Cognitive Science 30 (2006) 95–120

A Model of Plausibility

Louise Connell[a], Mark T. Keane[b]

[a]*Cognition & Communication Research Centre, Division of Psychology, Northumbria University*
[b]*School of Computer Science and Informatics, University College Dublin*

Received 9 May 2004; received in revised form 26 August 2005; accepted 26 August 2005

Abstract

Plausibility has been implicated as playing a critical role in many cognitive phenomena from comprehension to problem solving. Yet, across cognitive science, plausibility is usually treated as an operationalized variable or metric rather than being explained or studied in itself. This article describes a new cognitive model of plausibility, the Plausibility Analysis Model (PAM), which is aimed at modeling human plausibility judgment. This model uses commonsense knowledge of concept–coherence to determine the degree of plausibility of a target scenario. In essence, a highly plausible scenario is one that fits prior knowledge well: with many different sources of corroboration, without complexity of explanation, and with minimal conjecture. A detailed simulation of empirical plausibility findings is reported, which shows a close correspondence between the model and human judgments. In addition, a sensitivity analysis demonstrates that PAM is robust in its operations.

Keywords: Psychology; Cognition; Reasoning; Plausibility; Computer simulation; Symbolic computational modeling

1. Introduction

Every day, in many different scenarios, we judge the plausibility of things, whether we are reflecting on the plot quality of the latest disaster movie or listening to a child claim that the cat left those muddy boot prints on the floor. The pervasiveness of plausibility is reflected in the many different cognitive contexts in which it has been studied. In memory research, plausibility is used as a kind of cognitive shortcut in place of direct retrieval from long-term memory, especially when verbatim memory has faded (e.g., Reder, 1982; Reder & Ross, 1983; Reder, Wible, & Martin, 1986). In comprehension, it has been proposed to speed the interpretation of ambiguous sentences (Pickering & Traxler, 1998; Speer & Clifton, 1998; Traxler & Pickering, 1996) and constrain the understanding of novel compounds (Costello & Keane, 2000, 2001).

Correspondence should be addressed to Louise Connell, Cognition & Communication Research Centre, Division of Psychology, Northumbria University, Newcastle upon Tyne, NE1 8ST, United Kingdom. E-mail: louise.connell@northumbria.ac.uk

In thinking, it has been shown to support commonsense reasoning (Collins & Michalski, 1989), induction (Smith, Shafir, & Osherson, 1993), and the solution of arithmetic problem solving (e.g., Lemaire & Fayol, 1995). However, this very pervasiveness seems to have made plausibility harder to explain and to model, as it is typically treated in the literature as merely an operationalized variable (i.e., ratings of "goodness" or plausibility) or as an underspecified subcomponent of some other phenomenon.

The typical treatment of plausibility across cognitive science has meant that empirical research on plausibility is somewhat fragmented and computational models of plausibility are rare. However, one common thread runs throughout the literature, namely, the shared assumption that plausibility judgment involves some assessment of *concept–coherence;* that is, how well a particular scenario conceptually coheres with prior knowledge (e.g., Collins & Michalski, 1989; Connell & Keane, 2004; Johnson-Laird, 1983; Reder, 1982). In this article, we review the notion of concept–coherence in plausibility and describe previous attempts to capture plausibility computationally. We then present a cognitive model of human plausibility judgment, the Plausibility Analysis Model (PAM), and describe how it evaluates plausibility by determining the concept–coherence of a scenario. This computational model is evaluated by comparing it to human data in a detailed simulation, showing how PAM parallels people's judgments of plausibility. A sensitivity analysis of PAM is then described in the following section, and we discuss how its parameters are both necessary and cognitively motivated in modeling plausibility judgments. Finally, the model is discussed with respect to its wider implications for cognitive science.

2. Plausibility and concept–coherence

Although plausibility has not been well explained in the existing literature, there is a rough consensus that it has something to do with the coherence of concepts based on prior knowledge. This view holds that some concept, scenario, event, or discourse is plausible if it is conceptually consistent with what is known to have occurred in the past (e.g., Collins & Michalski, 1989; Johnson-Laird, 1983; Reder, 1982; Rehder, 2003a). For example, a small-winged creature that does not fly and yet still builds nests in trees might be considered a less plausible bird than a large-winged creature that does not fly and builds nests on the ground. According to Rehder (2003a, 2003b; see also Keil, 1989; Murphy & Medin, 1985), the plausibility of category membership can be viewed as a function of how well the object's features cohere with one another, according to prior knowledge of causal relations between category features. Thus, even though the latter creature has only one feature that is typical of birds (has wings), and the former has three (small size, has wings, nests in trees), the latter creature seems more plausible to us because the combination of nonflying and ground nesting is conceptually consistent with our prior knowledge of birds.

As well as plausibility of category membership, the concept–coherence approach can also be applied to plausibility of event scenarios. To judge plausibility by this account involves, first, drawing on relevant prior knowledge to make the necessary inferences and, second, somehow assessing if the scenario is a good match to what has been experienced in the past (either directly or vicariously). For example, if you were judging the plausibility of the scenario,

"The balloon landed on the pin and burst," you might make the inference that the pin *caused* the balloon to burst. Then you might judge the scenario to be plausible, because the phenomenon of balloons bursting when they hit sharp objects fits your past experience. In contrast, if you were judging the plausibility of the sentence, "The balloon landed on the pin and melted," it is difficult to connect the events with a suitable inference, though you could create a causal connection based on the balloon landing on an extremely hot pin, thus causing it to melt. In all likelihood, this scenario will be judged to be less plausible, because the inferred connection—involving some conjecture about the possible heat of the pin making the balloon melt—does not fit well with your past experience.

Black, Freeman, and Johnson-Laird (1986) examined this concept–coherence view of plausibility by showing that the plausibility–coherence of a story depended on suitable inferences being found between its sentences. They reordered the sentences in a story to disrupt causal dependencies, but held referential continuity constant, and they found that the judged plausibility of the story decreased as people's ability to make bridging inferences was disrupted. Indeed, other studies have shown that people monitor more than just causal continuity when reading and that they also track temporal, spatial, motivational, and other factors (Zwaan, Magliano, & Graesser, 1995; Zwaan & Radvansky, 1998).

Recently, Connell and Keane (2002, 2003, 2004) showed that different types of inference are reflected in differential plausibility ratings for sentence pairs describing simple events. For instance, they found that scenarios inviting causal inferences (such as the balloon-bursting scenario previously mentioned) were judged more plausible than those that failed to invite obvious causal inferences (such as the melting scenario previously mentioned). Furthermore, the causal scenarios were also found to be more plausible than sentence pairs that invited simple attributal inferences (e.g., Y specifies an attribute of X), which in turn were judged to be more plausible than inferences of temporal succession (Y happens after X). These studies provide specific concrete evidence that plausibility is influenced by the coherence of a situation, as shaped by the type of inferences made.

3. Capturing plausibility computationally

Plausibility has been used in theoretical and computational models across a wide variety of fields, such as reasoning (Collins & Michalski, 1989), conceptual combination (Costello & Keane, 2000; Lynott, Tagalakis, & Keane, 2004), and computational linguistics (Lapata, McDonald, & Keller, 1999). However, there is little consensus regarding the definition and use of plausibility, and in many cases, plausibility is simply implemented as an operationalised metric. For example, Collins and Michalski (1989) discussed plausible reasoning, but by this they merely meant reasoning based on inferences supported by prior experience; they did not characterize plausibility judgments per se. On the other hand, Friedman and Halpern (Friedman & Halpern, 1996; Halpern, 2001; see also, Shafer, 1976) created what they termed *plausibility measures,* but this is not intended to be a model of human plausibility judgment. Rather, the measures constitute a mathematical metric of uncertainty for use in fuzzy logic, of limited utility in modeling the psychology of plausibility.

Indeed, we know of only one computational model that deals directly with human plausibility judgments, the C^3 model of conceptual combination (Costello & Keane, 2000). In their constraint theory of conceptual combination, Costello and Keane (2000, 2001) identified plausibility as a key constraint in generating interpretations for novel noun–noun compounds. They argued that some interpretations of novel compounds were more acceptable than others by virtue of their plausibility: For example, in the novel compound *shovel bird,* the interpretation *"a bird which uses a shovel to dig for food"* is less acceptable than *"a bird with a flat beak it uses to dig for food"* for reasons of plausibility. The C^3 model, which instantiates constraint theory, computes plausibility as part of the process of generating interpretations by counting the features of the interpretation that overlap with stored concept instances. In the *shovel bird* example, the second interpretation receives a higher plausibility score because there are several stored instances of birds having beaks of a particular shape; in contrast, the first interpretation has a lower plausibility score because there are no stored instances of a bird using a tool. Thus, C^3 models plausibility as the degree to which the features of an interpretation overlap with prior knowledge. Although the C^3 model essentially adheres to the concept–coherence view of plausibility, it is narrowly focused on conceptual combination. Calculating the plausibility of concepts by counting feature overlap is quite different, both cognitively and computationally, from calculating the plausibility of discourse scenarios that describe events.

