the elements in the target mapped to the previous source, they are all (with the exception of the current probe) prohibited from mapping to the new source. Instead, the new source drives inference of new elements in the target's evolving explanation: The child infers that the earth rotates away from the sun, causing the sun to appear to move (Figure 6d).

If we provide Persephone with a mechanism for iteratively elaborating an explanation, we must also provide it with a criterion for when to stop. One possibility is that it could halt when it has created a causal chain with an ultimate cause it believes to be plausible. Another is that it could interpret a failure to retrieve any additional information as a sign that the explanation is "complete," in the sense that it has explained as much as it can.

Each of these approaches has its own problems. Using a plausible ultimate cause requires having some a priori justification for what makes a cause plausible; at this juncture, such justification would rest more on the modeler's taste than on actual data. Using retrieval failure as the criterion requires that the model *can* fail to retrieve, even when marginally similar propositions still exist in long term memory. However, this approach has the virtue of being agnostic with respect to the particular contents of any given explanation; instead, it would extend the general, similarity-based retrieval algorithm that the model already uses. Accordingly, this is the approach Persephone adopts: Its chance to retrieve a particular state in long-term memory is proportional to the similarity of that state to the retrieval cue. The retrieval algorithm scales the probability of retrieval so that the total probability of retrieving anything from long-term memory is less than 1, allowing for the possibility of no retrieval.[4]

4. The model must remember its explanations for future use

Explanation creates useful causal models. In order to make those models readily available for future tasks, Persephone must save the results of its reasoning. Its architecture makes this straightforward: A new explanation is an episode (i.e. an "analog") in LISAese that can be saved to LTM and retrieved in the future just like any other episode. In order to avoid filling its long-term memory with failed explanations (i.e. targets about which it never made inferences), it only saves episodes that include at least one inferred unit.

IV. Model predictions

[4] Specifically, $p[s] = s_T / (k + S_T)$. This states that the probability of retrieving state *s*, p[*s*], equals the similarity of *s* to the target probe, s_T, divided by the sum of all the similarities of the states in long-term memory to the target, S_T, and a constant probability of non-retrieval, k.

The architecture as previously described leads to a few broad classes of predictions.

1. The primacy of knowledge

The first general prediction of this approach is that the structure of an explanation will be strongly driven by content, i.e. what the reasoner knows. The proposed component processes of explanation - retrieval, mapping, and inference - are all simple. The approach assumes that these processes are applied in a relatively invariant order across explanation episodes. In other words, the process of explanation is the same for any explanation; only the contents of the explanations will differ. In turn, these contents depend less on the basic algorithmic operations that assemble them than they do on the contents of the long-term memory from which they are retrieved.

Thus, the approach makes the negative prediction that the model will produce a range of plausible, human-like explanations without resorting to special mechanisms to make those explanations long or short, to make them rich or impoverished; to make them consistent or inconsistent (more on this below), to determine when an explanation is complete, or to explain why, in some cases, people make no explanations (i.e. causal inferences) at all (think of the student who "explains" a principle by restating the principle in slightly different words). The model's retrieval mechanism will probabilistically retrieve items to the extent that they are similar to the target probe. This could result in the retrieval of long causal chains, of individual cause-effect fragments, or nothing at all. Regardless of what it retrieves, the model will attempt to map it onto the target situation and make appropriate inferences where warranted. In short, the model's mechanisms will be indifferent both to the topic of reasoning and to whether the goal is explanation or analogy.

2. The ubiquity of analogy

This prediction flows directly from prediction one. Given that the general, knowledge-based mechanisms of the model will be those of analogy, the modeling approach predicts that explanation will have the same strengths and limitations as analogy. For example, with regard to working memory capacity limits, explanations that require representing a state that exceeds the capacity of working memory will not be generated, even if the reasoner has the requisite knowledge. It should resemble analogy in the nature of retrieval (primarily object and first-order relation-based), the nature of mapping (primarily role-based) and inference (licensed only when the mapping is good enough (Lassaline, 1996). [5]

3. Contingent development of preferences

The model's long-term memory is episodic. New instances of reasoning will create new episodes that the model will add to its episodic memory. The inevitable consequence is that the model's preference for one class of explanation over another can be contingent on chance retrievals during early reasoning episodes. Consider the situation in which the model knows two propositions about light sources: Light sources produce less light when they are far away, and light sources produce less light when they are covered. When asked to explain what happens to the sun at night, the model will have an equal chance of retrieving either proposition. If it

[5] Note that this capacity limit applies to state representations, not causal links, which, as previously established, do not depend on working memory capacity.

retrieves the first, it will make the inference that the sun moves far away at night; if it retrieves the second, it will make the inference that the sun is covered (say, by clouds) at night. In either case, the results of this reasoning will be saved in long term memory. The next time the model is asked this question, it will know three facts that are relevant. Two will support a single class of explanation, and one will contain items that are highly similar to items in the question (such as the sun). Once the model has embarked on an explanatory trajectory, that trajectory will tend to persist. This predicts that explanations will tend to persist, and change only slowly in the face of new information, a prediction that is consistent with the behavior of children (Vosniadou & Brewer,1994).

