CHAPTER 23

Explicativity: A Mathematical Theory of Explanation with Statistical Applications (#1000)

By *explicativity* is meant the extent to which one proposition or event explains why another one should be believed. Detailed mathematical and philosophical arguments are given for accepting a specific formula for explicativity that was previously proposed by the author with much less complete discussion. Some implications of the formula are discussed, and it is applied to several problems of statistical estimation and significance testing with intuitively appealing results. The work is intended to be a contribution to both philosophy and statistics.

1. INTRODUCTION

By *explicativity* I mean the extent to which one proposition or event F explains why another one E should be believed, when some of the evidence for believing E might be ignored. Both propositions might describe events, hypotheses, theories, or theorems. For convenience I shall not distinguish between an event and the proposition that states the event. In practice usually only putative explanations can be given and this is one reason for writing "should be believed" instead of "is true," but explanation in the latter sense can be regarded as the extreme case where belief is knowledge.

The word "explanatoriness" is not used here because it is defined in the *Oxford English Dictionary* as a quality, where "explicativity" is intended to be quantitative as far as possible. Also it has a more euphonic plural.

The concept of explicativity can be thought of as a "quasiutility," which is a substitute for utility, preferably additive, when ordinary utility is difficult to judge. The condition of additivity for quasiutilities is necessary to justify the maximization of their expected values (#618). The need for at least a rough measure of explicativity arises in pure science more obviously than in commerce where utilities can often be judged in financial terms. But if a measure of explicativity is proposed in general terms it should make sense whatever the field of application. One such field consists of the estimation of statistical parameters since any such estimate can be regarded as a hypothesis that helps to explain observations. Examples of statistical estimation and of significance testing will be given in this paper.

The topic of explicativity belongs to the mathematics of applied philosophy. The present account is based on (#599, **#846**, Good, 1976) and goes much further, though it does not cover everything on the topic in the previous publications.

The advantage of the mathematics of philosophy over classical philosophy is that a formula can be worth many words. The topic is mathematical because it depends on probability. In this respect explicativity resembles some explications for information, weight of evidence, and causal propensity, and it will be convenient to list these explications first, without details of their derivations.

It may be possible sometimes to invert our approach, and to use explicativity inequalities to aid us in our probability judgments.

2. NOTATION

Let A, B, C, E, F, G, H, J, K, sometimes with subscripts or primes, usually denote propositions, or events, or hypotheses, etc. For example, E often denotes an event and *also* the proposition that asserts that the event "obtains." Conjunctions, disjunctions, and negations are denoted by &, v, and a vinculum [macron] respectively. I shall not distinguish between hypotheses, theories, and laws.

Let $P(\mathrm{E}|\mathrm{H})$ denote the probability of E given H or assuming H. Similarly let $P(\mathrm{H})$ denote the initial probability of H and let $P(\mathrm{H}|\mathrm{E})$ denote its final probability. Often $P(\mathrm{H})/P(\mathrm{H}')$ is less difficult to judge than $P(\mathrm{H})$ and $P(\mathrm{H}')$ separately. In practice all probabilities are conditional so that $P(\mathrm{E}|\mathrm{H})$, $P(\mathrm{H})$ and $P(\mathrm{H}|\mathrm{E})$ are abbreviations for $P(\mathrm{E}|\mathrm{H}\ \&\ \mathrm{G})$, $P(\mathrm{H}|\mathrm{G})$, and $P(\mathrm{H}|\mathrm{E}\ \&\ \mathrm{G})$, where G is some proposition, usually complicated, that is taken for granted. It will sometimes be left to the reader's imagination to decide whether any probability mentioned is physical, logical, or subjective. We shall assume the usual axioms of probability whichever of these interpretations of probability is intended.

The *information concerning a proposition* A *provided by another proposition B, given G throughout,* is denoted by $I(\mathrm{A{:}B}|\mathrm{G})$ and is defined by

$$I(\mathrm{A{:}B}|\mathrm{G}) = \log\frac{P(\mathrm{B}|\mathrm{A}\ \&\ \mathrm{G})}{P(\mathrm{B}|\mathrm{G})} = \log\frac{P(\mathrm{A}|\mathrm{B}\ \&\ \mathrm{G})}{P(\mathrm{A}|\mathrm{G})}\,. \tag{1}$$

(We shall not niggle about zero probabilities.) The base of the logarithms exceeds 1 and determines the unit in terms of which information is measured. For example, if the base is the tenth root of 10, the unit is the deciban, a word suggested by A. M. Turing in 1941 in connection with "weight of evidence." With base 2 the unit is the "bit." When G is taken for granted we write $I(\mathrm{A{:}B})$, and a similar abbreviation will be used for other notations. Sometimes $I(\mathrm{A{:}A})$ is denoted by $I(\mathrm{A})$ and (1) implies

$$I(\mathrm{A}) = -\log P(\mathrm{A}). \tag{2}$$

For a derivation of these formulae see, for example, p. 75 of #13 and #505. Mathematical expectations of (1) occur in Shannon's theory of communication (1948). Information has the additive property

$$I(\mathrm{A{:}B}\ \&\ \mathrm{C}) = I(\mathrm{A{:}B}) + I(\mathrm{A{:}C}|\mathrm{B}). \tag{3}$$

The *weight of evidence in favour of H_1 as compared with H_2 provided by E given G* is defined by

$$\begin{aligned} W(\mathrm{H}_1/\mathrm{H}_2{:}\mathrm{E}|\mathrm{G}) &= \log\frac{O(\mathrm{H}_1/\mathrm{H}_2|\mathrm{E}\ \&\ \mathrm{G})}{O(\mathrm{H}_1/\mathrm{H}_2|\mathrm{G})} \\ &= \log\frac{P(\mathrm{E}|\mathrm{H}_1\ \&\ \mathrm{G})}{P(\mathrm{E}|\mathrm{H}_2\ \&\ \mathrm{G})} \\ &= I(\mathrm{H}_1{:}\mathrm{E}|\mathrm{G}) - I(\mathrm{H}_2{:}\mathrm{E}|\mathrm{G}), \end{aligned} \tag{4}$$

where O denotes odds (the ratio of the probabilities of H_1 and H_2). Weight of evidence, which is the logarithm of a Bayes factor, has the additive property

$$W(\mathrm{H}_1/\mathrm{H}_2{:}\mathrm{E}\ \&\ \mathrm{F}) = W(\mathrm{H}_1/\mathrm{H}_2{:}\mathrm{E}) + W(\mathrm{H}_1/\mathrm{H}_2{:}\mathrm{F}|\mathrm{E}) \tag{5}$$

and of course we can condition on G throughout. If the disjunction H_1 v H_2 is also taken for granted, so that H_2 becomes $\overline{\mathrm{H}}_1$, the negation of H_1, then the notation $W(\mathrm{H}_1/\mathrm{H}_2{:}\mathrm{E})$ can be abbreviated to $W(\mathrm{H}_1{:}\mathrm{E})$.

For some literature on weight of evidence see Peirce (1878), #13, and numerous papers cited in **#846**.

