Constructing and Revising Commonsense Science Explanations: A Metareasoning Approach

Scott E. Friedman,[1] Kenneth D. Forbus,[1] and 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 commonsense science knowledge is an important challenge for Artificial Intelligence. This paper presents a system that revises its knowledge in a commonsense science domain by constructing and evaluating explanations. Domain knowledge is represented using qualitative *model fragments*, which are used to explain phenomena via *model formulation*. Metareasoning is used to (1) score competing explanations numerically along several dimensions and (2) evaluate preferred explanations for global consistency. 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 accurately represent student knowledge and that our system produces and revises a sequence of explanations similar those of the students.

Introduction

Constructing and revising explanations about physical phenomena and the systems that produce them is a familiar task for humans, but poses several challenges for cognitive systems. A subset of these challenges includes:

1. Representing knowledge about physical phenomena and dynamic systems
2. Organizing this knowledge such that gaps, misconceptions, and inconsistencies can exist, yet explanations are still coherent
3. Flexibly revising knowledge and explanations given new information

This paper presents an approach to addressing these challenges. We integrate several techniques: qualitative *model fragments* (Falkenhainer & Forbus, 1991) for domain knowledge representation, an explanation-based knowledge organization that allows multiple inconsistent explanations, and metareasoning for computing preferences over explanations and performing belief revision. We describe some promising results with a simulation that models students' explanations and belief revisions during a clinical interview about the changing of the seasons (Sherin et al., in review).

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 its domain knowledge, (2) uses temporal and qualitative reasoning over the scenario model to support the proposition, (3) numerically scores all resulting explanations using a cost function, and (4) analyzes preferred explanations for consistency. In our approach, metareasoning does not directly monitor the domain-level reasoning; rather, it inspects the explanations produced by domain-level reasoning and controls future domain reasoning by encoding preferences. The system organizes its explanations and model fragments using the explanation-based network of Friedman & Forbus (2010, 2011). By constructing a new explanation and encoding a preference for it over a previously-preferred explanation, the system effectively revises its set of preferred beliefs. This is a cognitive model of the psychological self-explanation effect (Chi, 2000), whereby people repair incorrect domain knowledge by constructing explanations.

We evaluate our simulation based on its accuracy and coverage of the students interviewed by Sherin et al. The experimenters cataloged the intuitive knowledge that each students used while explaining the changing of the seasons, including mental models and propositions regarding the earth, the sun, heat, and light. In each simulation trial, the system begins with a domain theory corresponding to a single student in Sherin et al., encoded using an extended OpenCyc[2] ontology. The system explains the changing of the seasons using this knowledge, resulting in an intuitive explanation like those described in Sherin et al. Like the student, the system is then given new information (e.g., Chicago's summer coincides with Australia's winter) which – in some trials – causes an inconsistency across preferred explanations and forces a revision. We compare the system's explanations and explanation revisions to those of the students in the initial study.

We begin by discussing Sherin et al.'s study, and then we review qualitative process theory and model formulation.

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

We then describe our approach and present simulation results. We close by discussing related work and future work.

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 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 their model. 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.

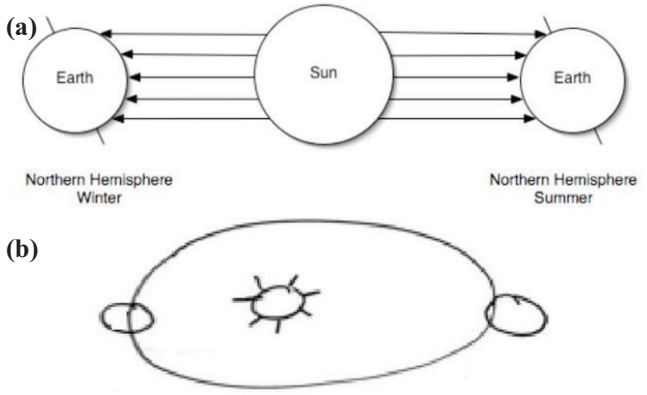

Figure 1. Two explanations of the seasons: (a) the scientific explanation, and (b) a common misconception sketched by an interviewee.