4. Consistency as a happy accident

Creating explanations and evaluating explanations are separate processes. If evaluating explanations is difficult, then why are explanations ever internally consistent? The answer predicted by the modeling framework proposed here is that the world is frequently consistent, so our knowledge of the world is also frequently consistent. Consistency is thus a happy accident: People build explanations using collections of propositions that are similar to one another, and similar propositions tend to cohere because they are learned by observing a world that coheres. However, there are no guarantees. If your physical experience tells you that the world is flat, and your teachers tell you that the world is round, you are – according to the model - perfectly capable of accommodating both of those propositions in an explanation that is internally inconsistent. The use of post-hoc evaluation is the main distinction between folk theories and scientific theories: The latter are laboriously checked by many parties, and end up being unusually consistent.

V. Unsolved problems

There are a variety of thorny issues related to explanation that I do not intend to solve in this work. However, I wish to acknowledge them for the sake of completeness. Most of these problems boil down to a single question: Where is the burden of proof?

> If an explanation is internally inconsistent, which inconsistent component should you remove? The question is undecidable within the context of the explanation.
>
> If an explanation or prediction contradicts some other piece of knowledge, which is in error? Once again, the question is undecidable.
>
> If you make a chain of inferences (backward for explanation, or forward for prediction), and you decide that the final result is incorrect, how far back the chain should you retreat? Where does the error lie?

All of these questions are formally undecidable; practically speaking, they all demand some kind of tiebreaker, in which propositions that command a higher degree of confidence win out over propositions that command a lower degree of confidence. In most cases, people will have to learn confidence information from the world (plausible candidates for confidence information are frequency, utility, causal power…). Therefore, the ability to evaluate explanations, both for

internal consistency and for their relationship to other knowledge, requires two mechanisms: A mechanism for learning confidence information from the world, and a mechanism for knowledge comparison and reorganization that uses that information. Both of these are monumental problems that I will save for another day.

References

Ahn, W., Brewer, W. F., & Mooney, R. J. (1992). Schema acquisition from a single example. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(2), 391-412.

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Baillargeon, R. (2004). Infants' physical world. *Current Directions in Psychological Science, 13*, 9–94.

Blaisdell, A. P., Sawa, K., Leising, K. J., & Waldmann, M. R. (2006). Causal reasoning in rats. *Science, 311*(5763), 1020-1022.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104*(2), 367-405.

Chi, M. T., Bassok, M., Lewis, M. W., & Reimann, P. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13*(2), 145-182.

Chi, M. T. H., de Leeuw, N., Chiu, M., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18*(3), 439-477.

Clement, J. (in press). Thought experiments and imagery in expert protocols. In L. Magnani, ed., *Model-based reasoning in science and engineering*, pp.1-16. London: King's College Publications.

Clement, J. (2003). Imagistic simulation in scientific model construction. *Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society, 25.* Mahwah, NJ: Erlbaum.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*, 87-114.

Doumas. L. A. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review, 115*, 1 - 43.

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

Goldman (1981). *The mouth of heaven: An introduction to Kwakiutl religious thought*. Huntington, NY: RE Krieger Pub. Co.

Halford, G. S. (2005). Development of thinking. In K.J. Holyoak & R.G. Morrison, eds., *The Cambridge handbook of thinking and reasoning.* New York, NY, US: Cambridge University Press.

Halford, G. S., Wilson, W. H., & Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral and Brain Sciences, 21*(6), 803-864.

Holyoak, K. J. (2005). Analogy. In K.J. Holyoak & R.G. Morrison, eds., *The Cambridge handbook of thinking and reasoning.* New York, NY, US: Cambridge University Press.

Holyoak, K. J., & Thagard, P. (1989). Analogical mapping by constraint satisfaction. *Cognitive Science, 13*, 295-355.

Holyoak, K. J., & Thagard, P. (1995). *Mental leaps: Analogy in creative thought*. Cambridge, MA: MIT Press.

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review, 104*, 427-466.

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review, 110*, 220-264.

