Jaakko Hintikka
Boston University

and

Ilpo Halonen
University of Helsinki

INTERPOLATION AS EXPLANATION

In the study of explanation, one can distinguish two main trends. On the one hand, some philosophers think — or at least used to think — that to explain a fact is primarily to subsume it under a generalization. And even though most philosophical theorists of explanation are not enchanted with this idea any longer, it is alive and well among theoretical linguists. We will call this kind of view on explanation a subsumptionist one. It culminated in Hempel's deductive-nomological model of explanation, also known as the covering law model.

On the other hand, some philosophers see the task of explaining a fact essentially as a problem of charting the dependence relations, for instance causal relations, that connect step by step the explanandum with the basis of the explanation, with "the given," as it is sometimes put. We will call this kind of approach an interactionist one. Among the most typical views of this kind, there are the theories according to which explaining means primarily charting the causal structure of the world.

What logical techniques are appropriate as the main tools of explanation on the two opposing views? The answer is clear in the case of explanation by subsumption. There the crucial step in explanation is typically a generalization that comprehends the explanandum as a special case. The logic of Hempelian explanation is thus essentially the logic of general implications (covering laws).

But what is the logic of interactionist explanation? Functional dependencies can be studied in first-order logic, especially in the logic of relations. But what kinds of results in the (meta)theory of first-order logic are relevant here? There undoubtedly can be more than one informative answer to this question. In this paper, we will concentrate on one answer which we find especially illuminating. It highlights the promise of Craig's interpolation theorem as a tool for the study of explanation.

What Craig's theorem says is reasonably well known, even though it has not found a slot in any introductory textbook of logic for philosophers. It can be stated as follows, using the turnstile – as a metalogical symbol to express logical consequence:

Craig's Interpolation Theorem. Assume that F and G are (ordinary) first-order formulas, and assume that

(i) F – G

(ii) not – ~F

(iii) not – G

Then there exists a formula I (called an interpolation formula) such that

(a) F – I

(b) I – G

(c) The nonlogical constants and free individual variables of I occur both in F and in G.

It is the clause (c) that gives Craig's interpolation theorem its cutting edge, for (a)-(b) would be trivial to satisfy without (c).

The nature and implications of Craig's interpolation theorem are best studied with the help of some of the other basic results concerning first-order logic. First, from the completeness of ordinary first-order logic it follows that because of (i) there exists an actual proof of $(F \supset G)$ (or a proof of G from F) in a suitable axiomatic formulation of first-order logic. From Gentzen's first *Hauptsatz* it follows that those suitable complete formulations include cut-free methods like

Beth's *tableau* method, which of course is only a mirror image of suitable cut-free variants of Gentzen's sequent calculus.

In this paper, we will use a particularly simple variant of the *tableau* method. It is characterized by the fact that in it no movement of formulas between the left and the right side of the *tableau* is allowed. The negation-free subset of the *tableau* rules can be formulated as follows where λ is the list of formulas on the left side of some *subtableau* and μ similarly the list of formulas on the right side:

(L.&) If $(S_1 \& S_2) \in \lambda$, add S_1 and S_2 to λ.

(L.$\vee$) If $(S_1 \vee S_2) \in \lambda$, you may start two *subtableaux* by adding S_1 or S_2, respectively, to λ.

(L.E) If $(\exists x)S[x] \in \lambda$ and if there is no formula of the form $S[b] \in \lambda$, you may add $S[d]$ to λ, where d is a new individual constant.

(L.A) If $(\forall x)S[x] \in \lambda$ and if b occurs in the same *subtableau*, you may add $S[b]$ to λ.

Right-hand rules (R.&), (R.$\vee$), (R.C) and (R.A) are duals (mirror images) of these rules. For instance,

(R.&) If $(S_1 \& S_2) \in \mu$, then you may start two *subtableaux* by adding S_1 or S_2, respectively, to μ.

Negation can be handled by the following rewriting rules.

Rewrite

(R.R)	$\sim\sim S$	as	S
	$\sim(S_1 \vee S_2)$	as	$(\sim S_1 \ \& \sim S_2)$
	$\sim(S_1 \ \& \ S_2)$	as	$(\sim S_1 \vee \sim S_2)$
	$\sim(\exists x)$	as	$(\forall x)\sim$
	$\sim(\forall x)$	as	$(\exists x)\sim$

By means of these rewriting rules, each formulate can effectively be brought to a negation normal form in which all negation signs are prefixed to atomic formulas or identities.

