Broadening Plausibility: A Sensitivity Analysis of PAM

Louise Connell (louise.connell@northumbria.ac.uk)
Division of Psychology, Northumbria University,
Newcastle upon Tyne, NE1 8ST, UK

Mark T. Keane (mark.keane@ucd.ie)
Department of Computer Science, University College Dublin,
Belfield, Dublin 4, Ireland

Abstract

The judgement of plausibility is severely under-specified in cognitive science despite its diverse uses in many cognitive tasks. Recently, a model of human plausibility judgement, called the Plausibility Analysis Model (PAM), has been proposed and has been shown to closely model human plausibility ratings of event scenarios. In the present study, we present a sensitivity analysis to explore PAM's robustness with a view to assessing its broader implications in cognitive science and cognitive modelling. Overall, this analysis shows that PAM is consistent with its underlying theory and is robust in a wide range of operational contexts, thus indicating that the model is well grounded in its characterisation of plausibility effects.

Introduction

People make consistent and constant use of plausibility judgements in everyday life for a variety of reasons, from assessing the quality of a movie plot, to determining guilt in a tabloid murder trial, to considering a child's excuse for a broken dish. Yet, plausibility remains poorly understood or explored in cognitive science. Recently, Connell and Keane (2003, in prep.) have advanced the Plausibility Analysis Model (PAM) as the first cognitive model of human plausibility judgements. In this paper, we consider the implications of this model in a broader context and illustrate the robustness of its performance with sensitivity analyses.

We know of very few cognitive models that make explicit use of plausibility to guide, for example, decision-making, problem solving or natural language understanding. Yet, people constantly seem to use plausibility judgements to guide diverse cognitive tasks. For example, people often use plausibility judgements in place of costly retrieval from long-term memory, especially when verbatim memory has faded (Lemaire & Fayol, 1995; Reder, Wible & Martin, 1986). Plausibility is also used as a kind of cognitive shortcut in reading, to speed parsing and resolve ambiguities (Pickering & Traxler, 1998; Speer & Clifton, 1998). In everyday thinking, plausible reasoning that uses prior knowledge appears to be commonplace (Collins & Michalski, 1989), and can even aid people in making inductive inferences about familiar topics (Smith, Shafir & Osherson, 1993). It has also been argued that plausibility plays a fundamental role in understanding novel word combinations by helping to constrain the interpretations produced (Costello & Keane, 2000; Lynott, Tagalakis & Keane, 2004). Many of these tasks have broad implications for models of cognition and underscore the centrality of plausibility. In this paper, we explore the computational aspects of our research program on plausibility. Specifically, we outline a computational model of plausibility and demonstrate its robustness as a model with an extensive sensitivity analysis.

Plausibility and the Knowledge-Fitting Theory

In the Knowledge-Fitting Theory of Plausibility, Connell and Keane (2003, in prep.) define plausibility judgements as being about assessing how well a scenario fits with prior knowledge. They show that the plausibility rating of a scenario depends upon its concept-coherence (i.e., the inference and prior knowledge used to connect the scenario's events). In addition, Connell and Keane (2004) have shown that the type of connection between a scenario's events influences its plausibility (see Table 1 for examples). People consider events linked by causal connections (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.

In the Knowledge-Fitting Theory, plausibility judgement spans two stages: comprehension (where a representation of the scenario is formed) and assessment (where this representation is analysed to ascertain its concept-coherence). The Knowledge-Fitting Theory holds that three key aspects of the representation interact to determine a scenario's concept-coherence: complexity, corroboration and conjecture. Briefly stated, as *complexity* increases, plausibility decreases. This, however, is tempered by the *corroboration* of the scenario, as even a very complex scenario will be 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. In essence, the most plausible scenarios are those with high concept-coherence.