Hummel, J. E., & Ross, B. H. (2006). Relating category coherence and analogy: Simulating category use with a model of relational reasoning. In *Proceedings of the Twenty Fourth Annual Conference of the Cognitive Science Society*.
Keil, F.C. (2006). Explanation and understanding. *Annual Review of Psychology. 57*, 227-254.
Kelemen, D. (1999). Function, goals and intention: Children's teleological reasoning about objects. *Trends in Cognitive Sciences, 3*(12), 461-468.
Lassaline, M. E. (1996). Structural alignment in induction and similarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(3), 754-770.
Leslie, A. M. (1994). Pretending and believing: Issues in the theory of ToMM. *Cognition, 50*(1-3), 211-238.
Lombrozo, T. (2006). The structure and function of explanations. *Trends in cognitive sciences, 10*(10), 464-470.
Lombrozo, T., & Carey, S. (2006). Functional explanation and the function of explanation. *Cognition, 99*(2), 167-204.
Morrison, R.G., Krawczyk, D.C., Holyoak, K.J., Hummel, J.E., Chow, T.W., Miller, B.L., & Knowlton. (2004). A neurocomputational model of analogical reasoning and its breakdown in Frontotemporal Lobar Degeneration. *Journal of Cognitive Neuroscience, 16*, 260-271.
Newell, A. & Simon, H.A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall, Inc.
Pearl, J. (2000). *Causality: Reasoning, models, and inference.* Cambridge, UK: Cambridge University Press.
Penn, D ., Povinelli, D., & Holyoak, K. J. (In Press). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences*.
Richland, L.E., Morrison, R.G., & Holyoak, K.J. (2006). Children's development of analogical reasoning: Insights from scene analogy problems. *Journal of Experimental Child Psychology, 94*, 249–273.
Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science, 26*(5), 521-562.
Sera, M., & Smith, L. B. (1987). Big and little: "Nominal" and relative uses. *Cognitive Development, 2*, 89–111.
Smith, L. B. (1989). From global similarities to kinds of similarities: The construction of dimensions in development. In S. Vosniadou & A. Ortoney (Eds.), *Similarity and analogical reasoning* (pp. 147–177). Cambridge, England: Cambridge University Press.
Sutton, R. S., & Barto, A.G. (1998). *Reinforcement Learning: An Introduction.* Cambridge, MA, US: The MIT Press.
Taylor, E., Landy, D.L., & Ross, B.H. (In preparation) Unpublished protocols on explanation generation.
Viskontas, I., Morrison, R., Holyoak, K. J., Hummel, J. E., & Knowlton, B. J. (2004) Relational integration, inhibition, and analogical reasoning in older adults. *Psychology and Aging*, *19*, 581 – 591.
Vosniadou, S., & Brewer, W. F. (1994). Mental models of the day/night cycle. *Cognitive Science, 18*(1), 123-183.

Figure 1. The structure of a single proposition

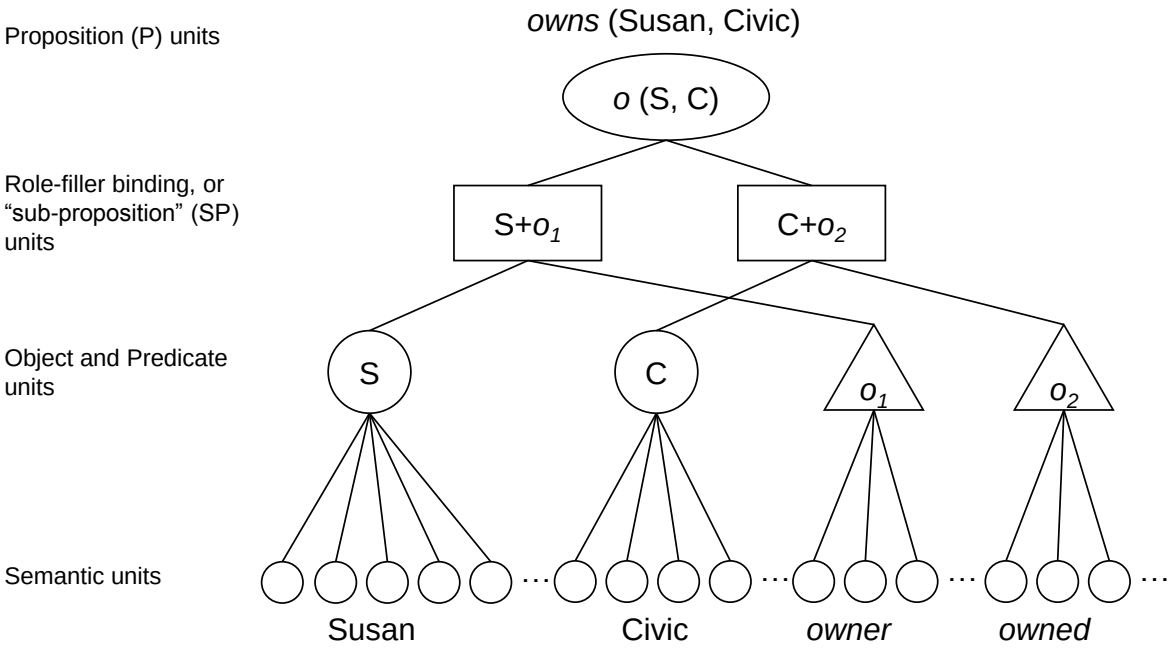


Figure 2. Mapping between analogs. The target states that Susan owns a Civic, while the source states that Bill wants to own a jeep.