A computational model of plausibility that can calculate the concept–coherence of scenarios and thus capture complex plausibility judgments would be applicable to a wide range of cognitive tasks, and would impact cognitive science in fields from reasoning to discourse comprehension. To this end, we now present PAM.

4. The Plausibility Analysis Model (PAM)

When people must make a plausibility judgment, they examine how well a particular scenario conceptually coheres with what they know about the world. In other words, to judge the plausibility of a scenario, it must be mentally represented, assessed, and its concept–coherence determined. Theoretically, we view concept–coherence as being about consistency with previous experience, as measured by the degree of fit between a given scenario and prior knowledge. Hence, our theory is called the knowledge-fitting theory (see also, Connell, 2004; Connell & Keane, 2003).

In the knowledge-fitting theory, plausibility judgments involve two main processing stages: a comprehension stage and an assessment stage. During the *comprehension stage,* a mental representation of the presented scenario is created from the verbal description and from the inferences made using prior knowledge (e.g., Gernsbacher, 1990; Kintsch, 1998; McKoon & Ratcliff, 1992; Singer, Graesser, & Trabasso, 1994). For example, to properly comprehend the scenario, *"The bottle fell off the shelf. The bottle smashed,"* a person must represent the events themselves (the bottle falling and the bottle smashing) and also use prior knowledge to infer that the bottle's fall *caused* it to smash. In this scenario, prior knowledge relevant to making this inference may include that bottles are often fragile, that shelves are located at a height, that fragile things often break when they hit the ground, and so on. Once the mental representation has been formed, the comprehension stage is complete. The *assessment stage* then takes over,

whereupon this mental representation is examined to determine its fit to prior knowledge (i.e., its concept–coherence). The knowledge-fitting theory considers a highly plausible scenario to be one that fits well with prior knowledge, whereas an implausible scenario is one that fits poorly, if at all. For example, the previously mentioned *bottle* scenario seems fairly plausible because the fall-caused-smashing inference fits well with what we know of the world: that is, it is corroborated in many ways by prior knowledge without needing a complex explanation of the events based on conjecture.

In terms of a theoretical model, the knowledge-fitting theory views a plausible scenario as one that fits prior knowledge

1. *using many different sources of corroboration.* That is, the scenario should have several distinct pieces of prior knowledge supporting any necessary inferences.
2. *without complex explanation.* That is, the scenario must be represented without relying on extended or convoluted justifications.
3. *using minimal conjecture.* That is, the scenario must be represented by avoiding, where possible, the introduction of hypothetical entities (i.e., no deus ex machina).

The knowledge-fitting theory holds that plausibility results from the theoretical function shown in Fig. 1. In this function, the plausibility of a scenario will drop as its implausibility rises. A scenario will be perfectly plausible only if its representation has minimal *complexity* and *conjecture,* and/or maximal *corroboration.* In other words, as complexity increases, plausibility decreases. This result, however, is tempered by the corroboration of the scenario, as even a very complex scenario will become plausible if it is corroborated by prior knowledge. In addition, the interaction of complexity and corroboration is affected by conjecture, as conjecture will make even the simplest, best-supported scenario seem less plausible. As we shall see, this account of plausibility has the advantage of being based on empirical findings and of being realized as a computational model.

In terms of a computational model, PAM is a computational implementation of the knowledge-fitting theory as applied to the judgment of plausibility in discourse (i.e., descriptions of events and happenings in the world). Much of the empirical work on plausibility judgment has examined descriptions of simple events presented as pairs of sentences (Connell & Keane, 2002, 2004), so much of the focus of our present account will be on this type of discourse. However, we later discuss how PAM would deal with the plausibility of more extended discourse.

Operationally, PAM takes sentence-pair inputs (e.g., *"The bottle fell off the shelf. The glass smashed"*) and outputs a plausibility rating (from 0 to 10) for the scenario described in the sentences (see Fig. 2). As in the knowledge-fitting theory, the plausibility judgment process is

$$plausibility = 1 - implausibility$$

$$implausibility = \frac{complexity}{corroboration - conjecture}$$

Fig. 1. Theoretical plausibility function of the knowledge-fitting theory. Perfect plausibility is reduced as complexity or conjecture increases, but is restored as corroboration increases.

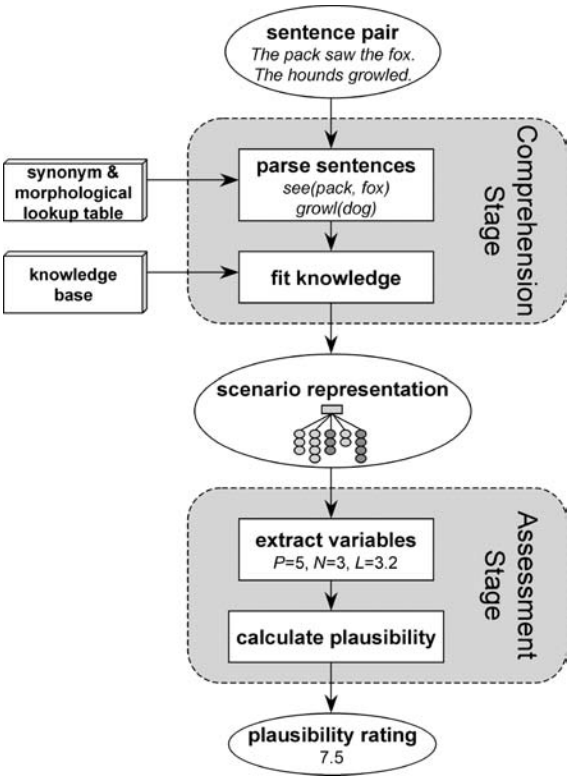

Fig. 2. Plausibility Analysis Model (PAM), showing comprehension and assessment stages.

marked in PAM by two main processing stages: a comprehension stage and an assessment stage. These stages are described in detail in the following sections.

4.1. The comprehension stage

The comprehension stage takes a sentence pair as input and outputs to the assessment stage a representation of the scenario described in those sentences. During comprehension, PAM parses the sentence pair into propositional form and makes appropriate inferences by fitting the scenario to relevant prior knowledge.

4.1.1. Parsing the sentence pair

To represent the scenario, PAM must break down each sentence into propositional form. First, each sentence is converted into a simplified form with the aid of a synonym and morphological lookup table. This process replaces words with their more common synonyms, singularizes plural nouns, and changes verbs into the present tense third-person singular form; for example, the sentences "the hounds growled" and "the dogs snarled" are both converted to the same simplified form "the dog growl." Next, the simplified sentences are parsed according to a set of basic grammatical rules and converted into propositions. To do this, PAM passes the

sentences through a simple syntactic parser that extracts verbs and adjectives to use as predicate names, and extracts nouns to use as arguments; for example, the simplified sentence "the dog growl" is converted into the proposition *growl*(*dog*). It should be noted that the ease with which PAM can break sentences into propositional form is due to the regular syntax of the sentence pairs. Although the automatic conversion of text into propositions is a nontrivial task (see Kintsch, 1998), this syntactic form of PAM's input lends itself quite well to automation.

4.1.2. Knowledge fitting

Once the sentences are in propositional form, PAM makes the inferences between the sentences by fitting their propositions to information in the knowledge base. PAM's knowledge base is organized as a predicate set, where each entity (noun) is defined as part of a type hierarchy, and each predicate (verb) is defined by the conditions of its constituent arguments in PAM's knowledge base.[1] For example, Table 1 shows how PAM's knowledge base structures the type hierarchy that gives rise to the definitions of the nouns *dog* and *fox.* In addition, Table 2 illustrates a simplified snapshot of PAM's knowledge base entry for the *growl* predicate, showing the various conditions that must be fulfilled for *growl*(*X*) to be true, including the argument conditions of any other predicates called. This predicate represents the idea that there are many conditions under which a thing *X* may growl—such as being in pain, being afraid of something, or being playful—but only some of these conditions will be fulfilled for particular values of *X.* Each predicate in the knowledge base (e.g., *growl, hunt, play*) was defined as broadly as possible, but the specific conditions for any given predicate are not critical to PAM's operation as the algorithm will operate over any set of knowledge represented in this style. In short, any knowledge base with this structure will be exploited by PAM's processes in the way described in the following.

