COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 42 (2018) 1110–1145

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12574

Representing, Running, and Revising Mental Models: A Computational Model

Scott Friedman,[a,b] Kenneth Forbus,[b] Bruce Sherin[c]

[a]*Smart Information Flow Technologies (SIFT), Minneapolis*
[b]*Qualitative Reasoning Group, Northwestern University*
[c]*School of Education and Social Policy, Northwestern University*

Received 15 May 2016; received in revised form 16 September 2017; accepted 17 October 2017

Abstract

People use commonsense science knowledge to flexibly explain, predict, and manipulate the world around them, yet we lack computational models of how this commonsense science knowledge is represented, acquired, utilized, and revised. This is an important challenge for cognitive science: Building higher order computational models in this area will help characterize one of the hallmarks of human reasoning, and it will allow us to build more robust reasoning systems. This paper presents a novel *assembled coherence (AC) theory* of human conceptual change, whereby people revise beliefs and mental models by constructing and evaluating explanations using fragmentary, globally inconsistent knowledge. We implement AC theory with Timber, a computational model of conceptual change that revises its beliefs and generates human-like explanations in commonsense science. Timber represents domain knowledge using predicate calculus and qualitative model fragments, and uses an abductive model formulation algorithm to construct competing explanations for phenomena. Timber then (a) scores competing explanations with respect to previously accepted beliefs, using a cost function based on simplicity and credibility, (b) identifies a low-cost, preferred explanation and accepts its constituent beliefs, and then (c) greedily alters previous explanation preferences to reduce global cost and thereby revise beliefs. Consistency is a soft constraint in Timber; it is biased to select explanations that share consistent beliefs, assumptions, and causal structure with its other, preferred explanations. In this paper, we use Timber to simulate the belief changes of students during clinical interviews about how the seasons change. We show that Timber produces and revises a sequence of explanations similar to those of the students, which supports the psychological plausibility of AC theory.

Keywords: Artificial Intelligence; Cognitive modeling; Conceptual change; Explanation; Qualitative reasoning; Commonsense science; Belief revision

Correspondence should be sent to Scott Friedman, SIFT, 319 1st Ave N., Suite 400, Minneapolis, MN 55401. E-mail: friedman@sift.net

1. Introduction

Constructing causal explanations about physical phenomena—and revising explanations in light of new information—is a ubiquitous process in our cognitive development and formal education. This plays an important part in the larger cognitive process of *conceptual change*, where new beliefs and representations are adopted in the presence of conflicting beliefs. Unfortunately, we lack large-scale computational models of the mental representations and processes involved in our conceptual change. This is not for lack of empirical results in the cognitive science literature: Solid research empirically documents students' misconceptions (e.g., Clement, 1982; diSessa et al., 2004; Hestenes et al., 1992; Ioannides & Vosniadou, 2002; McCloskey, 1983; Vosniadou & Brewer, 1992, 1994) and identifies instructional interventions that help repair them (e.g., Brown & Clement, 1989; Chi et al., 1994; Vosniadou et al., 2007). Psychological theories have been proposed to explain conceptual change (e.g., Carey, 1985, 2009; Chi, 2008; diSessa, 1993; Ohlsson, 2009; Posner et al., 1982; Vosniadou, 1994), but each has gaps, for example, not specifying how knowledge is represented or how competing theories coexist or how explanations are constructed.

Constructing theories and computational models of human conceptual change are grand challenges for Cognitive Science. We address three principle aspects of this challenge in this paper:

1. Representing people's knowledge about physical phenomena and dynamic systems.
2. Organizing this knowledge such that gaps, misconceptions, and inconsistencies can exist, yet explanations are still locally coherent (i.e., internally consistent and usable to explain new and previous phenomena).
3. Flexibly revising knowledge and explanations in light of new information.

This paper describes (a) a novel theory of human conceptual change, (b) an implementation of the theory with an integrated computational model, and (c) empirical results comparing the computational model to human novices learning about seasonal change. We describe these contributions next and relate them to other theories of conceptual change.

1.1. Outline of assembled coherence theory

We present the *assembled coherence* (AC) theory of human mental models and conceptual change. AC theory hypothesizes that people's mental models consist of interlocking pieces, including spatial relations, temporal relations, qualitative influences, and a hierarchical ontology of categories. These pieces are globally incoherent (i.e., they do not necessarily share common assumptions or representations; Friedman & Forbus, 2010, 2011) and globally inconsistent (i.e., two or more beliefs might be logically contradictory; Friedman & Forbus, 2010). People make sense of the world by assembling subsets of fragmentary knowledge into one or more coherent mental models that explain particular

behaviors, each with a distinct set of assumptions and causal structure. People evaluate competing explanations for likelihood, simplicity, and agreement with other beliefs. They then identify their best rationale for the phenomenon, which we call a *preferred explanation*. They associate the mental model (i.e., assumptions and causal structures) of the preferred explanation(s) with the phenomenon being explained, and they retain this association for subsequent reuse. Consequently, people may be aware of multiple coherent models and mechanism-based explanations for different phenomena, but they may regard them with different degrees of preference, and these preferences can change over time. When asked to explain a phenomenon or solve a problem they have explained before, they utilize fragments—or the entirety—of the mental model associated with the preferred explanation for that phenomenon.

Under AC theory, people have a *reuse bias* to leverage assumptions and causal structure from previous, preferred explanations within new explanations (Friedman & Forbus, 2010, 2011). This increases coherence (i.e., reduces complexity, reduces the number of assumptions, and increases shared causal structure) across all preferred explanations for different phenomena. If someone has already committed to a mental model that has productively explained and predicted diverse phenomena, it has high practical utility. The reuse bias promotes coherence across the learner's knowledge, but it also leads to *entrenchment* in mental models and assumptions, proportional to how pervasively they support preferred explanations. This means that merely suggesting an alternative account of the world is unlikely to promote deep conceptual change in a learner; rather, conceptual change involves (a) recognizing that their current account is incoherent, (b) recognizing that the new account increases coherence relative to the old model, and (c) incrementally shifting to the new account by retrospectively explaining phenomena with the new account.

Assembled coherence theory shares theoretical commitments with other theories of conceptual change. From the AC theory perspective, constructing explanations for the sake of internal sensemaking (e.g., Chi, 2000) involves assembling elementary knowledge elements (e.g., diSessa, 1993; diSessa & Sherin, 1998) with respect to central epistemic commitments (e.g., Vosniadou & Brewer, 1992, 1994), and this promotes metacognitive evaluation of mental models for consistency and coherence (e.g., Vosniadou, 2007). Under AC theory, misconceived mental models that productively explain and predict the world (despite their incorrectness) will be more resilient to change, all else being equal (Smith, diSessa, & Roschelle, 1994). Like Carey's (2009) theory of conceptual change, AC theory involves building multiple coherent accounts of the world and transitioning from one to another.

Assembled coherence theory relies on knowledge representation and reasoning formalisms from Artificial Intelligence as a theoretical account of how people reason about continuous physical systems and assemble this knowledge into explanations. People represent continuous processes (e.g., orbiting, rotation, heat transfer) and causal relationships between quantities (e.g., *the closer something is to a heat source, the greater its temperature*). AC theory draws from qualitative process (QP) theory (Forbus, 1984), reviewed in Section 2.2, for representation and reasoning formalisms for continuous processes and

quantities like these. This provides an account of how people reason about continuous changes in the world and how those changes propagate to other phenomena. AC theory relies on compositional modeling (Falkenhainer & Forbus, 1991), reviewed in Section 2.3, as an account for how people assemble these knowledge components into reusable mental models and explanations. Using qualitative models and QP theory to simulate human-like mental models in physical domains is not a new idea; this was an initial motivator for qualitative physics research (Forbus, 1984; Forbus & Gentner, 1997).

Assembled coherence theory makes novel theoretical commitments. For instance, AC theory hypothesizes that the reason people learn by explaining—called the *self-explanation effect* (Chi, 2000; Chi et al., 1994)—is due to the learner assembling fragmentary knowledge into more coherent, preferable, reusable mental models. More generally, AC theory hypothesizes that fragmentation and coherence—which have historically been opposing perspectives (diSessa et al., 2004; Ioannides & Vosniadou, 2002)—are actually different sides of the same coin. Furthermore, AC theory makes representational and computational commitments about how fragmentary knowledge is represented, assembled into coherent aggregates, evaluated, and revised. We describe a computational implementation of AC theory, and then we revisit AC theory in Section 6.

1.2. Outline of the computational model and experiment

Our computational model TIMBER (for Transforming Models & Beliefs via Explanation & Reflection) implements the AC theory of conceptual change using Artificial Intelligence techniques. TIMBER represents people's domain knowledge using (a) qualitative model fragments (e.g., Falkenhainer & Forbus, 1991; Rickel & Porter, 1997) to describe continuous processes (Forbus, 1984) and complex entity relationships and (b) predicate calculus expressions to describe entity categories, spatial relations, temporal relations, and ordinal relations. TIMBER uses explanations to organize knowledge: It records explanatory structure, competing explanations, and preferences over explanations. It uses an abductive model formulation algorithm and a meta-level cost function (Friedman, 2012) to model how people construct and evaluate explanations, respectively. Its cost function assigns zero cost for assumptions and inferences already used in previous, preferred explanations, resulting in a bias to reuse existing explanatory structure. It uses a greedy restructuring algorithm to support incremental belief revision.[1]

TIMBER has been used to simulate students' conceptual change in commonsense science domains including physics (Friedman & Forbus, 2010), biology (Friedman & Forbus, 2011), and the day–night cycle (Friedman, Barbella, & Forbus, 2012). As TIMBER constructs and installs new preferences over explanations, it revises the set of beliefs that it will subsequently be used to explain phenomena, but it still retains the beliefs and explanatory structure that it no longer prefers. This models the psychological self-explanation effect (Chi, 2000) whereby people repair erroneous mental models by constructing explanations.[2]

This paper describes promising results using TIMBER to simulate students' explanations and belief revisions during a clinical interview about the changing of the seasons (Sherin,

Krakowski, & Lee, 2012). To explain a proposition such as *Chicago is hotter in its summer than in its winter*, TIMBER performs the following operations:

1. Construct a causal qualitative model to justify the proposition, using model fragments and axioms in domain knowledge.
2. Identify competing explanations within the proposition's justification structure.
3. Numerically score competing explanations using a cost function, taking previous explanations and beliefs about other phenomena into consideration.
4. Select and record a preferred explanation.
5. Opportunistically revisit and revise preferred explanation(s) for previously explained phenomena to further reduce cost (i.e., increase global coherence).

We evaluate TIMBER based on its ability to simulate the students interviewed by Sherin et al. about seasonal change. The experimenters cataloged the intuitive knowledge that each student used while explaining the changing of the seasons, including mental models and propositions regarding the Earth, the sun, heat, and light. They also documented how students changed their account of the seasons during the course of the interview, when given new information. In each simulation trial, TIMBER begins with a domain theory corresponding to a single student in Sherin et al., encoded using an extension of the OpenCyc[3] ontology. TIMBER explains the changing of the seasons using this knowledge, resulting in an intuitive explanation like those described in Sherin et al. Like the student, TIMBER is then presented the information that Chicago's summer coincides with Australia's winter. In some trials, this information causes a high-cost inconsistency across preferred explanations, and TIMBER subsequently revises its explanation preferences to improve the cost. We compare TIMBER's explanations and explanation revisions to those of the students in the initial study.