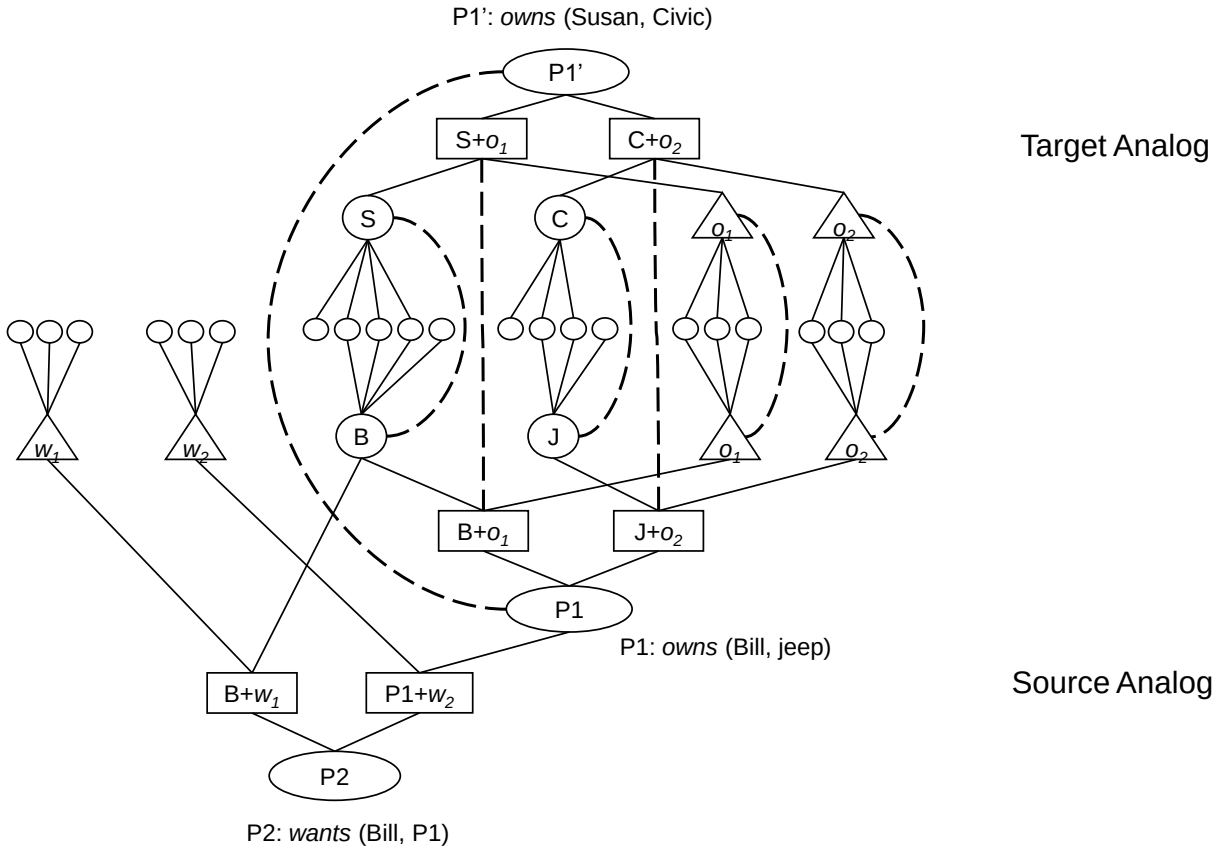


Figure 3. What would a causal model looked like if it was represented using relational roles? The causal model states that when the sun goes behind the hills, that causes the sun to disappear, which causes night [*goes-behind* (sun, hills) ➔ *disappear* (sun) ➔ *night* (earth)].