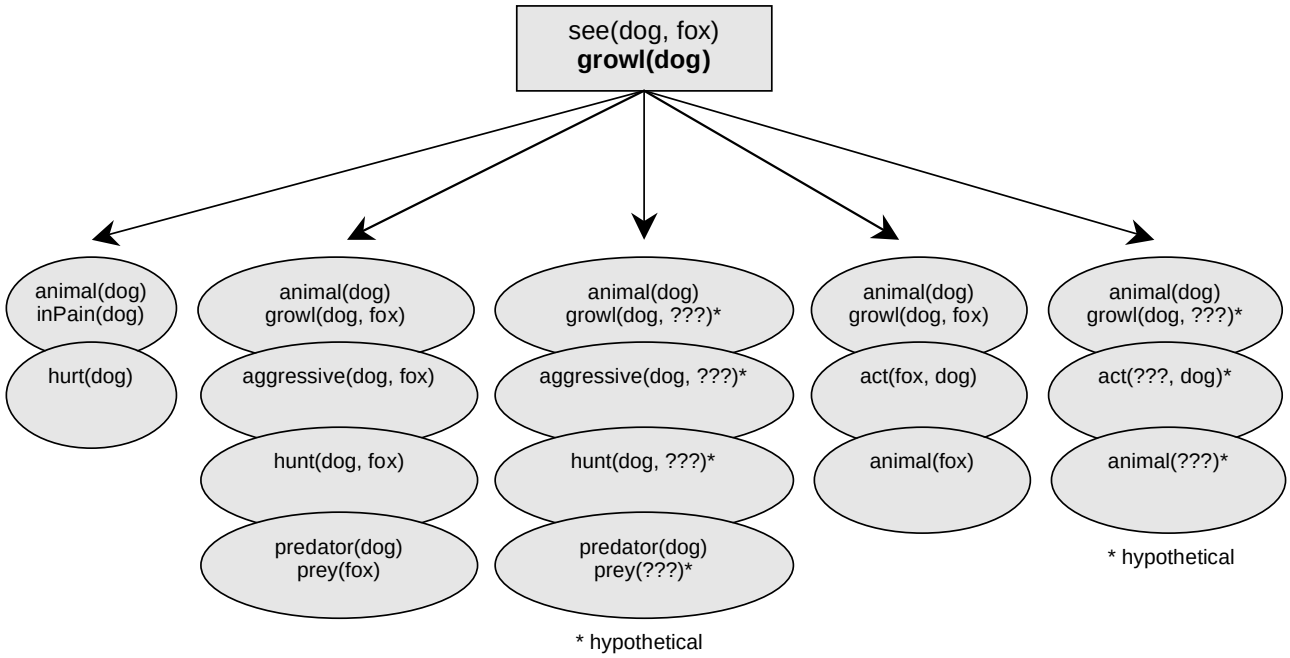

Figure 1: Form of scenario representation created by PAM in the comprehension stage for the scenario "*The pack saw the fox. The hounds growled.*" – it is then analysed in the assessment stage to extract variables and determine plausibility.

The Plausibility Analysis Model (PAM) is a computational implementation of the Knowledge-Fitting Theory. In the next section, we outline PAM's workings and describe how its key plausibility variables are extracted and used in its plausibility assessment function. In the remainder of the paper, we elaborate a sensitivity analysis of PAM. This sensitivity analysis shows that PAM's various components each play a necessary and important role in its utility as a robust model of plausibility.

PAM: The Plausibility Analysis Model

PAM is a computational implementation of the Knowledge-Fitting Theory of Plausibility, detailed elsewhere (Connell & Keane 2003, in sub.). The model takes sentences as inputs and outputs a plausibility rating (from 0 – 10) for the scenario described in the sentences. PAM implements both the comprehension and assessment stages of the Knowledge-Fitting Theory, using its knowledge base to model concept-coherence and provide judgements of plausibility that reflect those made by people.

Comprehension Stage The role of the comprehension stage is to create a conceptual representation of the scenario. To do this, PAM uses a simple parsing mechanism to break down each sentence into propositional form, and then makes the inferences between the sentences by fitting their propositions to information in the knowledge base.