We begin by discussing research in commonsense science, and then we review Sherin et al.'s study and the knowledge representation and reasoning techniques used in TIMBER. We then describe our approach and present simulation results. We close by discussing related work and future work.

1.2.1. Research on commonsense science

Simulating how people reason about physical phenomena such as the motion of a tossed ball, the boiling of a pot of water, or the changing of the seasons, is important for at least two reasons. The first reason is relatively obvious: In order to function as humans in the world—and to communicate with other humans about the world—we need to learn and reason about these physical phenomena.

The second, more subtle reason is that in cognitive science, research on what has been called "commonsense science" has played a uniquely central role. *Commonsense science* refers to knowledge of the natural world that is gained outside of formal science instruction. This includes knowledge gained from direct experience: tossing balls, watching pots of water boil, and feeling the warmth of direct sunlight. It also includes culturally derived knowledge, such as the information we gain from conversation and reading. Its importance has long been recognized in AI (e.g., Hayes, 1978).

Formal science instruction has become closely related to commonsense science in that commonsense science knowledge is adapted or replaced as formal scientific knowledge is gained. This image of formal science education has, in many contexts, come to be taken as a central example of how learning happens across disciplines. Consequently, research on commonsense science has taken on additional importance.

Even with this intense focus, there remain fundamental disagreements about the nature of commonsense science knowledge. On one side of the debate is the *theory theory*. According to this view, commonsense science knowledge is coherent, much in the way that the theories of scientists are coherent. An implication of this view is that students' commonsense science (i.e., their theories) must be replaced by instruction. On the other side of the debate is the *knowledge-in-pieces* perspective. In this view, commonsense science consists of a large number of fragments that are assembled, in a context-dependent manner, to explain physical phenomena. Educationally, the same fragments that support incorrect intuitive explanations could be leveraged and reused to support formal scientific knowledge.

In educationally oriented cognitive science, attempts to resolve this debate have been largely empirical: Students are given tests, or interviewed, and their responses are examined for coherence. There have been few attempts to create cognitively precise theories of commonsense science reasoning, and there have been even fewer efforts to build computationally explicit models of the type of commonsense science reasoning at the heart of these debates. That is the goal of this paper: to model the reasoning of students in a commonsense science interview setting.

From the perspective of AC theory, fragmentation (from the knowledge-in-pieces perspective) and coherence (from the theory theory perspective) describe the same knowledge system at different granularities: Fragmented models and beliefs can be fashioned into larger, coherent explanatory structures, and these structures can be evaluated and manipulated at a larger granularity to improve global coherence. This is the central principle of AC theory and its computational model TIMBER.

1.2.2. How seasons (and explanations) change

Sherin et al.'s study—and our TIMBER simulation thereof—focuses on how people explain and understand the changing of the seasons. 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 (Lelliott & Rollnick, 2010; Sherin et al., 2012). 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. After the student elaborated on their initial explanation, the interviewer would introduce, when appropriate, challenges to the student's explanation. For example, if the student's initial explanation 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 alerted them to an inconsistency in their account, and they subsequently revised their explanation. In this way, a student might transition among various intuitive explanations during an interview. Sherin et al. includes a listing of conceptual knowledge used by the students during the interviews, including propositional beliefs, general schemas, and fragmentary mental models.

The scientifically accurate explanation of the seasons depends on the fact that the Earth's axis of rotation is tilted. As the Earth orbits the sun, the axis of rotation always points in the same direction. When a hemisphere is inclined toward the sun, it receives more direct sunlight per unit area than when pointed away, which results in warmer and cooler temperature, respectively. While 12/21 students mentioned that the Earth's axis is tilted, only six of them used this fact in an explanation, and none of these were fully accurate. This is an important characterization: Some students knew *parts* of the correct explanation (e.g., the Earth's tilt), but they were unable to assemble this with *causal mechanism* knowledge to produce a coherent mental model to explain seasonal temperatures. Instead, students frequently explained that the Earth is closer to the sun during the summer and farther during the winter (Fig. 1).

The interview transcript from the student Angela[4] is listed in the online supplementary material, courtesy of Sherin et al. Angela begins by explaining that the Earth is closer to the sun in the summer than in the winter, and seasons change as the Earth approaches and retreats from the sun throughout its orbit. This *near-far* explanation is illustrated by a student sketch in Fig. 1. When the interviewer asks Angela if she has heard that Australia experiences its winter during Chicago's summer, and whether this is a problem for her explanation, Angela sees that her explanation is problematic. She eventually changes her answer by explaining that the spin of the Earth changes the seasons: The parts of the Earth that face the sun experience their summer, while the parts that face away experience winter. We call this the *facing* explanation. Other students used the near-far

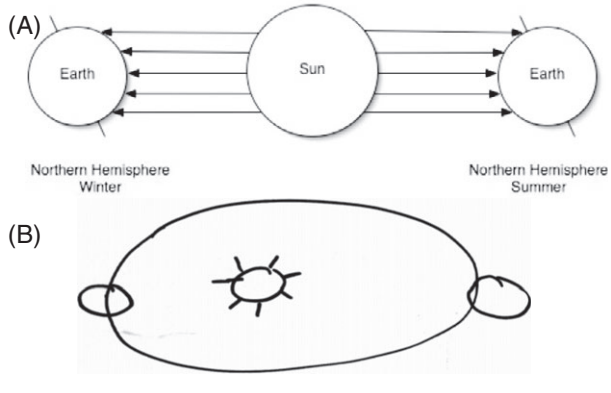

Fig. 1. Common misconception of seasonal change sketched by an interviewee.

explanation and the facing explanation, and many students mentioned idiosyncratic knowledge (e.g., they had seen a picture of a sunny day in Antarctica) that influenced their explanations.

Our TIMBER simulation models how Angela and other students in Sherin et al. create explanations of dynamic systems from fragmentary knowledge. We vary the knowledge in our model to simulate different students, some of whom have incomplete knowledge (e.g., lacking knowledge of causal mechanisms), and others who have misconceptions. Although the students in Sherin et al. were not given the correct explanation, we include a simulation trial that has access to the requisite knowledge to show that the model can simulate both novice and expert reasoning. We next review the qualitative modeling techniques used in TIMBER.

2. Background

We next review qualitative reasoning, QP theory, and compositional modeling, which are TIMBER's methods of representing and assembling conceptual knowledge.

2.1. Qualitative reasoning

"Quantity" is not synonymous with "number." A quantity (e.g., the volume of lemonade in a pitcher) may be assigned a numerical, unit-specific value (e.g., 12 fluid ounces) at a specific time, but people can very effectively reason about quantities without numbers. For instance, we might infer the volume of lemonade in a pitcher with an ordinal relationship such as "less than the volume of the pitcher" or with a qualitative label such as "a lot," based on anchors within our space of experiences (Paritosh, 2004). We can reason about causality in a similar non-numeric fashion. For example (quantities in italics), we know that if we increase the *angle of the pitcher*, the *height of the pitcher lip* will decrease. Once the lip decreases below the *height of the lemonade*, the liquid will begin flowing, and as we increase the *angle of the pitcher*, we will increase the *rate of flow*. This simple qualitative reasoning used "increase" and "decrease" to capture the sign of nonzero quantity derivatives over time, and "below" to capture ordinal relationships between values of two quantities. People can thereby qualitatively reason about continuous quantities, rates of processes, directionalities of change, and ordinal relationships (i.e., greater than, less than, equal to) between them. Previous work provides methods for representing and reasoning about processes (e.g., Forbus, 1984) and devices (e.g., de Kleer & Brown, 1984), and simulating systems provided this knowledge (e.g., Kuipers, 1986).

Novices and experts alike often reason with incomplete and imprecise qualitative knowledge, especially in situations of informational uncertainty (Trickett & Trafton, 2007). Consider the incorrect near-far novice explanation of how the seasons change (Fig. 1): the Earth orbits the sun along an elliptical path and is closer to the sun in the summer than in the winter. This mental model includes no numbers, but relies on

quantities (e.g., the Earth's distance from the sun, the Earth's temperature) and relations between quantities (e.g., the Earth's temperature increases as its distance to the sun decreases). This is qualitative reasoning. We next review methods for representing, assembling, and reasoning with qualitative models.

2.2. Knowledge representation conventions

We built TIMBER using an extended OpenCyc ontology, and we use OpenCyc conventions to describe TIMBER's knowledge and algorithms in this paper. We review relevant knowledge representation conventions here.

The models, inferences, and input knowledge of TIMBER are all relational statements and composites thereof. A statement such as **(greaterThan (Temp mycoffee) (Temp myspoon))** has a *relation* (i.e., **greaterThan**) that traditionally begins with a lowercase letter and precedes the *arguments* (i.e., **(Temp mycoffee)** and **(Temp myspoon)**). In this example, both arguments are *function* terms. Functions can refer to specific dimensions or properties of *entities* such as **mycoffee** and **myspoon**. Taken together, the statement asserts a greater than relationship between the temperatures of two entities. In TIMBER, relational statements describe all spatial, temporal, and quantitative information.

Some statements, such as **(isa myspoon Teaspoon)**, assert category membership with the **isa** predicate: The entity **myspoon** is a member of the **Teaspoon** category.[5] An entity can be a member of more than one category. Categories are organized into a lattice: If an entity is a member of a category, it is implicitly a member of all ancestors of that category. The same is true for relations: If a relation is asserted, all ancestor relations hold implicitly.

2.3. Qualitative process theory

Qualitative process (QP) theory (Forbus, 1984) provides a vocabulary for representing changes in physical systems. Under QP theory, only *processes* cause changes in a physical system. For our example of pouring lemonade in the previous section, processes include the tilting of the pitcher and the flow of lemonade.

QP theory defines two kinds of qualitative causal relationships between quantities. *Direct influences* positively or negatively constrain the derivative of a quantity by the rates of processes. For example, for a process of liquid flow L,

(i+ (Mass destination) (Rate L))
(i- (Mass source) (Rate L))

Direct influences are additive; if water is flowing into a bathtub at the same rate that water is flowing out of it via the drain, the mass of the water in the tub has a derivative of zero (i.e., it is not changing). Direct influences are causal: they describe exactly how a process directly affects some aspect of the world.

The other type of influence is *indirect influences*, also known as *qualitative proportionalities*, because they propagate the direct effects of processes. These provide partial information about monotonic functional dependence, for example,

(qprop (Level water) (Mass water))
(qprop- (Level water) (Width (Container water)))

The above **qprop** statement asserts that, all else being equal, when the mass of water in a container increases, it causes the water level to rise (and if mass decreases, level will fall). The above **qprop-** statement asserts that as the width of the container increases (think of an inflatable swimming pool), the water level will fall (and if width decreases, level will rise). Unlike direct influences, these indirect influences are algebraic constraints where changes accumulate over time. For both types of influences, closed world assumptions must be made to reason about causal effects.

2.4. Compositional modeling

Model fragments (Falkenhainer & Forbus, 1991) represent physical or conceptual entities (e.g., the asymmetrical path of a planet's orbit) and processes (e.g., a planet approaching and retreating from its sun along that path, as in Fig. 1). Modeling the common misconception in Fig. 1 involves several such model fragments. Fig. 2 shows two model fragment types used in the simulation: the conceptual model fragment

```
ConceptualModelFragmentType AstronomicalHeating
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))
```

Fig. 2. AstronomicalHeating (top) and Approaching-PeriodicPath (bottom) model fragment types.