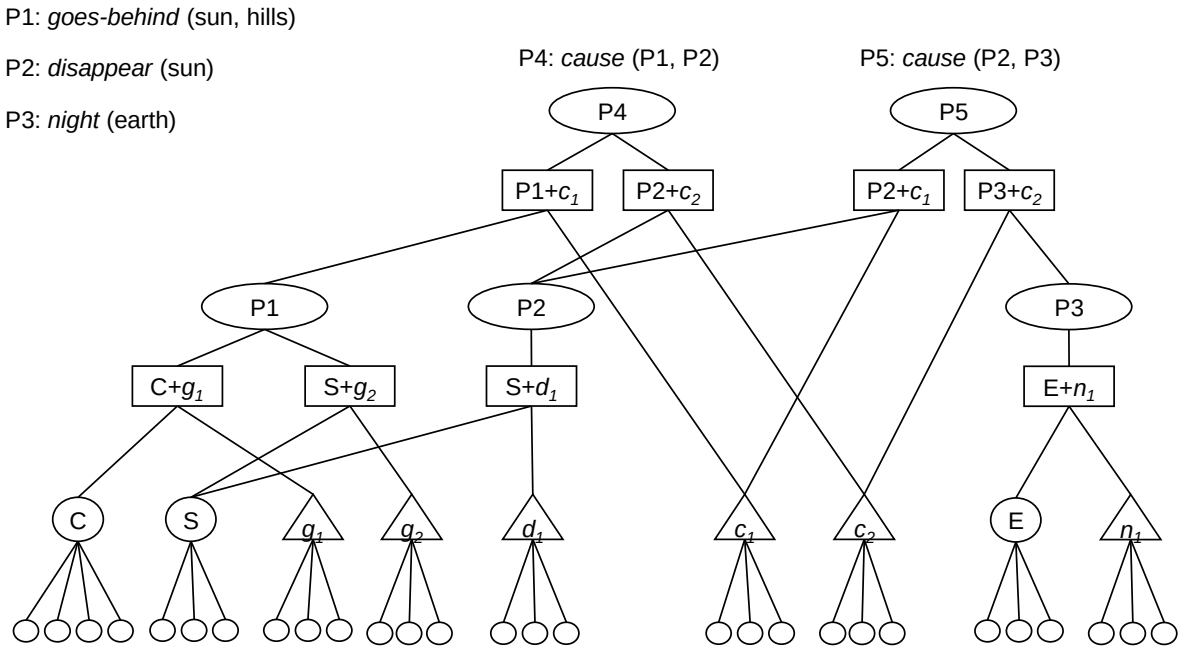


Figure 4. A causal chain composed of groups. *C* denotes a Cause state, *E* denotes an Effect state, and *L* denotes a link. Object, predicate, and semantic units have been removed for clarity. This is the same causal model as Figure 3 [*goes-behind* (sun, hills) ➔ *disappear* (sun) ➔ *night* (earth)].

