ANALYSIS OF CONDITIONAL EXPRESSIONS

Kenneth A. Presting

A thesis submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Arts in the Department of Philosophy.

Chapel Hill

2006

Approved by:

Keith Simmons

Thomas Hofweber

William Lycan

ABSTRACT

Kenneth A. Presting

Analysis of Conditional Expressions

(under the direction of Keith Simmons)

What David Lewis proved in 1976 was stronger than he realized. Not only can no system of logic can have a conditional connective with non-trivial conditional probability, but also no probability space can have even a single non-trivial conditional event. However, Lewis' definition of conditional connective is flawed, and does not apply to his original target, the Stalnaker/Thomasson C2 logic. Lewis assumed a property which Stalnaker's system does not have – McGee's export-import law.

Modal models of Stalnaker's C2 exist for every first-order model. Stalnaker's corner connectives, when interpreted as Lycan-style quantified conditionals, do have non-trivial conditional probability. Interpreting propositions as indicator functions instead of sets of possible worlds, the modern Kolmogorov theory of conditional expectation opens new possibilities for simultaneously modeling both objective and subjective probability as the expectation of truth. I use the new interpretation to defend Lycan's theory of conditionals against an objection from Dorothy Edgington.

TABLE OF CONTENTS

I. Prologue

Analytic philosophy is full of formulas – symbolic renderings of abstract assertions. Not every issue of every journal exhibits symbolic expressions. Sometimes we are permitted the luxury of simply reading one another's thoughts as expressed in our native tongue. But certainly no reader of analytic philosophy is ever surprised to see the device of the displayed line, and bare letters (perhaps from a foreign alphabet) where most readers would demand normal words.

Some of us, from time to time, take up the project of investigating the uses of those symbolic expressions themselves. This can occur from a sort of mathematical impulse, to understand the consequences of the formal rules which describe the symbols and their relations. Or there can be an empirical motivation, such as in Chomsky's development of formal grammars as models of human syntax.

The present essay will present the results of a formal investigation which seems to me to have interest not only in the technical arena or the empirical arena, but more generally in the regimentation of language – the philosophical project of giving reliable forms for the discussion of otherwise confusion-prone topics.

Not just philosophical discussions, but often human conversations in practical life, must address issues which are not issues of immediate, present facts, but instead are issues of future, past, or otherwise distant reality. It is a common experience to find significant utility in the anticipation of otherwise unknown events using information gleaned from the immediate environment, or past experience. We all have so much to gain from using our immediate perceptions to infer the presence (or imminent arrival) of other objects which we do not yet observe, it is no exaggeration to say our survival depends on it. The employment of perceptions to generate expectations and estimates is neither specifically human nor linguistic. It is characteristic of what we commonly call 'intelligent behavior' and we all observe it around us every day, among animals as well as among our human fellows.

A more specifically human phenomenon, and necessarily a phenomenon which can only be observed in a language, is the verbalization of reasons and reasoning as such – discussions which are not about the causal connections between events, the psychological relations between thoughts, or the functions of objects in systems, but instead whether assertions are consistent and properly ordered. All organisms display adaptive responses to present events. In many cases the complexity of the adaptation and the subtlety of the available stimuli lead us to describe some animals as 'anticipating.' But in human speech we observe a phenomenon which has notoriously resisted reduction to causal and material terms. Humans, it is generally assumed, are conscious symbol users, and conscious of their use of symbols. We are not only aware that we use words as symbols, we also use words to discuss, explain, and criticize each other in our use of words. We fault each other of overreaching for (or failing to recognize) consequences, as well as outright inconsistency.

There is a certain problem which guides the present work, but which I will neither discuss directly, nor attempt to solve. Quite a bit of human communications, including scholarship and science, is on the topic of concrete events, objects, and their relations. But those topics do not account for the whole of language use. Another large category of discussion deals with more abstract problems of how to manage discussions – which terms we should use, how to decide when those terms apply to events or objects, etc. All language users must, from time to time, ask one another about words, the meaning of words, and the use of words. Phrases such as "what do you mean," "what are you talking about," and "why do you say that" are idiomatic in English and no doubt have correlates in every living language. But there is another type of discussion which is at a third remove from concrete experience, and it is within this third level where we conduct scholarly discussion of logical consistency and inferential ordering among assertions. In these discussions, not only is our concrete experience with our native language weighted against the abstract topic at hand, but also evokes a sense of familiar security which is, I fear, positively misleading.

Any philosophical discussion of either logical consistency or logical inference in natural language is necessarily conducted in a context which is semantically closed. That is, when we take up the question of which assertions are consistent with others, or which propositions follow from others, we must both refer to expressions and predicate semantic relations of those expressions. In fact, everything we want to say about our language and its terms will be expressed in this very language, given that academic discourse is nowadays conducted in vernacular, rather than (e.g.) Latin.

Just to make my concerns concrete, let me present a few examples. The first sort of problem I have in mind is a "Conditional Liar" paradox, along the lines of:

(1) If this conditional is true,

then its consequent is false

or

(2) If the consequent of this conditional is true,
then its antecedent is false

Dialectically, it's easy to set these examples aside on the grounds that they are simply versions of the familiar Liar paradox, and no account of conditional semantics need take special account of them. But there is a closely related paradox which has been widely cited as a counter-example to an influential theory of conditional semantics, viz:

(3) If Reagan was working for the KGB,
We'll never know it.

I say (3) is like (1) and (2) because "knowing it" as used in the consequent requires "it" – the antecedent – to be true. All three of these conditionals create no logical problem simply by being true, or even by having a true antecedent. Apparently, it is the hope of adding to knowledge by means of inference which makes (3) a problem. The other two are clearly more narrow technical problems. Many debates in recent literature depend on examples involving obtaining information which might be interpreted as referring either to its own truth, or to the truth, reliability, etc of background assumptions. One reason why I have been interested in probability for quantified and modal logic is to provide a means for clarifying the various roles of self-reference, semantic closure, and conditional semantics in puzzles involving conditionals.

Another problem was reported more than a century ago in Charles Dogdson's "What Achilles said to the Tortoise," and thereafter discussed repeatedly. None of us can restrain ourselves from asserting that certain statements follow from certain others, without abandoning the enterprise of reasoning. Once we allow ourselves also to question a claim that a given inference is valid, we will have a new question for every answer. Therefore to explicitly identify all the assumptions behind any conclusion, one might require an infinite number of citations. This conclusion is not a *reductio* by infinite regress, but rather a demonstration that a language whose inference scheme is semantically closed with respect to entailment cannot also have the logical property of compactness, which is required by finite minds.

I raise these questions to give the reader some idea why I have chosen a particular scope for the following discussion of conditionals. My strategy, in view of the range of difficulties which seem to arise in a naturalistic context, is to begin instead from the wholly artificial perspective of formal languages. Whether such a beginning entails a permanent limitation on the ultimate theory is the subject of the next section.

II. Introduction – Inferences and Relations

Singular forms of the relational argument were understood as early as Plato, but the following version including generalization was not known to be formalizable in the days of C. S. Peirce and A. DeMorgan, who gave us the following example:

All mammals are animals

Therefore, any head of a mammal is a head of an animal

The critical innovation in this field is due to G. Frege, who introduced the use of predicate symbols with multiple variable positions, along with quantification operators:

$$(\forall x)(\mathrm{Mammal}(x) \supset \mathrm{Animal}(x))$$

$$\text{Therefore, } (\forall y)(\ (\exists x)(\mathrm{Mammal}(x)\ \&\ \text{Head-of}(x,y)) \supset (\exists z)(\mathrm{Animal}(z)\ \&\ \text{Head-of}(x,z))\ )$$

Clearly, the quantifiers in the conclusion have crucial roles, but the distinction between existential and universal quantifiers was already known to Aristotle. It was because Aristotle's theory of syllogism employed only terms identifying a category, rather than an abstract relation with explicit variables to show the connection between terms and instances, that Frege's innovation was needed. Of course, expressions with multiple variables were known in mathematics for more than a century before Frege (consider the role of equations and polynomial expressions in algebra, and vector functions in physics and calculus) so we should keep in mind that Frege's contribution was to logic as such, rather than the wider field of formalized reasoning.