The *causal support for E provided by F,* or the *propensity of F to cause E,* denoted by $Q(\mathrm{E{:}F})$, where E and F denote events, is defined (#223B) by the equation

$$Q(\mathrm{E{:}F}) = W(\overline{\mathrm{F}}{:}\overline{\mathrm{E}}|\mathrm{U}\ \&\ \mathrm{L}), \tag{6}$$

the weight of evidence against F if E did not occur, given the state U of the universe just before F occurred, and also given all true laws L of nature. This quantitative explication of causal propensity is basically consistent with the requirements of Suppes (1970) which, however, are only qualitative. The relationship between this monograph and #223B is discussed in #754.

The need for mentioning U in (6) is exemplified by the fact that seeing a flash of lightning is not an important cause of hearing loud thunder soon afterwards. Both events were caused by a certain electrical discharge. Equally, the thunder is not *explained* by the visual experience of lightning. On the other hand seeing the lightning does explain why one *believes* that thunder will soon occur; whereas hearing thunder is a good reason for believing that the lightning flash previously occurred. The experiences are thus valid reasons for prediction and retrodiction respectively.

If F occurs after E, it turns out that $Q(E{:}F) = 0$. This is because U "screens off" E from F under usual assumptions about the nature of time. This notion of "screening off" is explained in more detail by Reichenbach (1956, pp. 201-205) and herein, p. 216. It is analogous to a Markov chain property.

One potential value of measuring causal tendency quantitatively is for the apportioning of credit and blame, as is done, for example by the British Admiralty if two ships collide, though without using (6), and would be done more generally in the courts of justice if they thoroughly deserved their name.

3. PHILOSOPHICAL ASPECTS

There is a large and interesting literature on the philosophy of explanation (for example, Mill, 1843/1961; Hempel, 1948/65; Braithwaite, 1953; Popper, 1959; Nagel, 1961; Scheffler, 1963; Kim, 1967; Rescher, 1970; Salmon, 1971; and numerous further references in these publications). The present account is succinct but is intended to be full enough for the reader to see how the statistical examples fit into the philosophical background. Also I believe that the philosophical discussion contains some new ideas.

The following terminology is fairly standard: what is to be explained or partially explained is called the *explanandum,* and what explains it the *explanans.*

There are at least three main categories of explanation, with various subcategories. They correspond roughly to the questions "what," "how," and "why."

(1) *Explaining "what," or semantic explanation:* answering the question "What do you mean?"

(1.1) *Dictionary definition.*

(1.2) *Philosophical explication:* extraction of more consistent and precise meaning or meanings by analytic consideration of the usage of words by "good" authors. This definition involves an implicit iterative "calculation" because we should say what is meant by a good author.

(2) *Explaining "how," or descriptive explanation:* answering the question "How is this object constructed?"

(2.1) in Nature;

(2.2) in manufacture.

(3) *Explaining "why," or causal (and probabilistic causal) explanation*

(3.1) The explanandum is an event (or the proposition describing an event).

(3.2) The explanandum is a class of events.

(3.3) The explanandum is a scientific law.

(3.4) Explaining why the explanandum should be believed (*to some extent*), when some knowledge supporting this belief, apart from the explanans itself, might be ignored. (For example, we may "know" E is true and still demand an explanation.) Here the explanans is a (partial) cause of *belief* in E rather than a cause of E itself, though it might be both. (Observing the shadow of an elephant can explain why we believe an elephant is present; whereas observing an elephant can explain *both* why the shadow is there *and* why we believe the shadow should be there.) An explanation of this kind might be a prediction or a retrodiction, or a reasoned argument, or a mixture of two or three of these activities. We might have called this kind of explanation "diction" if this word had not been preempted, and anyway a "dictionary" deals with category (1.1). A retrodiction is always a "belief-type" of explanation, rather than a causal type, if it is assumed that causes always precede their effects. I shall make this assumption in this paper though I am not dogmatic about it (see ##882, 1322A).

(3.4.1) The explanandum is a mathematical or logical theorem and the explanans is a proof or heuristic argument. Sometimes an incomplete proof is a better explanation of why a theorem is true than a complete proof. For example, if AOB is a triangle with a right-angle at O, and if a perpendicular is dropped on AB from O, then the three triangles now present all have the same shape so that their areas are proportional to the squares of corresponding linear dimensions. This explains *why* Pythagoras's theorem is true in the sense that the proof is not artificial.

Sometimes "teleological explanation," in which future goals are mentioned, is regarded as forming an additional category, but, unless we allow for precognition, and we shall not do so, this category is not distinct from categories (3.1)-(3.3). This fact is well known. For example, a homing missile, though it acts purposefully, obeys the usual laws of physics. It is its own present *prediction* of the future that affects it, not the future itself.

The present work is an exercise in applied philosophical explication (category 1.2) and its subject matter is category (3). Headings (1.1) and (2) are ignored. The explication of explanation in category (3) often depends on *dynamic* or *evolving* probabilities which can be changed by *reasoning alone* as in a game of chess, and not by new empirical observations. This notion may superficially appear fancy, and is usually overlooked, but I am convinced that it is essential (see especially #938). This is obvious when the explanation comes under category (3.4.1), though the above example concerning Pythagoras's theorem shows that the notion of mathematical explanation cannot be fully captured in terms of probabilities alone. We shall soon see that physical explanation also requires something extra.

Dynamic probabilities are also required for the rest of category (3), as shown in ##599, **846**. For example, to give the argument in outline, the apparent motions of the planets (event E), as projected upon the celestial sphere, had their dynamic probabilities enormously increased, in ratio, when it was noticed that the motions are implied by the inverse square law H of gravitation. This was because the inverse square law had, for most scientists, a non-negligible prior probability, owing to its simplicity and to the analogy of light emerging from a point source, and because it explained why objects like apples fall. That apples behave in some respects like planets is an example of what William Whewell called the "consilience of laws": see Kneale (1953, pp. 364-366). Thus $P(E)$, which exceeds $P(E|H)P(H)$, is much greater than the original value of $P(E)$. This would be true even without bringing in "apples" or the consilience of laws, so that our argument is distinct from Whewell's and Kneale's, and has a somewhat clearer need for the notion of dynamic probability.

To explain why a physical event E occurred is to explain what caused it or tended to cause it, and this requires explicit or implicit reference to a causal chain or causal network that leads to E over some time interval of appreciable duration t. The longer the duration t the fuller the explanation. A causal network cannot be described without at least implicit reference to laws of nature. This shows that probabilities alone, without reference to physical structure, cannot fully capture the notion of physical explanation. Again, if E is itself a law of nature, an explanation of it must be in terms of yet other laws of nature. These will often be more general than E, though explanations by analogy are also possible, and then the explanans might consist of laws no more general than E. Thus, whether the explanandum denotes an event (or set of events) or a law of nature, the explanans will involve laws of nature, and this is a view that has been adopted by many philosophers of science since Mill (1843) or earlier. An immediate consequence of this view is that an event E cannot be regarded as an explanation of itself, since we need $t > 0$, but if you have knowledge that E is true, then this of course fully explains your *belief* in E. Usually in practice our explanations are only putative and only explain beliefs, for real causal networks are enormously complex. Accordingly, the explanation of beliefs will be our main topic.

Sometimes the laws of nature that form part of the explanation of E are taken for granted because of their familiarity. For example, when we say that a window-pane broke because Tom threw a stone at it, we are taking for granted that glass panes are liable to break when hit by fast-moving hard objects that are not too small. Thus a law of physics is here implicit in the explanation. As another example, we might say that it is bad for Ming Vases to leave them unsupported in mid-air.