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 simulation models (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 the qualitative modeling techniques used in this study.

Background

Simulating human reasoning about dynamic systems makes several demands on knowledge representation. First, it must be capable of representing 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, 1984; Forbus & Gentner, 1997). We next review model fragments and model formulation, which are our system's methods of representing and assembling conceptual knowledge, respectively.

Compositional Modeling & QP Theory

Compositional modeling (Falkenhainer & Forbus, 1991) uses *model fragments* to represent entities and processes, e.g., as the asymmetrical path of 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*, which constrain the derivatives 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.

```
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.

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 facts, such as
   `(spatiallyDisjoint PlanetEarth TheSun)`
   `(isa PlanetEarth AstronomicalBody)`
2. A domain theory ***D*** that contains Horn clauses and model fragment types, such as `Approaching-PeriodicPath`.
3. A target assertion to explain, such as
   `(greaterThan`
   `(M (Temp Chicago) ChiSummer)`
   `(M (Temp Chicago) ChiWinter))`[3]

The model formulation algorithm proceeds by recursively finding all direct and indirect influences *i* relevant to the target assertion, such that either (a) $\boldsymbol{S} \wedge \boldsymbol{D} \models i$ (i.e., ***S*** and ***D*** entail the influence) 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 $\boldsymbol{S} \wedge \boldsymbol{D} \models$ `(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 that include the participant binding `?heated` to `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 later stages of explanation construction must avoid activating inconsistent combinations of model fragments created here.

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 this explanation process next.

Learning by Explaining

Just as people learn from self-explanation (Chi, 2000), our system's explanation-based network changes after explaining a fact. Here we describe our approach to explanation construction, specifically: (1) explanation-based organization of domain knowledge; (2) metareasoning for computing a total preferential pre-order (i.e., some explanations may be equally preferred) over competing explanations; and (3) inconsistency handling to preserve global coherence across preferred explanations.

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 of the system's explanations for seasonal temperature differences 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, using model fragments from Figure 2. The network contains three tiers:

Domain knowledge. The bottom tier contains beliefs from the domain theory. This includes propositional domain beliefs (e.g., $f_{0\text{-}2}$), model fragment types (e.g., $m_{0\text{-}4}$), and target beliefs requiring explanation (e.g., $f_{21\text{-}22}$).

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 formulation, as described in the previous section, provides the majority of the justification structure in Figure 3. Addi-

[3] 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))

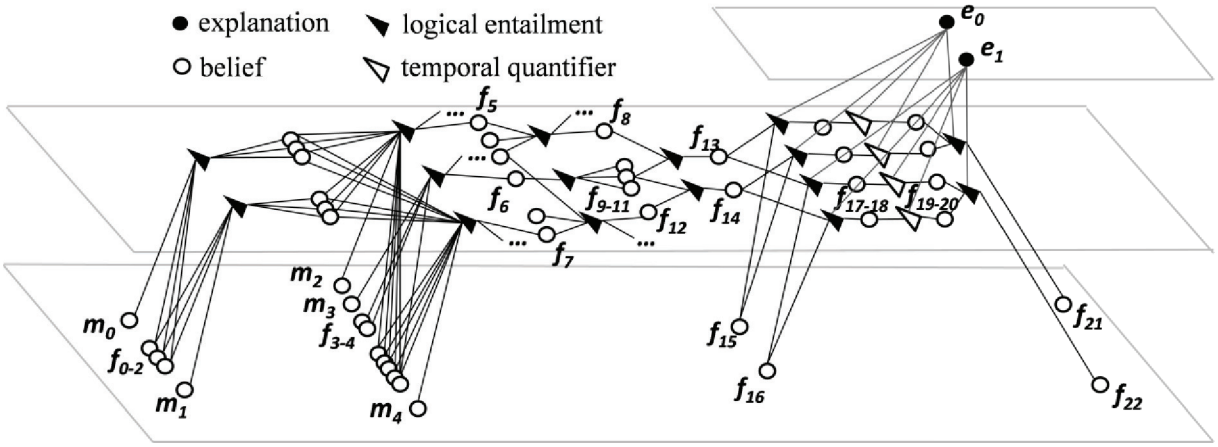