To represent a particular scenario, PAM must check the conditions of each proposition as it is defined in the knowledge base. For example, the propositional form of the scenario "The pack saw the fox. The hounds growled." is *see*(*pack, fox*), *growl*(*dog*). In representing this scenario, PAM must first check the predicate *see* in the knowledge base to determine if its arguments meet the conditions specified. The *see* predicate requires that its first argument be an animal (i.e., something must be an animal to see). As the definition of *pack* shows that it contains *dogs,* and the type hierarchy for *dog* shows that it is an *animal,* the first condition of *see* is met. Also, the *see* predicate requires that its second argument must be a nonabstract entity (i.e., something must be

Table 1
PAM's knowledge base entry (simplified) for the type hierarchies of the entities *dog* and *fox*

Predicate	Conditions
entity (X)	→ animate(X)
animate (X)	→ animal(X)
animal (X)	→ predator(X)
animal (X)	→ prey(X)
predator (dog)	
predator (fox)	
prey (fox)	

Table 2
PAM's knowledge base entry (simplified) for the *growl* predicate, showing the conditions that must be fulfilled for *growl* (and any other predicate called) to be true for argument *X*

Predicate	Conditions			
growl(X)	→ animal(X)			
	inPain(X)	→ hurt(X)		
	→ animal(X)			
	growl(X, Y)	→ afraid(X, Y)	→ human(X)	
			phobia(X, Y)	
			→ hunt(Y, X)	→ predator(Y)
				prey(X)
		→ aggressive(X, Y)	→ hunt(X, Y)	→ predator(X)
				prey(Y)
			→ act(Y, X)	→ animal(Y)
		→ act(Y, X)	→ animal(Y)	
		→ not(human(X))		
		play(X, Y)	→ not(predator(X))	
			entity(Y)	
			not(prey(Y))	

nonabstract to be seen). Because the type hierarchy of *fox* shows that it is an animal and not an abstract entity, the second condition of the *see* predicate is met. The way in which each condition is met is listed, and if all conditions are fulfilled, PAM returns this list as a path.

When the first proposition has been represented, PAM moves on to processing the second proposition, *growl*(*dog*) and searches for ways to meet the conditions of the *growl* predicate. Fig. 3 shows the paths that PAM finds for this proposition; for example, the second path represents the ideas that the dogs are growling because they are growling at the fox, because they are hunting it, because dogs are predators and foxes are prey. Some of the conditions in the growl *predicate* lead to other predicates that have their own conditions attached, such as *hunt*(*dog*), which requires that *dog* must be a predator and that the *fox* of the first sentence must be prey. More often than not, there are several paths in the knowledge base that could be followed to fulfill the conditions of a particular predicate, and PAM will record all these alternative paths (shown in Fig. 3). Sometimes, a path may involve conjecture; that is, the path contains a condition that could only be fulfilled by assuming the existence of a hypothetical entity not explicitly mentioned. For example, the dogs may *growl* at something else other than the fox, but that would involve assuming the arrival on the scene of some other creature. PAM also records these hypothetical paths and marks them as such. This representation of alternative paths can be conceived of as PAM's way of modeling group behavior in plausibility judgment: Rather than limiting its operation to the representation to a single path that one individual may consider, PAM represents the set of paths that a group may consider and averages out the differences.

The scenario is therefore represented by PAM in the form shown in Fig. 3: consisting of several distinct paths, each of which consists of a set of one or more conditions. There is no hard-coded distinction between different types of inference (e.g., causal, temporal); PAM simply tries to build a path by drawing in whatever information is necessary to fulfill the condi-

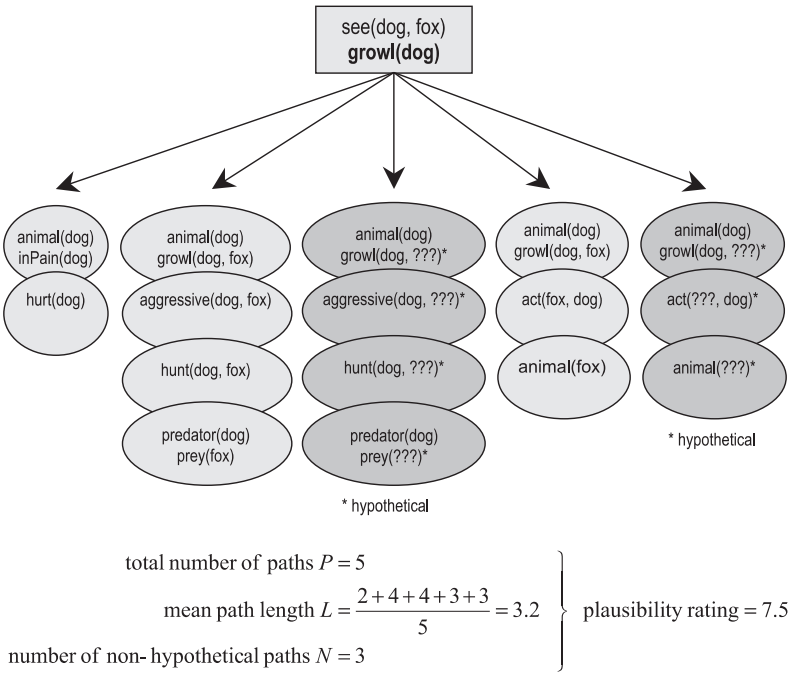

Fig. 3. Form of scenario representation created by PAM in the comprehension stage for the scenario, "The pack saw the fox. The hounds growled". It is then analyzed in the assessment stage to extract variables and determine plausibility (see sample values).

tions in the predicate. The structure of this representation is analyzed in the assessment stage to determine its concept–coherence and is used by PAM in calculating the plausibility rating for the sentence pair.

4.2. The assessment stage

Once a scenario has been comprehended, the representation is taken as input to the assessment stage, which outputs a plausibility judgment in the form of a rating between 0 (*not plausible*) and 10 (*completely plausible*). PAM's analysis extracts three main variables from the representation (see Fig. 4)[2] and uses them to calculate plausibility by applying a function that

$$plausibility\ rating = 10 \times (1 - implausibility)^2$$

$$implausibility = \frac{\frac{L}{L+1}}{P + \frac{N}{P}}$$

complexity

corroboration - conjecture

Fig. 4. PAM's formula for plausibility ratings (P = total number of paths, N = number of nonhypothetical paths, L = mean path length).

ascertains the quality of the knowledge fit (i.e., the scenario's concept–coherence). As shown in Fig. 4, the formula PAM uses to calculate plausibility relates directly to the theoretical plausibility function outlined in Fig. 1.

4.2.1. Computing the key components of the theory

The three main components of the theory (i.e., corroboration, complexity, and conjecture) have specific correlates in the key variables of the plausibility function used to assess the constructed representation.

1. Total number of paths (*P* capturing *corroboration*). This component is quantified as the number of different paths in the representation. It reflects the number of different ways the given sentence pair's predicate conditions can be met in the knowledge base.
2. Mean path length (*L* capturing *complexity*). This component is quantified as the sum of all path lengths in the representation (i.e., all conditions across all paths) divided by *P.* It reflects the average count of how many different conditions must be met per path.
3. Number of nonhypothetical paths (*N* capturing *conjecture*). This component is quantified as the number of paths whose conditions do not contain a hypothetical argument. It reflects the number of paths that could be constructed without needing to assume the existence of something not explicitly mentioned.

Each of these variables is motivated by the underlying theory, and contributes to plausibility in much the same way as its theoretical counterpart. For example, the mean path length *L* represents the *complexity* of the inferential connection, as complex inferences are considered less plausible than simple inferences. In addition, the total number of paths *P* and the number of nonhypothetical paths *N* are important to modeling the plausibility judgments of a group of people: Together, they represent the prior knowledge *corroboration* of the variety of ways in which the events in the scenario may be plausibly connected and the *conjecture* that lowers the value of such connections and makes the scenario less plausible.

4.2.2. Calculating plausibility

To arrive at a plausibility rating for the scenario, PAM uses the three variables previously mentioned to ascertain its concept–coherence. The asymptotic function in Fig. 4 returns a plausibility rating between 0 (*not plausible*) and 10 (*completely plausible*). The number of paths (*P*) ranges from [0, infinity], and high *P* values mean higher plausibility because there are more possible ways that the scenario can be represented. The mean path length (*L*) ranges from [1, infinity], and high *L* values mean lower plausibility because elaborate requirements must be met to represent the scenario. Finally, the number of nonhypothetical paths (*N*) ranges from [0, *P*], and high *N* values mean higher plausibility because the scenario can be represented without assuming the existence of entities that may not be present. As already mentioned, the form of PAM's plausibility function is asymptotic. The main reason for this functional form is to cause the rating formula to approach its asymptote of 10 as the corroboration variable *P* approaches its infinite upper limit. This effectively implements an implicit threshold on *P,* allowing low values to exert a strong influence on plausibility, but preventing high values from engulfing the effects of the other variables. For example, a scenario with $P = 20$ and one

with $P = 100$ are both well corroborated by prior knowledge and are both very plausible, but there should not be a large difference in their plausibility ratings. Although many functions could potentially implement the assertions about plausibility that underlie the knowledge-fitting theory, the asymptotic function shown in Fig. 4 was found to correspond particularly well with human judgments.

