How do the seasons change?
Creating & revising explanations via model formulation & metareasoning

Scott E. Friedman[1], Kenneth D. Forbus[1], & Bruce Sherin[2]

[1]Qualitative Reasoning Group, Northwestern University
2133 Sheridan Road, Evanston, IL, 60208 USA

[2]Learning Sciences, Northwestern University
2120 Campus Drive, Evanston, IL, 60208 USA

{friedman, forbus, bsherin}@northwestern.edu

Abstract

Reasoning with incomplete or potentially incorrect knowledge is an important challenge for Artificial Intelligence. This paper presents a system that revises its knowledge about dynamic systems by constructing and evaluating explanations. Conceptual knowledge is represented using compositional *model fragments*, which are used to explain phenomena via *model formulation*. Metareasoning is used to (1) score the resulting explanations numerically along several dimensions and (2) evaluate preferred explanations for global consistency. Global inconsistencies cause the system to favor alternative explanations and thereby change its beliefs. We simulate the belief changes of several students during clinical interviews about how the seasons change. We show that qualitative models reasonably represent student knowledge, and that our system revises its beliefs in a fashion similar to the students.

1 Introduction

Constructing and revising explanations about physical phenomena and the systems that produce them is a familiar task for humans, but an important challenge for Artificial Intelligence. Cognitive science research has shown that learning is aided by self-directed explanation, which helps the learner repair incorrect conceptual knowledge [Chi, 2000]. This paper applies this self-explanation principle to qualitative reasoning systems to (1) model human explanation and belief revision in a conceptual reasoning domain and (2) demonstrate the flexibility that such an approach offers for autonomous learning systems.

Our system uses qualitative *model fragments* [Falkenhainer & Forbus, 1991] to represent domain knowledge. To explain a proposition (e.g. Chicago is hotter in its summer than in its winter) the system (1) performs *model formulation* to create a scenario model from a domain theory of model fragments and propositions, (2) uses temporal and qualitative reasoning over the scenario model to support the proposition, (3) numerically scores all resulting explanations, and (4) analyzes the best explanations for consistency. The system organizes its explanations and model fragments using the knowledge-based network of Friedman & Forbus [2010, 2011].

We simulate results from the cognitive science literature [Sherin *et al.*, in review] that characterize how students explain the changing of the seasons with intuitive knowledge in clinical interviews. Sherin et al. catalogs various units of intuitive knowledge that students use while explaining the changing of the seasons, including mental models and propositions regarding the earth, the sun, and quantities such as heat and temperature. According to the *knowledge-in-pieces* theory [diSessa et al., 2004], these fragmentary units of knowledge are assembled into larger explanations to make sense of other beliefs and observations. Our system is not a cognitive model of the knowledge-in-pieces view per se, but our results indicate that it can construct humanlike mental models and explanations from fragmentary knowledge.

In each simulation trial, the system begins with a subset of the fragmented intuitive knowledge described by Sherin et al., pertaining to a single student, encoded using an extended OpenCyc[1] ontology. The system explains the phenomena using this knowledge, resulting in an intuitive explanation like those described in the literature. Like the interviewees, the system is then given new information (e.g. Chicago's summer coincides with Australia's winter) which causes a potential inconsistency across preferred explanations. We compare the system's explanations and explanation revisions to those of the students in the initial study.

We begin by discussing the learning science study that characterizes student reasoning about the changing of the seasons, and then we review qualitative process theory and model formulation. We then describe our approach and present the results of our simulation. We conclude by discussing related research and future work.

1.1 How seasons (and explanations) change

Most people have commonsense knowledge about the seasons, but the scientifically-accepted explanation of how seasons change poses difficulty even for many scientifically-literate adults [Sherin *et al.*, in review]. This makes it an

[1] http://www.opencyc.org/

Friedman, S. E., Forbus, K. D. & Sherin, B. (2011). How do the seasons change? Creating & revising explanations via model formulation & metareasoning. Proceedings of the 25th International Workshop on Qualitative Reasoning. Barcelona, Spain.

interesting domain to model belief change about dynamic systems and commonsense science reasoning.

Sherin *et al.* interviewed 21 middle-school students regarding the changing of the seasons to investigate how students use commonsense science knowledge. Each interview began with the question “Why is it warmer in the summer and colder in the winter?” followed by additional questions and sketching for clarification. If the interviewee’s initial mental model of seasonal change did not account for different parts of the earth experiencing different seasons simultaneously, the interviewer asked, “Have you heard that when it’s summer [in Chicago], it is winter in Australia?” This additional information, whether familiar or not to the student, often lead them to identify an inconsistency in their explanation and reformulate an answer to the initial question. Consequently explanation revision frequently occurred during the course of the interview, where students encountered and recombined beliefs to arrive at a new explanation. The interviewer did not relate the correct scientific explanation during the course of the interview, so the students transitioned between various intuitive explanations. Sherin *et al.* includes a master listing of conceptual knowledge used by the students during the interviews, including propositional beliefs, general schemas, and fragmentary mental models.