Either way, it was understandable that Frege and (soon) Russell would have hopes for bringing new fields of language and argument under the umbrella of formal regimentation, using the new resource of terms with multiple variables. Later in the twentieth century, there has been another line of research, into syntactic features of language, with even more encouraging results for those who hope for formal understanding of speech. I refer of course to Chomsky's invention of formal grammars. Chomsky grammars are precise mathematical abstractions, whose properties may be studied by means of rigorous theorems. Yet these grammars may also be observed empirically to describe the features of a wide range of human languages, across many differences of culture and history. Generality on this scale is unprecedented in the field of social science. The most basic principles of economics have little to say where there is no money or trade, but what humans do not speak?

What we learn from the success of Chomsky grammars, as well as Chomsky's successors and competitors, is that at least in its syntactical aspect, reasonable scientists who are familiar with the complexities of the phenomenon in question, may rationally undertake the project of formal ly representing the entire human phenomenon of speech. Given the expansion of understanding which has occurred after the introduction of Chomsky grammars, students of language would (as a group) be derelict if at least some of them did not pursue the formalization of grammar as if that tactic could well encompass all of grammar. After all, how else could the limits of formalized theories become known?

Some readers will be aware that limitations have already been discovered in the formal study of grammar. The discovery of limitations in syntactic theory was almost immediate upon the invention of the discipline. Some grammatical systems, such as the regular grammars, efficiently discriminate correct from incorrect strings, but cannot make complex distinctions (such as parentheses nested to an arbitrary depth). Other grammatical systems, such as context-free or context-sensitive grammars, are less efficient but can discriminate more complex phrase structures. More general systems of grammatical rules, while precise and unambiguous in their significance and application, define languages which are undecidable, in the sense that a sequence of rule-governed tests for the correctness of an expression might never reach a conclusion.

The study of how formal representations of syntax are limited has proven to be as interesting as the positive results showing how much can be represented within formal syntax. One reason is because the limitations of formal syntaxare precisely co-extensive with independently discovered limitations in other abstract fields where rules, patterns, procedures or calculations are studied. Some of the first results in this field compared computer programs to grammatical systems, and found that every system of grammar corresponds to a computer program (i.e. to recognize the syntactically correct expressions in that grammar. Equally, every computer program corresponds to a system of formal grammar. The standard results in this field are now in the basic curriculum of computer science. Many different formal approaches to the concept of rule-governed procedure have been compared to Chomsky grammars with the stunning result that every known form of formal rule can be expressed as a rule of grammar. Everything from rules of arithmetical calculation to rules of solving equations in algebra or calculus, rules for geometrical constructions, rules of proof in logic – they are all reducible to one or another kind of grammar[1].

For example, consider a simple arithmetical function such as:

f(x) = 2x

This function can be represented by a set of ordered pairs, as may any function. In this case we could write an infinite family of ordered pairs in the form:

{0,0) (1,2) (2,4) (3,6) . . . }

Each parenthetical expression in this sequence is a string of symbols from the finite alphabet S = {'(', ')', ',', '1', '2', …, '0'}. A formal grammar could easily generate all and only the ordered pairs (7) which satisfy the equation (6). That is to say, the arithmetical relation of doubling a number is convertible into a grammatical property of symbol-strings. Arithmetical properties are equally translatable into questions of the provability of theorems in first-order logic. The reverse translation of logical properties into arithmetical relations was first undertaken by Kurt Gödel for the purposes of his famous work on the undecidability of arithmetic logic.

There is a well-known assertion, that the reducibility of rules to grammar which has been observed heretofore will continue to hold – that there will never be any form of calculation, procedure or rule which cannot be translated into a grammar. This thesis is due to Alonzo Church and Alan Turing. It is generally accepted, not least because there is a lively community of researchers who are trying to disprove it, without success. Confidence in Church's thesis depends on both the continued pursuit of counterexamples, and the ongoing failure of that pursuit. The epistemic status of Church's thesis would be a fine example of reflective equilibrium if there were more equilibrium among the various reflections on the principle. Any extension of computing technology has the potential for significant social and economic impact. Substantial resources are expended on research into novel paradigms of computation. This state of affairs is an interesting epistemic model for *a priori* investigation.

Because of its unique epistemic status, I propose to use Church's thesis as a point of departure for the study of semantics. The most important conclusion to derive at the outset, is that the general semantic project of giving formalized accounts of meaning (or truth, etc) for sentences can take some comfort in having a clear limit on the complexity of its problem, at least on the syntactic side. Simply put, we know what the grammatically correct sentences will look like, in the sense of having mathematically provable constraints on systems of grammar. Even better, we know that the complexity of sentences will never grow beyond the range of classically axiomatizable theories and recursively computable functions. Theories and functions are exactly the devices which Frege and Russell had in mind for their formal theories of meaning, so when we put together the linguistic discoveries of Chomsky with the Church-Turing thesis we have a strong endorsement of Frege's and Russell's semantic project.

Now we can turn to theannounced subject of conditionals. Frege/Russell semantics has had limited success with natural language conditionals, and limited consensus on where success might be found. Some of the best informed and most respected researchers on the semantics of conditionals have been led so far from the Frege/Russell paradigm as to argue that conditionals are not propositions and have no truth values. If this were true, it would be an important discovery. Among other important consequences, it would mark a significant departure from the logical atomism of the *Tractatus*, perhaps toward a richer view of meaning and language. Perhaps such a richer view is to be found in the direction of the *Philosophical Investigations*. That seems to be the direction taken by Grice and Jackson. Perhaps what is needed is a richer ontology. That seems to be the direction taken by Lewis and Stalnaker. Perhaps a richer set of truth values is required. That is the approach of Lucasewicz, and also the probability theorists such as Adams. Perhaps philosophical concern with conditionals should be satisfied with understanding their role in inference (more on this topic later). Truth-conditional semantics may have found its Waterloo in the conditional. After all, if a project goes on and on, with limited success, it is reasonable to ask again how we got into it in the first place. In this introduction I have provided a sketch of such an investigation, and concluded that Frege/Russell semantics is not just the only game in town, it is the only game.

While in a certain sense it is gratifying to find a long-term project endorsed again, in another sense it might have been a relief to have an excuse to abandon the old grind in favor of something fresh. But here we are – for all of the difficulties of finding a Frege/Russell theory which accounts for the meaning (or the use) of conditionals, I see no reason to conclude that we will find any other devices which will do better. The classical predicate calculus has no less representational complexity than any other formal scheme. Nor can we conclude that the problem of conditionals is beyond the range of logical theory or formal calculation – for example, in a special domain of social convention or pragmatic hermeneutics. I'm afraid that the familiar tools of formal representation will have to suffice for us – we won't find anything elsewhere which is both effective and reliable. The alternative is to offer our semantic investigations as evidence of a counter example to Church's Thesis. That would be a stirring discovery, but I prefer to work in a more conservative direction.

Having (tentatively, or better, methodologically) concluded that novelty is unlikely to be of much help in finding a basis for a formal semantics of conditionals, I am led to propose a theory of my own for the formal representation of the meaning of conditionals, in terms of their truth-values as well as their role in inference. In anticipation, I will offer a remark of Aristotle's, viz., "truth is like the proverbial door which none can fail to hit." (*Metaphysics*, bk. II pt. 1) It seems to me that the past century of research on conditionals has been largely on target, not only well-motivated but deeply revealing. Just about everyone has hit the door, and has the bruises to show for it. The recent encyclopedic works of Edgington and Bennett show how much has been learned about conditionals.