AstronomicalHeating, and the process **Approaching-PeriodicPath**. A type of model fragment can be uniquely defined by its sets of participants, constraints, assumptions, conditions, and consequences. We describe these using the model fragments in Fig. 2 as an example, where terms preceded by a question mark are variables that can be filled by observed or assumed entities.

Participants are the entities involved in the phenomenon, such as **?heater** in **AstronomicalHeating**. All participants have a *variable term* (e.g., **?heater**), a *type*, and a *role* in the model fragment. Participant **?heater** has a type **HeatSource**, so the proposition **(isa TheSun HeatSource)** must be true for **TheSun** to fill the **?heater** participant role. Participant **?heater** has role **providerOf** within **AstronomicalHeating**, so **(providerOf AH-inst TheSun)** would be true of any **AstronomicalHeating** instance **AH-inst** where **TheSun** is **?heater**. Entities in the scenario (e.g., **TheSun**) are bound to participant variables (e.g., **?heater**), using a *binding list* such as {⟨**?heater**, **TheSun**⟩, ⟨**?heated**, **PlanetEarth**⟩}.

Constraints are statements about the participants that delimit the model fragment's existence. When the constraints hold, an *instance* of the model fragment type is inferred as a distinct entity, given the participant bindings. For example, if **(spatiallyDisjoint TheSun PlanetEarth)** is true of **HeatSource** instance **TheSun** and **AstronomicalBody** instance **PlanetEarth**, then the participant roles and the constraints hold. Consequently, a new model fragment instance can be created with bindings {⟨**?heater**, **TheSun**⟩, ⟨**?heated**, **PlanetEarth**⟩}.

Modeling assumptions describe the granularity, perspectives, and approximations of the model fragment. These help determine the model fragment's relevance, since the behavior of a single physical phenomenon (e.g., light from the sun reaching the Earth) can be described at multiple granularities (e.g., waves or particles).

Conditions are statements about a model fragment's participants that delimit its behavioral scope, such as **(active ?movement)** in **Approaching-PeriodicPath**. When all conditions of a model fragment instance hold over the participants, the instance is *active*. These differ semantically from model fragment constraints (defined above): If the constraints are satisfied but the conditions are not, an instance of a model fragment exists, but it is not active.

Consequences (*S*) are statements that describe a model fragment instance's behavior when it is active. For example, one consequence of **AstronomicalHeating**—which is only inferred when the process is active—is that the temperature of **?heated** increases as the distance from **?heater** to **?heated** decreases, all else being equal.

This technique of instantiating and activating model fragments is known as *model formulation* (Falkenhainer & Forbus, 1991). Model formulation occurs in a logical context, called a *scenario*, containing a partial description of the phenomena to be modeled, such as propositional facts about **HeatSource** entities, **AstronomicalObject** entities, and spatial relations over them. Model fragments are stored within a *domain theory* comprised of model fragments and scenario-independent beliefs. Model formulation produces a *scenario model* consisting of one or more model fragment instances. One model fragment instance may serve as a participant of another, so the scenario model may have a nested structure.

3. TIMBER

Here, we describe the TIMBER computational model, using the results from our simulation of Sherin et al.'s Angela subject, discussed above, as an extended example. For the Angela trial, TIMBER starts with a set of model fragments for both the near-far explanation and the facing explanation, since Angela constructed both of these explanations during the interview without learning these models from the interviewer.

Using the Angela example, we describe TIMBER with a focus on (a) organization of explanations and domain knowledge; (b) an abductive model formulation algorithm for building scenario models from model fragments; (c) metareasoning for computing a total preferential pre-order[6] over competing explanations; (d) incorporating new, credible knowledge; and (e) handling inconsistencies to increase global coherence across preferred explanations.

3.1. Organizing explanations and domain knowledge

In TIMBER, explanations and their constituent beliefs are organized in a network that supports metareasoning and conceptual change. A portion of a network from TIMBER's Angela trial is shown in Fig. 3. The legend of Fig. 3 labels the key beliefs (facts and model fragments) for reference (terms labeled *f* are facts and terms labeled *m* are model fragments), but the specific beliefs are not yet important. We describe the network with respect to this example. To improve readability, we lay out the network on three tiers. We describe them from bottom to top.

3.1.1. Bottom (domain knowledge) tier

The bottom tier of the network in Fig. 3 contains domain knowledge, including model fragments and beliefs that are supported by observation or instruction. Domain knowledge can serve as *premises*: They need no justification and are believable independently of explanations they support. Fig. 3 plots a small subset of the propositions and model fragments used in TIMBER's Angela trial.

3.1.2. Middle (assembly) tier

The middle tier plots inferences—including assumptions, assembled model fragment instances, and logical assertions—that TIMBER generates while constructing explanations as well as their justifications (displayed as right-pointing black triangles). This is based on the justification structure network of a traditional truth-maintenance system (Forbus & de Kleer, 1993). The antecedents of a justification are adjacent on its left, and its consequences are adjacent on its right. Each justification represents a logical inference: Believing the antecedents is sufficient for believing the consequences. The assumptions and justifications in Fig. 3 do not represent all of TIMBER's inferences; this is a fraction of the network. Unlike the bottom tier, the inference nodes on this tier are not directly supported by observation or instruction; they are inferred during the explanation construction

Legend

f_0	(isa earthPath EllipticalPath)	f_9	(active AH-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	(isa ProximalPoint ModelFragment)	f_{12}	(i+ (Dist TheSun PlanetEarth) (Rate RPP-inst))
m_1	(isa DistalPoint ModelFragment)	f_{13}	(increasing (Temp PlanetEarth))
m_2	(isa Approaching-Periodic ModelFragment)	f_{14}	(decreasing (Temp PlanetEarth))
m_3	(isa AstronomicalHeating ModelFragment)	f_{15}	(qprop (Temp Australia) (Temp PlanetEarth))
m_4	(isa Retreating-Periodic ModelFragment)	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 AH-inst AstronomicalHeating)	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))

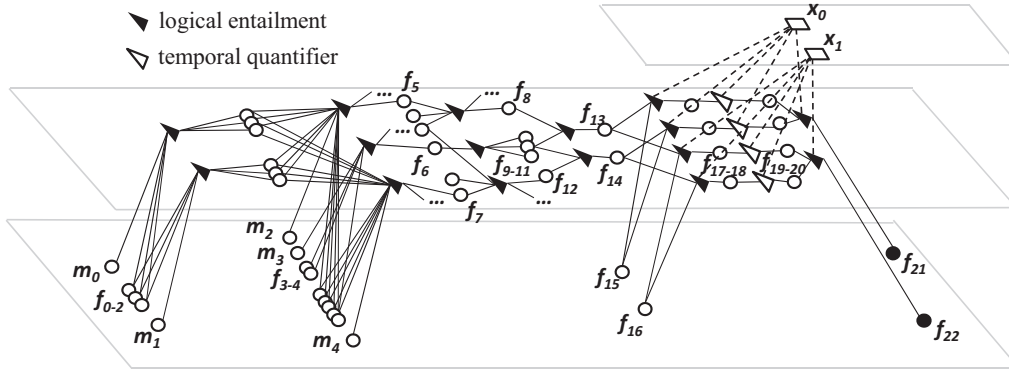

Fig. 3. A knowledge-based network of explanations (top tier), assembly (middle tier), and domain theory (bottom tier). Explanations justify seasonal change in Australia (x_0) and Chicago (x_1). Only key beliefs are labeled.

process. If a belief on this tier is subsequently observed or supported by instruction, it is moved to the bottom tier.

Some beliefs are explicitly quantified in specific states using temporal quantifiers represented as white triangles in Fig. 3. Consider the temporal quantifier that justifies f_{20} with f_{18} in Fig. 3. This states that TIMBER believes f_{20} (i.e., **(holdsIn (Interval ChiSummer ChiWinter) (decreasing (Temp Chicago)))**) so long as the belief f_{18} (i.e., **(decreasing (Temp Chicago))**) and all beliefs justifying f_{18} hold within the state **(Interval ChiSummer ChiWinter)**. This compresses the explanation structure: without these temporal quantifiers, we would have to store each belief b left of f_{20} as **(holdsIn (Interval ChiSummer ChiWinter) *b*)**. We can decompress the explanation into this explicitly quantified notation without any loss of information, but we can also perform temporal reasoning without decompressing.

3.1.3. Top (explanation) tier

The top tier plots explanation nodes. Fig. 3 depicts a subset of all explanations constructed by TIMBER, plotted with quadrilateral nodes x_0 and x_1 on the top tier. Each explanation represents a unique *well-founded explanation* for some situation or belief. A well-founded explanation for a node n is any set of justifications $J = \{j_0, \ldots, j_k\}$ such that (a) node n is justified by j_k, (b) all antecedent beliefs of j_k—and of all other justifications in J—are justified by another justification in J, and (c) if any justification is removed, some justification in J will have an unsupported antecedent. This is based on the concept of *well-founded support* from the truth-maintenance system literature (Forbus & de Kleer, 1993).

Each explanation in TIMBER is uniquely defined with the tuple $\langle M, J, B \rangle$, where

- M is the *explanandum*: a set of one or more propositions explained by the explanation. In Fig. 3, the explanandum beliefs are at the right.
- J is a set of justifications that comprise a well-founded explanation for M. In Fig. 3, explanation nodes x_0 and x_1 have dashed lines to some of their justifications J, and other lines are omitted for clarity.
- B is the set of beliefs that comprise the explanation, including all antecedents and consequences of the explanation's justifications J. This includes domain knowledge (bottom tier) and inferences/assumptions (middle tier).

Based on these definitions, the network in Fig. 3 describes the shared structure, causal mechanisms, premises, and temporal quantification of explanations x_0 and x_1. In the following, we walk through TIMBER's simulation of the student Angela to describe how TIMBER constructs, evaluates, reuses, and revises explanations.

3.2 Creating explanations in TIMBER

Here, we describe how TIMBER explains Chicago's seasonal change and then subsequently explains Australia's seasonal change, opportunistically reusing beliefs from its explanation of Chicago's seasonal change.

3.2.1. Explaining Chicago's seasons

At the beginning of our Angela trial, we query the system for an explanation of why it is warmer in Chicago's summer than in its winter. TIMBER then builds a scenario model to justify the proposition via model formulation (defined in Section 2). Some model formulation algorithms perform exhaustive forward-chaining to instantiate all model fragments possible within a scenario (e.g., Forbus, 2010), while others back-chain from a target assertion and create scenario models that entail the assertion (e.g., Rickel & Porter, 1997). TIMBER's algorithm is an improvement of the back-chaining approach used in Friedman and Forbus (2010, 2011). It initializes a domain context D as a set of model fragments to use in its explanation. It initializes a scenario context S containing propositional beliefs, for example, from other previous explanations. S includes model fragment instances from previous explanations, in order to leverage already inferred causal mechanisms.

After initialization, TIMBER begins model formulation with the procedure ***justify-ordinal-proposition***, shown in Fig. 4 (Fig. 5). This takes an ordinal (e.g., **greaterThan**) proposition *m* that requires an explanation.[7] The algorithm directly justifies the belief by formulating a compositional qualitative model that logically entails the ordinal proposition. For our example, we use the following proposition that states that the temperature of Chicago in greater in its summer than in its winter:

(greaterThan (M (Temp Chicago) ChiSummer) (M (Temp Chicago) ChiWinter))
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).