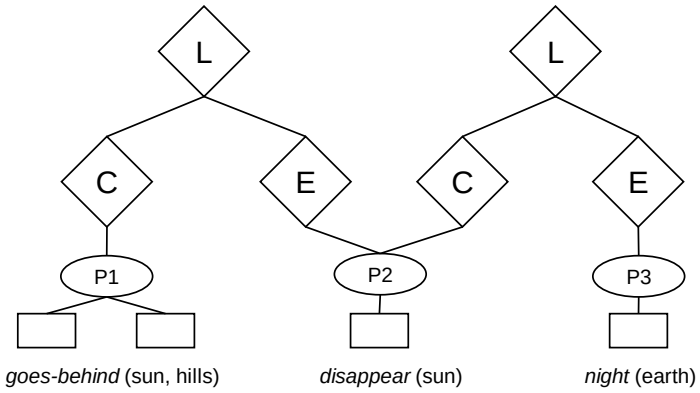


Figure 5a. Analogical inference of a causal model

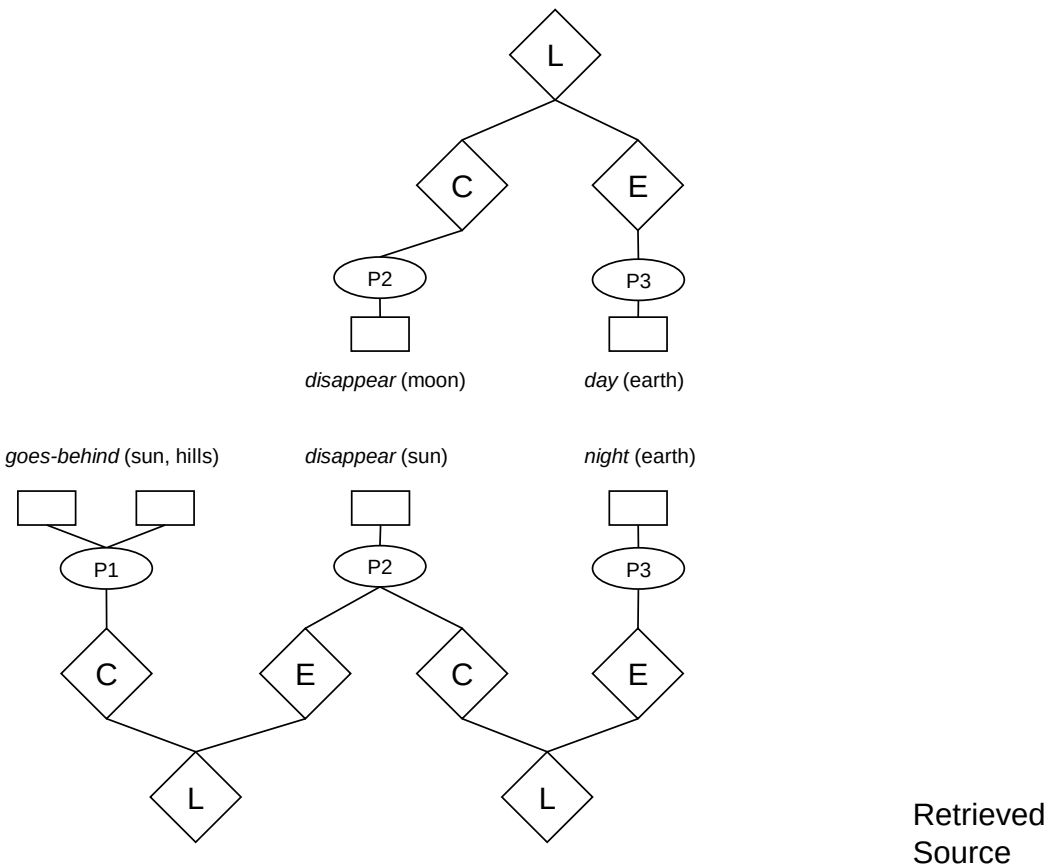
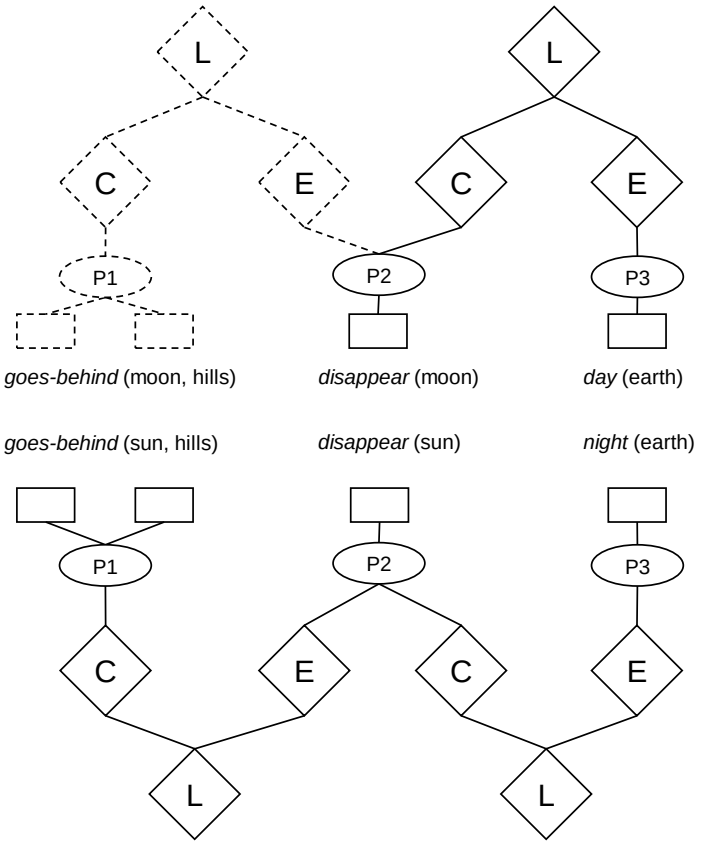


Figure 5b.

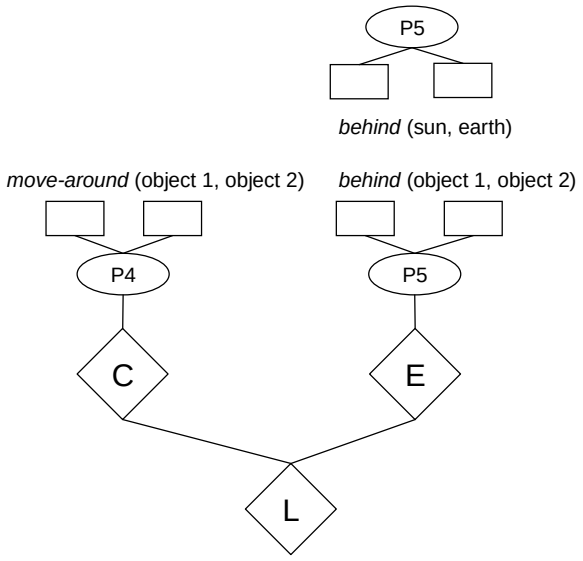
Target



Retrieved Source

Figure 6a. Elaboration of an explanation using multiple causal fragments

Target



Retrieved
Source

Figure 6b.