Roughly put, my project is to bring together several existing strands of research to show that the proper Fregean device for representation of conditionals in natural language is not any one propositional connective, but instead the family of relations which are clustered around the meta-linguistic implication relation or "gate" which is used in two forms, the proof-theoretic single gate " $\vdash$ " and the model-theoretic double gate " $\models$ ". More precisely, I claim that the success of the logical system we have inherited via Frege and Russell is largely due to the parallel structure exhibited among the two gate relations, the material conditional connective, and (here is a bit of novelty) the theory of statistical inference based on conditional expectation.

Few philosophers (and almost no analytic ones) earn their degree without seeing the soundness and completeness theorems of Gödel and Henkin, which are summarized by:

$$\models \psi \text{ iff } \vdash \psi.$$

Slightly less familiar, but no less important is the deduction theorem:

$$\phi \vdash \psi \text{ iff } \vdash \phi \supset \psi$$

which Bill Lycan paraphrases[2] into English as:

"Assuming P, Q" is equivalent to "If P then Q"

and I like to compare to Ramsey's famous footnote which is so often misquoted that I will reproduce it here:

> If two people are arguing 'If p will q?' and are both in doubt as to p, they are adding p hypothetically to their stock of knowledge and arguing on that basis about q … If either party believes non-p for certain, the question ceases to mean anything for him except as a question about what follows from certain laws or hypotheses.
> ("General Propositions and Causality", 1929)

In other words, an argument itself expresses a conditional assertion. If one has no commitment to the premises, then one has no obligation to the conclusion. The theorems which relate inference to entailment, and conditionals to proofs, are equivalence theorems. If the Ramsey footnote is comprehensible[3] at all, then it means both that conditional expressions assert inferences, and that inferences assert conditionals.

We should be clear that Ramsey has in mind not just propositional calculus when he mentions "if p then q". The subject of his essay is "General Propositions" and his comment holds just as true for reasoning about quantified conditional propositions, (as he called them, using proper Fregean terminology, "propositional functions"), as about Boolean truth functions. Recall that in a first-order proof of a generalization, the essential step is typically UG, "universal generalization," which allows the addition of a quantifier to a derived expression derived (subject to a certain special constraint). What I want to point out now is, even when proving formulas in an interpreted first-order language, the quantifier-free expressions are all Boolean propositions, with truth values coming straight from truth tables. In particular, any conditional expressions are material conditionals. The rest of this essay will hinge on the relation between the propositional material conditional, and the quantified general conditional. But we won't appreciate the differences unless we begin by noticing that the proof theory the quantified conditional depends on the material conditional.

The same is true for the semantic evaluation of quantified conditionals. If we want to know the truth of

$(x)(Fx \supset Gx)$

then we are asking exactly to know the truth for every x of the propositional function:

$Fx \supset Gx$

One of the important lesson philosophers learn in Logic 101 is that none of these results would hold regarding any connective other than the material conditional. To have compositional syntax mirrored in both a compositional semantics as well as compositional proof theory, we need the material conditional and *modus ponens*, respectively, in the molecular formulas and inferences.

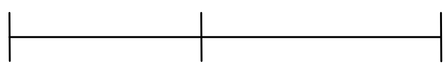
There remains one important analogy which has been anticipated to hold between formal reasoning and conditional expressions. In uncertain reasoning, should the proper degree of belief in a conditional expression match the degree of conditional probability for the consequence given the hypothesis? Alan Hájek calls this thesis "CCCP," the conditional construal of conditional probability. Jonathan Bennett calls it simply "The Equation."

David Lewis first showed in 1974 that a straightforward identification of the credibility of a conditional expression A→B with the obvious conditional probability P(B|A) is not possible. Lewis was especially concerned to refute the claim of Robert Stalnaker to have analyzed conditional expressions with his modal logic C2. Lewis' famous argument is mistaken.

III. Conditional Probability Diagrams

Before presenting Lewis' argument itself, I will prove a stronger result within probability theory, without reference to any logical connective. The proof will use a system of geometrical diagrams for representing conditional probabilities which will be helpful immediately, and also in subsequent discussions.

David Sanford has made the useful observation that the probabilities of events can be efficiently represented by a divided line segment of unit length, for example:

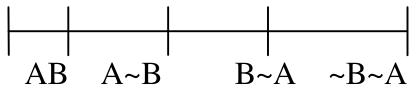


In this diagram we see a partition of all possibilities into two logically related events, called A and its complement ~A. This simple segment diagram neatly displays two features of a probability measure, simultaneously in one image.

(I) Probability is *normalized*. The whole segment has length 1

(II) Events are *complemented*. The dividing point which shows the end of one subsegment also shows the beginning of the remainder.

Adding another event to the segment diagram is also instructive. In a probability space generated by two events A and B, we can represent the four atomic events specified by conjunctions with the same segment divided three times, as follows:



Now we can see a few more features:

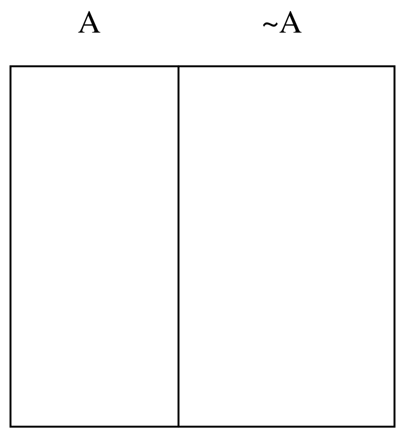
(III) Probability is *disjointly additive*. Subsegments which do not overlap combine to create increasingly probable events, whose probability is the total length of the subsegments.

(IV) *Order* and *Contiguity* are insignificant in a probability space. Either of the segments above might be drawn with subsegments in any order. As long as the subsegment lengths are the same, the probability represented is the same.

(V) Probability is a *dimensionless scalar* quantity. Probabilities for any two (disjoint) events, however simple or complex, may always be added.

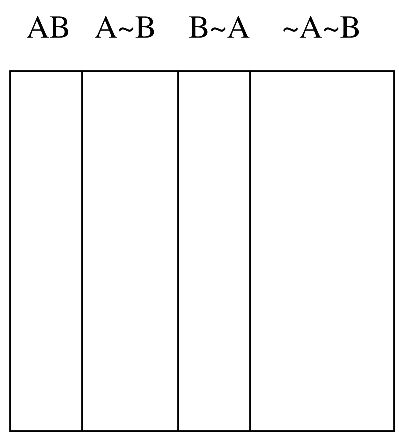
The more familiar Venn or Euler diagrams which depict events as regions in a plane somewhat obscure the crucial role of complementation, and also the dimensionless nature of the probability measure. I would like now to introduce a diagram which is a composite of features from the familiar Venn style with the Sanford style. These new diagrams will highlight the multiplicative relationships among probabilities, which are definitive of conditional probability.