Figure 3: A knowledge-based network of explanations (top tier), justification structure (middle tier), and domain theory (bottom tier). Explanations e_0 and e_1 justify seasonal change in Australia (e_0) and Chicago (e_1). Only key beliefs are labeled.

tional 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. Belief nodes at this tier are conditions and consequences of model fragment instances that are not believed independently.

Explanations. The top tier plots explanations (e.g., e_1). Each explanation can be uniquely defined as $\langle \boldsymbol{J}, \boldsymbol{B}, \boldsymbol{T} \rangle$, where $\boldsymbol{J}$ is a unique set of justifications with beliefs $\boldsymbol{B}$ that provide *well-founded support* for the target belief(s) $\boldsymbol{T}$ (e.g., $\{f_{22}\}$), such that $\boldsymbol{J}$ is free of cycles and redundancy. Two explanations could have the same beliefs $\boldsymbol{B}$ and target belief(s) $\boldsymbol{T}$, but differ in their justifications $\boldsymbol{J}$. Note that both e_0 and e_1 in Figure 3 contain all justifications left of $f_{8\text{-}12}$: edges are omitted for clarity. Each explanation node also refers to a logical context where the set $\boldsymbol{B}$ of all of the antecedent and consequent beliefs of $\boldsymbol{J}$ are believed. Consistency within each explanation's beliefs $\boldsymbol{B}$ is enforced during explanation construction, whereas consistency *across* certain explanations (e.g., $\boldsymbol{B_0} \cup \boldsymbol{B_1}$) is tested and enforced via different methods, discussed below. In sum, each explanation is an aggregate of well-founded justification structure $\boldsymbol{J}$ that clusters the underlying domain knowledge $\boldsymbol{B}$ into a coherent subset. The system's granularity of consistency is at the explanation-level rather than the KB-level.

4.2 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:

- $\boldsymbol{e_1}$: The earth retreats from the sun for Chicago's winter and approaches for its summer (shown in Figure 3).
- $\boldsymbol{e'_1}$: 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-to-winter interval without returning to its prior value somewhere in the remainder of the state cycle, summer-to-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 (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.

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 cost $\mathbf{C}(e_i)$ of each explanation e_i, and sorting by cost. The cost is computed via the following equation:

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

Each explanation's cost starts at zero and incurs a cost 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 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 $best\text{-}xp(b_t)$. However, this ordering was computed by analyzing each explanation in isolation. It therefore does not account for inconsistency across explanations, which we discuss next.

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 $best\text{-}xp(f_{22})$. 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. As above, we use notation $\boldsymbol{B_i}$ 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:

$\boldsymbol{B_0} \wedge \boldsymbol{B_1} \nvDash \bot$.

After ρ is known, e_0 and e_1 are inconsistent:

$\boldsymbol{B_0} \wedge \boldsymbol{B_1} \wedge \rho \vDash \bot$.

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

$\boldsymbol{B_0} \vDash$
```
(holdsIn
    (Interval AusSummer AusWinter)
    (decreasing (Temp PlanetEarth)))
```