In deterministic physics a specific event E can sometimes be explained by some boundary conditions B, including initial conditions, combined with differential equations that describe a general law, L. Then B & L explains E, but sometimes, as in the example just given, we call B the explanation when L is taken for granted. The division of an explanation into a *contingent* part and *general laws* is not restricted to physics.

It is difficult to specify sharply whether one law is more general than another. Nagel (1961, pp. 37-42) makes a valiant attempt which he does not regard as fully successful, and I shall here merely point out the relevance of the matter to statistical problems. Suppose that a random scalar or vector x has a probability density function $f(x|\theta)$, where θ is a parameter which is also a scalar or vector. The distribution determined by $f(x|\theta)$ is a "law" in the sense that it says something about a *population* of values of x, and it is often *called* a law (see, for example, Jeffreys, 1939/61). Any proposition of the form $\theta \epsilon \Theta$ (some set of possible values of θ) is a disjunction of laws, and can again reasonably be called a law. Note that θ must be fixed before x can take on a specific value so the time direction is appropriate. If θ itself is regarded as a random scalar or vector containing hyperparameters, as in hierarchical Bayesian techniques (see, for example, ##**26**, 398, **1230**), then a specification of a constrained set of values for these hyperparameters could reasonably be regarded as a law that is more general than $\theta \epsilon \Theta$. For it can be regarded as a proposition about a superpopulation. And similarly for hyper-hyperparameters, etc. A law of the form $\theta \epsilon \Theta$ is a somewhat primitive form of explanation because it does not give detailed information about the structure of the (probabilistic) causal network that leads to an observed value of x, but we cannot usually demand more from statistical estimation procedures. In this example there is no contingent part in the explanans, whereas in regression problems the value of the concomitant ("independent") variable is contingent, when regarded as part of the explanation of a specific value of the dependent variable, whereas the equation of the regression line is lawlike.

There is an intimate relationship between explanation and causation. The broken window was both caused and explained by Tom's naughty behavior. This relation can be formalized to some extent in probabilistic terms: if $P(E|B \& L) > P(E|L)$ then B is a probabilistic cause, and a putative partial explanation of E, when the law L is taken for granted. On the other hand, if $P(E|B \& L) > P(E|B)$, then L is a putative partial explanation of E, but hardly a probabilistic cause, when B is taken for granted. So causation and explanation are related but are not identical (see also §9).

We shall denote by $\eta(E\!:\!F|G)$ the explicativity or explanatory power of F with respect to E, given background information G, and shall arrive at a formula for it, based on some desiderata. Here F may or may not include general laws. This notation interchanges the positions of E and F as used in ##599, 846. The reason for the reversal is that it is more consistent with the notation Q for causal propensity. For grasping the notation we may read $\eta(E\!:\!F|G)$ from left to right as "the explainedness of E provided by F given G," so that the colon can be pronounced "provided by" whether we are talking about information I, weight of evidence W, causal support Q, or explicativity η. (Having two names "explicativity" and "explainedness" for the same thing is analogous to calling $P(E|H)$ both a probability of E and a likelihood of H.) By calling G "background information" we mean that it is assumed to be true and that it has already been taken into account for helping to explain E. (See Desideratum (iii) in §4.) There may also be further evidence G′, such as direct evidence that E is true, which is deliberately ignored and is omitted from our notation.

We shall assume that $\eta(E\!:\!F|G)$ depends only on various probabilities, and we shall not incorporate those requirements that are necessary for regarding F as a partial explanation of E and which do not depend on these probabilities. Thus $\eta(E\!:\!F|G)$ will denote a putative explicativity when F is a putative explanation of E (given G) and will otherwise denote something more general. In fact it will be a measure of *the degree to which F explains why you should believe E, given G all along, and disregarding evidence for E that is not provided by F and G.* We shall call η "explicativity" in all cases although "dictivity" might be preferred. (See the remark about "diction" under category [3.4].) The name is less important than that η should measure something of interest.

Some philosophers claim, with some justification, that F cannot be a (probabilistic) explanation of E unless F is true. But in practice F can perhaps never be known to be true, even in pure mathematics, so that in this paper we shall regard nearly all explanations as only putative. In practice we talk about "explanations" without saying "putative" each time, and accordingly we sometimes put "putative" in parentheses or omit it.

We regard explanations as good or bad depending in part on whether the probability of the explanans is high or low. Let us then allow the explicativity $\eta(E:F)$ to depend on $P(F)$. When F is assumed to be known to be true let us use the somewhat hypallagous expression *informed explicativity.* An informed explicativity is of course an extreme case of a (putative) explicativity.

As an example of the distinction between (putative) explicativity and informed explicativity let us again consider the broken window (event E). The hypothesis F that Tom threw a stone at it has more (putative) explicativity than that the Mother Superior did so (hypothesis F_{MS}). For we believe that Tom is naughtier than the Mother Superior as well as being a better shot. On the other hand, if we *saw* the Mother Superior throw the stone vigorously, F_{MS} would have very high *informed* explicativity.

By using the expression "informed explicativity" we do not wish to imply that the whole causal network preceding E is known; we mean only that F becomes known to be true, but is not taken for granted in advance. The informed explicativity of F with respect to E might be high and yet it might turn out that F is not part of the true explanation of E after all.

Both a (putative) explicativity and its extreme case, an informed explicativity, are intended to be measures of the explanatory power of F with respect to E relative to the knowledge that we (or "you") have, and that knowledge will seldom include the certainty of F. We can only hope to measure the extent to which our beliefs about F explain why we should believe E (imagining E to be unobserved). Under this interpretation it is not necessary that F should precede E chronologically; and $\eta(E:F|G)$ will sometimes measure the predictivity or retrodictivity of F with respect to E, or some mixture. Again, if F is a "law," it need have no position in time, and it might be used for prediction, retrodiction, or putative explanation of E.

Since we regard informed explicativity as an extreme case of (putative) explicativity, we do not need a separate notation for it. It will be merely a matter of putting $P(F|G) = 1$ or $P(F) = 1$ in whatever formula we use for $\eta(E:F|G)$ or $\eta(E:F)$.

We conclude this philosophical background with one further property of explanation. Most philosophers believe that an explanation should be based on all relevant evidence apart from the evidence G′ that is deliberately ignored such as the direct observation of E. With our notation $\eta(E:F|G)$ this would mean that F & G must contain all evidence relevant to E, apart from G′. In practice, when we are estimating an explicativity, we must make do with the evidence that appears to us to be sufficiently relevant.

4. THE DESIDERATA AND EXPLICATION FOR EXPLICATIVITY

As a preliminary to proposing some desiderata for explicativity, let us consider a naive approach and an early historical approach to explanation.

Perhaps the most naive suggestion is that E is explained by H if H logically implies E. This is neither a necessary nor a sufficient condition for H to be a good explanation of E. For example, the hypothesis $0 = 1$ logically implies everything and in particular it implies E, but $0 = 1$ is an extremely poor (putative) explanation of anything! Nor does it help to append some irrelevant laws of nature so as to make the explanans lawlike. So we need something less naive. Let us recall a little history.