Consider again the simple probability segment diagram from above, with only one division. Now, instead of drawing it as a segment of unit length, let us draw it as a square whose sides have unit length, and therefore also has unit area:



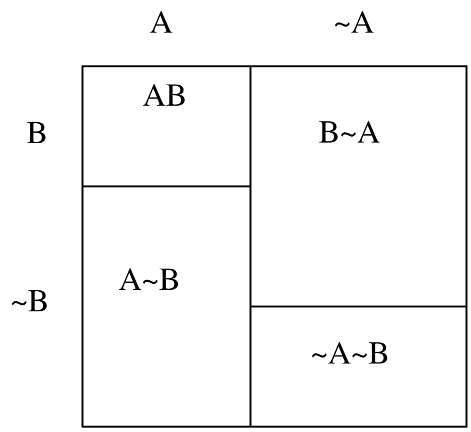
In the new square diagram, each rectangular subregion has a proportion of the total area equal to the corresponding segment's proportion of the total length. There are two ways to see this fact, geometrically and arithmetically. Geometrically, the square has been vertically divided in the same proportion as the segment, so the two areas have the same ratio as the lengths. The arithmetical relation is more significant. Every rectangle's area is its length times its width. Given that the top side of the square diagram simply is the segment diagram from before, then the numerical width of each rectangle is the probability value from the original segment diagram. Since the square diagram has unit lengths on each of its sides, the length of each rectangular region is exactly one unit. Therefore the area of the rectangles, being the unit length times the width, is numerically equal to the width – that is, the length of the short side. Only a dimensionless quantity like probability could have this feature, that one number would systematically represent both a length and an area at the same time

To see that the square diagrams are quite general[4], observe that any segment diagram is immediately translatable into a square diagram. We simply draw a square with the given segment diagram as its top side, and then extend each division point down across the area of the square. For example, to translate the four-section segment diagram from above we would draw:



Again, each rectangle has an area numerically equal to the width of its short side. Because the area of the square is normalized to one unit, just as the length of the side is normalized, the areas of the regions inside the square retain the nature of dimensionless scalar quantities. In familiar plane geometry, it would be meaningless to add an area to a length. The resulting quantity would be uninterpretable and useless for any further computations. But in the specialized context of normalized probability, lengths and areas retain the quality of dimensionless scalars.

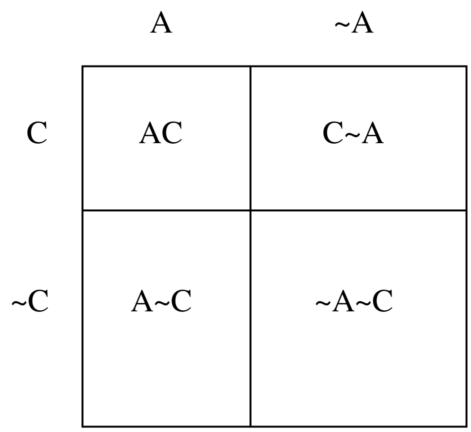
To represent conditional probability, we draw a square diagram not simply as an extension of the segment diagram, but instead make deliberate use of the two-dimensional nature of the square. Starting from the square diagram with the events A, ~A represented as vertical rectangles, we can represent the events B, ~B with arbitrary horizontal divisions, to show four conjunctions or "joint events":



Here, the vertical rectangle for event A has been divided into two events AB and A~B, in the same proportion[5] as the segment for event A was divided by its subevents in the segment diagram. The corresponding proportional division has been made within event ~A.

In this diagram, it is important to notice that the horizontal lines dividing B from ~B are not drawn as one continuous segment, horizontally across the page. This will happen in many cases as a result of constructing a square diagram for joint events. In this diagram, we have chosen the event-pair A,~A to be represented as contiguous rectangular shapes. If we were to draw this square diagram with B,~B depicted as horizontal rectangles, then we would be indicating a relationship between the events which might not always obtain.

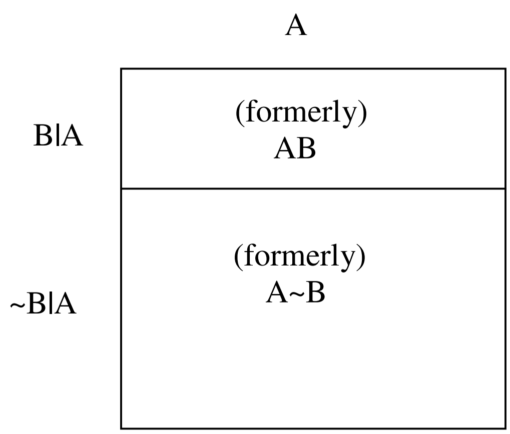
Only when the complementary event-pairs are *independent* can we draw a square diagram with all the basic ("generating") events depicted as rectangular regions. For example, suppose some event C is probabilistically independent of event A. Then we can draw the following diagram:



In contrast to the square diagram for A and B, this diagram shows both pairs A,~A and C,~C as rectangles. As a result, this square diagram has two segment diagrams as its top and its left sides. As we have noted above, both pairs of long rectangles A,~A and C,~C have the property that their areas are numerically equal to their widths. This creates a special relationship between the long vertical or horizontal rectangles, and the small rectangles formed by their intersections. The intersections of the regions are joint events denoted by the conjunctive expressions AC, A~C, ~AC, ~A~C. The probabilities of the joint events are given by their areas. Each joint event is itself a rectangle, and therefore its area is its length times its width. In the present diagram, the sides of each small rectangle are equal to the sides of the larger rectangles from the extension of the segment diagrams. Therefore the area for a joint event, say AC, is the numerical product of its sides, which are the probabilities of the basic events A and C. We have thus the product rule for the probability of independent events:

(VI) When events A,C are *independent*, the probability of the joint event is the product of the separate probabilities, $P(AC) = P(A) \cdot P(C)$

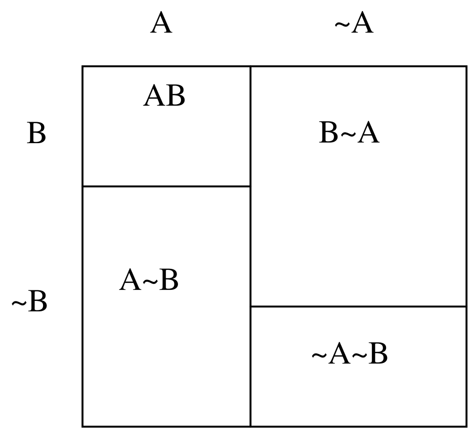
Now we can ask about conditional probability. Let's look at the diagram for A, B and ask, "What sort of diagram would correspond to learning that A has occurred while B is still unknown?" In other words, we want to draw a square diagram in which both B and ~B are represented as undetermined possibilities while ~A is no longer represented as a possibility. This is done in two steps, first by deleting from the square diagram everything outside the rectangle depicting the event A, then by re-normalizing the remaining rectangle, which is just "stretching" it until it again is a square. We can obtain a new probability space from the original, by including only the joint events consistent with A, then dividing by the fractional probability of A. Starting from the space of events above showing A and B, the result of conditionalizing on A is:



The new diagram shows just the joint events which are consistent with A, while the space as a whole has unit area, in virtue of the horizontal "stretching" or renormalizing, effected by dividing each probability by the fraction P(A). The horizontal lines in the new diagram each have length 1, because each horizontal length from the unconditioned diagram has been multiplied by the factor 1/P(A). In the unconditioned diagram, the horizontal divisions within the rectangle for event A all had length P(A). In the new diagram, the resulting length is P(A) · (1/P(A)), so of course each horizontal length is a unit.

The resulting square diagram appears to have a segment diagram as its left side, which is only partially correct. The segment diagram on the side of this square is not a segment diagram for the unconditioned events B, ~B. Nor is it a segment diagram for the joint events AB, A~B. Instead, it is a segment diagram for the "conditional" events B|A and ~B|A. The lengths down the left side are equal to the areas of the horizontal rectangles, which are (by definition) the conditional probabilities. After classical conditioning, the resulting square diagram will always have a segment diagram as is left side, originating from the joint events within the conditioning event, but now representing the conditional probabilities.

The renormalized diagram has lost information, which was present in the unconditioned diagram. Suppose now, instead of representing a change in belief brought about by learning that event A has in fact occurred, we want to represent only that we are thinking that A *might* occur or *might not*, and we need to think about the consequences of both A and ~A. For example, it might be up to us to choose between A versus ~A, and we are interested in the outcome of that choice. If this is our interest, we cannot use the classical conditioning diagram by itself, just because that diagram has no information about the consequences of ~A. Since we have seen how classical conditioning on an event results in a new square probability diagram, determined by proportions from the original, we can return to the generic diagram to see how conditional probability is already represented there.