$\boldsymbol{B_1} \vDash$
```
(holdsIn
    (Interval ChiWinter ChiSummer)
    (increasing (Temp PlanetEarth)))
```

The new information ρ creates a temporal intersection in which the two contradictory assertions `(increasing`

[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.

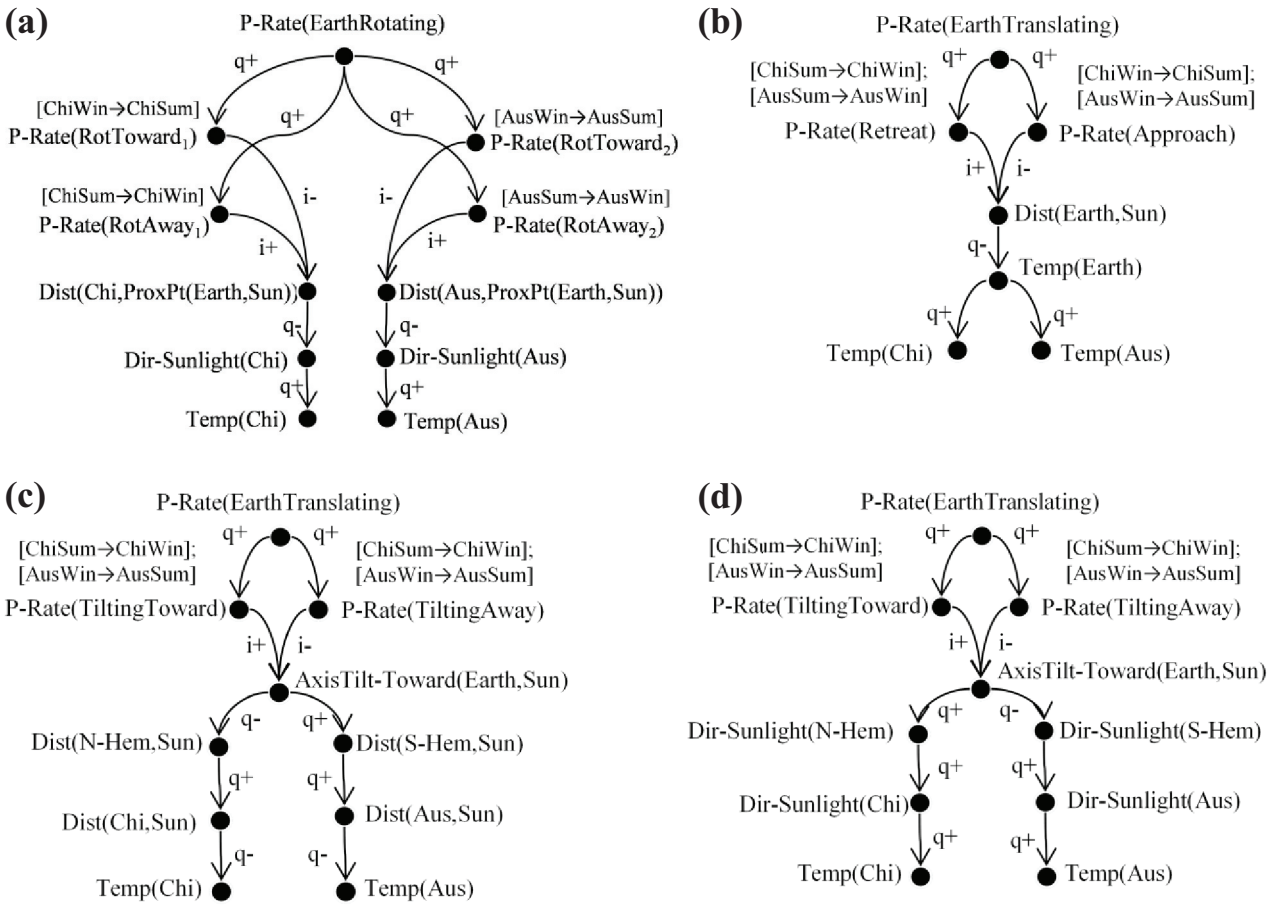

Figure 4: Influence graphs from explanations produced by the simulation. Edges describe qualitative (q+, q-) proportionalities and direct influences (i+, i-).

`(Temp PlanetEarth)` and `(decreasing (Temp PlanetEarth)` are believed. Consequently, e_0 and e_1 are 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 and e_1 must be removed as *best-xp*(b_t) for its target belief b_t.

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, as described 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 an 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 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 1a, 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 knowledge-includes: (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 cost of 56. This explanation is consistent with the opposite seasons fact, so no revision occurs.

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. To model Deidra and Angela's preference for the distance-based explanation, we used an additional ten-point cost on the belief `(qprop (Temp X) (Sunlight X))`. Under these parameter settings, the system creates 36 explanations,[5] and computes a preference for the explanation in graph (b), with a cost of 56. The system also created the explanation for graph (a) with a cost of 66. When presented with the opposite seasons fact, the system (like Deidra and Angela) changes its preferred explanation to that in graph (a) due to an inconsistency across *best-xp* explanations. Modeling individual differences in preferences is an important consideration, as discussed below.

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 costs 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 Sherin et al. 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. 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 more monitoring of the model formulation process.

Related Work

Like the system describe here, other cognitive systems extend and revise their knowledge by constructing or evaluating explanations. We discuss several related systems.

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 there being 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.

Other systems construct explanations using abduction. For example, Molineaux et al. (2011) determines the causes of failures through abductive explanation. Abduction increases the agent's knowledge of hidden variables, and consequently improves the performance of planning in partially-observable environments. Similarly, ACCEL (Ng & Mooney, 1992) creates multiple explanations via abduction, and it uses simplicity and set-coverage metrics to determine which is best. When performing diagnosis of dynamic systems, ACCEL makes assumptions about the state of components (e.g., a component is abnormal or in a known fault mode), and minimizes the number of assumptions used. By comparison, our explanation evaluation is more complex – the system can assume any model fragment condition, some assumptions (e.g., quantity changes) are more expensive

[5] The increased number of explanations is due to the belief that both distance and sunlight affect temperature.