According to the translation by Charlton (1970, p. 10), Aristotle said " . . . it is better to make your basic things fewer and limited, like Empedocles." In the early fourteenth century the "doctor invincibilis," William of the village of Ockham in Surrey said "plurality is not to be assumed without necessity." This sentiment had been previously emphasized by John of the village of Duns in Scotland who has often been thought, apparently incorrectly, to have been William of Ockham's director of studies (Anon., 1951; Moody, 1967). The saying that "entities should not be multiplied without necessity," though apparently never expressed quite that way by William of Ockham, has come to be known as "Ockham's razor." For a detailed history, but with the Latin untranslated, see Thorburn (1918).

A more modern interpretation of the Duns-Ockham razor is that, of two hypotheses H and H′, both of which explain E, the simpler is to be preferred (see, for example, Margenau, 1949). But the hypothesis $0 = 1$ is simple, at least in the sense of brevity, so we need to sharpen the razor some more. The next improvement is that if H and H′ both imply E, then the hypothesis with the larger initial probability is preferable. In nearly all applications the judgment of whether $P(H) > P(H')$ is subjective or personal, although different people often agree about a specific judgment. Note that if $P(H) > P(H')$, and H and H′ each imply E, then $P(H|E) > P(H'|E)$, that is, the final probability of H exceeds that of H′. One advantage of this way of interpreting Ockham's razor is that it rules out impossible explanantia such as the hypothesis $0 = 1$.

Whereas the initial probability of a hypothesis has something to do with its simplicity the relationship is not obvious, and if we express all our formalism in terms of probabilities we do not need to refer explicitly to simplicity or complexity. In #599 I defined the complexity of a proposition H as $-\log P(H)$, but I retracted this in #876. There is more than can be and has been said on the relationship between complexity and probability, but to avoid distraction we discuss this matter in appendix A.

What if the two hypotheses H and H′ do not logically imply E but merely increase its probability, so that

$$P(\mathrm{E|H}) > P(\mathrm{E}) \text{ and } P(\mathrm{E|H'}) > P(\mathrm{E})?$$

Is H a better explanation of E than H′ if $P(\mathrm{E|H}) > P(\mathrm{E|H'})$? Not necessarily if $P(\mathrm{H}) < P(\mathrm{H'})$. Some compromise is required, to be discussed later.

Let us assume the following desiderata. (i) The explicativity of H with respect to E, denoted by $\eta(\mathrm{E:H})$, is a function of at most 52670 variables, namely all probabilities of the form $P(\mathrm{A|B})$ where A and B run through all the propositions that can be generated from E and H by conjunctions, disjunctions, and negations, and where each of these probabilities is not necessarily equal to 0 or 1. It is not important to check that 52670 is the correct number because an equivalent assumption is that $\eta(\mathrm{E:H})$ depends at most on $P(\mathrm{E})$, $P(\mathrm{H})$, and $P(\mathrm{E \;\&\; H})$. (ii) If K and F have nothing to do with H and E then $\eta(\mathrm{E \;\&\; F:H \;\&\; K})$ depends only on $\eta(\mathrm{E:H})$ and $\eta(\mathrm{F:K})$. (iii) $\eta(\mathrm{E:H|H})$ does not depend on E or H (in fact you can reasonably call it zero). (iv) $\eta(\mathrm{E:H})$ increases with $P(\mathrm{E|H})$ if $P(\mathrm{E})$ and $P(\mathrm{H})$ are fixed. (v) $\eta(\mathrm{H:H}) \geqslant \eta(\mathrm{T:T})$ where T is a tautology. (vi) $\eta(\mathrm{T:H}) \leqslant \eta(\mathrm{T:T})$ (because a tautology needs no explanation).

Then it can be proved [see appendix B] that $\eta(\mathrm{E:H})$ must be some increasing function of $I(\mathrm{E:H}) - \gamma I(\mathrm{H})$ where γ does not depend on the probabilities and where $0 < \gamma < 1$ (see appendix B). Since the main purpose is to put explicativities in order we may as well take $\eta(\mathrm{E:H}) = I(\mathrm{E:H}) - \gamma I(\mathrm{H})$. Moreover this choice converts (ii) into the strictly additive property

$$\eta(\mathrm{E \;\&\; F:H \;\&\; K}) = \eta(\mathrm{E:H}) + \eta(\mathrm{F:K}) \tag{7}$$

(when K and F have nothing to do with H and E), and this justifies us in regarding $\eta(\mathrm{E:H})$ as a proper quasi-utility. Various forms of $\eta(\mathrm{E:H})$ are:

$$\eta(\mathrm{E:H}) = I(\mathrm{H:E}) - \gamma I(\mathrm{H}) \tag{8}$$

$$= \log P(\mathrm{E|H}) - \log P(\mathrm{E}) + \gamma \log P(\mathrm{H}) \tag{9}$$

$$= I(\mathrm{E}) - I(\mathrm{E|H}) - \gamma I(\mathrm{H}). \tag{10}$$

We must adjust equation (9), when dynamic probabilities are relevant, as a formula for "dynamic explicativity," $\eta_{\mathrm{D}}(\mathrm{E:H})$, namely

$$\eta_{\mathrm{D}}(\mathrm{E:H}) = \log P_1(\mathrm{E|H}) - \log P_0(\mathrm{E}) + \gamma \log P(\mathrm{H}). \tag{9 D}$$

Here $P_0(\mathrm{E})$ is the initial probability of E, judged *before* H is brought to your attention, whereas $P_1(\mathrm{E|H})$ is the conditional probability of E given H *after* H is brought to your attention. When H is a good simple theoretical explanation of E, as in the example of the inverse square law, it can easily happen that $P_1(\mathrm{E|H})$ $P(\mathrm{H})$, which is equal to $P_1(\mathrm{E \;\&\; H})$, is much larger than $P_0(\mathrm{E})$. When dynamic probabilities are relevant it is ambiguous to omit the subscripts 0 and 1 from the notations, but sometimes it may not be too misleading to write $\eta(\mathrm{E:H})$ instead of $\eta_{\mathrm{D}}(\mathrm{E:H})$. For, in ordinary linguistic usage, the inverse square law is called simply an "explanation" of the planetary motions. It happens to be a dynamic explanation in both senses of "dynamic."

A few exercises, extracted from **#846**, are:

$$\eta(\mathrm{E:0 = 0}) = 0, \tag{11}$$

$$\eta(\mathrm{E:0 = 1}) = -\infty, \tag{12}$$

$$\eta(\mathrm{E \;\&\; F:H}) = \eta(\mathrm{E:H}) + \eta(\mathrm{F:H|E}) + \gamma I(\mathrm{E|H}), \tag{13}$$

a modified additivity property. If H and L are mutually exclusive then H v L has less explicativity for E than does H if and only if

$$\frac{P(\mathrm{E|L})}{P(\mathrm{E|H})} < \left[1 + \frac{P(\mathrm{L})}{P(\mathrm{H})}\right]^{1-\gamma} - 1. \tag{14}$$

For example, when $P(\mathrm{H}) \approx P(\mathrm{L})$ and $\gamma = ½$, the right side is 0.414.