By comparing the generic A, B diagram to the diagram after conditioning, we can see that any rectangle which is stretched sideways will grow in the sideways direction, but never in the vertical direction. The height of regions is unchanged by the horizontal renormalization. We have already noted that when rectangles have their long side equal to one unit in length, the area of the rectangle is numerically equal to the length of its short side. Here is the central observation for square diagrams: the length of these short sides is already the conditional probability of the small rectangle, conditioned on the event depicted in the vertical rectangle. Thus the conditional probabilities are observable in the generic diagram, prior to stretching or renormalizing. To have segments in the diagram whose lengths are these conditional probabilities depends on the partition of the diagram by vertical rectangles. Such a partition is always possible for any conditioning event. This is a consequence of the classical "law of total probability," which shows both the role of conditionalization on an event and its complement, as well as the re-weighting of the conditional probabilities by the unconditional probabilities. The total probability law for the case of two complemented events is:

(1) $P(B) = P(B|A)*P(A) + P(B|\sim A)P(\sim A)$

and, for the complement of B:

(2) $P(\sim B) = P(\sim B|A)*P(A) + P(\sim B|\sim A)P(\sim A)$

We can interpret the total probability formula in terms of the square diagrams. Each conditional expression, such as P(B|A) or P(B|~A), denotes the vertical length of a subsegment in the segment diagram, down the left side of a joint event, such as AB or B~A. The weighting factor P(A) or P(~A) is the width of the vertical rectangle containing the joint event. The total probability formula says that the area of an event is the sum of disjoint rectangles, each from a distinct vertical stripe.

Whenever we see in a theorem of probability theory that two probabilities are multiplied, it is implicit that the events are independent. Each term in the total probability formula is a product of probabilities, and in the square diagram we also see that each term is also the area of a rectangular region. In other words, the depiction of vertical rectangles in the square diagram shows the sense in which each joint event is independent of the rectangle containing it. The joint events are *conditionally* independent of the vertical rectangles which contain them.

We can see in the diagrams that vertical rectangles, stripes across the square, represent the antecedents or "denominators" of conditional probability. We also saw that (when organized into vertical stripes) events represented by rectangles have probability equal to their area, and conditional probability equal to their height. We also saw that for the long vertical rectangles themselves, the length of the short side is equal to the area. We saw that in the diagram of classical conditional probability, there does appear to be an event which corresponds to a conditional expression. The next section will explore the limits of this relationship.

IV. Lewis' Proof in Diagrams

The modified argument presented here uses only concepts from probability theory, thereby showing that Lewis' result is a fact of probability theory itself, rather than a limitation on the possible relations between probability and logic.

Suppose a probability space contains at least three events A, B, C, their complements, and all their Boolean combinations. We want to ask whether B is a "conditional event" which means an event whose probability is a conditional probability, in a strong sense. From Lewis (1976) we get two formal criteria for being a conditional event:

Definition. If $P(A) > 0$, B is a conditional event for C given A, iff both

CE1) $P(B) = P(C|A)$, and

CE2) $P(B|E) = P(C|AE)$ for any event E where $P(AE) > 0$

A word of explanation is in order for these two criteria. First, while CE1 is a natural property to expect from any event called "conditional," it still might be satisfied by an event B which had no logical, causal, or statistical connection to A or C. Condition CE2 solves this problem by requiring that B has the property of conditional equivalence to C given A, even when both are further qualified by any event E. In fact, CE1 is implied by CE2, when the additional event E is taken to be the universal event Ω.

(Jackson gives a good motivation for principle CE2 in terms of rational belief, in *Conditionals*. Blackburn's proof of Lewis' theorem also emphasizes the role of CE2.) CE2 can also be motivated by comparing the "|" symbol of conditional probability to conditional connectives in logic, where principles such as the following (called *importation*) often hold:

(3) $P \supset (Q \supset R) \Leftrightarrow (P \& Q) \supset R$

Ultimately we will see that CE2 is of limited applicability, but we need to understand its consequences, which are summarized in the following:

Theorem. Let A, B, C be events in a probability space <Ω, *F*, P>, with neither of AC or A~C having probability zero. If B is a conditional event of C given A, then we have

1) $P(AC) = P(A)\cdot P(C)$

2) $P(B{\sim}C) = P(C{\sim}B) = 0$

3) $P(B) = P(C)$

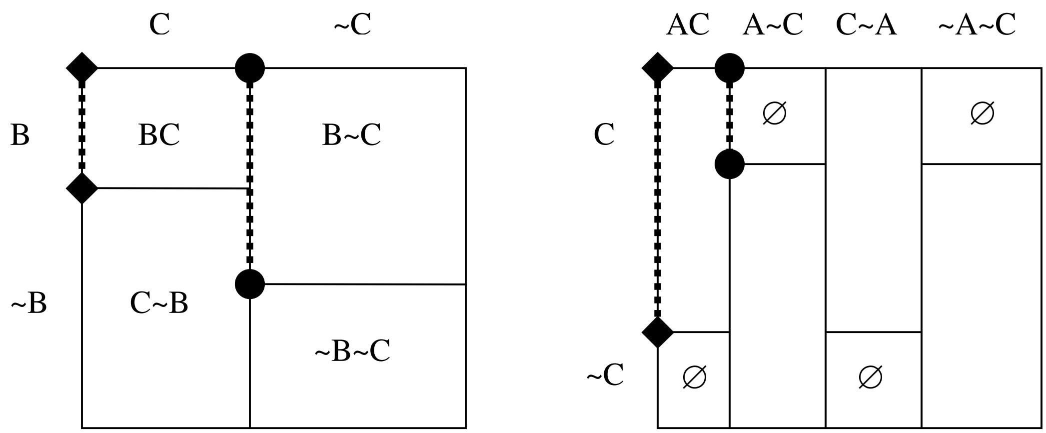
Proof:

Since B is a conditional event for C given A, it must satisfy CE2 for all values of E, in particular, when E is C itself, and when E is ~C. That is, substituting into CE2, we have the two equations:

$P(B|C) = P(C|AC)$

$P(B|{\sim}C) = P(C|A{\sim}C)$

We can draw and mark two square diagrams to record the information expressed in these two equations. We need two diagrams because two different families of denominators are used in the two equations, the family {C, ~C}for conditioning event B, and the family {AC, A~C, C~A, ~C~A} for conditioning event C. We can represent both equations by using a distinct congruence mark for each equation. The dashed lines with corresponding patterns (on the highlighted segments) indicate congruent lengths, much as hatch marks are used in figures for plane geometry:



The right hand diagram has certain regions which must be empty because they represent logically impossible events, and are therefore marked with the empty set symbol "∅". These impossible joint events must have probability zero. The hypotheses of our theorem ensure that the events AC and A~C have positive probability, and therefore the stripes representing those events must have positive width. A rectangle with non-zero width and zero area must have zero height, which means all regions labeled as empty must have zero heights (and their complements must have unit height.)

In these diagrams, the diamond-ended dashed lines appearing on event BC and event AC indicate the first equation, P(B|C) = P(C|AC). The circle-ended dashed lines, appearing on events B~C and A~C, indicate the second equation, P(B|~C) = P(C|A~C). We can read two of our intended conclusions off of these diagrams immediately. The picture shows

(4) P(C|AC) = 1 and

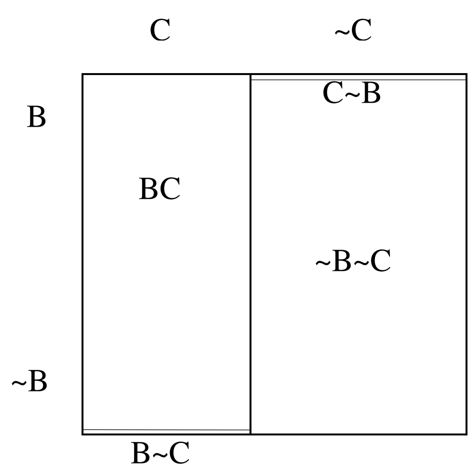
(5) P(C|A~C) = 0

which follows simply from observing which areas are empty. From the indicated congruences between the left and right diagrams, we also have:

(6) P(B|C) = P(C|AC) = 1 and

(7) P(B|~C) = P(C|A~C) = 0.