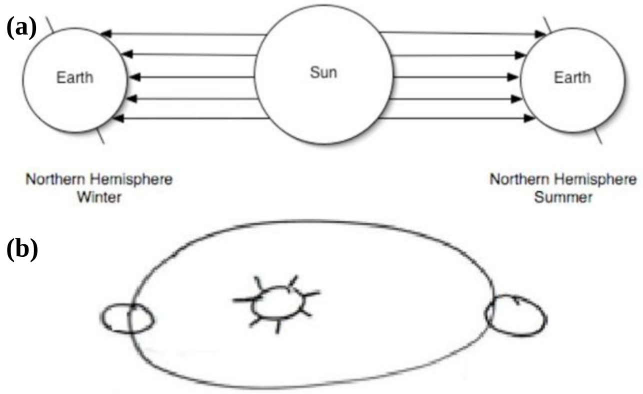

Figure 1. Two explanations of the seasons: (a) the scientific explanation, and (b) common misconception sketched by an interviewee.

The scientifically accurate explanation of the seasons (Figure 1a) is that the earth’s axis of rotation always points in the same direction throughout its orbit around the sun. When the northern hemisphere is inclined toward the sun, it receives more direct sunlight than when pointed away, which results in warmer and cooler temperature, respectively. While 12/21 students mentioned that Earth’s axis is tilted, only six of them used this fact in an explanation, and none of these were scientifically accurate. Students frequently explained that the earth is closer to the sun during the summer and farther during the winter (Figure 1b).

Our intent is to computationally model (1) how people create explanations of dynamic systems from fragmentary knowledge and (2) how explanations are revised after encountering contradictory information. Though the students in Sherin *et al.* were not given the correct (Figure 1a) explanation, we include a simulation trial that has access to the knowledge required for the correct explanation. This demonstrates that the system can construct the correct explanation when provided correct domain knowledge. We next review qualitative modeling and model formulation as it relates to simulating the reasoning involved in this study.

2 Background

Simulating humanlike reasoning about dynamic systems makes several demands on knowledge representation. First, it must be capable of representing ambiguous, incomplete, and incorrect domain knowledge. Second, it must represent processes (e.g. orbiting, rotation, heat transfer) and qualitative proportionalities (e.g. the closer something is to a heat source, the greater its temperature). Our system meets these demands by using qualitative process (QP) theory [Forbus, 1984]. Using qualitative models and QP theory to simulate humanlike mental models in physical domains is not a new idea: this was an initial motivator for qualitative physics research [Forbus & Gentner, 1997; Forbus, 1984]. We next review model fragments and model formulation, which are our system’s methods of representing and assembling conceptual knowledge, respectively.

2.1 Model Fragments & QP Theory

Model fragments [Falkenhainer & Forbus, 1991] can represent entities and processes, e.g. as the asymmetrical path of

```
ConceptualModelFragmentType RemoteHeating
Participants:
 ?heater HeatSource (providerOf)
 ?heated AstronomicalBody (consumerOf)
Constraints:
 (spatiallyDisjoint ?heater ?heated)
Conditions: nil
Consequences:
 (qprop- (Temp ?heated) (Dist ?heater ?heated))
 (qprop (Temp ?heated) (Temp ?heater))

QPProcessType Approaching-PeriodicPath
Participants:
 ?mover AstronomicalBody (objTranslating)
 ?static AstronomicalBody (to-Generic)
 ?path Path-Cyclic (alongPath)
 ?movement Translation-Periodic (translation)
 ?near-pt ProximalPoint (toLocation)
 ?far-pt DistalPoint (fromLocation)
Constraints:
 (spatiallyDisjoint ?mover ?static)
 (not (centeredOn ?path ?static))
 (objectTranslating ?movement ?mover)
 (alongPath ?movement ?path)
 (on-Physical ?far-pt ?path)
 (on-Physical ?near-pt ?path)
 (to-Generic ?far-pt ?static)
 (to-Generic ?near-pt ?static)
Conditions:
 (active ?movement)
 (betweenOnPath ?mover ?far-pt ?near-pt)
Consequences:
 (i- (Dist ?static ?mover) (Rate ?self))
```

Figure 2: *RemoteHeating* (above) and *Approaching-PeriodicPath* (below) model fragment types.

a planet's orbit, and the processes of approaching and retreating from its sun along that path (Figure 1b), respectively. For example, modeling the common misconception in Figure 1b involves several model fragments. Figure 2 shows two model fragment types used in the simulation: the conceptual model fragment `RemoteHeating`, and the process `Approaching-PeriodicPath`. Both have several components: (1) *participants* are the entities involved in the phenomenon; (2) *constraints* are relations that must hold over the participants in order to *instantiate* the model fragment as a distinct entity; (3) *conditions* are relations that must hold for the instance to be *active*; and (4) *consequences* are relations that hold when the instance is active.

QP theory's notion of influence provides causal relationships that connect quantities. Figure 2 provides examples. The relations `i+` and `i-` assert *direct influences*, i.e. constraints on the derivative of quantities. In this example, `(Dist ?static ?mover)` will be decreasing and increasing by `(Rate ?self)` while an instance of `Approaching-PeriodicPath` is active. Further, the relations `qprop` and `qprop-` assert monotonic *indirect influences*. In Figure 2, the `qprop-` relation asserts that all else being equal, decreasing `(Dist ?heater ?heated)` will result in `(Temp ?heated)` increasing.