5. THE CHOICE BETWEEN HYPOTHESES

More important than assigning an explicativity to a single hypothesis, with respect to E, is deciding which of two hypotheses H and H′ has the greater explicativity and by how much. Then the term $\log P(\mathrm{E})$ in (9) is irrelevant, because it is mathematically independent of the hypotheses. Let us denote H v H′ by J and take it for granted, as is permissible when we are choosing between H and H′. Denote by $\eta(\mathrm{E:H/H'|J})$ or $\eta(\mathrm{E:H/H'})$ the amount by which the explicativity of H exceeds that of H′, or 'the explainedness of E provided by H as against H′ (given J)'. Then

$$\eta(\mathrm{E:H/H'}) = \eta(\mathrm{E:H}) - \eta(\mathrm{E:H'}) \tag{15}$$

$$= W(\mathrm{H/H':E}) + \gamma \log O(\mathrm{H/H'}) \tag{16}$$

$$= \log O(\mathrm{H/H'|E}) - (1-\gamma)\log O(\mathrm{H/H'}) \tag{17}$$

$$= (1-\gamma)W(\mathrm{H/H':E}) + \gamma \log O(\mathrm{H/H'|E}). \tag{18}$$

Equation (18) has an interesting interpretation. It exhibits the excess in explicativity of H over its negation as a compromise between two extremes, the weight of evidence on the one hand and the final log-odds on the other. The former of these extremes ($\gamma = 0$) corresponds to the philosophy of "letting the evidence speak for itself" (as advocated by some in the Likelihood Brotherhood), and the latter ($\gamma = 1$) to that of preferring the hypothesis of maximum final probability. Neither of these two philosophies is tenable as we may see clearly by an example, although their implications are reasonably judged to be good enough in some circumstances.

Let E denote the proposition that planets move in ellipses, let H denote the inverse square law of gravitation, and K that there is an elephant on Mars. If we took $\gamma = 0$ we'd find that $\eta(\mathrm{E:H \;\&\; K}) = \eta(\mathrm{E:H})$, in other words that the explicativity of H would be unaffected by cluttering it up with an improbable irrelevant elephant. Thus the size of γ depends on how objectionable we regard it to have clutter, or to "multiply entities without necessity."

The case $\gamma = 0$ of (8), namely the mutual information between E and H, was proposed independently as an explication of explanatory power by Good (1955) and Hamblin (1955), both in relation to Popper's writings. The fact that it did not allow for clutter was pointed out in #599, and this explication was therefore called *weak* explanatory power. In our present terminology it is the "informed (putative) explicativity" of H. Expected amounts of information of this kind, and of the effectively more general notion of weight of evidence . . . , were related to statistical physics by Gibbs (1902, chap. XI) and Jaynes (1957), and to non-physics statistical practice by, for example, Turing in 1941 (see #13), Jeffreys (1946), Shannon (1948), Good (1950/53), Kullback & Leibler (1951), Rothstein (1951), Cronbach (1953), #77, Lindley (1956), Jaynes (1957, 1968), Kullback (1959), ##322, 524, Tribus (1969), #755, over thirty other publications by the present writer, and in several publications by Rothstein and by S. Watanabe.

Next suppose we take $\gamma = 1$, then $\eta(E:H)$ would reduce to $\log P(H|E)$ and there would be no better hypothesis than a tautology such as 1 = 1. This shows, as in Bayesian decision theory, that it is inadequate to choose the hypothesis of maximum final probability as an unqualified principle.

So we must take $0 < \gamma < 1$. There may not be a clearly best value for the "explicativity parameter" γ, but $\gamma = ½$ seems a reasonable value. It exactly "splits the difference" between the two extreme philosophies just mentioned, and is also the simplest permitted numerical constant.

The *sharpened razor* is the recommendation to choose the hypothesis that maximizes the explicativity with respect to E, or for all known evidence. It differs from a central theme of Popper's philosophy, namely that a useful theory is one that is of low (initial) probability and highly testable. Certainly high checkability is a desirable feature of a theory, and, *if a theory turns out to have a high final probability*, then a low initial probability is desirable because it shows that the theory was informative. But Popper's philosophy does not allow for final probabilities.

It is of interest to consider how much more explicative E itself is than H, relative to E,

$$\begin{aligned}\eta(E:E/H) &= \eta(E:E) - \eta(E:H)\\ &= \gamma\log P(E) - \gamma\log P(H) - \log P(E|H),\end{aligned} \tag{19}$$

or, when dynamic probabilities are used,

$$\eta_D(E:E/H) = \gamma\log P_0(E) - \gamma\log P(H) - \log P_1(E|H). \tag{19 D}$$

For $\eta(E:E/H)$ we have the following theorem:

When dynamic probabillties are not used, there is no more explicative proposition relative to E than is E itself; in symbols $\eta(E:E/H) \geqslant 0$, that is,

$$\eta(E:E) \geqslant \eta(E:H). \tag{20}$$

Equality occurs only if $P(E|H) = P(H|E) = 1$. The corresponding result for dynamic explicativities is false.

Proof. The right side of (19) can be written

$$(1-\gamma)\log P(H) + \gamma\log P(E) - \log P(E \,\&\, H).$$

Since $P(E \,\&\, H)$ exceeds neither $P(E)$ nor $P(H)$, this expression is at least as large as both $(1-\gamma)[\log P(H) - \log P(E)]$ and $\gamma[\log P(E) - \log P(H)]$ and must therefore be non-negative. It vanishes only if $P(H) = P(E) = P(E \,\&\, H)$, that is, only if $P(E|H) = P(H|E) = 1$, which for practical purposes means that E and H are logically equivalent.

That the theorem is false for dynamic explicativities is clear from the example of the planetary motions and the inverse square law. The dynamic explicativity $\eta_1(E:H)$ can exceed, equal, or "subceed" $\eta_0(E:E)$.

When $\gamma = ½$ we have, when we do not use dynamic probabilities,

$$\eta(E:E/H) = \log\frac{[P(E)P(H)]^{½}}{P(E \,\&\, H)} \tag{21}$$

which is symmetrical in E and H, just as $I(E:H)$ is, though a closer analogue is $I(E:E) - I(E:H) = I(E|H)$ which is not symmetrical. (Of course it can be forcibly symmetrized by writing $I(E|H) + I(H|E)$.) If we accept the value $\gamma = ½$, (21) could be called the *mutual* explicativity "distance" between E and H, by analogy with the name "mutual information" for $I(E:H)$. It equals 0 if H = E and ∞ if $H = \overline{E}$, and resembles $I(E|H)$ in this respect. Symmetry in E and H is an elegant property but it is not a compelling desideratum. The triangle inequality is not satisfied, but it may be of interest that

$$\begin{aligned}\eta(E:E/F) + \eta(F:F/G) - \eta(G:G/E) &= I(E|F) + I(F|G) - I(E|G)\\ &= I(G|F) + I(F|E) - I(G|E),\end{aligned} \tag{22}$$

so that the "triangles" for which the triangle inequality is valid are the same for the functions $(\lambda E)(\lambda F)\eta(E:E/F)$ and $(\lambda E)(\lambda F)I(E|F)$(in Alonzo Church's λ notation).