While we used empty set signs above to represent regions with zero area, we can show the same facts more graphically by drawing the zero-area regions using doubled lines, which visually suggests that they are logically present, but insignificant in probability. If we use the information in (6) and (7) to redraw the left hand diagram, relating B to C, we obtain the following:



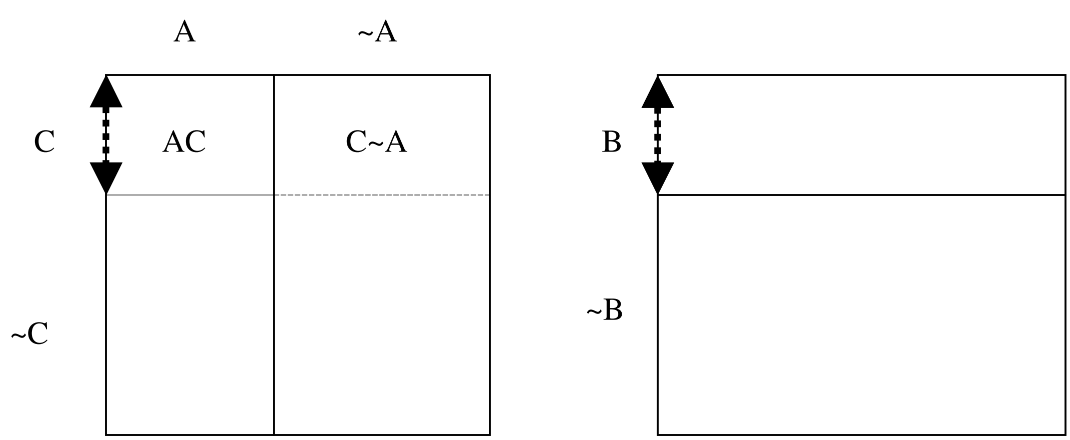
because P(B|C) = 1 is the length of the CB section of the stripe for C, and P(B|~C) = 0 is the length of the B~C section of the stripe for ~C. By noting the zero height of the joint events C~B and B~C, we have demonstrated

(8) P(C~B) = P(B~C) = 0

which is assertion (2) of the theorem. By noting the negligible differences between the joint event BC and each of the events B and C we obtain assertion (3):

(9) P(B) = P(C)

To conclude we willprove assertion (1) . We must draw a diagram representing the relation between A and C, with A, ~A along the top edge to show conditioning on A. If we can show that in such a diagram, the horizontal lines must be collinear, then we will have shown A and C are probabilistically independent. We will represent CE1, using two diagrams, one showing P(B) and the other showing P(C|A):



The hatched vertical lines indicate congruent segments, as required by CE1:

(CE1) P(B) = P(C|A)

The short side of event B in the right diagram is numerically equal to the area of the horizontal rectangle, since the long horizontal side has length 1. In the left diagram, the vertical side of joint event AC is the value of the conditional probability P(C|A), so we see in the diagram the relation asserted in equation CE1. The dotted horizontal line in the ~C stripe represents the fact we want to prove. Now we simply observe that the rectangular describing event B and the rectangle drawn with the dotted line in the left diagram have corresponding sides congruent, and therefore equal area. From this and the result that P(B) = P(C) we can conclude that the dotted line is in fact the boundary of C~A, which shows events A and C are independent.

This concludes the demonstration.

V. Non-Trivial Conditional Probabilities

Some of the most recalcitrant philosophical problems of the last century present us a with a sentence which has the logical form of a generalization, then in some way ask us to estimate our degree of belief in the sentence. Or, perhaps, we are to consider how to increase our degree of belief. Take for example, Nelson Goodman's uncertainty over whether, on a certain date, all the coins in his pocket were silver. Goodman assumed that the only way to answer his question was to examine each coin but then, as Goodman observed, the resulting observation no longer has the epistemic nature of a generalization. Or consider Carl Hempel's wondering whether all ravens are black. Hempel was concerned with how to increase one's confidence in the truth of a generalization. In each of these famous problems the presence of a universal quantifier is central to the philosophical difficulty. Also in each problem, issues of probability are either explicit or close at hand. Yet, none of the literature discussing these classic problems raises an issue which was identified in Frank Ramsey's famous paper, "General Propositions and Causality" (1990). The study of generalizations is fundamental to the two problems raised above, and to many other deep issues in the philosophy of science. We offer here an early, but still striking result obtained in the course of a broader investigation into the probability of generalizations.

A prominent feature of Ramsey's seminal work was the conjecture that an indicative conditional proposition, 'If F, then G', would have a probability equal to the conditional probability of ‘G’, on the assumption that ‘F’. This thesis is sometimes credited to Ernest Adams, who gave indirect support for it in his 1975 work on non-truth-functional conditionals. Around the same time, Lewis’s work showed that if a probability measure should give any truth-functional schema the conditional probability of its components, then that probability could take only four distinct values instead of the usual continuous interval. Recent work has extended, strengthened, and simplified the triviality results (Eells 1994 and Milne 2003).

We avoid Lewis’s triviality results by changing the context from a purely Boolean propositional logic to a first-order logic, where the Ramsey claim is borne out in almost its original form. Ramsey took quantified assertions to express not propositions, but instead propositional functions. We shall calculate conditional probabilities for general conditional propositions under a non-trivial measure; thereby demonstrating how a conventional application of probability theory to first-order models can assign a probability to a quantified conditional formula, as a function of its predicates’ extensions.

Before we present our general proof, it will be useful to provide a particular instance of our method. Suppose an urn contains nine blocks of two shapes and three colors, thus:

Spheres: 2 Yellow, 1 Red, 3 Blue

Cubes: 1 Yellow, 1 Red, 1 Blue

If one were to perform the experiment of drawing a single block from this urn, one would, under common assumptions of uniformity, assign a probability of 2/3 to the event of drawing a sphere, 1/3 to the event of drawing something yellow, and 1/9 to the event of drawing the yellow cube. But instead of just one, let us draw a number of shapes from the urn, without replacement. Assume we do so in a way that assigns equal probability to each subset of the objects in the urn[6]. Since we are considering a collection of objects rather than an individual, we may address ourselves to general assertions regarding the set of objects we select: What is the probability that the objects selected are all yellow? What is the probability that within a selection there exists a sphere?

Though we will soon part company with his analysis, we follow Gaifman (1964) in taking the existential quantifier as primitive. By a common intuition one may say 'There is an F' truly, when and only when the collection under discussion contains at least one member satisfying the open formula 'Fx'. Collections satisfying the universal are then defined by the usual double complementation, '$\forall$x(Fx)' iff '~$\exists$x~(Fx)'. For example, just as every particular object is either yellow or not yellow, so in every collection either there is a non-yellow object or there is not. If no object is non-yellow, then we say all are yellow[7].