2.2 Model Formulation

Given a domain theory described by model fragments and a relational description of a scenario, the process of *model formulation* automatically creates a model for reasoning about the scenario [Falkenhainer & Forbus, 1991]. Our approach uses a back-chaining algorithm (similar to [Rickel & Porter, 1997]) to build scenario models. The algorithm is given the following as input:

1. Scenario description ***S*** that contains relations over entities, e.g.:
   `(spatiallyDisjoint PlanetEarth TheSun)`
   `(isa PlanetEarth AstronomicalBody)`
2. A domain theory ***D*** that contains Horn clauses and model fragment types, e.g. `Approaching-PeriodicPath`.
3. A target assertion to explain, e.g.:

```
(greaterThan
  (M (Temp Chicago) ChiSummer)
  (M (Temp Chicago) ChiWinter))
```
[2]

The model formulation algorithm proceeds by recursively finding all direct and indirect influences i relevant to the target assertion, such that either (a) $\boldsymbol{S} \wedge \boldsymbol{D} \vDash i$ or (b) i is a non-ground term consequence of a model fragment within ***D*** that unifies with a quantity in the target assertion. For example, if $\boldsymbol{S} \wedge \boldsymbol{D} \vDash$ `(qprop (Temp Chicago) (Temp PlanetEarth))`, the algorithm finds influences on `(Temp PlanetEarth)`, e.g. the consequence of `RemoteHeating` `(qprop- (Temp ?heated) (Dist ?heater ?heated))`, provided `?heated` is bound to `PlanetEarth`. Model formulation then occurs via back-chaining, instantiating all model fragments provided the participant variable binding `?heated` → `PlanetEarth`. The algorithm works backwards recursively, instantiating model fragments as necessary to satisfy unbound participants of `RemoteHeating`.

The product of model formulation is the set of all potentially relevant model fragment instances. This set includes model fragments that are mutually inconsistent, e.g. an `Approaching-PeriodicPath` instance and a `Retreating-PeriodicPath` instance for `PlanetEarth`. The process of constructing explanations needs to avoid activating inconsistent combinations of model fragments, and be sensitive to any logical contradictions that arise from their consequences.

Thus far, we have described how our system represents its domain theory and assembles scenario models. Next, the system must activate these models and analyze their assumptions and consequences in contexts representing distinct qualitative states to explain how quantities (e.g. `(Temp Chicago)`) change across states (e.g. `ChiWinter` and `ChiSummer`). We discuss the rest of the explanation process next.

3 Learning by Explaining

Just as people learn from self-directed explanation [Chi, 2000], our system's knowledge-level *epistemic state* changes after explaining a fact. This section describes our system's epistemic state and approach to explanation construction, specifically: (1) explanation-based organization of conceptual knowledge; (2) metareasoning for computing a total preferential pre-order over competing explanations; and (3) inconsistency handling across explanations to preserve global coherence across preferred explanations.

3.1 Explanation-based knowledge organization

In our system, domain knowledge is organized in a knowledge-based tiered network as in Friedman & Forbus [2010, 2011]. Figure 3 shows a small portion of the network, with two explanations constructed by the system seasonal change in Australia (e_0, justifying f_{21}) and Chicago (e_1, justifying f_{22}). These encode part of the popular novice model illustrated in Figure 1b. Several beliefs and model fragments in Figure 3 are labeled for reference, e.g. to Figure 2. The network contains three tiers:

Conceptual knowledge. The bottom tier contains beliefs from the domain theory. This includes relational domain knowledge (e.g. $f_{0\text{-}2}$), model fragment types (e.g. $m_{0\text{-}4}$), and target beliefs requiring explanation (e.g. $f_{21\text{-}22}$). From a knowledge-in-pieces standpoint, these are the component pieces of knowledge.

Justification structure. The middle tier plots justifications (triangles) that connect antecedent and consequent beliefs. Justifications include (1) logical entailments, including model fragment instantiations and activations, and (2) temporal quantifiers that assert that the antecedents – and their antecedents, and so forth – hold within a given state. Model

[2] The `M` operator from QP theory denotes the measurement of a quantity at a state (e.g. `(Temp Chicago)`) within a given state (e.g. `ChiSummer`).