6. REPEATED TRIALS

Sometimes E can be defined as a compound event, or time series, which describes the probabilistic outcome $E_1 \,\&\, E_2 \,\&\, \ldots \,\&\, E_N$ of an experiment performed "independently" N times under essentially similar circumstances. If N is large, the frequencies of the various outcomes settle down, with high probability, to a distribution. A hypothesis H that predicts this distribution has an expected explicativity gain per observation, as compared with another hypothesis H′, and this gain tends in probability to

$$(1-\gamma)\mathcal{E}\{W(H/H':E)|H\}, \tag{23}$$

which is proportional to the expected weight of evidence per observation. The second term in (18) gets divided by N and so contributes nothing to the limiting value (23). Thus, for "repeated trials," the application of the notion of explicativity to statistics will lead to the same results as when (expected) weight of evidence is used as a quasiutility, as in numerous publications cited earlier. In particular, if H asserts the true physical probability density $p(x,y)$ of two random variables, whereas hypothesis H′ asserts that the density is $p(x)q(y)$, then $\eta(\mathrm{E}:\mathrm{H}/\mathrm{H}')/N$ tends in probability to

$$(1-\gamma)\int\int p(x,y)\log\frac{p(x,y)}{p(x)q(y)}\,dx\,dy \tag{24}$$

which is $1-\gamma$ times the "rate of transmission of information" concerning x provided by y and can of course be expressed in terms of three entropies. This formula can be used in the choice of an experimental design. The factor $1-\gamma$ is irrelevant for this purpose: see Cronbach (1953), #77, and especially Lindley (1956). Thus, in this application, the value of γ does not matter.

Greeno (1970), unaware of these references, suggested rate of transmission of information as an explication for explanatory power. We see from the above argument how this proposal is deducible from the notion of explicativity, and even from the earlier (Good, 1955; Hamblin, 1955) special case of weak explanatory power (informed explicativity), when E denotes an "infinitely repeated trial."

7. PREDICTIVITY

As we have seen, a probabilistic prediction of the result of an experiment or observation is a special case of a putative explanation, being made before the experimental result occurs. In these circumstances it is natural to measure the *predictivity* of a hypothesis as the mathematical expectation of the putative explicativity, the expectation being taken over the population of possible outcomes. It is appropriate to take expectations of η rather than of some monotonic function of η because of the additive property (7).

The explicativity of H, per trial, with respect to repeated trials, as given by (23), is formally nearly the same as predictivity, owing to the law of large numbers.

For a theory with a wide field of possible applications, the notion of predictivity is necessarily vague; but it might be defined as the expected explicativity over all future observations with discounting of the future at some rate. The concept is important in spite of its vagueness.

For experimental design, predictivities (expected explicativities) are natural quasiutilities. This fact can be regarded as an explication in hindsight why entropies occur in the work of Cronbach (1953) and Lindley (1956). In virtue of these two publications it is not necessary to consider experimental design further here. Instead, we work out in detail only examples of estimating parameters in a distribution law, after observations are taken. In this estimation problem entropies do not occur because expectations are not taken. Hypothesis testing can be regarded as a special case of parameter estimation (and vice versa).

8. "COLLATERAL" INFORMATION VERSUS BACKGROUND INFORMATION

Consider the propositions

E: Jones won the Irish Sweepstake,

H: Jones bought a ticket in this lottery,

and for the sake of simplicity assume that

$$P(\mathrm{H}) = 2^{-8},\ P(\mathrm{E}|\mathrm{H}) = 2^{-20}, \text{ and therefore } P(\mathrm{E}) = P(\mathrm{E}\ \&\ \mathrm{H}) = 2^{-28}.$$

Then, if $\gamma = ½$, we have $\eta(\mathrm{E}:\mathrm{H}) = 8 - 8/2 = 4$ bits. If we knew all along that H was true we would have $\eta(\mathrm{E}:\mathrm{H}|\mathrm{H}) = 0$, meaning that H cannot help to explain E if we have already taken H into account. But in another sense, if we discover that H is true we raise the probability of H to 1, and the explicativity of H with respect to E, which is now "informed" explicativity, is $I(\mathrm{E}:\mathrm{H}) = 8$ bits. Thus, for the sake of completeness, it is convenient to have a notation for the explicativity of H when its probability is conditional on some *collateral* information K. Let us use a semicolon to mean "given the collateral information." Then we have

$$\eta(\mathrm{E}:\mathrm{H};\mathrm{K}|\mathrm{G}) = \log P(\mathrm{E}|\mathrm{H})\ \&\ \mathrm{G}) - \log P(\mathrm{E}|\mathrm{G}) + \gamma\log P(\mathrm{H}|\mathrm{K}\ \&\ \mathrm{G}) \tag{25}$$

where we have included G for greater generality. In particular,

$$\eta(\mathrm{E}:\mathrm{H};\mathrm{H}) = I(\mathrm{E}:\mathrm{H}). \tag{26}$$

Background information is taken for granted in computing all the probabilities, whereas collateral information affects only the probability of the explanans H and is not taken into account when computing the probability of the explanandum E. Of course $\eta(\mathrm{E}:\mathrm{H};\mathrm{H})$ is the informed explicativity of H. No special terminology for $\eta(\mathrm{E}:\mathrm{H}|\mathrm{H})$ is proposed because it necessarily vanishes.

The notation $\eta(\mathrm{E}:\mathrm{H};\mathrm{K})$ or $\eta(\mathrm{E}:\mathrm{H};\mathrm{K}|\mathrm{G})$ helps to formalize the familiar situation in which an explanans H is strengthened by having its *own* probability increased by evidence K. For example, when we discover that Tom was at the scene of the crime, the probability is increased that he threw a stone at the window. Explicativity depends on the explanandum, the explanans, the collateral information, and the background information. We have

$$\eta(\mathrm{E}:\mathrm{H};\mathrm{K}) = \eta(\mathrm{E}:\mathrm{H}\ \&\ \mathrm{K}) \text{ if and only if } P(\mathrm{E}|\mathrm{H}\ \&\ \mathrm{K}) = P(\mathrm{E}|\mathrm{H}). \tag{27}$$

9. THE QUANTITATIVE DISTINCTION BETWEEN EXPLICATIVITY AND CAUSAL PROPENSITY

In our lottery example the explicativity of the ticket-purchase, with respect to E, is appreciable (whether the explicativity is "informed" or not), although

$P(E|H)$ is small. There is a distinction between (putative) explicativity and causal propensity: the purchase of the ticket did not do much to *cause* E although it was a necessary condition for it. If Jones had not won the sweepstake, it would have been negligible evidence against his having bought a ticket, so, according to (6), the causal propensity of the purchase is small. Similarly, if Ms Aksed is hit by a small meteorite when out walking, we would not blame her and accuse her of suicidal tendencies. Her decision to go for a walk was a necessary condition for the disaster, but if she had not been hit by a meteorite, it would have been negligible evidence that she was indoors when the meteorite fell. The insurance company would call the incident an Act of God.

10. APPLICATIONS TO STATISTICAL ESTIMATION AND SIGNIFICANCE TESTING

. . . [The eight pages omitted here show that $\eta(E:H)$ can be applied to statistics with entirely sensible results. This confirms the reasonableness of η as an explication of explicativity.]