PAM's knowledge base is organised 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, the scenario "The pack saw the fox. The hounds growled." in propositional form is *see(pack, fox), growl(hounds)*. To represent this scenario, PAM must check the conditions of each proposition as it is defined in the knowledge base. The *see* predicate requires that its first argument is an *animal* (i.e., something must be an animal in order to see), and since 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 the *see* predicate is met. Also, the *see* predicate requires that its second argument is a non-abstract entity (i.e., something must be non-abstract in order to be seen). Since 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* (see Figure 1).

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. Figure 1 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 which 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

[1] All entries in the knowledge base were added in a "blind" fashion; that is, each entity and predicate was defined as thoroughly as possible without reference to the sentence pairs that make up the simulations. In total, this resulted in a knowledge base consisting of several hundred entities and predicates.

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 Figure 1). 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. In this respect, PAM models group behavior in plausibility judgement; rather than limit 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. Indeed, it is the fundamental point of PAM that plausibility is based on some assessment of these diverse inferential possibilities between events.

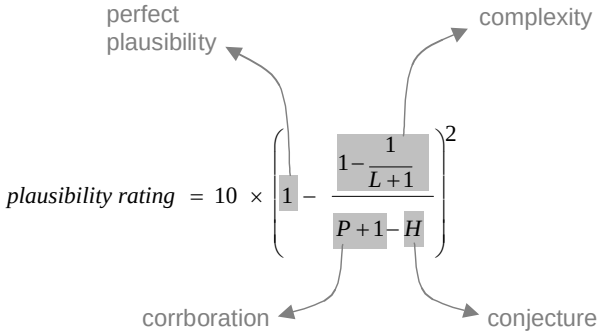

Figure 2: PAM's formula for plausibility ratings (P = total number of paths, H = proportion of hypothetical paths, L = mean path length).

Assessment Stage When the comprehension stage is completed, it is the role of the assessment phase to analyse the structure of the path representation to calculate the plausibility of the scenario. PAM's analysis extracts three main variables from the representation (see Figure 2) and uses them to calculate plausibility by applying a function that finds the quality of the knowledge fit (i.e. the scenario's concept-coherence).

1. Total Number of Paths (P). This is quantified as the number of different paths in the representation. It reflects the number of different ways the sentence conditions can be met in the knowledge base, and relates theoretically to the *corroboration* of the scenario by prior knowledge.

2. Mean Path Length (L). This 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 conditions must be met per path, and relates theoretically to the *complexity* of the scenario's explanation.

3. Proportion of Hypothetical Paths (H). This is quantified as the number of paths that contain a condition with a hypothetical argument, divided by P. It reflects the proportion of all paths that contain a condition that was only met by assuming the existence of something not explicitly mentioned, and relates theoretically to the *conjecture* involved in inferring connections between a scenario events.

Each of these variables is motivated by the underlying theory (see Connell & Keane, in sub.), and contributes to plausibility. 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* is important to modelling the plausibility judgements of a group of people, because it represents the prior knowledge corroboration of the variety of ways in which the events in the scenario may be connected. Finally, the hypotheticality variable *H* is also important, because it represents how conjecture makes any scenario less plausible.

It has been demonstrated in simulations (Connell & Keane, 2003, in sub.) that using this approach, PAM's performance is close to human judgements. In these simulations, PAM's output was compared to human responses by running the model on the same sentence pairs presented to human participants in experiments reported by Connell and Keane (2004). Across a wide range of scenarios, PAM's ratings were shown to correlate highly with human plausibility judgements (r=0.78, r^2=0.61, p<0.0001, N=60). In addition, Table 1 gives the mean ratings for scenarios that invite different types of inference, comparing those produced by people to those ratings produced by PAM. Both people and PAM rated events linked by *causal* connections (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), and lastly, scenarios containing *unrelated* events are rated least plausible of all.