We will abbreviate $\sim(a = b)$ by $(a \neq b)$.

As derived rules (construing $(S_1 \supset S_2)$ as a rewritten form of $(\sim S_1 \vee S_2)$ we can also have

(L.⊃) If $(S_1 \supset S_2) \in \lambda$, add $\sim S_1$ or S_2 to λ., starting two *subtableaux*

(R.⊃) If $(S_1 \supset S_2) \in \mu$, add $\sim S_1$ and S_2 to μ.

For identity, the following rules can be used:

(L.self =) If b occurs in the formulas on the left side, add (b =b) to the left side.

(L.=) If S[a] and (a = b) occur on the left side, add S[b] to the left side.

Here S[a] and S[b] are like each other except that some occurrences of a or b have been exchanged for the other one.

(R.self =) and (R.=) are like (L.self =) and (L.=) except that = has been replaced by its negation ≠.

As was stated, it can be shown that if F – G is provable in first-order logic, it is provable by means of the rules just listed. A proof means a *tableau* which is closed. A *tableau* is said to be closed if and only if the following condition is satisfied by it:

There is a bridge from each open branch on the left to each open branch on the right.

A bridge means a shared atomic formula.

A branch is open if and only if it is not closed.

A branch is closed if it contains a formula and its negation.

In the interpolation theorem, we are given a *tableau* proof of F – G. The crucial question is how an interpolation is found on the basis of this *tableau*. For the purpose, it may be noted that because of the assumptions of the interpolation theorem there must be at least one branch open on the left and at least one on the right. It can also be assumed (ii) that all the formulas are in the negation normal form.

Then an interpolation formula can be constructed as follows:

Step 1. For each open branch on the left, form the conjunction of all the formulas in it that are used as bridges to the right.

Step 2. Form the disjunction of all such conjunctions.

Step 3. Beginning from the end (bottom) of the *tableau* and moving higher up step by step, replace each constant introduced (i) from the right to the left by an application of (U.1) on the left or (ii) from the left to the right by an application of (E.1) on the right by a variable, say x, different for different applications.

In the case (i), moving from the end of the *tableau* upwards, prefix (∃x) to the formula so far obtained. In the case (ii), prefix (∀x) to the formula so far obtained.

It can be proved that the formula obtained in this way is always an interpolation formula in the sense of Craig's theorem. We will not carry out the proof here. Instead, we will illustrate the nature of the interpolation formula by means of a couple of examples.

Consider first the following closed *tableau* and the interpolation formula it yields:

(1.1)	$(\forall x)L(x,b)$	(1.2)	$(\forall y)(L(b,y) \supset m = y) \supset (m = b)$	
	Initial premise		Ultimate conclusion	
(1.6)	$L(b,b)$	(1.3)	$\sim(\forall y)(L(b,y) \supset m = y)$	
	from (1.1) by (L.A)		from (1.2) by (R.$\supset$)	
		(1.4)	$m = b$	
			from (1.2) by (R.$\supset$)	
		(1.5)	$(\exists y)(L(b,y) \,\&\, (m \neq y))$	
			from (1.3) by rewrite rules	
		(1.7)	$L(b,b \,\&\, (m \neq b)$	
			from (1.5) by (R.E)	
		(1.8)	$L(b,b)$	(1.9) $m \neq b$
			from (1.7) by (R.&)	from (1.7)
			bridge to (1.6)	by (R.&)
				closure by
				(1.4),(1.9)

The interpolation formula is

(1) $L(b,b)$

This example may prompt a *déjà vu* experience in some of our readers. If you interpret $L(x,y)$ as *x loves y*, b as *my baby*, and m as *me*, we get a version of the old introductory logic book chestnut. The initial premise says

(2) Everybody loves my baby

and the conclusion

(3) If my baby only loves me, then I am my baby.

Textbook writers cherish this example, because it gives the impression of expressing a clever, nontrivial inference. In reality, their use of this example is cheating, trading on a mistranslation. In ordinary usage, (2) does not imply that my baby loves him/herself. Its correct translation is therefore

(4) $(\forall x)(x \neq b \supset L(x,b))$

which does not any longer support the inference when used as the premise. Hence the inference from the initial premise to the ultimate conclusion turns entirely on the premise's implying $L(b,b)$. This is what explains the inference. Predictably, $L(b,b)$ is also the interpolation formula.