It is useful to illustrate the contribution that each variable makes to the plausibility rating function. This can be done by creating a three-dimensional space (one dimension for each variable) made up of a range of each variable's possible values and calculating the resulting plausibility rating for each combination of values. A sample of this space can be seen in Fig. 5, showing the plausibility ratings that PAM generates for an increasing number of paths (P) and for increasing path complexity (L). In addition, the best case (no paths hypothetical) and worst case (all paths hypothetical) values for the number of nonhypothetical paths N are shown as two separate surfaces. All plausibility ratings calculated by PAM fall into the range shown in Fig. 5. For example, a set of four (nonhypothetical) paths with a mean length of three will have a rating of 7.1 out of 10, whereas a set of three paths (again with a mean length of three) will have a rating of 6.5 out of 10. If one of those three paths were hypothetical, then the rating would drop to 6.1 out of 10.

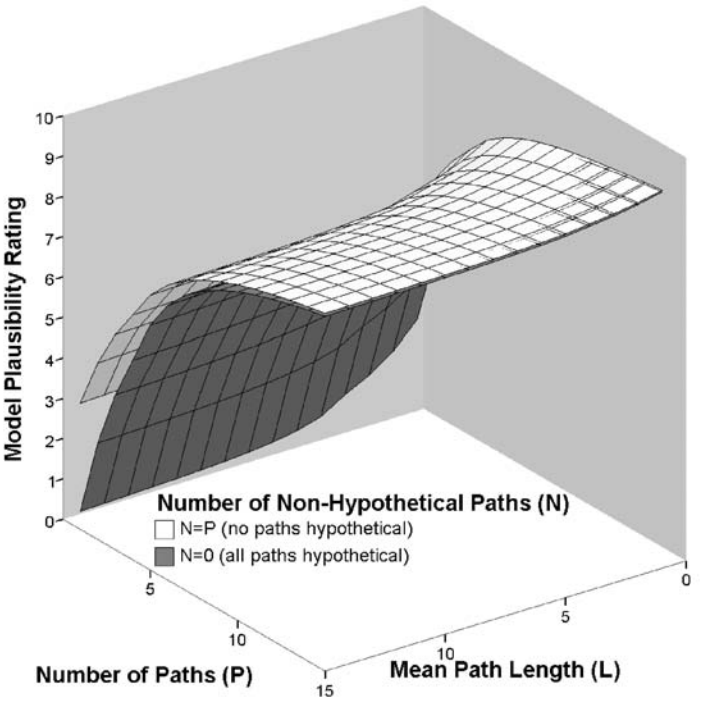

Fig. 5. Three-dimensional illustration of PAM's plausibility rating function for the variables P and L, with N's maximum and minimum values shown as separate surfaces. Note how the impact of L and N decreases as values for P increase.

4.3. Modeling different types of inference

PAM is designed to model and represent a number of different inference types, yet without making any presumptive hard-coded distinctions between them. In other words, because PAM returns every path that finds corroboration in its knowledge base, any given scenario may be represented with a diverse variety of causal, attributal, and temporal explanations. For example, Fig. 3 shows the paths returned for the scenario, "The pack saw the fox. The hounds growled." Although this scenario is correctly described as causal (Connell & Keane, 2004), not all explanations generated by people and not all paths returned by PAM are necessarily providing causal connections between the two events in the scenario. For example, to say that the hounds growled because they were in pain is actually a temporal explanation of events ("The hounds growled because they were in pain." merely happens in time after "The pack saw the fox"), but this is just as valid a way of representing the scenario as any causal explanation (e.g., the hounds growled at the fox because they were hunting it). Attributal inferences are dealt with in the same way as causal and temporal inferences. For example, take the attributal scenario, "The pack saw the fox. The hounds were fierce." Fig. 6 illustrates how this scenario is represented in PAM, showing how the knowledge base corroborates attributing ferocity to hounds. Again, a purely attributal explanation is possible—that the hounds were fierce because animals are sometimes just fierce—but also possible is the causal explanation that the hounds were fierce because they were hunting the fox. In this way, PAM can model scenarios that are preclassified as causal, attributal, temporal, or unrelated (i.e., no explanation exists), simply by fitting each scenario to prior knowledge in whatever way possible. In the following simulation, we examine how well PAM's plausibility ratings for different inference types correlate with human ratings for the same scenarios.

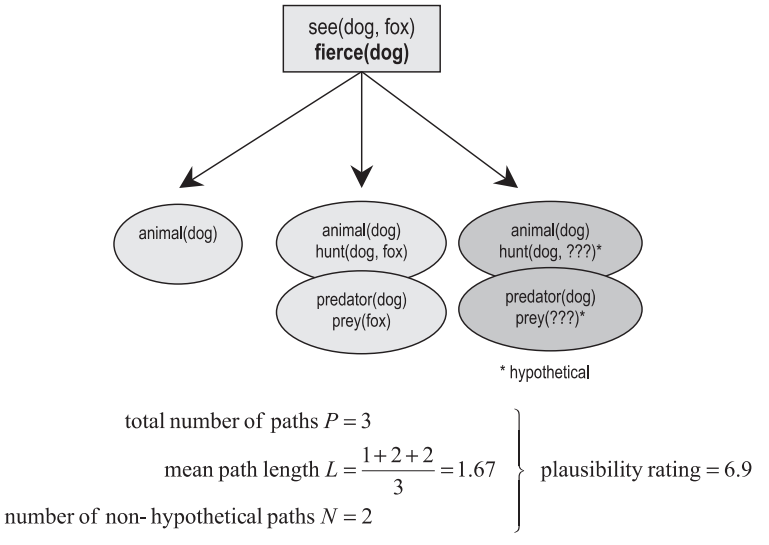

Fig. 6. Form of scenario representation created by PAM in the comprehension stage for the scenario, "The pack saw the fox. The hounds were fierce." It is then analyzed in the assessment stage to extract variables and determine plausibility (see sample values).

5. Simulation

To evaluate the model, we compared PAM's output to human responses in a simulation that examines PAM's plausibility ratings against the ratings produced by people in the experiments reported by Connell and Keane (2004). In this simulation, the model was run on the same sentence pairs presented to the human participants. As previously noted, the knowledge base used in PAM was built in a "blind" fashion; each of the individual predicates was defined using simple definitions of argument conditions, without checking possible path lengths that might emerge from combining these words in a sentence. Such knowledge bases will always be a crude approximation of the knowledge of a particular individual, but they should be closer to the aggregate knowledge that a group of participants bring to the task. The critical point was that the knowledge base was not modified in any iterative way to fit the data (see Costello & Keane, 2000, for a discussion of wider methodological issues). In addition, the test sentence pairs used in this simulation represented a different subset of materials to those used to test PAM's performance during the construction of the model, allowing us to test the generalizability of the model.

Connell and Keane's (2004) Experiment 1 manipulated the concept–coherence of their materials by creating scenarios with different inferential connections between their events (see Table 3). They found that people consider scenarios with causal connections between events (e.g., event *Y* was caused by event *X*) to be the most plausible, followed by events linked by the assertion of a previous entity's attribute (e.g., proposition *Y* adds an attribute to entity *X*), followed by events linked by temporal connections (e.g., event *Y* follows event *X* in time). Lastly, and perhaps more obviously, people consider scenarios containing unrelated events to be the least plausible of all.

This study shows that plausibility judgments are sensitive to the conceptual coherence of the different inference types involved in representing event descriptions. The concept–coherence of the scenario, and hence its perceived plausibility, is greatest in the causal pairs where a simple inference can be made with direct corroboration from prior knowledge. The concept–coherence, and hence plausibility, is lowest in the unrelated pairs where complex inferences and assumptions have to be made to connect the events (which, indeed, may fail to be corroborated by prior knowledge at all). Ranged in between are the attributal and temporal pairs, largely distinguished by the greater amount of complexity and conjecture involved in temporal inferences. If PAM is an accurate model of human plausibility judgment, then this same trend of de-

Table 3
Sample sentence pairs for each inference type (from Connell & Keane, 2004)

Inference Type	Sample Sentence Pair
Causal	The dress snagged on a nail. The silk ripped.
Attributal	The dress snagged on a nail. The silk was priceless.
Temporal	The dress snagged on a nail. The silk glittered.
Unrelated	The dress snagged on a nail. The silk shrank.

creasing plausibility ratings should be evident across inference types as causal > attributal > temporal > unrelated.