We describe TIMBER's model formulation algorithm using the above inputs, given a domain theory (*D*) and scenario description (*S*) capable of modeling the misconception in Fig. 1.

When ***justify-ordinal-proposition*** is called on the belief that Chicago is warmer in its summer than its winter, TIMBER binds *q* to **(Temp Chicago)**, s_1 to **ChiSummer**, and s_2 to **ChiWinter**. It then queries to determine whether (a) **ChiWinter** is after **ChiSummer** and whether (b) **ChiSummer** is after **ChiWinter**. Since both are true, the beliefs $f_{19–20}$ in Fig. 3 are encoded to justify the proposition. Next TIMBER must model *how* (**Temp**

Justifying ordinal propositions with qualitative models

global domain *D*, **scenario** *S*

procedure ***justify-ordinal-proposition*** **(proposition** *m***)**
// here *m* is of the form `(greaterThan (M <q> <s1>) (M <q> <s2>))`
let *q*, s_1, s_2 = **quantity-of**(*m*), **state-1-of**(*m*), **state-2-of**(*m*)
if query *S* **for** `(after` s_2 s_1`)` **then:** ***justify-quantity-change***(*q*, **i-**)
if query *S* **for** `(after` s_1 s_2`)` **then:** ***justify-quantity-change***(*q*, **i+**)

procedure ***justify-quantity-change*** **(quantity** *q*, **direction** *d***)**
// Find direct and indirect influences of *q*
// d is either **i+** or **i-**
instantiate-fragments-with-consequence**((qprop** *q* `?x)` **)**
instantiate-fragments-with-consequence**((qprop-** *q* `?x)` **)**
instantiate-fragments-with-consequence**(** (*d q* `?x)` **)**
let I_i = **query** *S* **for qprops on** *q*. // results are in form **(qprop/qprop-** *q* **?x)**
for each *i* **in** I_i**:**
let q_i = **influencing-quantity**(*i*)
let d_c = *d* **if (direction-of-influence**(*i*) **== qprop) else opposite**(*d*)
justify-quantity-change(q_i, d_c)

Fig. 4. Pseudo-code for back-chaining model formulation.

Abductive model formulation

global domain *D*, **scenario** *S*

procedure ***instantiate-fragments-with-consequence*** **(proposition** *p***)**
  let *F* = **query** *D* **for model fragments with some consequence that unifies with** *p*
  for each *f* **in** *F*:
    for each consequence *c* **of** *f* **that unifies with** *p*:
      let *B* = **bindings-between**(*c*, *p*)
      abductive-mf-instantiation(*f*, *B*)

procedure ***abductive-mf-instantiation*** **(modelfrag** *m*, **bindings** *B***)**
  // Bindings *B* may be incomplete.
  // Find participant collections $\{\langle slot_0, coll_0\rangle, \ldots, \langle slot_n, coll_n\rangle\}$ of *m*.
  let P_m = **participants-of**(*m*) **substituted by** $\langle slot, ent\rangle \in B$
  // Find the constraints of *m*.
  let N_m = **constraints-of**(*m*) **substituted by** $\langle slot, ent\rangle \in B$
  // Recursively instantiate participant model fragments.
  for each $\langle slot, coll\rangle$ **in** P_m **where** *coll* **is a Model Fragment type**:
    // Using local constraints N_m, find bindings for the recursive call.
    let N_f = **ground statements in** N_m **that:**
      1. have a participant role of *coll* **as its predicate.**
      2. have *slot* **as a first argument.**
    let B_f = **bindings between participant slots of** *coll* **and entities in** N_f
    // Make a recursive call to instantiate the participant.
    abductive-mf-instantiation(*coll*, B_f)
  // Compute participant bindings for modelfrag *m* in *D*, including incomplete ones.
  let *InstanceBindings* = **query** *D* **for bindings of** $P_m \wedge N_m$
  for each *I* **in** *InstanceBindings*:
    // Assume the existence of all unknown participants.
    let *UnkParticipants* = $\{\langle slot, ent, coll\rangle \in I : \mathbf{variable}(ent)\}$
    for each $\langle slot, ent, coll\rangle$ **in** *UnkParticipants*:
      let *e* = **new-skolem-entity**(*e*, *coll*)
      set *I* = **replace** $\langle slot, ent, coll\rangle$ **with** $\langle slot, e, coll\rangle$ **in** *I*
    // Add the constraints, conditions, consequences, and roles to the scenario model.
    instantiate-model-fragment(*m*, *I*)

Fig. 5. Pseudo-code for back-chaining model formulation and abductively instantiating model fragments by assuming the existence of participant entities.

Chicago) decreases between **ChiSummer** and **ChiWinter** and *how* it increases between **ChiWinter** and **ChiSummer**. It achieves this via two subsequent invocations:

justify-quantity-change((**Temp Chicago**), **i-**)
justify-quantity-change((**Temp Chicago**), **i+**)

Notice that these invocations make no mention of **ChiWinter** and **ChiSummer**. This is because the system is building a model of the *mechanisms* by which the temperature of Chicago might increase and decrease. These beliefs and causal mechanisms are explicitly quantified in specific states using temporal quantifiers (white triangles in Fig. 3).

The above invocation ***justify-quantity-change***((**Temp Chicago**), **i-**) first instantiates all model fragments in D that contain a consequence that unifies with one of the following patterns:

- **(qprop (Temp Chicago) ?x)**: **?x** influences Chicago temperature.
- **(qprop- (Temp Chicago) ?x)**: **?x** inversely influences Chicago temperature.
- **(i- (Temp Chicago) ?x)**: Process rate **?x** directly decreases Chicago temperature.This locates all mechanisms that can directly (i.e., **i-**) or indirectly (i.e., **qprop** and **qprop-**) decrease Chicago's temperature, and instantiates qualitative models accordingly.

In its Angela trial, TIMBER finds the influence **(qprop (Temp Chicago) (Temp PlanetEarth))** in its domain theory (plotted as f_{16} in Fig. 3), asserting that the temperature of Chicago will decrease if the temperature of the Earth decreases. It next attempts to justify the Earth's decrease in temperature **(decreasing (Temp PlanetEarth))**, plotted as f_{14} in Fig. 3. This results in the recursive invocation:

justify-quantity-change((**Temp PlanetEarth**), **i-**)

In this recursive invocation, TIMBER invokes ***instantiate-fragments-with-consequence*** (Fig. 5) to locate and instantiate mechanisms that affect decrease the Earth's temperature. The procedure finds **AstronomicalHeating** with relevant consequences, so it invokes ***abductive-mf-instantiation*** (Fig. 5) for model **AstronomicalHeating**, with {⟨**?heated**, **PlanetEarth**⟩}, so the **?heater** participant is unbound. This procedure searches for and instantiates all **AstronomicalHeating** instances conforming to the partial binding. This instantiates a single, complete model fragment instance with participant bindings {⟨**?heated**, **PlanetEarth**⟩, ⟨**?heater**, **TheSun**⟩}, producing the statements $f_{9–11}$ in Fig. 3, including the model fragment instance's consequences:

(qprop- (Temp PlanetEarth) (Dist TheSun PlanetEarth))
(qprop (Temp PlanetEarth) (Temp TheSun))

When the procedure next searches for influences of **(Temp PlanetEarth)**, it will find these statements and justify the Earth's cooling with an increase in **(Dist TheSun PlanetEarth)** or a decrease in **(Temp TheSun)**. This makes another recursive invocation of ***justify-quantity-change*** to justify an increase in **(Dist TheSun PlanetEarth)**. This subsequently composes a **Retreating-Periodic** instance whose rate increases the Earth's distance to the sun (statement f_{12} in Fig. 3) during part of its orbit around the sun.

We have described how TIMBER justifies Chicago's decreasing temperature. It justifies Chicago's *increase* in temperature in an analogous fashion, using some of the same model fragment instances (e.g., the same **AstronomicalHeating** instance) and some new model fragments, including an **Approaching-Periodic** instance whose rate decreases the Earth's

distance to the sun (statement f_8 in Fig. 3). This justifies the Earth's increase in temperature (statement f_{13} in Fig. 3), and downstream, Chicago's increase in temperature.

TIMBER justifies Chicago's seasonal change using mechanisms in its domain knowledge, but the justification structure may contain multiple, competing explanations. TIMBER creates a unique explanation node (e.g., x_1 in Fig. 3) for each well-founded explanation of the explanandum. In its simulation of Angela, TIMBER constructs multiple explanations for Chicago's seasons, only one of which (x_1) is shown in Fig. 3. Consider the following simplified explanations in English:

- $\boldsymbol{x_1}$: The Earth retreats from the sun for Chicago's winter and approaches for its summer (shown in Fig. 3).
- $\boldsymbol{x_2}$: The sun's temperature decreases for Chicago's winter and increases for its summer.
- $\boldsymbol{x_3}$: The sun's temperature decreases for Chicago's winter, and the Earth approaches the sun for its summer.
- $\boldsymbol{x_4}$: The Earth retreats from the sun for Chicago's winter, and the sun's temperature increases for its summer.

Explanations $\{x_1, x_2, x_3, x_4\}$ compete with each other to explain the seasons. However, x_2, x_3, and x_4 are all problematic. Explanations x_3 and x_4 contain asymmetric 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 rest of the cycle. Explanation x_2 is not structurally or temporally problematic, but D contains no model fragments that can describe the process of the sun changing temperature. Consequently, these changes are *assumed* rather than justified by processes. Assumed quantity changes are problematic because they represent unexplainable changes in a system. These are also problematic under the *sole mechanism assumption* (Forbus, 1984), which states that all changes in a physical system are the result of processes.[8] Just as we have analyzed and discredited TIMBER's explanations $x_{2\text{-}4}$ that compete with explanation x_1, TIMBER analyzes its explanations automatically, as we describe next.

3.3. Cost-based epistemic preferences

The cognitive science literature has characterized factors that impact people's judgments of explanations, including causal simplicity, coverage of observations, goal appeal, and narrative structure (Lombrozo, 2011). The Artificial Intelligence community has modeled some of these as a posteriori likelihood (Pearl, 1988), constraint satisfaction (Thagard, 2000), assumption counting (Ng & Mooney, 1992), and assumption cost (Charniak & Shimony, 1990). Unlike previous systems, TIMBER's evaluates explanations based on their credibility and causal complexity *across all preferred explanations*. This implements Occam's Razor globally, biasing TIMBER to choose explanations that cohere with others.

TIMBER's cost function numerically scores the *additional* complexity that an explanation would incur. It computes this by summing the cost of *epistemic artifacts* that would be incurred by preferring that explanation. Epistemic artifacts (hereafter "artifacts") are listed in Table 1 with their corresponding numerical costs. If an artifact, such as a model

fragment, is already used within another preferred explanation, the artifact incurs zero cost.

The artifacts listed in Table 1 each describe a different dimension of complexity or conflict: contradictions indicate logical conflicts; asymmetric and assumed quantity changes indicate systematic conflict; model fragments and model fragment instances indicate qualitative and quantitative complexity, respectively; assumptions indicate uncertainty complexity; and credibility indicates conflict with authority.

TIMBER computes each explanation's cost by identifying and summing all additional artifacts that would be incurred if that explanation was preferred. Consequently, when TIMBER simulates the Sherin et al. interviews, artifacts already incurred by explaining Chicago's seasons incur no additional cost to explain Australia's seasons. TIMBER sorts explanations by cost, where lower cost is better, and chooses the lowest-cost explanation as the preferred explanation for the explanandum. Artifacts grow monotonically with preferred explanations, and the only way to *remove* an artifact is to remove a preferred explanation or replace it with another.