Consider next another example of a closed *tableau* and its interpolation formula.

2.

(2.1)	$(\forall x)((A(x) \& C(x)) \supset B(x))$ IP			(2.3)	$(\forall x)(Ex \supset (A(x) \supset B(x)))$ UC
(2.2)	$(\forall x)((D(x) \vee {\sim}D(x)) \supset C(x))$ IP			(2.4)	$(E(\beta) \supset (A(\beta) \supset B(\beta)))$
(2.9)	$(A(\beta) \& C(\beta)) \supset B(\beta)$			(2.5)	${\sim}E(\beta)$
(2.10)	${\sim}(A(\beta) \& C(\beta))$	(11)	$B(\beta)$	(2.6)	$(A(\beta) \supset B(\beta))$
			bridge	(2.7)	${\sim}A(\beta)$
(2.12)	${\sim}C(\beta)$	(2.13)	${\sim}A(\beta)$	(2.8)	$B(\beta)$
			bridge		
(2.14)	$(D(\beta) \vee {\sim}D(\beta) \supset C(\beta)$				
(2.15)	${\sim}(D(\beta) \vee \neg D(\beta))$	(2.16)	$C(\beta)$		
			closure		
(2.17)	${\sim}D(\beta)$				
(2.18)	${\sim\sim}D(\beta)$				
	closure				

The interpolation formula is, as one can easily ascertain

(5) $(\forall x)({\sim}Ax \vee Bx)$

In this example, the first initial premise says that whatever in both C and A is also B. But the second initial premise entails that everything is C anyway. Hence the cash value of the initial premises is that if anything is A, it is also B.

The conclusion says that if any E is A, then it is B. But any A is B anyway, in virtue of the initial premises. Hence the component of the premises that suffices to imply the conclusion is

(6) $(\forall x)(A(x) \supset B(x))$

which is equivalent with the interpolation formula (5).

Again, the interpolation formula expresses the reason ("explanation") why the logical consequence relation holds.

Such examples bring out vividly the fact that interpolation theorem is already by itself a means of explanation. Suitable normalized interpolation formulas show the deductive interface between the premise and the conclusion and by so doing help to explain why the deductive relationship between them holds.

This result has remarkable consequences. For one thing, some philosophers of mathematics have discussed the question as to whether there can be explanations in mathematics. This question can now be answered definitively in the affirmative. The fact that mathematical arguments proceed purely deductively is no barrier to explanation, for it was just seen that there can be explanations of deductive relationships, too. For instance, in the sense involved here, the unsolvability by radicals of the general equation of the fifth degree is explained by the fact that a certain group which "from the outside" can be characterized as the Galois group of the general equation of the fifth degree, is symmetrical.

But in what sense does the interpolation theorem offer an analysis of functional dependencies? Here the way the interpolation formula is constructed (see above) provides an answer. We can think of the premise F (and its consequences on the left side) and of the conclusion G (together with its consequences on the right side) as each describing a certain kind of configuration. Each quantifier in the interpolation formula comes from an application of a general truth (expressed on the left by a universal sentence and on the right by an existential sentence) to a individual that is before the application occurred only on the other side. Since there is no other traffic between the two sides of a *tableau*, these applications are the only ones which show how the two configurations depend on each other so as to bring about the logical consequence relation between them. No wonder that they are highlighted by the interpolation formula, and no wonder that the interpolation formula in a perfectly natural sense serves to explain why the consequence relation holds.

The character of the crucial instantiation steps as turning on functional dependencies can be seen even more clearly when existential quantifiers on the left and universal quantifiers on the right are replaced by the corresponding Skolem functions. Then one of the crucial instantiation steps will express the fact that an ingredient of (individual in) one of the two configurations depends functionally of an ingredient of (individual in) the other configuration. Moreover, these are all the functional dependencies that have to be in place for the logical consequence to be valid.

Essentially the same point can be expressed by saying that the interpolation formula is a summary of the step-by-step functional dependencies that lead from the situation specified by the premises to the situation described by the conclusion. This role by the interpolation formula can be considered as the basic reason why it serves as a means of explanation.