5.1. Method

5.1.1. Materials

The materials for this simulation consisted of 60 sentence pairs, with manipulations of concept–coherence, from the two experiments reported in Connell and Keane (2004; see Appendixes A and B in that article for a full listing of materials). For each experiment, a set of base sentence pairs was created, and then the second sentence was modified to produce different inferential variants of it. Connell and Keane's (2004) Experiment 1 used four inference types (causal, attributal, temporal, and unrelated), whereas Experiment 2 focused on two (causal and attributal). The causal pairs were designed to invite a causal inference by using a second sentence (S2) that was a reasonably direct causal consequence of the first sentence (S1; e.g., "The dress snagged on a nail. The silk ripped"). The attributal pairs invited an attributal inference by using an S2 that referred to an attribute of its subject in a way that was not causally related to S1 (e.g., "The dress snagged on a nail. The silk was priceless"). The temporal pairs invited a temporal inference by using an S2 that could occur in the normal course of events, regardless of the occurrence of S1 (e.g., "The dress snagged on a nail. The silk glittered"). The unrelated pairs used an S2 that described an event that was unlikely to occur in the normal course of events and had no obvious causal link to S1 (e.g., "The dress snagged on a nail. The silk shrank").

Word frequency was controlled across inference types using word frequency counts from the British National Corpus.[3] In addition, two other factors were used in the creation of Connell and Keane's (2004) materials (noun type in Experiment 1, word coherence in Experiment 2), but neither of these factors affected plausibility ratings; because they are not relevant to our present purposes they will not be discussed further.

Thus, each sentence pair used in this simulation had one of four inference types connecting the sentences (causal, attributal, temporal, and unrelated). There were more causal and attributal sentence pairs than temporal and unrelated pairs, due to the unequal distribution of inference types across Connell and Keane's (2004) experiments, with 22 causal, 20 attributal, 9 temporal, and 9 unrelated sentence pairs.

5.1.2. Procedure

The procedure used for human participants is detailed in Connell and Keane (2004). Briefly, participants read instructions that explained the 0 to 10 plausibility scale (0 being *not at all plausible* and 10 being *highly plausible*). They were asked to read each sentence pair and to rate how plausible they found the scenario described in the sentences. They were asked to take their time over each decision and not to alter any answers already marked down. Each sentence pair was presented on a separate page with a marked space for participants to note their 0 to 10 plausibility rating. For the purposes of this simulation, the mean plausibility rating for each sentence pair (in each inference condition) was used.

The procedure used in the computational simulations involved presenting PAM with each natural language sentence pair and recording the plausibility rating returned; PAM outputs each rating (0–10) rounded to one decimal place.

Table 4
Mean plausibility ratings per inference type as produced by PAM and by participants (Connell & Keane, 2004), on a scale from 0 (*implausible*) to10 (*very plausible*)

Inference Type	Model Plausibility Rating	Human Plausibility Rating
Causal	8.3	7.8
Attributal	6.1	5.5
Temporal	5.5	4.2
Unrelated	1.5	2.0

5.2. Results

The simulation shows that PAM's output accurately reflects the product of human plausibility judgments. Inference-type effects on plausibility ratings are accurately modeled, with plausibility decreasing from causal > attributal > temporal > unrelated.

Table 4 gives the mean ratings per condition compared to the human responses from Experiment 1 in Connell and Keane (2004). PAM's estimates correlate strongly with participants' mean plausibility judgments per scenario, and regression analysis suggests that the model could be used as a successful predictor of human plausibility ratings ($r = .776$, $r^2 = .603$, $N = 60$, $p < .0001$). The relation between model output and participant means for each scenario is shown in Fig. 7's scatter plot, with each inference type distinguished.

Furthermore, PAM's estimates reveal the same response patterns found for human plausibility judgments, with causal scenarios attracting the highest plausibility ratings, followed by attributal, temporal, and finally unrelated scenarios. Table 5 shows a sample scenario for all four inference types, along with PAM's variable values and rating and the mean human rating for these particular scenarios. The same downward trend is found from causal > attributal > temporal > unrelated in both the model and human ratings. An analysis of variance of PAM's ratings for all scenarios showed that this effect of inference type is reliable, $F(3, 56) = 115.644$, $p < .0001$, mean square error = 0.943.

5.3. Discussion

This simulation confirms PAM's ability to model the judgment of plausibility. Inference-type effects are modeled by extracting three variables from the scenario representations formed by PAM (number of paths, mean path length, and number of nonhypothetical paths). However, PAM does not distinguish between the different types of inferences that may connect sentences; there is no hard-coded differentiation in either the knowledge base or the model framework. Yet the model produces distinctly different plausibility ratings for different types of inference. So how does this happen? The answer lies in how each inference type tends toward certain values for each of the extracted variables.

Table 6 illustrates the how each inference type tends toward high or low values for each of the extracted variables (number of paths *P,* mean path length *L,* and number of nonhypothetical paths *N*). According to PAM's plausibility rating function, the most plausible scenario will

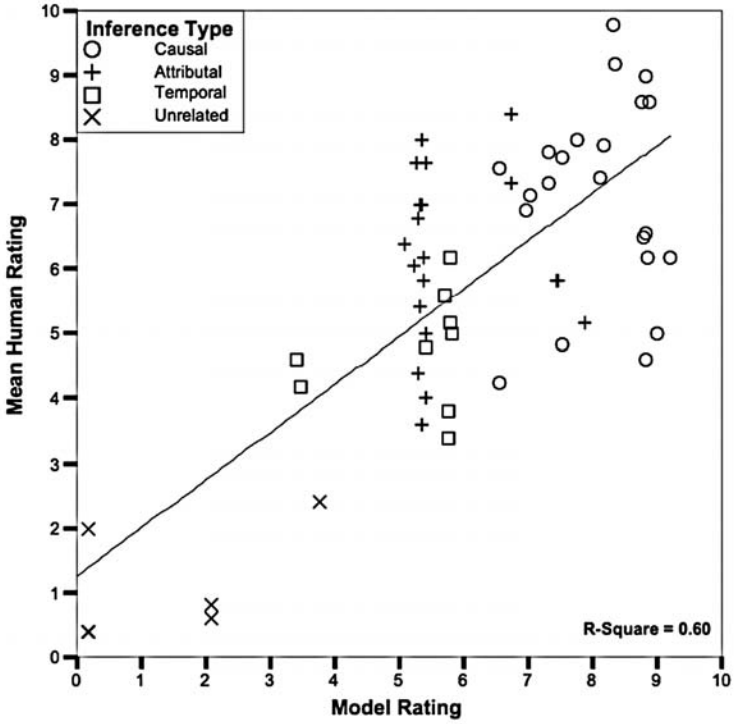

Fig. 7. Scatter plot of relation between plausibility ratings produced by PAM and by participants (r = .776), with each sentence pair distinguished by inference type.

have a high number of paths, a low path length, and a high number of nonhypothetical paths. First, let us consider the interaction of the two path-number variables shown in Table 6 (number of paths *P* and number of nonhypothetical paths *N*). Briefly stated, this results in causal scenarios being rated as most plausible, because their large number of paths gives them a high rating even though many are hypothetical. Attributal scenarios are rated with medium plausi-

Table 5
Sample scenarios for each inference with PAM's propositional form and variable values, along with model and human (Connell & Keane, 2004) plausibility ratings for those scenarios

		Model Variables				
Inference Type	Scenario	*P*	*L*	*N*	Model Rating	Human Rating
Causal	The waitress dropped the cup. The cup smashed.	11	3.6	2	8.7	9.8
Attributal	The waitress dropped the cup. The cup was delicate.	1	1.0	1	5.6	6.8
Temporal	The waitress dropped the cup. The cup glistened.	2	3.0	0	3.9	4.2
Unrelated	The waitress dropped the cup. The cup floated.	0	0.0	0.0	0.0	2.0

Table 6
Means (and standard deviations) per inference type for the variables extracted from PAM's representation. High *P* and *N* values increase plausibility ratings, whereas high *L* values decrease plausibility

	Extracted Variable					
	Number of Paths (*P*)		Mean Path Length (*L*)		Number of Nonhypothetical Paths (*N*)	
Inference Type	*M*	*SD*	*M*	*SD*	*M*	*SD*
Causal	10.68	(6.96)	2.82	(0.65)	5.00	(6.06)
Attributal	1.55	(1.00)	1.07	(0.21)	1.45	(0.89)
Temporal	2.11	(0.33)	2.33	(0.50)	1.33	(1.00)
Unrelated	0.89	(0.60)	1.39	(0.93)	0.00	(0.00)

bility, because they have a very low number of paths (even if few are hypothetical). Temporal scenarios are also rated with medium plausibility because, although they have more paths than attributal scenarios, many of them tend to be hypothetical. Last, unrelated scenarios are rated as least plausible because they have a small number of paths, all of which are hypothetical. Thus, based on the variables *P* and *N,* causal inferences are the most plausible, attributal and temporal inferences are tied with medium plausibility, and unrelated inferences have the lowest plausibility. Taking the other variable (mean path length *L*) into account, attributal scenarios become more plausible than temporal scenarios because of their much shorter path length. This gives rise to the causal > attributal > temporal > unrelated trend in plausibility ratings seen in the simulation. However, it is interesting to observe the variability within each inference type as well as the mean trend: For example, PAM rated certain temporal scenarios as more plausible than some attributal scenarios, and certain attributal scenarios as more plausible than some causal scenarios (see Fig. 7). Such variability in PAM's plausibility ratings parallels the variability seen in the human experiments and demonstrates the model's success in capturing the range of group behavior.