Legend

f_0	(isa earthPath EllipticalPath)	f_9	(active RH-inst)
f_1	(spatiallyDisjoint earthPath TheSun)	f_{10}	(qprop- (Temp PlanetEarth) (Dist TheSun PlanetEarth))
f_2	(isa TheSun AstronomicalBody)	f_{11}	(qprop (Temp PlanetEarth) (Temp TheSun))
m_0	*ProximalPoint*	f_{12}	(i+ (Dist TheSun PlanetEarth) (Rate RPP-inst))
m_1	*DistalPoint*	f_{13}	(increasing (Temp PlanetEarth))
m_2	*Approaching-PeriodicPath*	f_{14}	(decreasing (Temp PlanetEarth))
m_3	*RemoteHeating*	f_{15}	(qprop (Temp Australia) (Temp PlanetEarth))
m_4	*Retreating-PeriodicPath*	f_{16}	(qprop (Temp Chicago) (Temp PlanetEarth))
f_3	(isa TheSun HeatSource)	f_{17}	(increasing (Temp Chicago))
f_4	(spatiallyDisjoint TheSun PlanetEarth)	f_{18}	(decreasing (Temp Chicago))
f_5	(isa APP-inst Approaching-PeriodicPath)	f_{19}	(holdsIn (Interval ChiWinter ChiSummer) (increasing (Temp Chicago)))
f_6	(isa RH-inst RemoteHeating)	f_{20}	(holdsIn (Interval ChiSummer ChiWinter) (decreasing (Temp Chicago)))
f_7	(isa RPP-inst Retreating-PeriodicPath)	f_{21}	(greaterThan (M (Temp Australia) AusSummer) (M (Temp Australia) AusWinter))
f_8	(i- (Dist TheSun PlanetEarth) (Rate APP-inst))	f_{22}	(greaterThan (M (Temp Chicago) ChiSummer) (M (Temp Chicago) ChiWinter))

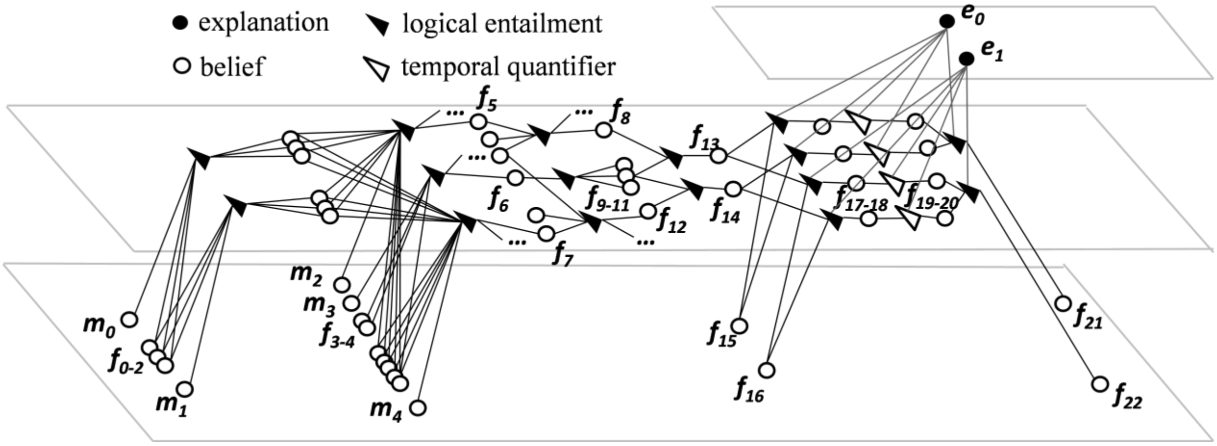

Figure 3: A knowledge-based network of explanations (top tier), justification structure (middle tier), and domain theory (bottom tier). Explanations e_0 and e_1 justify seasonal change in Australia (e_0) and Chicago (e_1).

formulation, as described in the previous section, provides the majority of the justification structure in Figure 3. Additional justifications and intermediate beliefs are computed after model formulation (e.g. temporal quantifiers, increasing and decreasing assertions, qprop assertions entailed by the domain theory) to connect the target beliefs ($f_{21,22}$ in Figure 3) to upstream justification structure.

Explanations. The top tier plots explanations (e.g. e_1). Each explanation contains a unique set of justifications that provide *well-founded support* for the target belief (e.g. f_{22}), such that the justification structure is free of cycles and redundancy. Note that both e_0 and e_1 in Figure 3 contain all justifications left of f_{8-12}, but the edges are omitted for clarity. Each explanation node also refers to a logical context where all of the antecedents and consequences of its component justifications are believed. Consistency within each explanation is enforced during explanation construction, whereas consistency *across* certain explanations is tested and enforced via different methods, discussed below. In sum, each explanation is an aggregate of well-founded justification structure that clusters the underlying domain knowledge into a productive and consistent subset. The system's granularity of consistency is at the explanation-level rather than the KB-level.

3.2 Competing explanations

The two explanations in Figure 3 use a scenario model similar to Figure 1b to justify the seasons changing in both Australia and Chicago. However, there frequently exist multiple, *competing* well-founded explanations for a target belief. For example, provided the RemoteHeating instance RH-inst (asserted via f_6, Figure 3) and its f_{11} consequence (qprop (Temp PlanetEarth) (Temp TheSun)), the system also generates additional justification structure for the changing of Chicago's and Australia's seasons: (Temp TheSun) increases between each region's winter and summer and decreases likewise. This additional justification structure (not depicted in Figure 3) results in three additional well-founded explanations (nodes) in the system for Chicago's seasons, and three analogous explanations for Australia's seasons, for a total of four explanations each:

- ***e_1***: The earth retreats from the sun for Chicago's winter and approaches for its summer (shown in Figure 3).
- ***e'_1***: The sun's temperature decreases for Chicago's winter and increases for its summer.
- ***e'_2***: The sun's temperature decreases for Chicago's winter, and the earth approaches the sun for its summer.
- ***e'_3***: The earth retreats from the sun for Chicago's winter, and the sun's temperature increases for its summer.