The numerical costs listed in Table 1 were determined empirically to maximize accuracy of cognitive simulation of the students in Sherin et al.'s (2012) study, and most of these artifacts and costs have also been used to simulate students reasoning about the day–night cycle (Friedman et al., 2012). We do not believe this list of artifacts is complete, and we discuss opportunities for expanding and refining these in future work.

3.3.1. Explaining Australia's seasons

At this point, TIMBER has constructed and chosen a preferred explanation for Chicago's seasonal change. We next query TIMBER for an explanation of why Australia is warmer in its summer than in its winter. This again invokes ***justify-ordinal-proposition*** to construct

Table 1
Epistemic artifacts used in our simulation, including numerical costs and conditions for existence

Artifact: Cost	Artifact Belief Constituents
Contradiction: 100	Any set of inconsistent beliefs *B* such that no proper subset thereof is inconsistent
Asymmetric quantity change: 40	A quantity change in an explanation *x* that does not have a reciprocal quantity change in a cyclical state-space
Assumed quantity change: 30	A quantity increase or decrease that has no influencing process or influence. Processes are the mechanisms of change in a physical system (Forbus, 1984), so the lack of an influencing process is an anomalous behavior
Model fragment: 4	Beliefs of form (**isa *mf* ModelFragment**), where *mf* is a model fragment, e.g., **AstronomicalHeating**
Model fragment instance: 2	A belief of form (**isa *inst mf***) where *inst* is the instance name and *mf* is the model fragment type, e.g., (**isa mfi0 AstronomicalHeating**)
Credibility: [−1,000, 0)	A belief communicated from another source. The utility (i.e., *negative* cost) of the artifact is proportional to the credibility of the source, e.g., the student Angela states that she learned about Earth's orbit from second grade (see online supplementary material)

explanations for Australia's seasons. When TIMBER chooses among competing explanations for Australia's seasons using the cost function, the cost of each explanation is influenced by the explanations it presently prefers (e.g., x_1 preferentially explains Chicago's seasons). As described above, TIMBER's zero-cost reuse of existing artifacts biases it to choose a low-cost near-far explanation for Australia's seasons (x_0 in Fig. 3) that shares most of its causal model with the preferred explanation for Chicago's seasons (x_1 in Fig. 3).

3.4. Comparing TIMBER explanations to student explanations

At this point, we want TIMBER to describe the mechanisms that cause seasonal change and temperature change. Sherin et al. do not give the interviewees a pretest or posttest; rather, they ask the student to explain it freely. Generating causal explanations in English is outside the scope of this research, so we have TIMBER describe causal models using influence graphs as illustrated in Fig. 6. Given one or more explanations, TIMBER automatically constructs an influence graph by (a) creating a vertex for each quantity described in the explanation and (b) creating a directed edge for every influence described in the explanation. In the case of Fig. 6, TIMBER graphs the two preferred explanations where Australia's seasons and Chicago's seasons are jointly explained with the same mechanisms.

The majority of the influence graph in Fig. 6 describes continuous causal mechanisms common to both explanations. The only explanation-specific components are the temperatures of Chicago and Australia and their qualitative proportionalities to the temperature of the Earth. This illustrates how TIMBER reuses knowledge across explanations and explains phenomena with existing causal structure, thereby implementing the reuse bias of AC

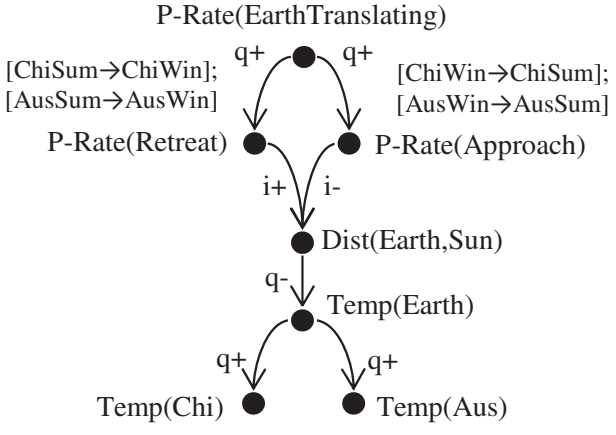

Fig. 6. An influence diagram of the near–far explanation of both Chicago's (Chi) and Australia's (Aus) seasons. Nodes are quantities and edges describe positive and negative direct influences (i+, i−) and indirect influences (q+, q−). Bracketed ranges quantify process activity.

theory. Even though explanations exist as separate compositions in AC theory (and in the computational model TIMBER), they share significant fragmentary knowledge.

3.5. Accommodating new, credible information in TIMBER

Thus far, we have described how TIMBER constructs and computes preferences for the two explanations plotted in Fig. 3: one for how Chicago's seasons change (x_1) and another for how Australia's seasons change (x_0). Other explanations for Chicago's and Australia's seasons exist in TIMBER, but they are not preferred since they incur a greater cost.

In Sherin et al.'s study, recall that if a student's explanation did not account for different seasons in different parts on the Earth—like TIMBER's presently preferred explanations—the interviewer asked them whether they were aware that Chicago's winter coincided with Australia's summer. This caused some students, including Angela, to revise their explanation of seasonal change. We describe how TIMBER models these students by incorporating new information and subsequently revising its preferred explanations.

To begin, we add the *opposite seasons information* to TIMBER's domain knowledge as the following four statements:

(cotemporal ChiSummer AusWinter)
(cotemporal ChiAutumn AusSpring)
(cotemporal ChiWinter AusSummer)
(cotemporal ChiSpring AusAutmn)

These statements are from a trusted source, so each has a credibility artifact of cost −1,000 (recall that negative cost indicates positive utility). TIMBER will receive this high utility while these statements exist in its preferred knowledge.

After adding these statements to domain knowledge and receiving high utility, TIMBER searches for contradictions across its preferred explanations (e.g., x_0 and x_1 in Fig. 3) and preferred domain knowledge. This uses domain-general rules for detecting logical inconsistencies (i.e., believing a belief and its negation), ordinal inconsistencies (i.e., greater-than and less-than conflicts), and derivative inconsistencies (i.e., when a quantity is simultaneously increasing and decreasing).

In the ongoing Angela example, consider the near-far Australia explanation x_0 with constituent beliefs B_0 and the near-far Chicago explanation x_1 with constituent beliefs B_1. The following statements—among many others—are included in these belief sets:

B_0 includes the statement *between Australia's summer and winter, the Earth cools*:

(holdsIn (Interval AusSummer AusWinter) (decreasing (Temp PlanetEarth)))

B_1 includes the statement *between Chicago's winter and summer, the Earth warms*:

(holdsIn (Interval ChiWinter ChiSummer) (increasing (Temp PlanetEarth)))

Before the opposite seasons statements were incorporated, B_0 and B_1 were globally consistent. Afterward, however, TIMBER infers that **(Interval AusSummer AusWinter)** coincides with **(Interval ChiWinter ChiSummer)** and the Earth's temperature is

believed to increase and decrease simultaneously. TIMBER detects that this contradiction creates a contradiction artifact over the following four statements:

(cotemporal ChiSummer AusWinter)
(cotemporal ChiWinter AusSummer)
(holdsIn (Interval AusSummer AusWinter) (decreasing (Temp PlanetEarth)))
(holdsIn (Interval ChiWinter ChiSummer) (increasing (Temp PlanetEarth)))

In all, TIMBER detects four contradictions between these explanations due to simultaneous increase and decrease in the Earth's temperature and simultaneous increase and decrease in Earth's distance to the sun. Artifacts are created for these four contradictions, incurring a total cost of 400.

TIMBER uses contradiction artifacts as triggers to change its preferred explanations and preferred domain knowledge. Since TIMBER will use preferred explanations and domain knowledge to explain new phenomena, revising these preferences also revises its usable beliefs. There is no guarantee that TIMBER will find a lower cost preference assignment, so these contradictions may persist across explanations and credible domain knowledge indefinitely.

TIMBER's belief revision procedure is called ***restructure-around-artifact***, shown in Fig. 7. Given an artifact (e.g., contradiction), it attempts to reduce cost by removing one or more of the beliefs supporting the artifact. For the Angela example, the procedure identifies domain knowledge supporting the contradiction, including the two **cotemporal** statements, and it identifies explanandums whose explanations support the contradiction, including Chicago's seasonal change and Australia's seasonal change. For each supporting belief in preferred domain knowledge, it computes whether revoking the belief's preferred status will lower the overall cost. Revoking preference for **(cotemporal ChiSummer AusWinter)** will remove all four contradictions for a cost reduction of 400, but it would lose the credibility benefit for a cost increase of 1000, so this is not desirable. The same is true of revoking preference for **(cotemporal ChiWinter AusSummer)**.

For each supporting phenomenon, TIMBER recomputes the lowest cost explanation. For example, changing Chicago's near-far explanation to the *facing* explanation described above removes preference for the beliefs that the Earth's temperature and the Earth–sun distance changes during Chicago's seasonal intervals. The facing explanation was not initially the lowest-cost explanation for Chicago's seasons, but these contradictions have since made the two near-far explanations much more costly.

When TIMBER changes its preferred explanation for Chicago's seasons to the facing explanation, it disables all four contradictions. TIMBER still processes the final explanandum (i.e., Australia's seasonal change) that initially supported the contradictions, and it computes a cost reduction by revising Australia's seasons to a facing explanation, because using the same model fragments, model fragment instances, and assumptions as Chicago's newly preferred explanation (i.e., the facing model) is less expensive. TIMBER then iterates through the same domain knowledge and explanandums again to compute additional revisions that reduce cost. Finding none, the belief revision procedure

Locally restructuring the Knowledge Base

function ***restructure-around-artifact*** **(artifact** $a = \langle t_a, B_a \rangle$**)**
// Find supporting ⟨proposition, explanation⟩ mappings.
let $M_a = \{\langle m, \langle J, B, M \rangle\rangle \in \mathbb{E}: (B_a \cap B) \neq \emptyset\}$
// Find supporting beliefs in the domain theory.
let $D_a = \mathbb{D}_a \cap B_a$
// Iterate until no further local revisions are made.
let $revised = true$
while *revised***:**
set *revised = false*
// Attempt to find a new explanation for proposition m_a
for each m_a **in** M_a:
// Find existing well-founded explanations for *m*.
let $X = \{\langle J, B, M \rangle \in \mathbb{X}: m_a \in M\}$
// Find the least cost explanation.
let $x = \min_{x \in X} \boldsymbol{explanation\text{-}cost}(x)$
// Make the least cost explanation the best explanation, if not already.
if $\langle m_a, x \rangle \notin \mathbb{E}$ **then**:
replace $\langle m_a, * \rangle$ **with** $\langle m_a, x \rangle$ **in** $\mathbb{E}$
set *revised = true*
// Attempt to remove beliefs from the domain theory.
for each *d* **in** D_a:
// If this belief can be retracted to reduce cost, retract it.
if ***retraction-savings***(*d*) > 0 **then**
// Remove *d* from adopted beliefs.
set $\mathbb{D}_a = \mathbb{D}_a - d$
set *revised = true*

Fig. 7. Algorithm for restructuring knowledge based on the presence of a high-cost artifact.

terminates. The procedure is guaranteed to converge because it only performs belief revision if cost can be reduced, and cost cannot be reduced infinitely. Restructuring is a greedy algorithm, so it is not guaranteed to find the optimal cost configuration of explanation preferences.