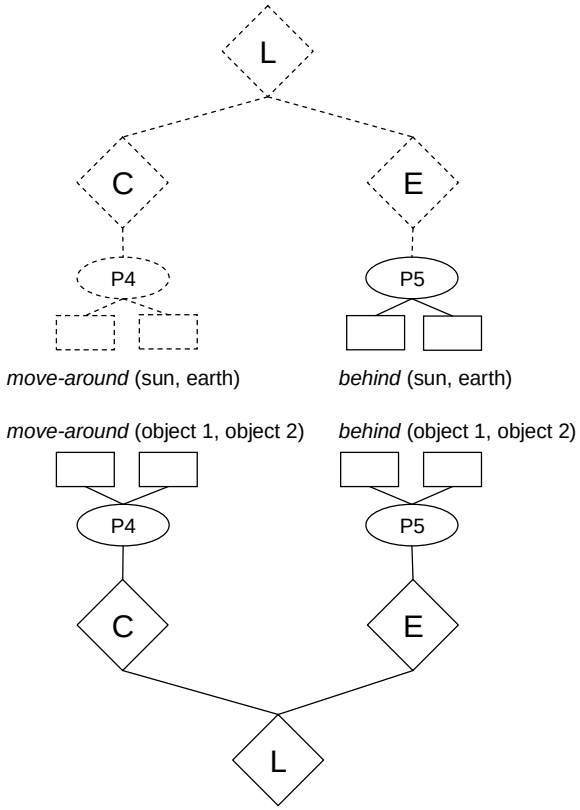

Target
L
C
E
P4
P5
move-around (sun, earth)
behind (sun, earth)
move-around (object 1, object 2)
behind (object 1, object 2)
P4
P5
C
E
L
Retrieved
Source

Figure 6c.