One possible worry about these results is that they tell us more about people's judgments of plausible sentences in a discourse than about plausible events in the world.[4] For various reasons, we do not share this concern and would assert that PAM is modeling people's judgment of events in the world (granted, as they are conveyed by discourse). First, we had previously examined, using these materials, whether distributional word co-occurrence (as measured by latent semantic analysis; Landauer & Dumais, 1997) had an impact on people's plausibility judgments. Word co-occurrence is an important discourse-sensitive metric that can predict the perceived "naturalness" of phrases (Lapata et al., 1999) and is likely to be correlated highly with any perceived naturalness of sentences (i.e., the plausibility of the discourse itself rather than of events in the world). We have repeatedly found no effects of such word co-occurrence on plausibility judgments (Connell & Keane, 2004). Second, there is evidence that, with suitable instruction, people can distinguish judgments of the naturalness of sentences from judgments of the plausibility of the events described in those sentences (see Gruber & Gibson,

Table 7
Sample novel scenarios modeled by PAM

Scenario	Propositional Form	Model Plausibility Rating
The man saw the girl. The man waved.	see (man, girl), wave (man)	8.6
The woman dropped the vase. The girl was afraid.	drop (woman, vase), afraid (girl)	8.3
The water wet the table. The table shone.	wet (water, table), shine (table)	6.9
The boy smashed the cup. The boy was angry.	break (boy, cup), angry (boy)	6.4
The girl waved at the cat. The cat boiled.	wave (girl, cat), boil (cat)	0.6

2004). In our experiments, people were explicitly instructed to judge the scenario described by the sentences.

5.4. Modeling novel scenarios

It is important to note that PAM is not restricted to modeling just those scenarios tested in this simulation. The nature of its knowledge base means that any scenario using known entities and predicates can have a plausibility rating estimated. For example, given the predicates *afraid, angry, boil, drop, see, shine, smash, wave,* and *wet,* it is possible to construct scenarios such as "The man saw the girl. The man waved" or "The woman dropped the cup. The girl was afraid." Table 7 shows some sample novel scenarios and propositions, along with the plausibility ratings produced by PAM. In each case, PAM has represented the scenario by fitting the propositions to predicates in its knowledge base and producing a plausibility rating that seems reasonably close to human intuition. For example, PAM rates the scenario, "The water wet the table. The table shone" as quite plausible (6.9 out of 10) because it found a reasonable fit for the scenario, including the causal explanation that the table shone because the spilled water reflected light and the attributal explanation that the table shone because its surface was a reflective material such as glass or varnish. These novel examples illustrate how PAM's knowledge base allows it to model many different types of scenario.

6. Sensitivity analysis

Sensitivity analysis is a useful tool in cognitive modeling, allowing the designer to examine if the model is consistent with its underlying theory and to test its robustness in a variety of operational contexts. PAM uses three key parameters in modeling plausibility judgment, which may invite the criticism that all of these variables are not really required to achieve predictive accuracy. If one or more of the variables (number of paths *P,* mean path length *L,* number of nonhypothetical paths *N*) is not making a significant contribution to PAM's performance, then a much more parsimonious model may exist for calculating plausibility. This state of affairs

could, in turn, have complexity implications for any future use of the model, as well as requiring some revision of the theory.

In any cognitive model, it is important that the key parameters of the model are those motivated by the theory and not those motivated simply by the need to make the model work (i.e., the so-called A|B distinction for cognitive models; Cooper, Fox, Farringdon, & Shallice, 1996). According to the A|B distinction, the separation of theory and implementation detail is required if the complex behavior displayed by a computational model is to be correctly attributed to theoretical, and not implementational, aspects of the specification. This separation is referred to by Cooper et al. as the above-the-line|below-the-line (or A|B) distinction, with theoretically motivated aspects being located above the line and implementation details located below the line. To justify being an "A" (theoretical) component, a variable must be critical to the behavior of the model as a whole, with variation of the variable yielding empirically measurable differences in behavior.

This section of the article describes a sensitivity analysis of PAM for its key variables of *P, L,* and *N.* In the first analysis, we perform an analysis of each variable's contribution to the plausibility rating function shown in Fig. 4. In the second analysis, we systematically vary the contribution of each variable to the plausibility function and examine whether PAM's replication of human plausibility judgment is robust. If all three variables are vital to plausibility estimation, as the knowledge-fitting theory holds, then varying their contribution to the plausibility function should result in a degradation of PAM's ability to simulate human performance. If no such degradation is visible, then the source of PAM's performance must lie elsewhere.

6.1. Analysis 1: Contribution of variables

As described earlier in the article, Fig. 5 illustrates the contribution that each variable makes to the plausibility rating function. The contribution of each variable can be analyzed by creating a three-dimensional space (one dimension for each variable) of each variable's possible values and calculating the resulting plausibility rating for each combination of values. A sample of this space is given in Fig. 5, showing PAM's plausibility ratings for an increasing number of paths (*P*) and for increasing path complexity (*L*), with separate surfaces for the best case (no paths hypothetical) and worst case (all paths hypothetical) values for the number of nonhypothetical paths *N.*

The relative contribution of each variable to PAM's plausibility function can then be determined by multiple-regression analysis. Using PAM's own plausibility formula (see Fig. 4) as the regression equation, the variables *P, L,* and *N* were applied as independent predictor variables to the set of plausibility ratings in the nonlinear-regression technique provided by *SigmaPlot* (2004). The resulting standardized beta coefficients represent standardized forms of the parameter estimates for each predictor variable (*P, L, N*) and can be used to compare the relative importance of each variable in generating PAM's plausibility ratings. Regression shows that the number of nonhypothetical paths *N* contributes most to PAM's plausibility function ($N\ \beta = 1.140$, $p < .0001$), with mean path length *L* not far behind ($L\ \beta = 1.000$, $p < .0001$). The total number of paths *P* is less important but is still a significant contributor ($P\ \beta = .333$, $p < .0001$). This analysis confirms that each variable (*P, L, N*) fulfills a necessary role in PAM's plausibility function, so we may now examine the robustness of the function's performance.

6.2. Analysis 2: Robustness of model

PAM reflects human performance in rating causal scenarios as the most plausible, followed by attributal, temporal, and unrelated scenarios. In general, we say that PAM's performance satisfies the data if this causal > attributal > temporal > unrelated trend is maintained.

In the following analysis, we test how sensitive PAM's performance is to changes in each variable's contribution. To do this, we re-run the simulation reported earlier, but systematically vary the weight of each variable in the plausibility rating function. We then examine the resulting correlations and whether the model performance satisfies the data. If PAM's modeling of plausibility ratings is indeed robust, then we should see the model's performance degrade as the variable weights change. It is important that performance degrades after a certain point (i.e., that there are certain parameter settings that do not fit the human data) because this serves to confirm that the theoretically motivated variables actually matter.

The results of the sensitivity analysis are shown in a series of three tables. Each table shows the systematic variation of two variables as they are weighted more lightly (1%–75%), unchanged (100%), or weighted more heavily (125%–200%). Each entry in the table shows the correlation score *r,* with human data for that combination of variable weights, and indicates by boldface whether those weights satisfy the data. Table 8 shows the results of PAM's sensitivity analysis for the variables *P* (number of paths) and *L* (mean path length). Table 9 shows the sensitivity analysis for the variables *P* and *N* (number of nonhypothetical paths), and Table 10 shows the sensitivity analysis for the variables *L* and *N.*

The sensitivity analysis shows us that there is a key region that satisfies the data, roughly corresponding to where weights for *P, L,* and *N* are between 50% and 150%. The total region that satisfies the data is indicated by the boldfaced areas in Tables 8–10. This is a reasonably large range of weightings and indicates that PAM's performance is robust and not hostage to a particular span of narrow parameter settings. The correlation between model and human data can also be seen to decrease as variable weights head toward extremes. Indeed, much lower

Table 8
Sensitivity analysis for variables total number of paths (*P*) and mean path length (*L*), showing correlation between model and human plausibility ratings

	Weight for *P*								
Weight for *L*	1%	25%	50%	75%	100%	125%	150%	175%	200%
1%	0.535	0.440	0.367	**0.343**	**0.331**	**0.324**	0.319	0.316	0.313
25%	–0.661	0.687	0.755	**0.743**	**0.702**	**0.656**	0.615	0.580	0.552
50%	–0.677	0.269	**0.772**	**0.789**	**0.765**	**0.728**	0.692	0.659	0.630
75%	–0.686	–0.250	**0.767**	**0.797**	**0.777**	**0.744**	0.712	0.682	0.656
100%	–0.692	–0.445	**0.752**	**0.796**	**0.776**	**0.746**	0.715	0.689	0.665
125%	–0.696	–0.524	**0.730**	**0.792**	**0.772**	**0.742**	0.714	0.689	0.667
150%	–0.699	–0.565	**0.703**	**0.787**	**0.767**	0.738	0.710	0.687	0.666
175%	–0.701	–0.590	**0.672**	**0.781**	**0.762**	0.733	0.707	0.684	0.664
200%	–0.703	–0.606	**0.640**	**0.775**	**0.757**	0.729	0.703	0.681	0.662

Note. Boldfaced values represent a region of weights that consistently satisfy the data (causal > attributal > temporal > unrelated) for all combinations of variables.

