Notre Dame Journal of Formal Logic
Volume XX, Number 2, April 1979
NDJFAM

A MODEL-THEORETIC ACCOUNT OF CONFIRMATION

JOHN PAULOS

Ever since Hempel's 1946 paper on confirmation, the subject and its paradoxes have attracted many investigators. The paper attempted to come to some reasonable understanding of confirmation, specifically to list some axioms or properties satisfied by the notion. Problems arose from even the most seemingly modest axioms and properties however. Our paper attempts both to provide a clearer understanding of confirmation and to "explain away the paradoxes". Notions from model theory provide us with a very useful and flexible framework for this project.

Before we embark on our model-theoretic construction, we will list here some of the axioms Hempel considered in his 1946 paper and some of the paradoxes they engender.

1) If evidence e confirms h and $h \rightarrow k$, then e confirms k.
2) If evidence e confirms k and $h \rightarrow k$, then e confirms h.
3) Any instance of a universal statement confirms the statement.
4) If e confirms h and $h \leftrightarrow k$, then e confirms k.

From 1 and 2 we can derive the following. Let h = the theory of relativity and k = the thermostat is above 80°. Then its being hot in this room tends to confirm k (on any intuitive understanding of confirmation) and thus tends to confirm $h \wedge k$ by 2 since $h \wedge k \rightarrow k$. Hence by 1) h is confirmed since $h \wedge k \rightarrow h$. Clearly something is wrong with 1 and 2 together. 2 seems especially suspect, but some (weakened) version of 2 is certainly often used and is necessary in the everyday practice of science. Even 1, as we shall see, is not always the case.

The following, the so-called raven paradox, derives from 3 and 4. A black raven confirms the hypothesis that all ravens are black by 3. Equivalent to this hypothesis is the hypothesis that all non-black things are non-ravens. Any non-black non-raven, a green table say, confirms this latter hypothesis. Hence by 4, observation of this green table also confirms that all ravens are black!

Received November 12, 1976

These paradoxes result, in part, from an attempt to impose the deductive technique on inductive problems. A natural place to look for clarification is the theory of probability, and a natural definition for "e confirms h" is the following.

Tentative Definition: e confirms h iff $\mathrm{P}(h/e) > \mathrm{P}(h)$.

There are three problems with this definition.

1) What sort of things are h and e?
2) On what σ-algebra of sets (classes) is the probability measure defined?
3) What criteria for assigning probabilities do we use?

We will in the sequel solve the first and the second problem and say something (more or less vacuous) about the third. We will show clearly how the paradoxes we constructed are resolved. Other anomalies (e can confirm h_1 and h_2 yet disconfirm $h_1 \wedge h_2$ or $h_1 \vee h_2$) will also be discussed and resolved. We begin our formal development

Definition 1 A language $\mathcal{L}$ of the form $K \cup \{\varepsilon\}$ where $K = \{c_j, R_j, f_j, Q_j\}$ is a finite collection of constant, relation, function, and sort symbols and where ε is a distinguished binary relation symbol. Sentences in $\mathcal{L}$ are built up in the usual inductive manner with clauses for $\neg$, $\vee$, and $\exists$.

We can also deal with languages of the form $\mathcal{L} = K_1 \cup K_2 \cup \{\varepsilon\}$ where the K_i are different formal languages. See [2]. We want, however, to characterize our models independently of the K_i. We do this by 1) taking all our models to be models of a certain set theory (**ZF** set theory for the sake of being definite) and by 2) interpreting the non-logical constants and relations of our languages K_i to be fixed elements of the universe. The intention is to think of the languages K_i as formal scientific languages. The symbol ε is added so that statements in any K_i as well as extra-linguistic observations can be described in some neutral formal language. Although we assume the universe to be set-theoretic, we do not assume that a K_i-speaker "thinks" in terms of sets, but only that an omniscient scientist could analyse the K_i-sentences in set-theoretic terms.

Definition 2 A scientific theory T expressed in a language $\mathcal{L} = K \cup \{\varepsilon\}$ is a set of sentences $T = \mathbf{ZF} \cup S \cup D$ where **ZF** stands for the Zermelo-Frankel axioms of set theory, S is a set of (scientific) sentences expressed in K and D is a finite set of sentences in $\mathcal{L}$ stipulating the set-theoretic type of the K-symbols.

As an example of a sentence from D we have $\forall y(y \in R \rightarrow y = \langle z_1, \ldots, z_n \rangle)$ stating that the element R is an n-ary relation. Similar sentences concerning constant, function, sort, and other relation symbols also appear in D. The semantics of a language $\mathcal{L} = K \cup \{\varepsilon\}$ is described in the following definition.