At the same time, we can see the limitations of Craig's interpolation theorem as a tool of an interactionist theory of explanation. A suitable interpolation formula I explains why G follows from F by showing how the structures specified by F interact with the structures specified by G so as to make the consequence inevitable. What it does not explain is how the internal dynamics of each of the two types of structures contributes to the consequence relation. It is for this reason that nonlogical primitives not occurring in both F and G cannot enter into I: they cannot play any role in the interaction of the two structures.

In the general theory of explanation the interpolation theorem can thus be used insofar as the explanation of an event, say one described by E, can be thought of as depending on two different things, on the one hand on some given background theory and on the other hand on contingent *ad hoc* facts concerning the circumstances of E.

Both the background theory and the contingent “initial conditions” specify a kind of structure. An explanation is an account of the interplay between these two structures.

Not surprisingly, the interpolation theorem turns out to be an extremely useful tool in the general theory of explanation. Here we can only indicate the most important result. (Cf. here Hintikka and Halonen, 1995.) In explaining some particular event, say that P(b), we have available to us some background theory T and certain facts A about the circumstances of the *explanandum*. The process of explanation will then consist of deducing the explanandum from T & A. If it can be assumed that T = T[P] does not depend on b and A = A[b] does not depend on P, we can apply (barring certain degenerate cases) in two different ways and to obtain two different interpolation formulas H[b] and I[P] such that

(7) $A[b] \text{ – } H[b]$

(8) $T[P] \text{ – } (\forall x)(H[x] \supset P(x))$

(9) P does not occur in H[b]

(10) $T[P] \text{ – } I[P]$

(11) $I[P] \text{ – } (\forall x)(A[x] \supset P(x))$

(12) b does not occur in I[P]

Even though we have not made any assumptions concerning explanation as subsumption, it turns out that as a by-product of a successful explanation we obtain not only one, but two covering laws, viz.

(13) $(\forall x)(H[x] \supset P(x))$

(cf. (8)) and

(14) $(\forall x)(A[x] \supset P(x))$

(cf. (11)). They have distinctly different properties even though they have been fallaciously assimilated to each other in previous literature. Neither of these covering laws can nevertheless be said to provide an explanation why the explanandum holds. These covering laws are by-products of successful explanatory arguments, not the means of explanation. Insofar as we can speak of *the* explanation here either the "antecedent conditions" H[b] or the "local law" I[P] can be considered as a viable candidate. This means of course that the notion of "*the* explanation" is intrinsically ambiguous. In particular, the reason why H[x] can claim to the status of an explanation is precisely what was pointed out earlier in this paper, viz. that it brings out the interplay of the situations specified by A[b] and $(T[P] \supset P(b))$.

From these brief indications it can already be seen that the interpolation theorem is destined to play a crucial role in any satisfactory theory of explanation. Why it can do so is explained in the earlier parts of this paper.

These remarks point also to an interesting methodological moral for philosophers of science. There is a widespread tendency to consider logical relationships from a purely syntactical matter, as relations between the several propositions of a theory or even between the sentences expressing such propositions. The entire structuralist movement is predicated on this assumption. This tendency is a fallacious one, however. The propositions considered in logic have each a clear-cut model-theoretical import. Spelling out this import can serve to clarify those very problems which philosophers of science are interested in, for reasons that can be spelled out (see Hintikka, forthcoming (b)). This model-theoretical meaning can be seen more directly when cut-free proof methods (such as the tree methods or the *tableau* method) are used, just as they are assumed to be used in the present paper.

Another example of how model-theoretical considerations can put issues in the philosophy of science is offered by Hintikka (forthcoming (a)).

REFERENCES

Craig, W. (1957), “Three Uses of the Herbrand-Gentzen Theorem in Relating Model Theory and Proof Theory,” *Journal of Symbolic Logic* vol. 22, p. 269-285.

Hintikka, J. (forthcoming (a)), “Ramsey Sentences and the Meaning of Quantifiers,” *Philosophy of Science*.

Hintikka, J. (forthcoming (b)), “The Art of Thinking: A Guide for the Perplexed.”

Hintikka, J. and I. Halonen (1995), “Semantics and Pragmatics for Why-questions,” *Journal of Philosophy* vol. 92, pp. 636-657.

Hintikka, J. and I. Halonen (forthcoming), “Toward a Theory of the Process of Explanation.”