11. FURTHER COMMENTS CONCERNING THE VALUE OF γ

If no other desiderata can be found for fixing γ, the value $\gamma = ½$ could often reasonably be adopted on grounds of maximum simplicity. This choice can itself be regarded as an application of a form of the Duns-Ockham razor (of higher type so to speak). Moreover there are many scientists who believe that the notion of simplicity is better replaced by that of elegance, or aesthetic appeal. For example, Margenau (1949) says "The physicist is impressed not solely by its [a theory's] far-flung empirical verifications, but above all by the intrinsic beauty of its conception which predisposes the discriminating mind for acceptance even if there were no experimental evidence for the theory at all." Again Dirac (1963) says " . . . it is more important to have beauty in one's equations than to have them fit experiment. . . . That is how quantum mechanics was discovered," and I believe Dirac expressed this view in conversation at least as early as 1940. From this point of view the value $\gamma = ½$ gains from the elegant symmetry of equation (21). . . . [As a discussion point, I believe that beauty is often a matter of simplicity arising out of complexity arising out of simplicity.]

12. SUMMARY

Philosophical aspects of explanation were discussed in §3 leading up to an informal definition of $\eta(E:F|G)$ and to the desiderata and exact explication of η in §4 in terms of probabilities or information. In §5 we showed the relevance of explicativity for a choice between hypotheses. In §6 we saw that if explicativity is used in experimental design it reduces in effect to expected weight of evidence or to rate of transmission of information. In §7 an informal quantification of predicitivity was suggested. In §8 it is pointed out that a distinction between background information and "collateral" information is necessary for formalizing a familiar aspect of explanation, so that η depends on four variables (apart from the evidence G' that is deliberately ignored: see §3). In §9 it is shown that explicativity and causal propensity can be quantitatively quite different, both in common parlance and in terms of the formalism. In §10 several examples of statistical estimation and significance testing are worked out in terms of explicativity with intuitively appealing results.

APPENDIX A. COMPLEXITY

Although an explication of simplicity or complexity is not required for that of explicativity, the latter depends on the initial probability of a proposition H and this probability surely depends to some extent on the complexity of H. For the complexity of the conjunction H & K of two propositions that are entirely independent is greater than the complexity of either of them separately, in any one's book, and is reasonably assumed to be the sum of the two complexities. If the complexity of H could be defined in terms of $P(H)$ alone then it would have to be $-\log P(H)$ as suggested in #599. But the two propositions $0 = 0$ and $0 = 1$ are about equally simple in my present judgment, though their probabilities are poles apart. So the complexity of H cannot be defined in terms of $P(H)$ alone. Fortunately this error in #599 did not undermine much else in that work. The error was admitted in #876, and on pp. 154-56, where attempts were made to improve the definition. It was proposed that the complexity of a proposition should be defined as the minimum value of $-\log p$ where $p = P(S)$ is the probability of some statement S of the proposition *regarded as a linguistic string* and the minimum is taken over all ways of expressing the proposition as a statement. Moreover, the language used must be one that is economical for talking about the topic in question.

A valid objection was raised against this definition by Peter Suzman, as mentioned in Good (1976b). Suzman asked whether the proposition that all caterpillars have chromosomes is more complex than that all dogs have chromosomes. My reply was to concede that these propositions are of (nearly?) *equal* complexity. Nor is it sufficient to modify the proposed definition of complexity, by defining $p(S)$ as the probability of the *syntactic structure* of S, nor by making the definition depend only on the number of dimensionless parameters in a law. For a parameter equal to 5.4603 is more complex than one that is equal to 2. Perhaps one cannot do much better than to define the complexity of a proposition as equal to the *weighted length* of the shortest way of expressing it, measured in words and symbols, where different weights should be assigned to different categories of words such as parts of speech. Perhaps the weights should be minus the logarithms of the frequencies of these *categories* of words (instead of using the frequencies of the individual words and symbols as such). This would reduce the problem to the specification of the categories.

A somewhat different ideal measure of the complexity of a scientific theory is the number of independent axioms in it (see, for example, Margenau, 1949), and I believe this is a useful rule of thumb. But it does not allow for the relative complexities of the axioms.

In practice, the beauty of a theory, rather than its simplicity, might be more important when estimating initial probabilities: see the quotations at the end of the main text. To fall back on beauty as a criterion is presumably to admit that the left hemispheres of the brains of philosophers of science have not yet formalized the intuitive activities of the right hemispheres.

Measurements of complexity or ugliness might help us to judge prior probabilities, but, if the prior probabilities could be adequately judged, the crutches of simplicity and beauty could be discarded. These crutches were not much used in the main text because our aim was to express explicativity in terms of probability.

APPENDIX B. THE FORM OF THE FUNCTION $\eta(E:H)$

. . . [A proof of (8) was given in this appendix.]

References

Discussion of Radioactive Decay using Likelihood Ratios

Proof. By A10 and A22, we may replace S by Q in T12, and drop the inequalities $p_1 \geqslant q_1$, $p_2 \geqslant q_2$. Let

$$\psi(\xi, \eta, x) = Q(1 - e^{\xi}, 1 - e^{\eta}, x(1 - e^{\xi}) + (1 - x)(1 - e^{\eta})),$$
$$p_1 = \exp \xi_1, \; q_1 = \exp \eta_1, \text{ etc.}$$

Then

$$\psi(\xi_1 + \xi_2, \eta_1 + \eta_2, x) = \psi(\xi_1, \eta_1, x) + \psi(\xi_2, \eta_2, x).$$

On putting $\eta_1 = \eta_2 = 0$, and provisionally regarding x as a constant, we get a well known functional equation whose only continuous solution is easily seen to be of the form

$$\psi(\xi, 0, x) = \xi \,.\, u(x),$$

where $u(x)$ is a function of x only. (The only other solutions are in fact non-measurable: see Hamel, 1905, or Hardy *et al.*, 1934, p. 96.) Likewise $\psi(0, \eta, x) = \eta \cdot w(x)$, where $w(x)$ is a function of x only. Therefore

$$\begin{aligned}\psi(\xi, \eta, x) &= \psi(\xi + 0, 0 + \eta, x)\\ &= \psi(\xi, 0, x) + \psi(0, \eta, x) = \xi u(x) + \eta w(x).\end{aligned}$$

Therefore

$$Q(p, q, xp + (1 - x)q) = u(x) \,.\, \log(1 - p) + w(x) \,.\, \log(1 - q).$$

T13 now follows from A5 combined with the equation

$$r = xp + (1 - x)q.$$

Q.E.D.

A23. *Consider a radioactive particle in a certain state, which I shall call the "white" state. In any time interval,* t, *it has probability* e^{-at} *of remaining in the white state throughout the interval if it starts the interval in that state. If it does not remain in the white state, then it proceeds to another state called here the "black" state, from which there is no return. Now let F be the event that the particle is in the white state at the start of an interval of duration T and let E be the event that it is in the white state at the end of this interval. Then we assume that, if F and E both occurred, χ(E:F) does not depend on the unit in terms of which time is measured.*

A24. *If F . E implies G, and F → G → E is a chain, then this chain is of the same strength as F → E.*

T14. $R(p, 0, r) = \nu(r/p) - k \,.\, \log p$,

where v(x) *is a non-negative analytic function of* x, *and* k *is a positive constant.*

Proof. Consider the radioactive particle described in A23. Let $P(\mathrm{F}) = x$. The degree to which F caused E is the limit of the strengths of finite chains obtained by breaking up the time interval $(0,T)$ into a "Riemann dissection" (see A9). Since g is a continuous function (A8) the resistances of these finite chains must also tend to a limit, which we may call the causal resistance from F to E. This must be some function of x, a, and T, say $R^*(x, a, T)$. By A23 we see that for

any positive constant, k, the resistance must be equal to $R^*(x, ka, T/k)$. Since this is independent of k it must be of the form of $R^*(x, aT)$.