Definition 3 The universe of a model is any set M large enough to contain an element corresponding to each object, relation, function, etc. in the "world". (It may contain extra elements). That is, we conceive of every

object, property, etc. in the "world" as being associated with an element in M. (Our ontology is intended to be very flexible and for our purposes here need not be made more precise.) The interpretation of each c_j, R_j, f_j, Q_j in $\mathcal{L}$ is a *fixed* element in M while the interpretation of ε can be *any* binary relation on M. Truth of a sentence in a model $\langle M, \ldots \rangle$ is defined in the usual inductive manner.

The last definition needed to complete our preliminaries follows.

Definition 4 An observation e is of the form $\langle \phi(x_1, \ldots, x_n), a_1, \ldots a_n \rangle$ where $\phi(x_1, \ldots, x_n)$ is a formula in the language of set theory, $\{\varepsilon\}$, and where $a_1, \ldots, a_n \varepsilon M$. The a_i need not be named by K.

An observation e is thus a sentence in the ε-diagram of model $\langle M, \ldots \rangle$. The definition of an "observation" e is extra-linguistic, independent of any particular K-language. Though most ordinary observations are expressible in the K-language, others will be expressible only in the ε-diagram of some model $\langle M, \ldots \rangle$.

A final dose of notation is still necessary. The class of all models of **ZF** is denoted simply by $\mathfrak{M}$. All those models associated with any e are denoted $\mathfrak{M}_e$; i.e., $\mathfrak{M}_e = \{\langle M, \ldots \rangle \mid \langle M, \ldots \rangle \models \phi(a_1, \ldots a_n)$, where $\phi(a_1, \ldots a_n)$ is the sentence in the ε-diagram expressing $e\}$. Not required is that every observation be true of the "real world", M_R. For a true observation e, however, we have $M_R \varepsilon \mathfrak{M}_e$. The class of all models of a scientific theory T is denoted $\mathfrak{M}_T$, all models of a sentence γ by $\mathfrak{M}_\gamma$. The notion of the expressibility and delimitability of observations e by K-sentences h is important. e is said to be expressible by a K-sentence h if $\mathfrak{M}_e = \mathfrak{M}_h$. e is said to be delimitable by the K-sentence h if $\mathfrak{M}_e \subseteq \mathfrak{M}_h$.

An example may be illuminating here. The observation, "That book is on the table", is expressed by $e = \{\langle x_1, x_2 \rangle \varepsilon x_3, a_1, a_2, a_3\}$ where a_1 is the element of M associated with the book a_2 is the element associated with the table, and a_3 the element associated with the relation "on". $\mathfrak{M}_e$ is the class of all models in which the book and the table stand in the binary relation "on".

Now we can answer the questions we posed ourselves at the beginning of this paper.

1) What sorts of things are h and e? h is any sentence in the language K and e is an observation (or set of observations).
2) What σ-algebra of sets (classes) shall we use to define our probability measure? The σ-algebra is defined on $\mathfrak{M}$, the class of all models of **ZF**.
3) How do we define a probability measure of $\mathfrak{M}$, the class of all models of the "world"? We assign *any* countably additive set function which satisfies the usual Kolmogorov axioms for probability.

Naturally for the assignment to be reasonable further restrictions are necessary and we may interpret probability as a frequency, subjective, or logical relation (or some combination thereof). In any case we refine our tentative definition as follows.

Definition 5 Given hypothesis h, observation e, and some probability measure P on the σ-algebra $\mathfrak{M}$, e confirms h iff $P(h/e) > P(h)$.

What does all this say about Hempel's original axioms and the paradoxes they engender? Before we answer this fully let's first develop a pictorial representation of the notions just introduced. $\mathfrak{M}$, the class of all models of **ZF** is denoted pictorially by a rectangle. Sentences γ in any K_i partition $\mathfrak{M}$ into 2 regions, the models of **ZF** in which γ is true and those in which it is false (Figure 1). An observation e is represented pictorially by $\mathfrak{M}_e$, the class of models in which the observation holds. (Figure 2)