The following calculations give answers to some probability questions we might ask about the subset we draw from the urn. Recall that for any set with N members, the set of its subsets is the power set, whose size is 2^N. For example, since three objects in the urn are yellow, the number of subsets containing no non-yellow objects is 2^3 or 8. Schematizing two predicates as: 'Sx': 'x is spherical' and 'Yx': 'x is yellow', we can calculate the following probabilities:

$$P\left[\forall x(Yx)\right] = \frac{2^3}{2^9} = 2^{-6} = \frac{1}{64}$$

$$P\left[\forall x(Sx)\right] = \frac{2^6}{2^9} = 2^{-3} = \frac{1}{8}$$

$$P\left[\forall x(Sx \;\&\; Yx)\right] = \frac{2^2}{2^9} = 2^{-7} = \frac{1}{128}$$

Thus we have values for the probability that our selected subset is entirely yellow, entirely spherical, or entirely both. We can use the values in the usual formula for conditional probability. The following calculations express, respectively, the conditional probability that our sample is all yellow given that it is all spherical, and that our sample is all spherical given that it is all yellow:

$$P\left[\forall x(Yx) | \forall x(Sx)\right] = \frac{P\left[\forall x(Sx \;\&\; Yx)\right]}{P\left[\forall x(Sx)\right]} = \frac{1/128}{1/8} = \frac{8}{128} = \frac{1}{16}$$

$$P\left[\forall x(Sx) | \forall x(Yx)\right] = \frac{P\left[\forall x(Sx \;\&\; Yx)\right]}{P\left[\forall x(Yx)\right]} = \frac{1/128}{1/64} = \frac{64}{128} = \frac{1}{2}$$

Next we calculate the probability of the corresponding quantified material conditionals. Note that five objects in the urn satisfy the open formula version of the material conditional, ‘$Sx \supset Yx$’, while eight satisfy its converse. Therefore the quantified expression will be satisfied by a number of collections equal to 2 raised to the respective exponent. We obtain the same probabilities again, but now by the direct calculations:

$$P\left[\forall x(Sx \supset Yx)\right] = \frac{2^5}{2^9} = 2^{-4} = \frac{1}{16}$$

$$P\left[\forall x(Yx \supset Sx)\right] = \frac{2^8}{2^9} = 2^{-1} = \frac{1}{2}.$$

Thus we see that an elementary conditional expression has a probability that is equal to the conditional probability of its consequent given its antecedent. It is easily verified that all the familiar laws of probability are satisfied as well, including Lewis’s postulates.

Seeing that all the probability theory here is standard, what distinguishes our analysis from the triviality results that have gone before? It is the use of two different but linked probability spaces for calculating the differing probabilities. One space is formed by the extensions of predicates as applied to simple objects, e.g., ‘Yellow’ or ‘Spherical’. The second space is the power set of the first[8]. Events in the first space are extensions of predicates, or Boolean combinations of predicates – the open formulas. Events in the second space are the extensions of quantified sentences, or Boolean combinations of sentences. The Lewis-style results apply to truth-functional compounds, but the quantified material conditional is neither a Boolean nor a truth-functional compound of its component predicates. Previous discussions of triviality have exclusively considered propositional logic, and thereby overlooked that the quantified conditional proposition can have a conditional probability without trivializing the entire measure.

Let us now consider a second example, which will clarify the assumptions we need for the theorem to follow. Suppose Professor Goodman had in his pocket, on that momentous day, not only a few coins, but also some other non-silver objects; brass keys perhaps. Suppose also that his coins are all dimes and quarters, except for one penny. Finally, suppose that he intends to determine whether or not all the coins in his pocket are silver, not by emptying his pocket and taking a census of its contents, but instead by taking a random sample. Unbeknownst to Professor Goodman, the penny in his pocket has become rather stuck to his front door key, which (for our purposes) means there is a high conditional probability that any sample which contains the penny will also contain the key, and vice versa. This is now a situation of systematic bias in the sampling, and therefore the relationship observed above with the urn will not obtain. If Professor

Goodman samples his pocket simply by reaching in and pulling out a subset of its contents[9], his observations of the conditional probability that a sample is all silver given that it is all coins will not be equal to the probability that in any sample, all the coins will be silver.

To see the problem, first observe that among all the subsets of all the objects, exactly half contain the front door key. Similarly, among the coins, exactly half of the subsets contain the penny. If there are no other correlations between objects and all the other objects have uniform probability, then the probabilities of the events "the key is in the sample" and of "the penny is in the sample" with both be one half. Nevertheless the joint probability, of finding both the key and the penny in a sample will not be the product, one fourth, as it would be if the separate events were independent. Instead, the probability of finding the penny together in a sample with the key is high, precisely because they are stuck together. This is the first sign of trouble.

What happens to the conditional probability that a sample will be all silver, given that it is all coins? Since the penny is just as stuck to the key as the key is stuck to the penny, any sample which includes the sticky copper penny will likely include also the front door key, and equally, the samples which include no keys are likely to exclude the penny. So the probability that a sample is all silver, given that it is all coins, is quite high. But the probability of the universal conditional, that in any sample all the coins will be silver, is not much affected by the correlation between the penny and the front door key. A sample which contains the penny stuck to the key fails just as much to satisfy "All the coins are silver" as a sample which contains a loose penny. Carefully avoiding the key/penny pair, while sampling from the pocket, might leave the probability of the conditional equal to the conditional probability. But we have assumed that Professor Goodman is unaware of this problem, so we should expect that his observations of the content of his pocket will diverge from the Ramsey relationship.

Unbiased random sampling is one of the bedrock requirements of sound experiment design and statistical analysis. We suggest that Goodman's arguments in *Fact, Fiction, and Forecast* (that only lawlike generalizations can be confirmed by their instances) appear rather less convincing when it is recalled that statistical sampling is often used to test accidental generalizations, but only in a context of careful experiment design, including randomization. Similar comments apply to Hempel's problem – it is only in a context of known experiment design that either positive or negative instances are relevant to estimation of probability. We cannot expand on these suggestions here, but our immediate objective is to show that attention to the unique characteristics of probability judgments regarding general propositions can reveal a new perspectives on otherwise refractory philosophical puzzles.

In the proof which follows next, we will assume a strong form of unbiased sampling, which comes from computing the probability of an event by simply counting cases[10]. Consider a first-order language *L*, interpreted into a non-empty finite domain U with N elements. We use the square brackets, as in '[Rx]' or '[∀x(Rx)]', to denote the extension of any expression, whether open or quantified.

Let $<U, \Sigma, P_0>$ be a Kolmogorov probability space (cf. Billingsley 1995), where U is, as above, the domain of interpretation for our language *L*, and Σ is a field of sets consisting of the extensions of the open formulas of *L*, with their intersections, unions, and complements. The measure $P_0(\cdot)$ corresponds to "base rate" phenomena[11], independent of the correlations found by random sampling, and is not relevant to the result.

Our second probability space is $<2^U, \Psi, P>$, where the domain Σ is, as above, the power set of the model domain U. In this space, the field of sets Ψ is generated not by the open formulas of *L*, but instead by the quantified sentences. The extension of a sentence '$\forall x(Fx)$' is given by:

$$[\forall x(Fx)] = \{ S \text{ in } 2^U \mid S \text{ is a subset of } [Fx] \}.$$

Since the new domain 2^U has 2^N elements, for any C in Ψ we define the probability measure P(C) as:

$$P(C) = \frac{\| C \|}{2^N}$$

Theorem.

Let *L* be a first-order language interpreted into a finite non-empty domain U, and let two probability spaces $<U, \Sigma, P_0>$ and $<2^U, \Psi, P>$ be defined as above. If 'Fx' and 'Gx' are any open formulas of *L*, then

$$P[\forall x(Fx \supset Gx)] = P[\forall x(Gx) \mid \forall x(Fx)].$$

Proof.