Explanations e_1 and $e'_{1\text{-}3}$ compete with each other to explain f_{22}. However, $e'_{1\text{-}3}$ are all problematic. Explanations e'_2 and e'_3 contain nonreciprocal quantity changes in a cyclic state space: a quantity (e.g. the sun's temperature) changes in the summer → winter interval without returning to its prior value somewhere in the remainder of the state cycle, summer → winter. Explanation e'_1 is not structurally or temporally problematic, but the domain theory contains no model fragments that can describe the sun changing its temperature. Consequently, the changes in the sun's temperature are assumed rather than justified by process instances, and this is problematic under the sole mechanism assumption[3] (Forbus, 1984). We have just analyzed and discredited system-generated explanations $e'_{1\text{-}3}$ which compete with explanation e_1 in Figure 3. The system performs metareasoning over its explanations to make these judgments automatically, which we discuss next.

3.3 Metareasoning & epistemic preferences

The tiered network and justification structure described above are stored declaratively within the KB as relational facts between beliefs and nodes. Consequently, the system can inspect and evaluate its own explanations to construct a total pre-order over competing explanations.

A total pre-order is computed by computing a numerical score **S**(e_i) of each explanation e_i, and sorting by score. The score is computed via the following equation:

$$S(e) = -\sum_{p \in P} cost(p) * |occurrences(p, e)|$$

Each explanation's score starts at zero and incurs a negative penalty for each occurrence of an artifact p_i in the explanation. Penalties are weighted according to the cost $cost(p_i)$ of the type of artifact, where costs are predetermined[4]. The artifacts computed and penalized by the system include:

- **Logical contradictions** (cost: 100) occur within an explanation when its beliefs entail a contradiction.
- **Asymmetric quantity changes** (cost: 40) are quantity changes that do not have a reciprocal quantity change in a cyclical state-space (e.g. in $e'_{2\text{-}3}$).
- **Assumed quantity changes** (cost: 30) are quantity change beliefs that have no direct or indirect influence justification.
- **Model fragment types** (cost: 4) are penalized to reward qualitative parsimony.
- **Assumptions** (cost: 3) are beliefs without justifications, that must hold for the explanation to hold.
- **Model fragment instances** (cost: 2) are penalized to reward quantitative parsimony.
- **Justifications** (cost: 1) are penalized to avoid unnecessary entailment.

Minimizing model fragment types and instances is a computational formulation of Occam's Razor. The resulting total pre-order reflects the system's preference across competing explanations, and the maximally-preferred explanation for the target belief b_t is marked *best-xp*(b_t). However, this ordering was computed by analyzing each explanation in isolation. It therefore does not account for inconsistency across explanations, which we discuss next.

3.4 Inconsistency across explanations

Ensuring consistency across explanations entails evaluating the union of their component beliefs. The system does not maintain consistency across all of its explanations – for instance, there is no need for consistency between two competing explanations (e.g. e_1 and e'_1 above) because only one can be asserted *best-xp*(f_{22}). Consequently, the system only checks for consistency across its best explanations for different target beliefs (e.g. e_0 and e_1 in Figure 3).

Inconsistencies are identified using logic and temporal reasoning. As mentioned above, each explanation is represented by a node in the network as well as its own logical context in which all of its constituent beliefs are asserted. We use notation B(e_i) to denote the set of beliefs asserted in the logical context of explanation e_i.

Consider the information Sherin *et al.* gives the students in the interview, "…when it is summer [in Chicago] it is winter in Australia." We can refer to this information as:

ρ = `(cotemporal ChiSummer AusWinter).`

Before ρ is known, explanations e_0 and e_1 in Figure 3 are consistent:

$$B(e_0) \wedge B(e_1) \nvDash \bot.$$

After ρ is known, e_0 and e_1 are inconsistent:

$$B(e_0) \wedge B(e_1) \wedge \rho \vDash \bot.$$

The new knowledge ρ causes several inconsistencies between explanations, because:

B(e_0) ⊨ `(holdsIn`
```
        (Interval AusSummer AusWinter)
        (decreasing (Temp PlanetEarth)))
```
B(e_1) ⊨ `(holdsIn`
```
        (Interval ChiWinter ChiSummer)
        (increasing (Temp PlanetEarth)))
```

The new information ρ creates a temporal intersection in which the two contradictory assertions `(increasing (Temp PlanetEarth)` and `(decreasing (Temp PlanetEarth)` are believed. Consequently, e_0 and e_1 are

[3] The agent might explicitly assume that an unknown, active, process is directly influencing the quantity, but such an assumption is still objectively undesirable within an explanation.

[4] The numerical penalties listed above are the system's default values, which were determined empirically and are used in the simulation presented here; however, they are stored declaratively, and are therefore potentially learnable.

inconsistent provided ρ, despite each being the preferred explanation for the seasons in Australia and Chicago, respectively. Inconsistent explanations cannot be simultaneously preferred by the system, so the inconsistency is recorded as metaknowledge and either or both of e_0, e_1 must be removed as *best-xp*(b_t) for its target belief b_t.