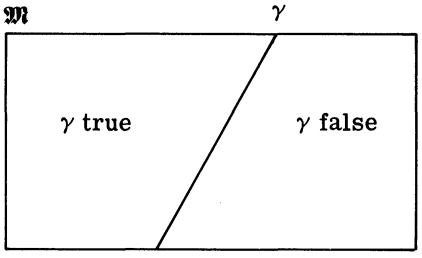


Figure 1

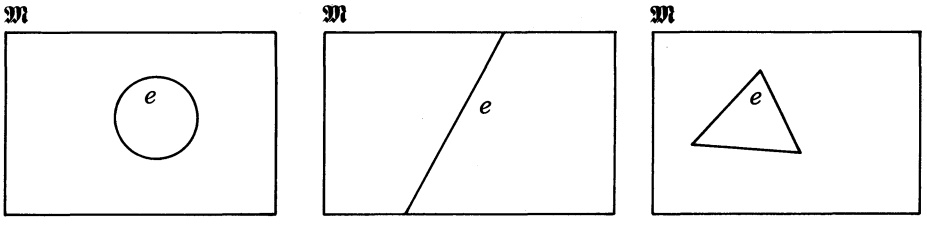


Figure 2

Consider Figure 3. If those models of

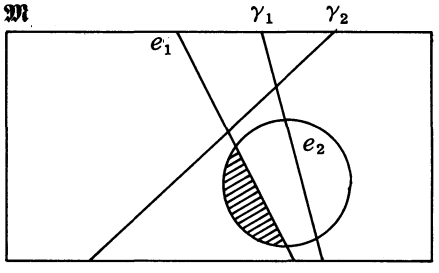


Figure 3

$\mathfrak{M}$ in which γ_1 and γ_2 are true are to the left of the lines marked γ_1 and γ_2, respectively, then γ_1 is true of e_1 while γ_2 is true of some models of $\mathfrak{M}_{e_1}$ and false of others. Moreover γ_2 is false of e_2 while γ_1 is true of some

models of $\mathfrak{M}_{e_2}$ and false of others. Given observations e_1 and e_2, i.e., $M_{e_1} \cap M_{e_2}$, we can conclude that γ_1 is true and γ_2 is false. As a second illustration consider Figure 4 where an observation e_1 is delimited

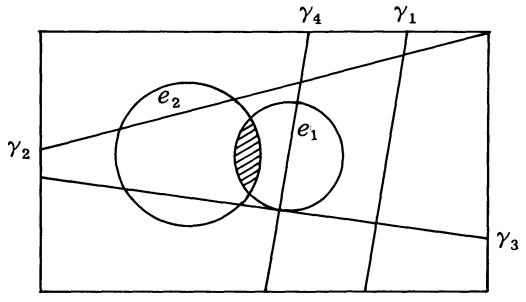


Figure 4

by sentences γ_1, γ_2, and γ_3. γ_4 is true of some models of $\mathfrak{M}_{e_1}$ and false of others. e_1 taken together with e_2 resolves γ_4. Stated differently e_1 determines the truth or falsity of γ_1, γ_2, and γ_3 but not that of γ_4. If the scientist can devise an experiment in which e_2 is a possible outcome, then the observation of e_2 determines the truth of γ_4. Hence e_1 and e_2 taken together decide γ_1, γ_2, γ_3, γ_4.

To illustrate pictorially properties of the notion of confirmation we will assume that the area of a region, $\mathfrak{M}_h$ or $\mathfrak{M}_e$ say, is proportional to its probability. Thus in Figure 5, a through d, we have that e verifies h, confirms h, disconfirms h, and falsifies h, respectively.

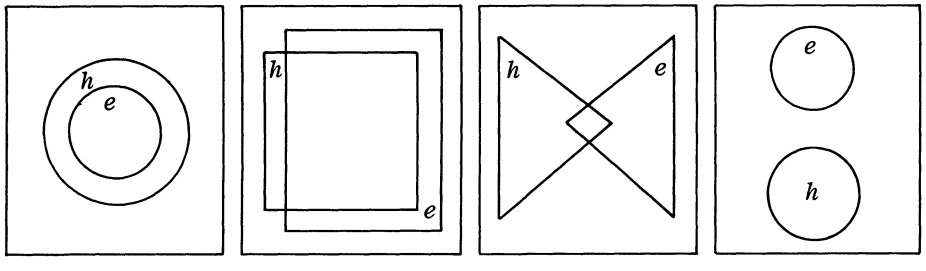
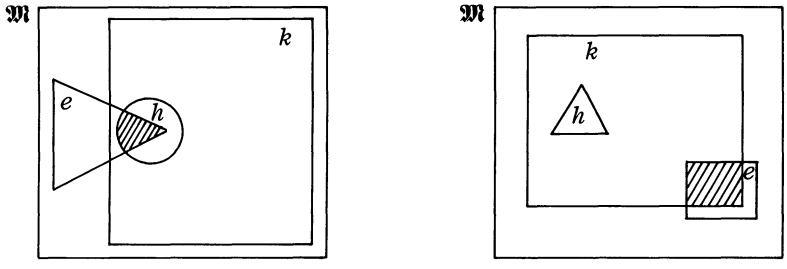