than others, and other aspects (e.g., model fragment types and instances) are penalized within explanations.

Previous research in AI has produced postulates for belief revision in response to observations. The AGM postulates (Alchourròn et al., 1985) describe properties of rational revision operations for a deductively-closed knowledge base of propositional beliefs. Katsuno and Mendelzon'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 is different: 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 forces the system to ensure consistency with other *best-xp* explanations.

Discussion

We have simulated how people construct explanations from fragmentary knowledge and revise them when provided new information. By changing the initial knowledge of the system, we are able to simulate different interviewees' commonsense science 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 cost function used by our system is domain-general, albeit incomplete. The cost function presented here analyzes explanations with regard to QP theory (e.g., quantity changes and process instances) plus some general properties of explanations (e.g., inconsistencies and assumptions). To be sure, other factors not addressed by our cost function are also important considerations for explanation evaluation: belief probability, belief pervasiveness, level of specificity, credibility of knowledge (and knowledge sources), and diversity of knowledge. We intend to expand our system to account for these dimensions and for individual differences in responses to instruction (e.g., Feltovich et al., 2001).

Our simulation provides evidence that our approach helps address the challenges of commonsense science reasoning listed in the exposition of this paper: (1) representing mental models; (2) constructing coherent explanations with inconsistent and incomplete knowledge; and (3) flexibly revising conceptual knowledge. We demonstrated these capabilities by modeling novices rather than experts, since expert knowledge is more consistent, more complete, and less prone to large-scale revision.

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 Socio-cognitive Architectures Program of the Office of Naval Research.

References

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

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. 161-238.

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

Feltovich, P., Coulson, R., & Spiro R. 2001. Learners' (mis)undestanding of important and difficult concepts: A challenge to smart machines in education. In K. D. Forbus & P. J. Feltovich (Eds.), *Smart machines in education* (pp. 349-375). Menlo Park, CA: AAAI/MIT Press.

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

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

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

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

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

Molineaux, M., Kuter, U., Klenk, M. 2011. What just happened? Explaining the past in planning and execution. *Proceedings of the Sixth International ExaCt Workshop.*

Ng, H. T., & Mooney, R. J. 1992. Abductive plan recognition and diagnosis: A comprehensive empirical evaluation. In *KR-92*: 499–508.

Sherin,B., Krakowski, M., Lee, V. R. in review. Some assembly required: how scientific explanations are constructed during clinical interviews.

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