4 Simulation

We implemented our system on the Companions cognitive architecture [Forbus *et al.*, 2009] and ran a series of trials to compare our system's explanations to those of students. In each trial, the system starts with a subset of knowledge pertaining to a student from Sherin *et al.*, but no explanations have been constructed. In terms of Figure 3, the starting state of the system is a series of nodes on the bottom (domain theory) tier of the network, but none elsewhere.

The individual differences of the students within the interviews involve more than just variations in domain knowledge. For example, some students strongly associate certain models and beliefs with the seasons (e.g. that Earth's axis is tilted) without knowing the exact mechanism. To capture this (e.g. in the "Angela" trial below), our system includes an additional numerical penalty over beliefs to bias explanation preference. We describe this further below.

After providing the system with fragmentary domain knowledge and numerical preferences, in each trial the simulation does the following:

1. Constructs explanations of the seasons changing in Chicago and Australia.
2. Diagrams preferred explanations using a quantity influence graph.
3. Incorporates the temporal facts relating Chicago's and Australia's seasons.
4. Reconstructs and diagrams the preferred explanations.

Before describing each trial, we review the explanations used by the system during simulation, illustrated as influence graphs in Figure 4. Graphs (a-c) reflect common student explanations found by Sherin *et al.*, and graph (d) is the scientific explanation in Figure 1a. Graph (a) explains that as the earth rotates, Chicago and Australia increase and decrease their distance from the proximal spot on the earth to the sun. This mediates their sunlight, and therefore, their temperature. This is an approximation of a popular student explanation, which states that regions that face the sun are

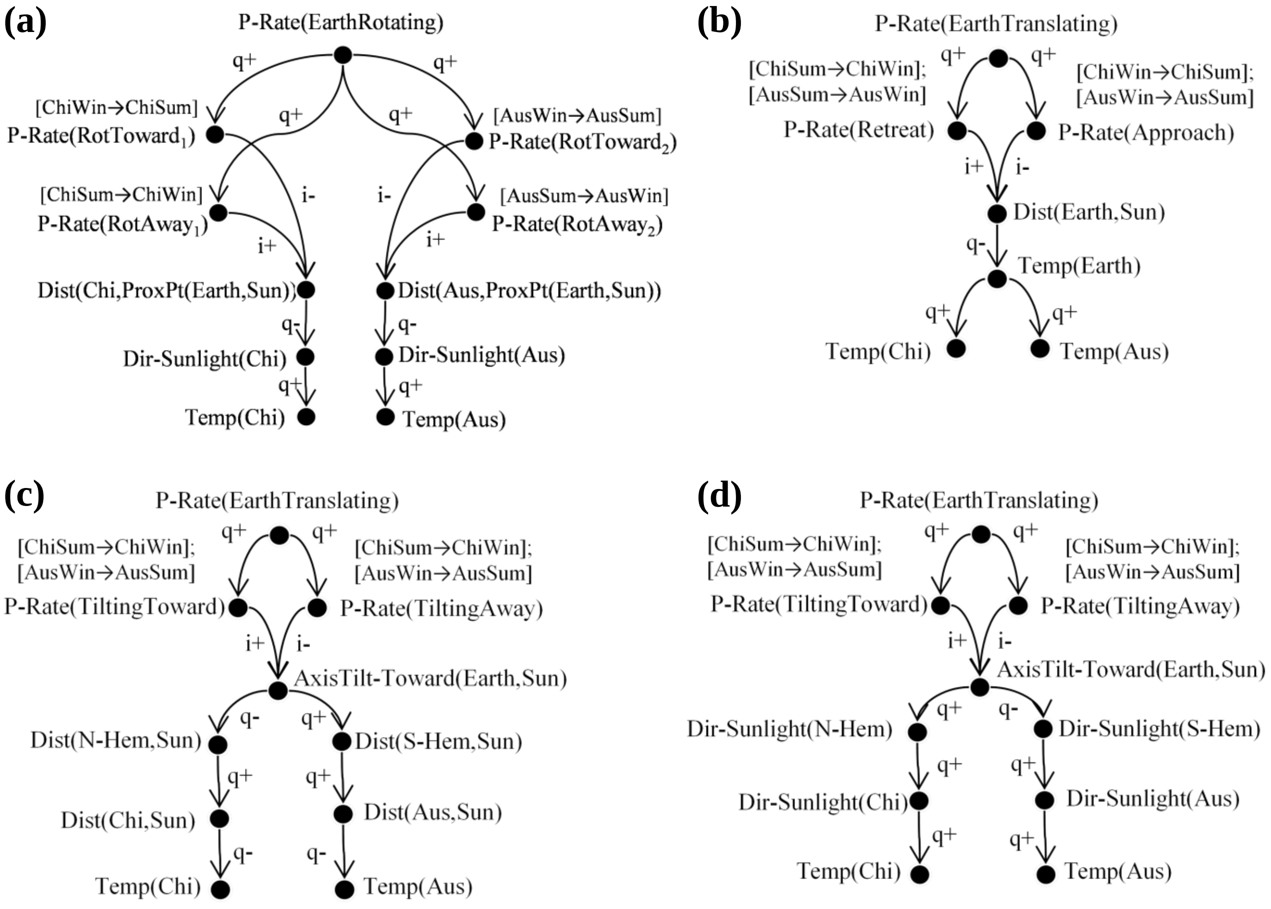

Figure 4: Quantity influence graphs from explanations produced by the simulation. Edges describe qualtative (q+, q-) proportionalities and direct influences (i+, i-).