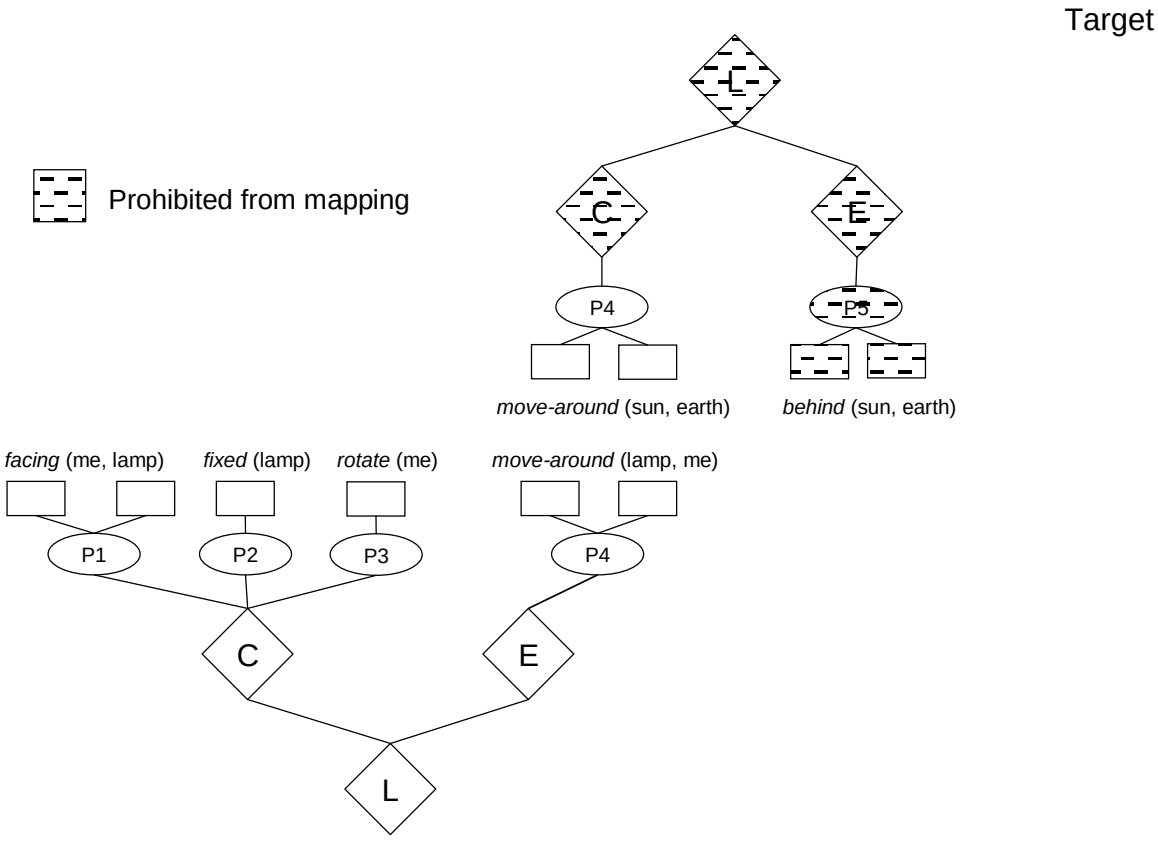
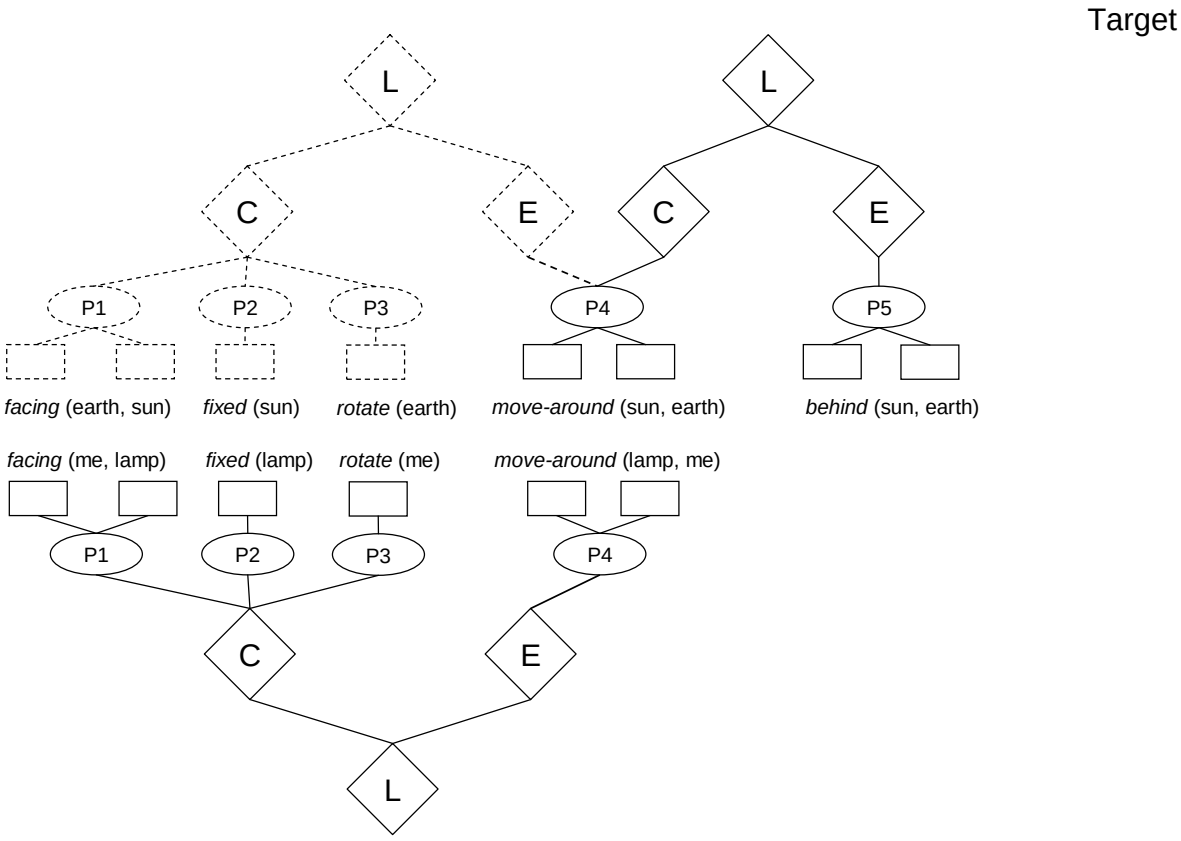

Target
Prohibited from mapping
L
C
E
P4
P5
move-around (sun, earth)
behind (sun, earth)
facing (me, lamp)
fixed (lamp)
rotate (me)
move-around (lamp, me)
P1
P2
P3
P4
C
E
L
Retrieved Source

Figure 6d.


Target
L
L
C
E
C
E
P1
P2
P3
P4
P5
facing (earth, sun)
fixed (sun)
rotate (earth)
move-around (sun, earth)
behind (sun, earth)
facing (me, lamp)
fixed (lamp)
rotate (me)
move-around (lamp, me)
P1
P2
P3
P4
C
E
L
Retrieved
Source