It is important to note that these effects emerge from the operation of PAM's plausibility function (see Figure 2) and that there is no hard-coded classification of the different scenarios in the input representations or the knowledge base. The reason that PAM produces distinctly different plausibility ratings for different types of inference is that each inference type tends towards certain values for each of the extracted variables. For example, the most plausible scenario will have a high number of paths, a low proportion of hypothetical paths (H), and a low path length (L), and the

Table 1: Mean plausibility ratings per inference type (showing sentence pair examples) as produced by participants and by PAM, on a scale from 0 (implausible) to 10 (very plausible).

Inference Type	Example Sentence Pair	Human Rating	Model Rating
Causal	The breeze hit the candle. The candle flickered.	7.8	8.3
Attributal	The breeze hit the candle. The candle was pretty.	5.5	6.1
Temporal	The breeze hit the candle. The candle shone.	4.2	5.5
Unrelated	The breeze hit the candle. The candle drowned.	2.0	1.5

interaction of these three variables produces the causal > attributal > temporal > unrelated ranking. It is the robustness of this interaction that we now turn to in the sensitivity analysis.

Sensitivity Analysis

Sensitivity analysis is a useful tool in cognitive modelling, allowing the designer to examine if the model is consistent with its underlying theory and to test the robustness of the model in a variety of operational contexts. PAM uses three key parameters in modelling plausibility judgements, inviting the criticism that that they are all not really required to achieve predictive accuracy. If one or more of the variables (number of paths *P*, mean path length *L*, proportion of hypothetical paths *H*) is not making a significant contribution to PAM's performance, then a much more parsimonious model may exist for computing plausibility. This state of affairs could, in turn, have complexity implications for any use of the model for other, related tasks.

Furthermore, as Cooper at al. (1996) have argued, 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 the A|B distinction for cognitive models). For these reasons, we performed a sensitivity analysis in which we systematically vary the contribution of each variable to the plausibility function to determine whether there was any resultant degradation in PAM's ability to simulate human performance.

Analysis 1: Contribution of Variables

First, it is useful to ascertain 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) of each variable's possible values, and calculating the resulting plausibility rating for each combination of values. An illustration of this space can be seen in Figure 3, showing PAM's plausibility ratings for an increasing number of paths (P) and for increasing path complexity (L), with separate planes for the best case (no paths hypothetical) and worst case (all paths hypothetical) values for the hypotheticality variable H.

The relative contribution of each variable to PAM's plausibility function can then be determined by applying a multiple nonlinear regression analysis to the set of plausibility ratings, using PAM's own plausibility formula (see Figure 2) as the regression equation, and observing the standardised regression coefficient *beta* for each of the predictor variables (*P*, *L* and *H*). Regression shows that the total number of paths *P* and mean path length *L* contribute equally to PAM's plausibility function (*P beta*=2.193, $p<0.0001$; *L beta*=2.193, $p<0.0001$) [2]. The proportion of hypothetical paths *H* is less important but is still a significant contributor (*H beta*=0.161, $p<0.0001$). This analysis confirms that each variable (*P*, *L*, *H*) fulfils a necessary role in PAM's plausibility function, and so we may now examine the robustness of the function's performance.

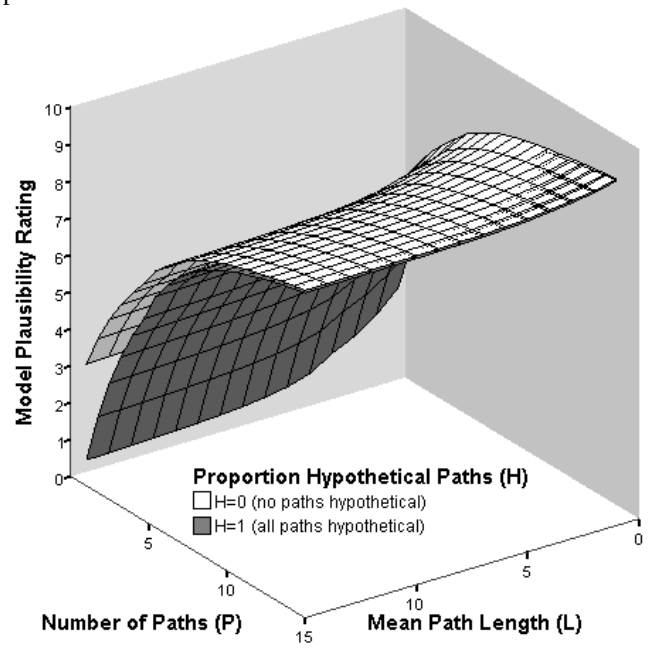

Figure 3: Three-dimensional illustration of PAM's plausibility rating function for the variables *P* and *L*, with *H*'s max (1) and min (0) values as separate planes. Note how the impact of *L* and *H* decreases as *P* values increase.