warmer than regions that do not. Graph (b) is the explanation sketched in Figure 1b, and plotted in Figure 3, and is the only one inconsistent with opposite seasons in Chicago and Australia. Graph (c) explains that as the earth translates, its tilt toward the sun increases and decreases. This mediates the distance to the sun from the earth's northern and southern hemispheres, which in turn affects their temperature and the regions within. Graph (d), modeled after the scientific explanation in Figure 1b, is analogous to (c), but references direct sunlight instead of distance to the sun. We describe three separate simulation trials, which model five students.

Ali & Kurt trial. The system's initial domain knowledgeincludes: (1) the earth rotates on a tilted axis, (2) temperature is qualitatively proportional to sunlight, and (3) the earth orbits the sun. However, there is no knowledge that each hemisphere is tilted toward and away during the orbit. Consequently, the system computes nine explanations, and computes a preference for the explanation shown in graph (a), with a score of -56. This explanation is consistent with the opposite seasons fact, so no revision occurs as a result.

Deidra & Angela trial. The system's initial domain knowledge includes: (1) the earth rotates; (2) the earth orbits the sun and is sometimes closer and sometimes farther; and (3) sunlight and proximity to the sun both affect temperature. The system creates 36 explanations[5], and computes a preference for the explanation in graph (b), with a score of -56. The system also created the explanation for graph (a) with a score of -66, due to an additional ten-point penalty on the belief `(qprop (Temp X) (Sunlight X))`. When confronted with the opposite seasons fact, the system (like Deidra and Angela) changes its preferred explanation to that in graph (a).

Amanda trial. The system's initial domain knowledge includes: (1) the earth orbits the sun; (2) the earth rotates on a tilted axis; (3) when each hemisphere is tilted toward the sun, it receives more sunlight and is more proximal to the sun, and (4) sunlight and proximity to the sun both affect temperature. In the interview, Amanda mentions two main influences on Chicago's temperature: (1) the distance to the sun due to the tilt of the earth, and (2) the amount of sunlight, also due to the tilt of the earth. Through the course of the interview, she settles on the latter. Amanda could not identify the mechanism by which the tilt changes throughout the year. We simulated Amanda twice: first with process models for `TiltingToward`, and `TiltingAway`, producing graphs (c) and (d) with scores -52 and -67, respectively, and second without these process models, which produced two similar graphs, but without anything affecting `AxisTiltToward(Earth,Sun)`.

By varying the domain knowledge and adding numerical biases in metaknowledge, the system was able to (1) construct several student explanations from the literature and (2) alter its preferred explanation similar to the way students did when confronted with an inconsistency. Further, in the Amanda trial, we provided additional process models to demonstrate that it could construct the correct explanation.

Our computational model provides a plausible account of how people might organize, represent, and combine domain knowledge into explanations. However, we believe that the simulation is doing much more computation than people to construct the same explanations – e.g. the system computed and evaluated 36 explanations in the Diedra & Angela trial. As described above, our system uses a back-chaining model formulation algorithm, followed by a complete meta-level analysis. At the algorithmic level, people probably use a more incremental approach to explanation construction, where they interleave meta-level analysis within their model-building operations. Such an approach would avoid reifying explanations that are known to be problematic (e.g. explanations $e'_{1\text{-}3}$ in section 3.2), but it would involve sophisticated monitoring of the model formulation process.

5 Related Work

ECHO [Thagard, 2000] is a connectionist model that uses constraint-satisfaction to judge hypotheses by their explanatory coherence. ECHO creates excitatory and inhibitory links between consistent and inconsistent propositions, respectively. Its "winner take all" network means that it cannot distinguish between no evidence for competing propositions versus balanced conflicting evidence for them. ECHO requires a full explanatory structure as its input. By contrast, our system generates its justification structure from fragmentary domain knowledge, and then evaluates it along several dimensions via metareasoning.

Creating and revising explanations is part of the larger cognitive process of conceptual change. INTHELEX [Esposito et al., 2000] is an incremental theory revision program that has modeled conceptual change as supervised learning. INTHELEX uses Datalog clauses as its knowledge representation, which might not suffice for explaining the behavior of dynamic systems, such as the simulation presented here. Furthermore, INTHELEX implements belief revision as theory refinement, so it revises its logical theories when it encounters an inconsistency, instead of reformulating explanations using existing knowledge.

Learning by creating explanations is an established method in Artificial Intelligence. Many systems that perform Explanation-Based Learning (EBL) [DeJong, 1993] create new knowledge by *chunking* explanation structure into a single rule [Laird *et al.*, 1987]. Chunking speeds up future reasoning by avoiding extra instantiations when a macro-level rule exists, but it does not change the deductive closure of the knowledge base, and therefore cannot model the repair of incorrect knowledge.