Figure 5

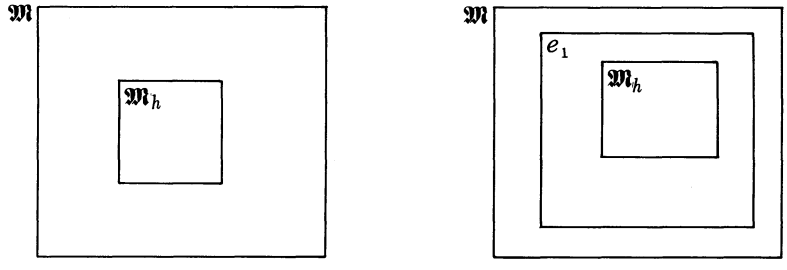
Getting back to Hempel, we see that Figure 6, a and b, shows that neither Hempel's axiom 1 nor his axiom 2 holds on our definition of confirmation.



(a) (b)

Figure 6

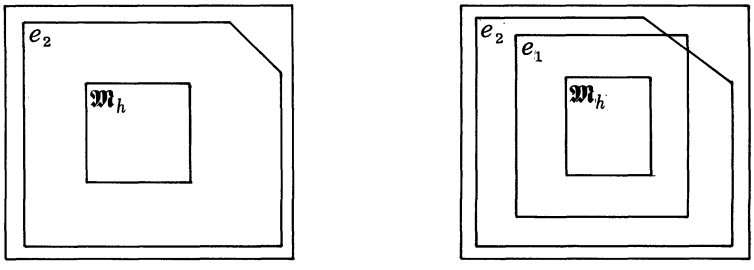
Now for the raven paradox. Let $h = \forall x(R(x) \rightarrow B(x))$, where R and B are predicates in some language K. Then $\mathfrak{M}_h$ is depicted in Figure 7a. Observation of a black raven, ($e_1 = \{x_1 \in x_2 \wedge x_1 \in x_3, a_1, a_2, a_3\}$ where a_1 is the element in M associated with the observed raven, a_2 is the element associated with the property of being black, and a_3 is the element associated with the property of being a raven), would confirm this hypothesis (Figure 7b).



(a) (b)

Figure 7

Observation of a non-black non-raven ($e_2 = \{x_1 \notin x_2 \wedge x_1 \notin x_3, a_1, a_2, a_3\}$ where a_1 is the element in M associated with some non-black non-raven, a green table say, and a_2, a_3 are as before) would also confirm h but not as much (assuming a "reasonable" assignment of probabilities). See Figure 8.

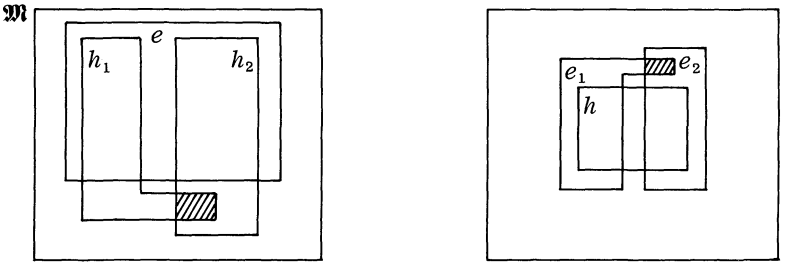


(a) (b)

Figure 8

This makes more precise the comment going back to Hossiasson that the raven paradox is due to the fact that non-black non-ravens are much more numerous than black ravens.

Figure 9a shows that e can confirm h_1 and h_2 individually yet disconfirm $h_1 \wedge h_2$. That two observations e_1 and e_2 can individually confirm a hypothesis h even though $e_1 \wedge e_2$ disconfirms h is shown by Figure 9b.



(a) (b)

Figure 9

In a similar manner one can find hypothesis h_1, h_2 such that e confirms each, yet disconfirms $h_1 \vee h_2$ as well as many other oddities. What of Hempel's axioms four of which we listed at the beginning of this paper? All (except for the purely deductive) fail as even Hempel began to anticipate in his 1964 paper. See [1]. There remain two problems with the notion of confirmation of course. One is that there may be a qualitative sense to the term, distinct and irreducible to our quantitative sense of the term, which does satisfy Hempel-like axioms. The other problem is that our explication of the quantitative sense of confirmation leaves unanswered the question of how to assign probabilities to the subclasses of M.

REFERENCES

[1] Hempel, C. G., "Studies in the logic of confirmation," in *Aspects of Scientific Explanation*, The Free Press, New York (1965).

[2] Paulos, J. A., "A model-theoretic explication of the theses of Kuhn and Whorf," Abstracted in *Notices of the American Mathematical Society* (1976); To appear in *Notre Dame Journal of Formal Logic*.

Temple University
Philadelphia, Pennsylvania