After TIMBER's belief revision in the Angela example, Chicago's and Australia's seasons are both explained by the facing model. The influence graph for both preferred explanations is shown in Fig. 8. Both explanations use **RotatingToward** and **RotatingAway** processes to explain change in temperature, the rates of which are qualitatively proportional to the rate of the Earth's rotation.

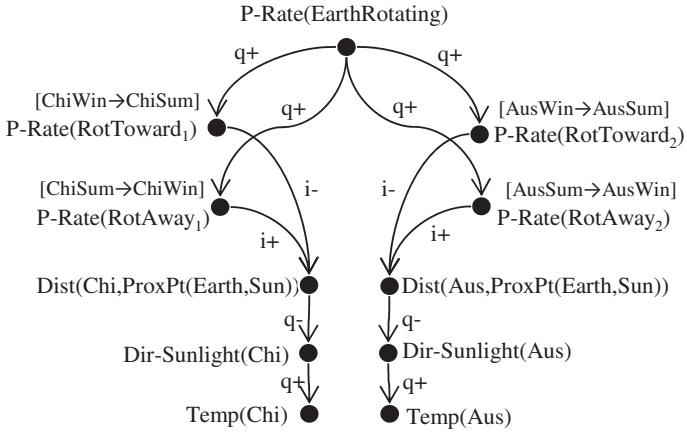

Fig. 8. An influence diagram of the facing explanation of both Chicago's (Chi) and Australia's (Aus) seasons.

Like the student Angela, TIMBER begins by explaining the seasons with a near-far explanation and ends the trial with a facing explanation. As described below, we simulate five of the students from Sherin et al.'s study, including Angela.

4. Simulation

We next compare simulation results of TIMBER against the interview transcripts of students in Sherin et al.'s (2012) study, describing our methodology for encoding student knowledge, our experimental setup, and our results.

For computationally validating AC theory and modeling students' commonsense science reasoning, we are interested in TIMBER's ability to (a) explain the changing of the seasons using causal models similar to the interviewed students, (b) identify inconsistencies in its causal models when given information from the interview about Chicago's and Australia's seasons, and (c) revise its causal models and explanations when given new information in similar ways as the interviewed students.

4.1. Method

The experimenters in Sherin et al. (2012) provided us with a spreadsheet describing domain-relevant beliefs and mental model components of the interviewed students. This included the following types of knowledge: topic-specific propositions (e.g., "*the sun is a source of heat and light*"); general causal schemas (e.g., "*closer/farther from a source provides greater/less effect of the source*" and "*facing/not-facing a source provides greater/less effect of the source*"); mental models (e.g., "*Earth orbits sun in an ellipse*" and "*Earth rotates*"); and ordinal or qualitative

beliefs (e.g., "*A is warmer than B*," "*A is far from the equator*," and "*A is warm*"). Sherin et al. described these knowledge components in English, so we constructed the corresponding formal knowledge representations by hand: We use model fragments to represent knowledge about continuous changes, Horn clauses to represent if-then knowledge, and predicate calculus statements to represent propositional beliefs. By using the OpenCyc ontology whenever possible, we reduce tailorability in our representations.

We implemented TIMBER using the Companion cognitive architecture (Forbus, Klenk, & Hinrichs, 2009). In each trial, TIMBER starts with hand-coded domain knowledge pertaining to one or more students from Sherin et al., but no explanations have been constructed. In terms of Fig. 3, the starting state of TIMBER consists only of the domain knowledge nodes on the bottom tier of the network. We first ask TIMBER to explain Chicago's and Australia's seasons, then we provide TIMBER with the opposite seasons information, and finally, we ask it to explain the seasons again, to assess how and whether it revised its explanations.

The individual differences in the students within the interviews involve more than just variations in domain knowledge. Some students strongly associate some beliefs with the seasons, for example, that the Earth's axis is tilted, without knowing the exact mechanism. We model gaps in mechanism knowledge by excluding model fragments. Other students start with a preference for one mechanism and then they transition to another. We model these initial explanatory biases, for example, in the Deidra and Angela trial below, using a priori credibility artifacts.

4.2. Results

Here, we describe the results of TIMBER's simulation of student interviews from Sherin et al. (2012) according to the methods described above. We describe specifics about inputs and results of each student group in each trial below.

4.2.1. Ali and Kurt trial

To simulate these two students, TIMBER's domain knowledge includes (a) the Earth rotates on a tilted axis, (b) temperature is qualitatively proportional to sunlight, and (c) the Earth orbits the sun. However, there is no knowledge that each hemisphere is tilted *toward* and *away* during the orbit. Consequently, TIMBER computes nine explanations for both Chicago and Australia, and computes a preference for the facing explanations shown in Fig. 8, with a cost of 56. This explanation is consistent with the opposite seasons information, so no revision occurs as a result: Like Ali and Kurt, TIMBER starts and ends with the facing explanation.

4.2.2. Deidra and Angela trial

TIMBER's initial domain knowledge includes (a) the Earth rotates, (b) the Earth orbits the sun and is sometimes closer and sometimes farther, and (c) amount of sunlight and distance from the sun both affect temperature. To model Deidra and Angela's preference

for the distance-based explanation, we used a credibility assertion on TIMBER's model fragments **Approaching-PeriodicPath** and **Retreating-PeriodicPath**, since Angela mentions that she learned about Earth's orbit behavior in second grade (see online supplementary material). Under these parameter settings, TIMBER constructs 16 explanations[9] and computes a preference for the near-far explanations graphed in Fig. 6, with a cost of 56. TIMBER also created facing explanations graphed in Fig. 8 with a cost of 66. The credibility artifact makes the near-far explanation preferable to the facing explanation. When confronted with the opposite seasons information, TIMBER, like the students Deidra and Angela, detects contradictions and revises its preferred explanation from the near-far explanations to the facing explanations.

4.2.3. Amanda trial

In Sherin et al.'s interview, Amanda mentions two influences on Chicago's temperature: (a) the distance to the sun due to the tilt of the Earth, and (b) the amount of sunlight, also due to the tilt of the Earth. By the end of the interview, she settles on the latter; however, she could not identify the *mechanism* by which the tilt changes throughout the year.

TIMBER uses the following domain knowledge to simulate Amanda: (a) the Earth rotates on a tilted axis, (b) when a hemisphere is tilted toward the sun, it receives more sunlight and is closer to the sun, and (c) sunlight and distance to the sun both affect temperature. TIMBER produced the two explanations in Fig. 9a (i.e., tilt affects temperature via distance from the sun) and Fig. 9b (i.e., tilt affects temperature via direct sunlight), where neither explanation includes a causal mechanism affecting the Earth's tilt. Fig. 9b was the final model that TIMBER—and Amanda—chose as a final explanation.

4.2.4. Amanda (correct explanation) trial

To demonstrate TIMBER's ability to generate a scientifically correct explanation, we repeated the Amanda trial with additional model fragments: (a) **TiltingToward**: a hemisphere of the Earth tilts toward the sun due to orbit along a tilted axis; and (b) **TiltingAway**: a hemisphere tilts away due to orbit along a tilted axis. TIMBER produced the graphs in Fig. 9c,d. The explanation in Fig. 9d is a simplified, scientifically correct model of seasonal change.

We have shown that TIMBER is able to (a) construct student explanations from Sherin et al. and (b) alter its preferred explanations similar to the way students did when confronted with an inconsistency. Furthermore, in the Amanda trial, we provided additional process models to demonstrate that TIMBER can construct a simplified correct explanation.

5. Related work

Like TIMBER, other cognitive systems extend and revise their knowledge by constructing or evaluating explanations. We discuss several lines of related research.

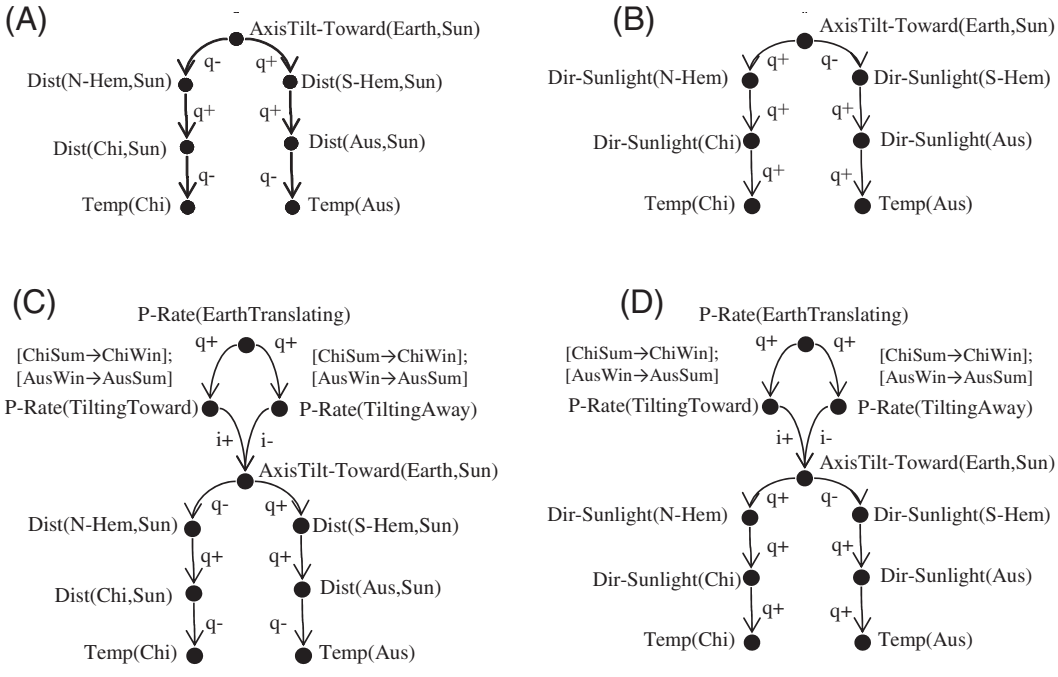

Fig. 9. Influence graphs for additional explanations produced by the simulation: (a) and (c) describe the tilt of the axis affecting each hemisphere's distance to the sun, thereby affecting temperature; (b) and (d) describe the tilt of the axis affecting direct sunlight to each hemisphere, affecting temperature. Explanations (c) and (d) contain a mechanism for the Earth's tilt, and (d) is a simplified, correct explanation of the seasons.

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, TIMBER generates its own justification structure from fragmentary domain knowledge, and then evaluates it along several dimensions via metareasoning.

Other systems construct explanations using abduction. For example, the system described in Molineaux, Kuter, and Klenk (2011) diagnoses failure through abductive explanation. Abduction increases the system's knowledge of hidden variables, and consequently it improves the performance of planning in partially observable environments. Similarly, ACCEL (Ng & Mooney, 1992) creates multiple explanations via abduction and then uses simplicity and set-coverage metrics to determine which is best. When diagnosing 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. TIMBER's explanation evaluation is more complex: It can assume any model fragment condition; some assumptions (e.g., quantity changes) are more expensive than others; and additional artifacts (e.g., model fragment types and instances) are penalized within explanations. In cost-based abduction (e.g., Charniak & Shimony, 1990; Santos, 1994) the goal is to find a least-cost proof (LCP) where each assumption has a weighted cost. Unlike traditional cost-based abduction, TIMBER's cost function evaluates an explanation's contents with respect to preferred knowledge *outside* of the explanation, allowing it to shift explanations and revise beliefs, thus changing its evaluation of future explanations.