Table 9
Sensitivity analysis for variables total number of paths (*P*) and number of nonhypothetical paths (*N*), showing correlation between model and human plausibility ratings

	Weight for *P*								
Weight for *N*	1%	25%	50%	75%	100%	125%	150%	175%	200%
1%	–0.692	–0.492	0.487	0.579	0.606	0.618	0.622	0.620	0.615
25%	–0.692	–0.495	0.641	0.682	0.674	0.663	0.652	0.641	0.630
50%	–0.692	–0.478	**0.731**	**0.749**	**0.724**	0.699	0.678	0.660	0.643
75%	–0.692	–0.460	**0.752**	**0.782**	**0.756**	**0.726**	0.699	0.676	0.655
100%	–0.692	–0.445	**0.752**	**0.796**	**0.776**	**0.746**	0.715	0.689	0.665
125%	–0.692	–0.432	**0.747**	**0.801**	**0.788**	**0.759**	0.728	0.699	0.674
150%	–0.692	–0.422	**0.741**	**0.801**	**0.795**	**0.769**	0.738	0.708	0.681
175%	–0.692	–0.414	0.736	**0.798**	**0.798**	**0.776**	0.746	0.716	0.688
200%	–0.692	–0.406	0.732	**0.795**	**0.799**	**0.781**	0.752	0.722	0.693

Note. Boldfaced values represent a region of weights that consistently satisfy the data (causal > attributal > temporal > unrelated) for all combinations of variables.

Table 10
Sensitivity analysis for variables mean path length (*L*) and number of nonhypothetical paths (*N*), showing correlation between model and human plausibility ratings.

	Weight for *L*								
Weight for *N*	1%	25%	50%	75%	100%	125%	150%	175%	200%
1%	0.329	0.646	0.647	0.625	0.606	0.591	0.581	0.573	0.567
25%	0.330	0.677	0.703	0.690	0.674	0.659	0.647	0.638	0.630
50%	**0.330**	**0.693**	**0.737**	**0.734**	**0.724**	0.712	0.702	0.693	0.685
75%	**0.331**	**0.699**	**0.755**	**0.761**	**0.756**	0.749	0.741	0.734	0.727
100%	**0.331**	**0.702**	**0.765**	**0.777**	**0.776**	**0.772**	**0.767**	**0.762**	**0.757**
125%	**0.331**	**0.703**	**0.769**	**0.785**	**0.788**	**0.787**	**0.784**	**0.781**	**0.777**
150%	**0.331**	**0.702**	**0.770**	**0.789**	**0.795**	**0.796**	**0.795**	**0.793**	**0.791**
175%	**0.331**	**0.701**	**0.770**	**0.791**	**0.798**	**0.801**	**0.801**	**0.800**	**0.799**
200%	0.332	**0.700**	**0.769**	**0.791**	**0.799**	**0.803**	**0.804**	**0.805**	**0.804**

Note. Boldfaced values represent a region of weights that consistently satisfy the data (causal > attributal > temporal > unrelated) for all combinations of variables.

(and even negative) correlations are observed when the variables *P, L,* and *N* are weighed at 1%, a weight so light as to almost remove the effect of that variable. It should be noted that Tables 8–10 only illustrate the interaction of two variables at a time, but all three variables were systematically tested. The highest correlation found for a combination of weights that satisfied the data was $r = .805$, where *P* was weighted at 100%, *L* at 175%, and *N* at 200%. This combination of variable weights represents the best fit of the model to this particular human data set; however, we do not wish to overfit the model to these data, so these weight values will not be adopted in PAM's plausibility function so as to preserve its generalizability to other data.

In addition, the sensitivity analysis also shows that PAM's key operations conform to the A|B distinction of cognitive models (Cooper et al., 1996). The variables used in calculating plausibility—number of paths *P,* mean path length *L,* and number of nonhypothetical paths *N*—have been shown to be critical to the behavior of the model as a whole. In this sense, all three variables are "A" components that are relevant to the theoretical rather than implementational aspects of the model.

Given the combined contribution of the variables *P* and *N,* it could be argued that expanding PAM's knowledge base could have a detrimental effect on the model's performance (i.e., that a larger knowledge base may contain a larger number of possible paths and may skew plausibility ratings). However, this issue is not of major concern. As seen in Fig. 5, plausibility ratings begin to level out with respect to increases in *P* as the rating asymptote of 10 is approached. Therefore, an effective threshold is already in place for the variable *P* that prevents high values from contributing disproportionately to the plausibility function. However, it may also be argued that a larger knowledge base may lead to a decrease in the number of nonhypothetical paths (*N*), which may also skew plausibility ratings. If this were found to be the case, PAM could preserve accuracy by implementing a specific threshold on the number of possible paths returned, which would also have the effect of limiting the number of admissible hypothetical entities. This would allow PAM to maintain its level of performance as its knowledge base grows. In models of analogy, Veale and Keane (1997) showed that the thresholding of inferential paths in a large knowledge base can effectively contain such combinatorial explosions and maintain system performance at acceptable levels.

7. General discussion

There are a number of novel achievements reported in this article. First, PAM is the first computational model that specifically and accurately addresses human plausibility judgment. Although there are many models of discourse comprehension that characterize the formation of inferences (McKoon & Ratcliff, 1992; Schank & Abelson, 1977; Singer et al., 1994), these models tend to finesse the specific characterization of plausibility. PAM models plausibility judgment by using a number of innovative techniques to capture the complex influences of concept–coherence that empirical work has shown to bear on plausibility (Connell & Keane, 2004).

Second, plausibility judgment is modeled as spanning two stages: comprehension (where a representation of the scenario is created) and assessment (where the representation is examined to determine how well the scenario fits prior knowledge). Theoretically, the knowledge-fitting theory separates the process of understanding the scenario from assessing its plausibility, and this separation allows PAM to define the concept–coherence of a given scenario as a function of the representation itself (i.e., as the degree of fit between the scenario and prior knowledge).

Third, PAM's comprehension stage uses a commonsense knowledge base to represent scenarios. This representation is based on an analysis of the requirements that must be met for a proposition to be true. Many of these requirements are based on what is intuitively regarded as common sense. For example, for an entity *X* to *melt,* one of the requirements is that *X* is not al-

ready *liquid.* For *X* to be a *liquid,* there is a further requirement that *X* is *nonabstract,* and so on. Although this generally precludes the use of figurative language in the sentence pairs that PAM takes as input, it would be possible to build up such a requirements set for future versions. In this way, PAM demonstrates how using a commonsense knowledge base allows a scenario to be represented with information that would normally form part of people's prior knowledge, unlike, for example, explicit information about probability distributions (e.g., Pearl, 1988; Shafer, 1976; Tenenbaum & Griffiths, 2001).

Fourth, the plausibility of a scenario is calculated in the assessment stage according to three key concept–coherence variables that reflect how well the scenario fits with prior knowledge. In the knowledge-fitting theory, a highly plausible scenario is one that fits prior knowledge (a) with many different sources of corroboration, (b) without complex explanation, and (c) with minimal conjecture. As one possible implementation of this theory, PAM shows that it is possible to realize these three factors computationally as (a) number of paths found, (b) mean path length, and (c) number of nonhypothetical paths. These variables are both theoretically motivated and essential to the plausibility rating function (as shown by sensitivity analysis) and allow PAM to model human plausibility judgment robustly and accurately. This explanation of concept–coherence is both more specific and broader reaching than any previous account. For example, it impacts on areas such as conceptual combination (Costello & Keane, 2000, 2001; Lynott et al., 2004), categorization (Rehder, 2003a), and argument evaluation (Smith et al., 1993) by describing plausibility as something more complex than just feature or proposition overlap. Also, it illustrates that plausibility does not just follow from the ability to make inference between events (Black et al., 1986), but rather is also dependent on the background knowledge that these inferences require.