Analysis 2: Robustness of Model

In the following analysis, we test how sensitive PAM's performance is to changes in each variable's contribution. 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. To perform the sensitivity analysis, we re-ran the simulation reported previously (Connell & Keane, 2003, in sub.) while systematically varying the weight of each variable in the plausibility rating function. We then examine the resulting correlations and whether the model's performance satisfies the data. If PAM's modelling 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

[2] Regression was performed through the origin because PAM's plausibility rating function has no constant term (i.e. no intercept).

Table 2: Sensitivity Analysis for variables Total Number of Paths (*P*) and Mean Path Length (*L*) showing correlation between model and human plausibility ratings [a].

Weight for *L*	Weight for *P*								
	1%	25%	50%	75%	100%	125%	150%	175%	200%
1%	0.435	0.432	0.366	0.343	0.331	0.324	0.319	0.316	0.313
25%	-0.662	0.710	0.772	0.745	0.702	0.658	0.619	0.586	0.557
50%	-0.677	0.123	0.784	0.786	0.765	0.735	0.703	0.673	0.645
75%	-0.686	-0.355	0.765	0.789	0.777	0.755	0.730	0.704	0.679
100%	-0.692	-0.495	0.731	0.782	0.776	0.759	0.738	0.716	0.693
125%	-0.696	-0.553	0.687	0.774	0.772	0.759	0.740	0.720	0.700
150%	-0.699	-0.584	0.639	0.764	0.767	0.756	0.740	0.722	0.703
175%	-0.701	-0.603	0.589	0.755	0.762	0.753	0.738	0.721	0.704
200%	-0.703	-0.616	0.541	0.746	0.757	0.749	0.736	0.720	0.704

Table 3: Sensitivity Analysis for variables Total Number of Paths (*P*) and Proportion of Hypothetical Paths (*H*) showing correlation between model and human plausibility ratings [a].

Weight for *H*	Weight for *P*								
	1%	25%	50%	75%	100%	125%	150%	175%	200%
1%	0.051	0.575	0.615	0.622	0.619	0.611	0.601	0.590	0.578
25%	0.146	0.639	0.670	0.670	0.660	0.647	0.633	0.618	0.604
50%	-0.015	0.680	0.720	0.717	0.704	0.686	0.668	0.650	0.632
75%	-0.521	0.639	0.752	0.756	0.743	0.725	0.704	0.683	0.662
100%	-0.692	-0.495	0.731	0.782	0.776	0.759	0.738	0.716	0.693
125%	0.212	-0.655	-0.409	0.770	0.798	0.788	0.769	0.748	0.724
150%	0.134	-0.174	-0.655	-0.385	0.791	0.806	0.794	0.776	0.754
175%	-0.006	-0.043	-0.676	-0.655	-0.378	0.803	0.811	0.798	0.780
200%	0.155	-0.363	-0.174	-0.637	-0.655	-0.376	0.809	0.812	0.800

Table 4: Sensitivity Analysis for variables Mean Path Length (*L*) and Proportion of Hypothetical Paths (*H*) showing correlation between model and human plausibility ratings [a].