Recent research focuses on evaluating explanations with respect to properties not implemented in TIMBER's cost function. For example, Licato, Sun, and Bringsjord (2014) describe methods for specifying and tuning a cognitive architecture's explanation preferences, for example, to prefer intentional explanations or mechanism-based explanations. Additional research has characterized the role of analogy and similarity for generating and evaluating explanations (e.g., Gentner & Markman, 1997; Hummel, Licato, & Bringsjord, 2014), including in scientific domains (e.g., Gentner et al., 1997; Nersessian, 2010; Thagard & Litt, 2012). Our TIMBER simulation of Sherin et al.'s student interviews does not involve analogy, but we have used analogy with TIMBER to transfer knowledge across domains and promote conceptual change (Friedman et al., 2012). This models people's tendency to retain highly contextualized misconceptions (Collins & Gentner, 1987; diSessa et al., 2004) despite also knowing scientifically correct models (Clement, 1982).

Creating and revising explanations is part of the larger cognitive process of conceptual change. INTHELEX (Esposito, Semeraro, Fanizzi, & Ferilli, 2000) is an incremental theory revision program that has modeled conceptual change as supervised learning. INTHELEX uses Datalog clauses as its knowledge representation, which is insufficient for explaining the behavior of dynamic systems, such as the seasonal change study 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 new and existing knowledge.

Learning by explaining 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 single rules (Laird, Newell, & Rosenbloom, 1987). Chunking speeds up future reasoning by avoiding operations when a chunked rule exists, but it does not alter the deductive closure or preferences over domain knowledge, so it cannot simulate belief revision.

Research in AI and philosophy has produced logical criteria for belief revision in response to observations. Alchourròn, Gärdenfors, and Makinson (1985) describe postulates of rational revision operations for a deductively closed propositional knowledge base, and Katsuno and Mendelzon's (1991) theorem equates these postulates to a revision mechanism based on total pre-orders over prospective knowledge bases. Unlike these globally consistent approaches, TIMBER does not maintain a globally consistent, deductively closed knowledge base; instead, TIMBER uses soft constraints and a greedy restructuring algorithm to attempt to approach consistency. This helps TIMBER model human-like

reasoning: Competing explanations and models may be simultaneously entertained and compared, but it has a bias for coherence within and across explanations and the mechanism-based models they contain.

6. Discussion

This paper presented the AC theory of how people represent, assemble, run, and revise mental models in commonsense science domains. We also described the TIMBER computational model of the theory and we presented empirical results of TIMBER simulating how students construct and revise explanations. Our simulation results provide empirical evidence of the following:

1. TIMBER's knowledge representations, including compositional qualitative models, can simulate students' knowledge in this commonsense science domain.
2. Fragmentary, globally inconsistent knowledge can be assembled into coherent models and then evaluated and manipulated as larger constructs.
3. When new information disturbs consistency or coherence, fragmentary domain knowledge can be reassembled into more preferred, globally coherent constructs.

These results suggest that AC theory and TIMBER provide a plausible computational account of how people's reason with—and repair—their mental models in commonsense science domains. 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.

We have previously used TIMBER to simulate students' reasoning and conceptual change when learning about the day-night cycle (Friedman et al., 2012), and we have used subsets of TIMBER to simulate students learning biology (Friedman & Forbus, 2011) and physics (Friedman & Forbus, 2010). This provides evidence that the representations and algorithms of TIMBER apply generally to commonsense science domains. We next revisit the key ideas of AC theory, and we discuss some opportunities for future work in making more adaptable cognitive models and autonomous learning systems using the methods in TIMBER.

6.1. Compositional coherence: In theory and in implementation

Unlike most other theories of conceptual change and mental models, AC theory has a working computational model capable of simulating students in commonsense science settings. Like any cognitive model, some properties of the computer model (i.e., TIMBER) do not reflect commitments of the theory (AC theory). We discuss these commitments here.

From a knowledge representation perspective, AC theory hypothesizes that people use qualitative, relational, and symbolic representations to describe space, time, entities, and processes. People reuse these symbolic relational structures individually (e.g., TIMBER's beliefs) or as operationalized aggregates (e.g., TIMBER's model fragments) to build and

rebuild mental models to explain and predict phenomena in the world. TIMBER uses the OpenCyc ontology, but other ontologies that are rich enough to support qualitative representations would work as well. AC theory is not committed to any *specific* predicate calculus statements or model fragments existing in a learner (cf. diSessa, 1993); rather, we reused relational beliefs and model fragments across TIMBER's simulation trials of different students based on available data.

From an algorithmic perspective, TIMBER models the following key cognitive processes of AC theory: (a) assembling fragmentary knowledge into coherent explanatory aggregates; (b) evaluating competing rationale to identify a preferred explanation for a phenomenon; (c) incorporating new knowledge via observation or instruction; and (d) revising preferences to increase global coherence. We discuss some of these processes in TIMBER that are less cognitively plausible, and hence targets for future work.

At present, we believe that TIMBER is doing much more computation than people to construct the same explanations. For example, TIMBER computed and evaluated 36 explanations in the Deidra and Angela trial. People probably use a more incremental approach to explanation construction, where they interleave meta-level analysis within their model formulation strategies. Using a narrower back-chaining algorithm in TIMBER would avoid reifying explanations (such as x_{2-4} described above) that are known to be structurally problematic or incomplete. In previous work (e.g., Friedman & Forbus, 2010, 2011), we demonstrate TIMBER using a model of similarity-based retrieval (Forbus, Gentner, & Law, 1995) to find similar problems and reuse the model fragments and beliefs from the corresponding preferred explanation(s). This further prevents TIMBER from generating a cognitively implausible and intractable number of explanations.

We see value in adding heuristic search to selectively reify explanations. For instance, Thagard (2007) suggests that explanations of greater depth (i.e., deeper justification structure to serve as rationale) have higher likelihood of correctness over time, so TIMBER might utilize a depth heuristic within a beam search through the abductive explanation space. Whether a depth-biased beam search will more accurately model human explanation construction is an empirical question.

TIMBER evaluates explanations with respect to its domain knowledge and other preferred explanations, using a numerical relative cost function. This causes TIMBER to favor explanations that cohere with preferred knowledge, thus modeling the reuse bias of AC theory. As noted earlier, the numerical costs were empirically derived and we do not believe the list of artifacts is complete. Since costs are expressed declaratively in the model, they might be learnable and/or adaptable over time. TIMBER does not currently simulate anomaly response strategies (Chinn & Brewer, 1998) or the development of metacognitive awareness thereof (Vosniadou, 2007).

We believe that learning by instruction involves reflecting on how the new information coheres or conflicts with existing knowledge. At present, TIMBER incorporates information (e.g., that Chicago and Australia experience opposite seasons), by adding it to domain knowledge and then using declarative heuristics to instantiate epistemic artifacts (e.g., for credibility and contradictions). We do not believe that people similarly identify all inconsistencies against new information, but Sherin et al.'s results

suggest that the students can easily detect when new information contradicts a recently used mental model.

TIMBER formulates mental model repair as a constrained optimization problem: it repairs its knowledge (i.e., revises epistemic preferences) to reduce cost (i.e., contradictions and complexity). Its reconstruction algorithm has the following computational consequences:

- The greedy algorithm biases TIMBER to retain as many existing preferences as possible, all else being equal. This conservatism in belief revision is not a new idea; it has been observed in both students and scientists (Chinn & Brewer, 1993) and has been proposed in the philosophical belief revision literature (e.g., Alchourròn et al., 1985; Doyle, 1991).
- By making a series of local cost reductions, TIMBER's reconstruction is an any-time, incremental, amortizable algorithm: It can partially reduce cost, stop, and resume restructuring later, retaining stability throughout (i.e., each phenomenon still has a preferred explanation). Incrementality is an important property, since human conceptual change is a prolonged process (Carey, 2009; Gentner et al., 1997).
- Contradictions are allowed in the knowledge base (i.e., they do not strictly prevent the adoption of new beliefs), but they are early targets for the restructuring algorithm, as demonstrated in our simulations.
- The conservative, incremental behavior of the restructuring algorithm helps TIMBER maintain tractability as it accrues knowledge and explanations over time. We believe these are important principles of models of human conceptual learning.

6.2. Social, emotional, and political considerations

Assembled coherence theory and its TIMBER computational implementation do not currently model social, emotional, and political considerations, aside from representing credibility of knowledge gained from other sources. Such factors are important for understanding cognition more broadly (Abelson, 1979), but we note that the cognitive science literature on conceptual change in science education, which we have focused on modeling, also ignores these factors.

Could AC theory and TIMBER be used to model conceptual change concerning emotionally charged topics? We note that qualitative models can be used to model at least some of the domains that are currently politically fraught (e.g., climate change), and they can be used more broadly in political reasoning (e.g., Forbus & Kuehne, 2005). Thagard and Findlay (2010a,b) use *emotional coherence* to explain people's difficulties in accepting new beliefs about climate change, evolution, and other emotionally charged topics, using emotional valence as a factor in belief revision. AC theory could potentially be extended to include emotional valence as epistemic artifacts with positive and negative costs, and thereby incorporated in the same conceptual change process. Valence for incoming information might be calculated by a version of appraisal theory (e.g., Wilson, Forbus, & McLure, 2013). We further note that existing

emotional coherence models do not themselves detect inconsistencies or construct explanations, and so something like TIMBER's model formulation algorithms could potentially provide new capabilities for such models.

6.3. Future work on TIMBER and AC theory

We see three lines of future work motivated by these results, and we discuss each in turn. First, we plan to explore TIMBER's capabilities in additional domains. In addition to expanding the catalog of epistemic artifacts, we can use TIMBER as a platform for modeling the effect of epistemic entrenchment (Alchourròn et al., 1985), level of specificity, source credibility, goal relevance, narrative structure (Lombrozo, 2011), individual differences in people's response to instruction (e.g., Feltovich, Coulson, & Spiro, 2001), and anomalous data (Chinn & Brewer, 1993).

Assembled coherence theory and TIMBER could potentially be applied within intelligent tutoring systems (ITS; e.g., Koedinger et al., 1997). ITSs automatically deliver customized feedback to a student based on his or her performance, using cognitive models of the domain and reasoning to understand what a student is doing, including qualitative models (e.g., de Koning et al. 2000). TIMBER could be used to find incoherence across a student's mental models and suggest examples that would help the student confront inconsistencies in his or her models.

Finally, TIMBER and AC theory provide architectural patterns for building more robust long-lived AI systems, since they abandon globally consistent knowledge stores in favor of a less constrained, highly contextualized knowledge organization strategy. In this framework, belief revision is the rule rather than the exception, and at any given time, the AI system may be attempting to increase coherence (i.e., reduce cost) in its causal models in multiple domains. This is especially relevant to AI systems that learn from reading, instruction, observation, and social interaction.

Acknowledgments

We thank anonymous reviewers from the Qualitative Reasoning Workshop and the AAAI Advances in Cognitive Systems Symposium for helpful comments on earlier manuscripts. This work was funded by the Northwestern University Cognitive Science Advanced Graduate Fellowship, the Socio-cognitive Architectures Program of the Office of Naval Research, and SIFT.