Now, by a continuity argument, we may generalize T9 to continuous chains, and hence deduce that, for any positive T and U we have

$$R^*(x, aT) + R^*(1, aU) = R^*(x, aT + aU).$$

By giving x the value 1 and subtracting from the equation with arbitrary x, we see that $R^*(x, aT)$ is of the form

$$R^*(x, aT) = \nu(x) + R^*(aT),$$

where, identically,

$$R^*(aT_1 + aT_2) = R^*(aT_1) + R^*(aT_2),$$

so that $R^*(aT)$ is of the form

$$R^*(aT) = k_1 aT.$$

Now, by repeated use of A24, we see that

$$R(p, 0, xp) = R^*(x, aT),$$

where $p = e^{-aT}$. Thus

$$R(p, 0, r) = \nu(r/p) - k \,.\, \log p.$$

Q.E.D.

T15. $$Q(p, q, r) = \log(1 - q) - \log(1 - p),$$
$$R(p, 0, r) = -\log p,$$

where the base of the logarithms may be taken as e. Q(p, q, r) *is mathematically independent of* r, *and may be abbreviated to* Q(p,q). *It can be written in other ways:*

$$Q(\mathrm{E}:\mathrm{F}|\mathrm{G}) = \log \frac{P(\bar{\mathrm{E}}|\bar{\mathrm{F}} \,.\, \mathrm{G})}{P(\bar{\mathrm{E}}|\mathrm{F} \,.\, \mathrm{G})} = \log \frac{O(\bar{\mathrm{F}}|\bar{\mathrm{E}} \,.\, \mathrm{G})}{O(\mathrm{F}|\mathrm{G})}$$
$$= W(\bar{\mathrm{F}}:\bar{\mathrm{E}}|\mathrm{G}) = -W(\mathrm{F}:\bar{\mathrm{E}}|\mathrm{G}),$$

the weight of evidence against F if E does not happen. More precisely, Q *is uniquely determined only up to a continuous analytic increasing transformation. Among all the explicata there is just one apart from a scale factor (choice of unit), for which theorems T9 and T11 are true.* We lose no real generality, and we gain simplicity, by choosing this explicatum.

Proof. By T13, T14, and A7, we have the identity

$$f(\nu(r/p) - \log p) = -u(r/p) \,.\, \log(1 - p).$$

Let $\nu(x) = y$, $-\log p = z$, $\log f(y + z) = \rho(y + z)$. Then $\rho(y + z)$ is of the form

$$\rho(y + z) = \rho_1(y) + \rho_2(z).$$

If $\nu(x)$ is not a constant, we can differentiate and deduce that $\rho(y)$ is a linear function of y, from which we can soon derive that $\log(1 - p)$ is a power of p.

Since this is false it follows that $v(x)$ is a constant, and hence also that $u(x)$ is a constant.

The theorem now follows from the remark that the choice of the base of the logarithms is equivalent merely to the choice of units of measurement of strength and resistance. We may call the units "natural," "binary," or "decimal," according as the base is e, 2, or 10. In this paper I shall use natural units. Possible names would be "natural causats" and "natural tasuacs."

Note that the explicatum for Q was by no means obvious in advance, nor was it obvious that all the desiderata could simultaneously be satisfied.

It is interesting to note that, if, contrary to most of the discussion, we assume E to be earlier than F, and if the universe has the "Markov" property, defined below, then the tendency of F to cause E is zero. This result may very well have been taken as a desideratum, but was in fact noticed only after the explicatum was obtained. By the Markov property is meant here that, for prediction, a complete knowledge of the immediate past makes the remote past irrelevant.

T16. *The relationship between* R *and* S *is symmetrical, namely*

$$R \geqslant 0, S \geqslant 0,$$
$$e^{-R} + e^{-S} = 1,$$

or equivalently,

$$R = -\log(1 - e^{-S}), \; S = -\log(1 - e^{-R}).$$

Further,

$$R(p, q, r) = \log(1 - q) - \log(p - q).$$

This is an immediate corollary of A7 and T15.

Thus the function f is its own inverse, g. It is tempting to permit negative and imaginary values because some of the formalism is faintly reminiscent of Feynman's formulation of quantum mechanics, but I shall not pursue this matter here.

T17. *If a chain consists of* n *links whose* p*'s and* q*'s are* (p_i, q_i), *where* $p_i \geqslant q_i$, *then its causal strength is*

$$-\log\left\{1 - \prod_i \frac{p_i - q_i}{1 - q_i}\right\}.$$

This follows from T16 and T9.

Before reading the proofs in the present section the reader will probably prefer to read the next two sections, in which some examples are given.

6. TWO-STATE MARKOV PROCESSES

The radioactive process described in Axiom 23 can be slightly generalized by permitting return from the black to the white state, with a parameter β corresponding to the α of the white-to-black transition. We have a two-state Markov

process with continuous time. The parameters α and β are of course both nonnegative. In the special case of the radioactive particle we have $\beta = 0$.

It can be shown that

$$Q(E{:}F) = \log[(\alpha + \beta e^{-(\alpha + \beta)T})/(\alpha - \alpha e^{-(\alpha + \beta)T})].$$

If the particle ever entered the black state during the time interval, T, the chain would be cut and the degree of causality would be zero. Assuming that this does not happen, we can calculate $\chi(E{:}F)$ by applying a Riemann dissection to the interval, so as to obtain a causal chain consisting of a finite number of events, and then proceed to the limit as the fineness of the dissection tends to zero. By applying T17 and A9 we find that

Alternative to Railton's D-N-P ⟸

$$\boxed{\chi(E{:}F) = -\log(1 - e^{-\alpha T}),}$$

which is mathematically independent of β.

For large T, both Q and χ are exponentially small, but Q is smaller than χ and is much smaller if β is large. This is reasonable since, if β is large, the initial state makes little difference to the probability of being in the white state at the end of the interval.

Note that χ is the degree to which being in the white state rather than in the black state at the end of the interval was caused by being in the white state rather than in the black state at the start of the interval. A similar explicit description can of course be given for Q.

7. PARTIALLY SPURIOUS CORRELATION

A well known pitfall in statistics is to imagine that a statistically significant correlation or association is necessarily indicative of a causal relationship. The seeing of lightning is not usually a cause of the hearing of thunder, though the two are strongly associated. Such associations and correlations are often described as "spurious," a better description than "illusory." They may also be *partially spurious*, and the explicata for Q and χ should help with the analysis of such things. Smoke and dust might be a strong cause of lung cancer, but smoking only a weak cause. Even so, the correlation between smoking and lung cancer may be high if there is more smoking per head in smoky districts. I mention this only as an example, and have not made a special study of this problem.

Note that

$$Q(E{:}F.G/\overline{F}.\overline{G}) = Q(E{:}G|\overline{F}) + Q(E{:}F|G),$$

so that the tendency to cause can be split into components, somewhat in the manner of an analysis of variance. For example, the tendency for lung cancer to be caused by smoking and living in a smoky district, as against not smoking and living in a clean district, is equal to the tendency through living in a smoky district, given no smoking, plus the tendency through smoking, given that the