Fifth, PAM is capable of producing different plausibility ratings for different inferential scenarios without making any explicit distinction between inference types. For example, there is no hard-coded differentiation in either the knowledge base or the model framework between causal relations and temporal relations, yet PAM (like people) rates causal scenarios as more plausible than temporal scenarios. By allowing the interaction of the three key variables to differentiate the plausibility ratings of causal, attributal, temporal, and unrelated scenarios, the model dispenses with many of the artificial complexities that would arise in the model if relational distinctions were made. Indeed, although PAM reflects the inference-type effects of people's plausibility ratings (decreasing plausibility from causal > attributal > temporal > unrelated scenarios), the nature of the model's plausibility function means that this trend is not absolute. For example, PAM, like many people, may give a poor causal scenario a lower plausibility rating than a good temporal scenario. Thus, it is clear that PAM's inference-type effects arise naturally from the concept–coherence of individual scenarios.

In its simulation, PAM was tested solely on the sentence pairs used in the experiments of Connell and Keane (2004), which raises some questions about the extendibility of the model to longer pieces of discourse. We see no obstacle, in principle, to the extension of the model to longer discourse. In the comprehension stage, each subsequent sentence would be folded into the representation in exactly the same way that the second sentence of the pair was processed, as there is no constraint on the size of the path-based representation. In the assessment stage, there is similarly no constraint on the size of the representation that could be subsequently analyzed. With longer pieces of discourse, there may nonetheless be a need for some additional

functionality. For example, we might want the distributional activation of regions from older sentences to decay over time, or we might need to optimize the analysis of the representation in some way, were it to grow to several hundred paths. However, these additions are fine-tunings of PAM, and they do not represent changes to its fundamental operations.

The PAM provides a computational account of how plausibility is judged and how concept–coherence is assessed. Given the pervasiveness of plausibility in many cognitive phenomena, the computational and theoretical issues raised here impact on many areas of research, such as conceptual combination, memory retrieval, discourse comprehension, and reasoning. In short, plausibility has been offered a clarity of definition that was previously absent from cognitive science.

Notes

1. All entries in the knowledge base were added in a "blind" fashion; that is, each entity and predicate were defined as thoroughly as possible in terms of argument conditions without reference to the original sentence pairs.
2. We would like to thank James Hampton for useful suggestions on simplifying the presentation of the plausibility function.
3. The British National Corpus contains 100 million words of British English, from both spoken and written sources, and is designed to represent a wide cross-section of modern British English use (see Aston & Burnard, 1998).
4. We would like to thank Nick Chater and an anonymous reviewer for raising this possibility.

Acknowledgments

This work has been funded in part by grants from the Irish Research Council for Science, Engineering and Technology, under the Embark Initiative to the first author and from Science Foundation Ireland under Grant No. 03/IN.3/I361 to the second author.

Some of the research described in this article was completed as part of a PhD dissertation submitted to University College Dublin (Connell, 2004) and has been previously reported at conferences of the Cognitive Science Society (Connell & Keane, 2002, 2003).

The authors are grateful to Dermot Lynott for valuable comments on earlier drafts of this article. We would also like to thank the members of the University College Dublin/Trinity College Dublin Cognitive Science Group for their feedback on this work.

References

Aston, G., & Burnard, L. (1998). *The BNC handbook: Exploring the British National Corpus with SARA*. Edinburgh, Scotland: Edinburgh University Press.

Black, A., Freeman, P., & Johnson-Laird, P. N. (1986). Plausibility and the comprehension of text. *British Journal of Psychology, 77*(1), 51–60.

Collins, A., & Michalski, R. (1989). The logic of plausible reasoning: A core theory. *Cognitive Science, 13,* 1–49.

Connell, L. (2004). *A cognitive theory and model of plausibility.* Unpublished doctoral dissertation, University College Dublin, Ireland.

Connell, L., & Keane, M. T. (2002). The roots of plausibility: The role of coherence and distributional knowledge in plausibility judgements. In J. D. Moore & K. Stenning (Eds.), *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society* (p. 998). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Connell, L., & Keane, M. T. (2003). PAM: A cognitive model of plausibility. In A. Markman & L. Barsalou, *Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society* (pp. 264–269). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Connell, L., & Keane, M. T. (2004). What plausibly affects plausibility? Concept–coherence and distributional word–coherence as factors influencing plausibility judgements. *Memory and Cognition, 32,* 185–197.

Cooper, R., Fox, J., Farringdon, J., & Shallice, T. (1996). A systematic methodology for cognitive modelling. *Artificial Intelligence, 85*(1–2), 3–44.

Costello, F., & Keane, M. T. (2000). Efficient creativity: Constraints on conceptual combination. *Cognitive Science, 24,* 299–349.

Costello, F., & Keane, M. T. (2001). Alignment versus diagnosticity in the comprehension and production of combined concepts. *Journal of Experimental Psychology: Learning, Memory and Cognition, 27,* 255–271.

Friedman, N., & Halpern, J. Y. (1996). Plausibility measures and default reasoning. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence, 2* (pp. 1297–1304). Menlo Park, CA: AAAI Press.

Gernsbacher, M. A. (1990). *Language comprehension as structure building.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Gruber, J., & Gibson, E. (2004). Measuring linguistic complexity independent of plausibility. *Language, 80*(3), 583–590.

Halpern, J. Y. (2001). Plausibility measures: A general approach for representing uncertainty. In B. Nebel (Ed.), *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence* (pp. 1474–1483). San Francisco: Kaufmann.

Johnson-Laird, P. N. (1983). *Mental models.* Cambridge, England: Cambridge University Press.

Keane, M. T., Ledgeway, T., & Duff, S. (1994). Constraints on analogical mapping: A comparison of three models. *Cognitive Science, 18,* 387–438.

Keil, F. (1989). *Concepts, kinds and conceptual development.* Cambridge, MA: MIT Press.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition.* Cambridge, England: Cambridge University Press.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis Theory of acquisition, induction and representation of knowledge. *Psychological Review, 104*(2), 211–240.

Lapata, M., McDonald, S., & Keller, F. (1999). Determinants of adjective-noun plausibility. In *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics.* (pp. 30–36). San Francisco: Kaufmann.

Lemaire, P., & Fayol, M. (1995). When plausibility judgments supersede fact retrieval—The example of the odd even effect on product verification. *Memory and Cognition, 23,* 34–48.

Lynott, D., Tagalakis, G., & Keane, M. T. (2004). Conceptual combination with PUNC. *Artificial Intelligence Review, 21,* 353–374.

McKoon, G., & Ratcliff, R. (1992). Inference during reading. *Psychological Review, 99,* 440–466.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92,* 289–316.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference.* Los Angeles: Kaufman.

Pickering, M. J., & Traxler, M. J. (1998). Plausibility and recovery from garden paths: An eye-tracking study. *Journal of Experimental Psychology: Learning, Memory and Cognition, 24,* 940–961.

Reder, L. M. (1982). Plausibility judgments vs. fact retrieval: Alternative strategies for sentence verification. *Psychological Review, 89,* 250–280.

Reder, L. M., & Ross, B. H. (1983). Integrated knowledge in different tasks: The role of retrieval strategy on fan effects. *Journal of Experimental Psychology: Learning, Memory and Cognition, 9,* 55–72.

Reder, L. M., Wible, C., & Martin, J. (1986). Differential memory changes with age: Exact retrieval versus plausible inference. *Journal of Experimental Psychology: Learning, Memory and Cognition, 12,* 72–81.

Rehder, B. (2003a). Categorization as causal reasoning. *Cognitive Science, 27,* 709–748.

Rehder, B. (2003b). A causal-model theory of conceptual representation and categorization. *Journal of Experimental Psychology: Learning, Memory and Cognition, 29,* 1141–1159.

Schank, R. C., & Abelson, R. (1977). *Scripts, goals, plans, and understanding.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Shafer, G. (1976). *A mathematical theory of evidence.* Princeton, NJ: Princeton University Press.

SigmaPlot (Version 9.0) [Computer software]. (2004). Richmond, CA: Systat Software.

Singer, M., Graesser, A. C., & Trabasso, T. (1994). Minimal or global inferences during reading. *Journal of Memory and Language, 33,* 421–441.

Smith, E. E., Shafir, E., & Osherson, D. (1993). Similarity, plausibility, and judgments of probability. *Cognition, 49*(1–2), 67–96.

Speer, S. R., & Clifton, C. (1998). Plausibility and argument structure in sentence comprehension. *Memory and Cognition, 26,* 965–978.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences, 24,* 629–640.

Traxler, M. J., & Pickering, M. J. (1996). Plausibility and the processing of unbounded dependencies: An eye-tracking study. *Journal of Memory and Language, 35,* 454–475.

Veale, T., & Keane, M. T. (1997). The competence of sub-optimal theories of structure mapping on hard analogies. In D. Koller & A. Pfeffer (Eds.), *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence* (pp. 232–237). San Francisco: Kaufmann.

Zwaan, R. A., Magliano, J. P., & Graesser, A. C. (1995). Dimensions of situation-model construction in narrative comprehension. *Journal of Experimental Psychology: Learning, Memory and Cognition, 21,* 386–397.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123,* 162–185.