Notes

1. Greedy algorithms are used to find good approximate solutions when finding an optimal solution would be computationally intractable, for example, require exponential computational resources. In exchange for not guaranteeing optimality,

greedy algorithms operate in polynomial time, which is important for cognitive modeling.

2. The simulation in Friedman and Forbus (2011) was able to explain 90% of the model changes of students by varying cost parameters in TIMBER's explanation evaluation process.
3. http://dev.cyc.com/ontology-development/.
4. Not her real name. All student names are pseudonyms, for privacy.
5. In OpenCyc, categories are modeled as *collections*, which can be thought of as the set of all things that satisfy that concept, although defined in a way that avoids the usual self-reference paradoxes of set theory, while preserving the intuitive semantics (i.e., one can think of the `isa` relation as indicating that something is a member of a category, and category inheritance as if it were a subset relationship).
6. In a preferential pre-order, some elements may be equally preferred (=), or of equal or greater preference (≥). TIMBER breaks ties by favoring earlier explanations, all else being equal.
7. TIMBER can also justify other types of propositions and entities, including events and processes (Friedman & Forbus, 2010, 2011), but this capability is not relevant for this simulation.
8. TIMBER might explicitly assume that an unknown, active, process is directly influencing the quantity, but such an assumption is still objectively undesirable within an explanation.
9. The increased number of explanations is due to the belief that proximity in addition to amount of sunlight affect temperature.

References

Abelson, R. P. (1979). Differences between belief and knowledge systems. *Cognitive Science*, *3*(4), 355–366.

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

Brown, D. E., & Clement, J. (1989). Overcoming misconceptions by analogical reasoning: Abstract transfer versus explanatory model construction. *Instructional Science*, *18*, 237–261.

Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: MIT Press.

Carey, S. (2009). *The origin of concepts*. New York, NY: Oxford University Press.

Charniak, E., & Shimony, S. E. (1990). Probabilistic semantics for cost based abduction. In *Proceedings of AAAI National Conference on Artificial Intelligence* (pp. 446–451).

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

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* (pp. 161–238). Hillsdale, NJ: Lawrence Erlbaum Associates.

Chi, M. (2008). Three types of conceptual change: Belief revision, mental model transformation, and categorical shift. In S. Vosniadou (Ed.), *Handbook of research on conceptual change* (pp. 61–82). Hillsdale, NJ: Erlbaum.

Chinn, C., & Brewer, W. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. *Review of Educational Research*, *63*(1), 1–49.

Chinn, C., & Brewer, W. (1998). An empirical test of a taxonomy of responses to anomalous data in science. *Journal of Research in Science Teaching*, *35*(6), 623–654.

Clement, J. (1982). Students' preconceptions in introductory mechanics. *American Journal of Physics*, *50*(1), 66–71.

Collins, A., & Gentner, D. (1987). How people construct mental models. In D. Holland & N. Quinn (Eds.), *Cultural models in language and thought* (pp. 243–265). Cambridge, UK: Cambridge University Press.

DeJong, G. (Ed.) (1993). *Investigating explanation-based learning*. Norwell, MA: Kluwer Academic Publishers.

de Kleer, J., & Brown, J. S. (1984). A qualitative physics based on confluences. *Artificial Intelligence*, *24*(1–3), 7–83.

de Koning, K., Bredweg, B., Breuker, J., & Wielinga, B. (2000). Model-based reasoning about learner behavior. *Artificial Intelligence*, *117*(2), 173–229.

diSessa, A. (1993). Towards an epistemology of physics. *Cognition and Instruction*, *10*(2–3), 105–225.

diSessa, A., Gillespie, N., & Esterly, J. (2004). Coherence versus fragmentation in the development of the concept of force. *Cognitive Science*, *28*, 843–900.

diSessa, A. A., & Sherin, B. L. (1998). What changes in conceptual change? *International Journal of Science Education*, *20*(10), 1155–1191.

Doyle, J. (1991). Rational belief revision. In J. Allen, R. Fikes and E. Sandewall (eds.), *Principles of knowledge representation and reasoning* (pp. 163–174). Los Altos, CA: Morgan Kaufmann.

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 & P. Mello (Eds.), *AI*IA99: Advances in artificial intelligence, Lecture Notes in artificial intelligence 1792* (pp. 214–225). Berlin: Springer.

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. (2010). Modeling amidst the microtheories. In *Proceedings of the 24th International Workshop on Qualitative Reasoning* (pp. 112–115). Portland, OR.

Forbus, K. D., Gentner, D., & Law, K. (1995). MAC/FAC: A model of similarity-based retrieval. *Cognitive Science*, *19*(2), 141–205.

Forbus, K., & de Kleer, J. (1993). *Building problem solvers*. Cambridge, MA: MIT Press.

Forbus, K., & Gentner, D. (1997). Qualitative mental models: Simulations or memories? Proceedings of the Eleventh International Workshop on Qualitative Reasoning, Cortona, Italy, June 3–6, pp. 97–104.

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

Forbus, K., & Kuehne, S. (2005). Towards a qualitative model of everyday political reasoning. Proceedings of the 19th International Qualitative Reasoning Workshop, Graz, Austria, May.

Friedman, S. (2012). Computational conceptual change: An explanation-based approach. Doctoral Dissertation. Northwestern University, Department of Electrical Engineering and Computer Science, Evanston, IL.

Friedman, S., Barbella, D., & Forbus, K. (2012). Revising domain knowledge with cross-domain analogy. *Advances in Cognitive Systems*, *2*, 13–24.

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

Friedman, S., & Forbus, K. (2011). Repairing incorrect knowledge with model formulation and metareasoning. Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Spain.

Gentner, D., Brem, S., Ferguson, R. W., Markman, A. B., Levidow, B. B., Wolff, P., & Forbus, K. (1997). Analogical reasoning and conceptual change: A case study of Johannes Kepler. *The Journal of the Learning Sciences*, *6*(1), 3–40.

Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist*, *52*(1), 45.

Hayes, P. J. (1978). The naïve physics manifesto. In D. Michie (Ed.), *Expert systems in the micro-electronic age* (pp. 242–270). Edinburgh, UK: Edinburgh University Press.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher*, *30*, 141–158.

Hummel, J. E., Licato, J., & Bringsjord, S. (2014). Analogy, explanation, and proof. *Frontiers in Human Neuroscience*, *8*, 867.

Ioannides, C., & Vosniadou, S. (2002). The changing meanings of force. *Cognitive Science Quarterly*, *2*, 5–61.

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

Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education*, *8*(1), 30–43.

Kuipers, B. (1986). Qualitative simulation. *Artificial Intelligence*, *29*(3), 289–338.

Laird, J., Newell, A., & Rosenbloom, P. (1987). SOAR: An architecture for general intelligence. *Artificial Intelligence*, *33*(1), 1–64.

Lelliott, A., & Rollnick, M. (2010). Big ideas: A review of astronomy education research 1974–2008. *International Journal of Science Education*, *32*(13), 1771–1799.

Licato, J., Sun, R., & Bringsjord, S. (2014). Using meta-cognition for regulating explanatory quality through a cognitive architecture. Proceedings of the 2014 Workshop on Artificial Intelligence and Cognition, Torino, Italy, pp. 27–38.

Lombrozo, T. (2011). The instrumental value of explanations. *Philosophy Compass*, *6*(8), 539–551.

McCloskey, M. (1983). Naïve theories of motion. In D. Gentner & A. L. Stevens (Eds.), *Mental models* (pp. 299–323). Hillsdale, NJ: Lawrence Erlbaum.

Molineaux, M., Kuter, U., & Klenk, M. (2011). What just happened? Explaining the past in planning and execution. Proceedings of the Sixth International ExaCt Workshop. Barcelona, Spain, July 16–22.

Nersessian, N. (2010). *Creating scientific concepts*. Cambridge, MA: MIT Press.

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

Ohlsson, S. (2009). Resubsumption: A possible mechanism for conceptual change and belief revision. *Educational Psychologist*, *44*(1), 20–40.

Paritosh, P. K. (2004). Symbolizing quantity. In Proceedings of the 26th Annual Meeting of the Cognitive Science Society.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo, CA: Morgan Kaufmann.

Posner, G. J., Strike, K. A., Hewson, P. W., & Gertzog, W. A. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. *Science Education*, *66*(2), 211–227.

Rickel, J., & Porter, B. (1997). Automated modeling of complex systems to answer prediction questions. *Artificial Intelligence*, *93*(1–2), 201–260.

Santos, E. (1994). A linear constraint satisfaction approach to cost-based abduction. *Artificial Intelligence*, *65*, 1–27.

Sherin, B. L., Krakowski, M., & Lee, V. R. (2012). Some assembly required: How scientific explanations are constructed during clinical interviews. *Journal of Research in Science Teaching*, *49*(2), 166–198.

Smith, J. P., diSessa, A. A., & Roschelle, J. (1993). Misconceptions reconceived: A constructivist analysis of knowledge in transition. *Journal of the Learning Sciences*, *3*(2), 115–163.

Thagard, P. (2000). Probabilistic networks and explanatory coherence. *Cognitive Science Quarterly*, *1*, 93–116.

Thagard, P. (2007). Coherence, truth, and the development of scientific knowledge. *Philosophy of Science*, *74* (1), 28–47.

Thagard, P., & Findlay, S. (2010a). Changing minds about climate change: Belief revision, coherence, and emotion. In E. Olsson & S. Enquist (Eds.), *Belief revision meets philosophy of science* (pp. 329–345). Dordrecht, the Netherlands: Springer Netherlands.

Thagard, P., & Findlay, S. (2010b). Getting to Darwin: Obstacles to accepting evolution by natural selection. *Science & Education*, *19*(6–8), 625–636.

Thagard, P., & Litt, A. (2012). Models of scientific explanation. In P. Thagard (Ed.), *The cognitive science of science: Explanation, discovery, and conceptual change* (Ch. 3, pp. 25–46). Cambridge, MA: The MIT Press.

Trickett, S. B., & Trafton, G. (2007). "What if. . .": The use of conceptual simulations in scientific reasoning. *Cognitive Science*, *31*, 843–875.

Vosniadou, S. (2007). The conceptual change approach and its re-framing. In S. Vosniadou, A. Baltas, & X. Vamvakoussi, (Eds.) *Reframing the conceptual change approach in learning and instruction* (pp. 47–62). Oxford: Elsevier.

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

Vosniadou, S., & Brewer, W. F. (1994). Mental models of the day/night cycle. *Cognitive Science*, *18*, 123–183.

Vosniadou, S. (1994). Capturing and modeling the process of conceptual change. In S. Vosniadou (Guest Editor). *Special Issue on Conceptual Change, Learning and Instruction*, *4*, 45–69.

Vosniadou, S., Skopeliti, I., & Gerakaki, S. L. (2007). Understanding the role of analogies in restructuring pro-cesses. In A. Schwering, U. Krumnack, K. U. Kuhnberger, & H. Gust (Eds.), *Analogies: Integrating multiple cognitive abilities, publications of the institute of cognitive science* (vol. 5, pp. 39–43). Osnabruck, Germany.

Wilson, J., Forbus, K., & McLure, M. (2013). Am I really scared? A multi-phase computational model of emotion. *Proceedings of the 2nd Conference on Advances in Cognitive Systems*, *289*, 304.

Supporting Information

Additional Supporting Information may be found online in the supporting information tab for this article:

Table S1. Transcript of an interview about the seasons.