Weight for *H*	Weight for *L*								
	1%	25%	50%	75%	100%	125%	150%	175%	200%
1%	0.311	0.525	0.595	0.615	0.619	0.617	0.614	0.610	0.606
25%	0.314	0.560	0.635	0.656	0.660	0.659	0.655	0.652	0.648
50%	0.317	0.602	0.680	0.700	0.704	0.702	0.698	0.694	0.690
75%	0.323	0.650	0.724	0.741	0.743	0.741	0.736	0.732	0.728
100%	0.331	0.702	0.765	0.777	0.776	0.772	0.767	0.762	0.757
125%	0.344	0.753	0.796	0.802	0.798	0.792	0.785	0.777	0.771
150%	0.369	0.793	0.815	0.809	0.791	0.765	0.734	0.701	0.667
175%	0.437	0.811	0.402	-0.179	-0.378	-0.461	-0.505	-0.532	-0.550
200%	-0.605	-0.625	-0.639	-0.648	-0.655	-0.659	-0.662	-0.665	-0.666

[a] Shaded areas represents region of weights that consistently satisfy the data for all combinations of variables.

by shading whether those weights satisfy the data. Table 2 shows the results of PAM's sensitivity analysis for the variables *P* (number of paths) and *L* (mean path length). Table 3 shows the sensitivity analysis for the variables *P* and *H* (proportion of hypothetical paths), and Table 4 shows the sensitivity analysis for the variables *L* and *H*.

The sensitivity analysis shows us that there is a key region that satisfies the data, roughly corresponding to where weights for *P*, *L* and *H* are between 50%-150%. The total region that satisfies the data is indicated by the shaded areas in Tables 2-4. 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 towards extremes. Indeed, much lower (and even negative) correlations are observed when the variables *P*, *L* and *H* 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 at all three variables were systematically tested. The highest correlation found for a combination of weights that satisfied the data was r=0.818 (r^2=0.669), where *P* was weighted at 150%, *L* at 75% and *H* 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 and hence these weight values will not be adopted in PAM's plausibility function so as to preserve its generalisability 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 proportion of hypothetical paths *H* – 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.

General Discussion

In this paper, we have performed a sensitivity analysis of PAM, the Plausibility Analysis Model. This analysis has shown us that the highlighted key variables in our plausibility function are, in fact, the key variables in modelling plausibility judgements. Furthermore, these variables are those motivated by the Knowledge-Fitting Theory. Having determined that the model is suitably robust as a characterisation of plausibility, we have fulfilled a necessary prerequisite for the broader application of PAM in a variety of other cognitive systems.

Given the large contribution of the variable *P*, 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 Figure 3, 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 increases in the proportion of hypothetical paths (*H*) 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. Indeed, parameters for this threshold could be grounded in empirical data of actual explanations given by people about event connectivity. This would allow PAM to maintain its level of performance as its knowledge base grows.

The factors of corroboration, complexity and conjecture were described by the Knowledge-Fitting Theory as being important to plausibility and were implemented computationally in PAM (Connell & Keane, 2003, in sub.). This paper has shown, in sensitivity analysis of the model, that PAM's plausibility function is indeed robust and that all three factors are vital to plausibility estimation. Any future models of human plausibility judgement, or models of cognitive tasks that utilise plausibility in a broader context, should take account of these three factors and the interactions between them.

Acknowledgments

This work was funded in part by a grant to the first author from the Irish Research Council for Science, Engineering and Technology under the Embark Initiative. Thanks also to Dermot Lynott for invaluable comments.

References

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

Connell, L. & Keane, M. T. (2003). PAM: A cognitive model of plausibility. *Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society,* 264-269.

Connell, L., & Keane, M. T. (2004). What plausibly affects plausibility? Concept-coherence & distributional word-coherence as factors influencing plausibility judgements. *Memory and Cognition, 32* (2), 185-197.

Connell, L., Keane, M. T. (in sub.). A model of plausibility. *Manuscript in submission.*

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

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

Lemaire, P. & Fayol, M. (1995). When plausibility judgments supersede fact retrieval: The example of the odd-even rule effect in simple arithmetic. *Memory and Cognition, 23,* 34-48.

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

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., 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.

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

Speer, S. R., & Clifton, C. (1998). Plausibility and argument structure in sentence comprehension. *Memory and Cognition, 26*(5), 965-978.