Previous research in AI has produced postulates for belief revision in response to observations or a sequence thereof. The AGM postulates [Alchourròn *et al.*, 1985] describe properties of rational revision operations for expansion, revision, and contraction of propositional beliefs within a deductively-closed knowledge base. Katsuno and Mendel-

[5] The increased number of explanations is due to the belief that proximity in addition to amount of sunlight affect temperature.

zon's [1991] theorem equates these postulates to a revision mechanism based on total pre-orders over prospective KB interpretations. Our system computes a total pre-order over competing explanations rather than over propositional belief sets. Consequently, the granularity of consistency of our approach differs from these accounts of belief revision: it does not ensure a consistent, deductively-closed KB, but it does ensure consistency across *best-xp* explanations. This permits a bounded consistency which enables us to model humanlike reasoning: competing explanations may be entertained, and choosing an explanation put pressures the system to ensure consistency with other *best-xp* explanations.

6 Discussion

We have simulated how people construct explanations from fragmentary knowledge and revise them, provided new information. By changing the initial knowledge of the system, we are able to simulate different interviewees' commonsense scientific reasoning regarding the changing of the seasons. Further, we demonstrated that the system can construct the scientifically-correct explanation using the same knowledge representation and reasoning approaches.

The numerical explanation scoring strategy used by our system is domain-general, albeit incomplete. The strategy presented here accounts for logical and causal patterns in an explanation (e.g. inconsistencies, assumptions, unjustified quantity changes) which constrain explanations to use a QP theory knowledge representation. This is not a serious constraint, as QP theory itself is domain-general. Regarding incompleteness, there are other patterns and artifacts of explanations that are not applicable in the seasons domain, but do apply in others: belief probability, epistemic entrenchment, level of specificity, credibility of knowledge (and knowledge sources), and diversity of knowledge. We intend to expand our system to account for these dimensions as we expand into other domains.

While our methods were sufficient to simulate several interviewees from Sherin *et al.*, we plan to increase our coverage by encoding more model fragments. We also intend to demonstrate the generality of our approach by applying it in other tasks, including learning via reading, instruction, and human interaction.

Acknowledgments

This work was funded by the Northwestern University Cognitive Science Advanced Graduate Fellowship and the Office of Naval Research.

References

[Alchourròn *et al.*, 1985] C. E. Alchourròn, P. Gärdenfors, and D. Makinson (1985). On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic, 50*: 510–530.

[Chi, 2000] Chi, M. (2000). Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. In R. Glaser (Ed.), *Advances in Instructional Psychology*, Hillsdale, NJ: Lawrence Erlbaum Associates. 161-238.

[DeJong, 1993] DeJong, G. (*Ed.*) (1993). *Investigating Explanation-Based Learning*. Kluwer Academic Publishers, Norwell, MA, USA.

[diSessa *et al.*, 2004] diSessa, A. A., Gillespie, N. M., & Esterly, J. B. (2004). Coherence versus fragmentation in the development of the concept of force. *Cognitive Science, 28*(6): 843-900.

[Esposito *et al.*, 2000] Esposito, F., Semeraro, G., Fanizzi, N., & Ferilli., S. (2000). Conceptual Change in Learning Naive Physics: The Computational Model as a Theory Revision Process. In E. Lamma and P. Mello (Eds.), *AI*IA99: Advances in Artificial Intelligence, Lecture Notes in Artificial Intelligence 1792*, 214-225.

[Falkenhainer & Forbus, 1991] Falkenhainer, B. & Forbus, K. (1991). Compositional modeling: Finding the right model for the job. *Artificial Intelligence*, 51: 95-143.

[Forbus, 1984] Forbus, K. (1984). Qualitative process theory. *Artificial Intelligence*, 24: 85-168.

[Forbus & Gentner, 1997] Forbus, K. & Gentner, D. (1997). Qualitative mental models: Simulations or memories? *Proceedings of the Eleventh International Workshop on Qualitative Reasoning.*

[Forbus *et al.*, 2009] Forbus, K., Klenk, M., & Hinrichs, T. (2009). Companion cognitive systems: Design goals and lessons learned so far. *IEEE Intelligent Systems*, 24(4): 36-46.

[Friedman & Forbus, 2010] Friedman, S., & Forbus, K. (2010). An integrated systems approach to explanation-based conceptual change. *Proceedings of the 25th Annual AAAI Conference on Artificial Intelligence*.

[Friedman & Forbus, 2011] Friedman, S. & Forbus, K. (2011). Repairing Incorrect Knowledge with Model Formulation and Metareasoning. *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*.

[Katsuno & Mendelzon, 1991] Katsuno, H., & Mendelzon, A. (1991). Propositional knowledge base revision and minimal change. *Artificial Intelligence, 52*: 263-294.

[Laird *et al.*¸1987] Laird, J., Newell, A., & Rosenbloom, P. (1987). SOAR: An architecture for general intelligence. *Artificial Intelligence*, 33(1): 1-64.

[Sherin *et al.*, in review] Sherin,B., Krakowski, M., Lee, V. R. (in review). Some assembly required: how scientific explanations are constructed during clinical interviews.

[Thagard, 2000] Thagard, P. (2000). Probabilistic Networks and Explanatory Coherence. *Cognitive Science Quarterly, 1*: 93-116.