Let the number of elements in [Fx] be the non-negative[12] integer N_f, and let the number of elements in [Gx] and [Fx & Gx] be, respectively, the non-negative integers N_g and N_{fg}. Since [Fx] and [Gx] are subsets of the domain of interpretation U, 2^U will contain all subsets of [Fx] and of [Fx & Gx]. By definition, the extension of the generalization [$\forall$x(Fx)] is just the set of S in 2^U which are subsets of [Fx]. Similarly for [$\forall$x(Fx & Gx)]. The cardinality of these extensions is then

$$\| [\forall x(Fx)] \| = 2^{N_f}$$

$$\| [\forall x(Fx \,\&\, Gx)] \| = 2^{N_{fg}} .$$

With these values, we can calculate the conditional probability by:

$$\begin{aligned} P[\forall x(Gx) \mid \forall x(Fx)] &= \frac{P[\forall x(Fx \,\&\, Gx)]}{P[\forall x(Fx)]} \\ &= \frac{2^{N_{fg}}/2^N}{2^{N_f}/2^N} \\ &= \frac{2^{N_{fg}}}{2^{N_f}} . \end{aligned}$$

We can also calculate the size of the extension [Fx $\supset$ Gx]. Since [~Fx] is disjoint from [Fx], this is given by:

$$\| [Fx \supset Gx] \| = \| [{\sim}Fx] \cup [Fx \,\&\, Gx] \| = (N - N_f) + N_{fg}.$$

Then the corresponding generalization has the cardinality:

$$\begin{aligned} \| [\forall x(Fx \supset Gx)] \| &= 2^{((N - N_f) + N_{fg})} \\ &= (2^{(N - N_f)})(2^{N_{fg}}) \\ &= (2^N)(2^{-N_f})(2^{N_{fg}}) , \end{aligned}$$

and therefore the probability

$$\begin{aligned} P\left[\forall x(Fx \supset Gx)\right] &= \frac{(2^N)(2^{-N_f})(2^{N_{fg}})}{2^N} \\ &= (2^{-N_f})(2^{N_{fg}}) \\ &= \frac{2^{N_{fg}}}{2^{N_f}} \end{aligned}$$

which is, as desired, the conditional probability we calculated above.

To conclude this section I will consider relational expressions and other formulas with more than one free variable. Our concept of probability for every sentence is always the probability that the sentence is true. Therefore our concept extends immediately to more complex sentences. Since the truth of closed relational expressions is defined by the usual Tarski semantics, every sentence has a truth value at every point in the space of samples. The extension of the sentence is the set of points where it is true, and the probability of the sentence is the measure of its extension. This much is easy – it is just like every other probability space.

Closed sentences define binary-valued random variables, since they have a truth value at every point in the sample space. Expressions with single free variables define a set-valued random variable, since the expression has a set as its extension at every sample point. This set-valued random variable also defines a real-valued random variable. This works by taking the P_0-measure of the formula's extension at every sample point. We may now take the expectation of this variable over the sample space. Having constructed both binary-valued and real-valued random variables from logical formulas alone, it seems reasonable to anticipate that more complex statistical structures may also be found from logical constructions.

VI. The Dialectical Status of Lewis' Proof

At this point we have some conflicting indications, which will take some time to sort out. We have seen that a strong form of the triviality thesis is provable with no reference to logic at all There is not only no conditional probability connective, but also there are no particular conditional events. But we have also seen that an apparently contrary result is also provable. We calculated non-trivial conditional probabilities directly, and proved they will recur systematically.

But however the space of samples is constructed, as a simple probability space or a power set space, it is still a Boolean algebra, all the familiar propositional connectives are present, and they all operate according to the classical truth tables. If that is the case, then triviality results should hold in the space of samples just as in any other space. To this puzzle we now turn. We shall see that the crucial difference between the space of samples and other probability spaces is expressed in the importation principle. To clarify the role of importation, we will begin by reviewing David Lewis' original proof.

The article published in 1976 contains a significant error, which will be instructive to diagnose. First let me reproduce the relevant text:

> (1) P(C/C) =df P(CA)/P(A), if P(A) is positive.
> [. . .]
> Suppose that → is interpreted in such a way that, for some particular probability function P, and for any sentences A and C,
> (6) P(A → C) = P(C/A), if P(A) is positive;
> Iff so, let us call → a *probability conditional for P*. Iff → is a probability conditional for every probability function in some class of probability functions, then let us call → a *probability conditional for* the class. And iff → is a probability conditional for all probability functions, so that (6) holds for any P, A, and C, then let us call → a *universal probability conditional*, or simply a *probability conditional*.

Observe that if $\rightarrow$ is a universal probability conditional, so that (6) holds always, then (7) also holds always [1]:

(7) $P(A \rightarrow C/B) = P(C/AB)$, if $P(AB)$ is positive

To derive (7), apply (6) to the probability function P' that comes from conditionalizing on B; such a P' exists if P(AB) and hence also P(B) are positive. Then (7) follows by several applications of (1) and the equality between P'(--) and P(--/B). In the same way, if $\rightarrow$ is a probability conditional for a class of probability functions, and if that class is closed under conditionalizing, then (7) holds for any [2] probability function P in the class, and for any A and C. (It does not follow, however, that if (6) holds for a particular probability function P, then (7) holds for the same P. [3])

First Triviality Result

Suppose by way of reduction that $\rightarrow$ is a universal probability conditional. Take any probability function P and any sentences A and C such that P(AC) and P(A~C) are both positive. Then P(A), P(C), and P(~C) are also positive. By (6) we have:

(8) $P(A \rightarrow C) = P(C/A)$

By (7), [4] taking B as C or as ~C and simplifying the right-hand side, we have:

(9) $P(A \rightarrow C/C) = P(C/AC) = 1$

(10) $P(A \rightarrow C/{\sim}C) = P(C/A{\sim}C) = 0$

For any sentence D, we have the familiar expansion by cases:

(11) $P(D) = P(D/C) \cdot P(C) + P(D/{\sim}C) \cdot P({\sim}C)$

In particular, take D as $A \rightarrow C$. Then we may substitute (8), (9), and (10) into (11) to obtain:

(12) $P(C/A) = 1 \cdot P(C) + 0 \cdot (P{\sim}C) = P(C)$

With the aid of the supposed probability conditional, we have reached the conclusion that if only P(AC) and P(A~C) both are positive, then A and C are probabilistically independent under P.

There is some confusion in the text around assertion (7). I will focus on points in the text identified by bracketed numbers 1-4. The use of "always" at point [1] is potentially ambiguous between specifying a universal quantifier for each variable named in the following formula (7) including the probability function P(·), vis-à-vis quantifying only the propositional variables A, B, C. The use of "any" at [2] seems to be a plain indication that a universal quantifier should bind the probability function variable "P(·)" in formula (7). Unfortunately, the parenthetical qualification containing point [3] seems to revoke that impression. It cannot be the case that both (6) and (7) are intended to be universally quantified over P(·), if there are instances of P(·) such that (6) is true of P(·) while (7) is not. It is not clear whether Lewis intended his qualifications for probability conditionals with less than universal scope to apply here. I will return below to the significance of the qualification at [3], but for now I will examine the use of (7) in the proof of the "First Triviality Result".

Whatever restriction might have been intended for (7), no restriction is acknowledged when that assertion is used in the triviality proof. The first line of the proof (at "Suppose") specifies that the arrow → is to represent a universal probability conditional. Then a probability function denoted by the symbol "P" is introduced, also without restriction. Assertion (8) is the first line in the derivation, and it is introduced by appeal to the definition of the universal probability conditional in (6). Immediately thereafter, substitution instances of the problematic assertion (7) are introduced at lines (9) and (10), citing (7) without qualification. Note that each of (9) and (10) use the same symbol "P" as used in line (8). Superficially, there is an outright neglect of the qualification at [3]. This is enough to raise the question: why make the qualification, if it is to be ignored?

But perhaps this is an uncharitable reading. Perhaps the re-use of an identical symbol-type "P" in (8), (9) and (10) is not intended to imply that the same probability function P is denoted in each line. After all, the when Lewis gives an outline for deriving (7) he mentions a second probability function P'. But this issue is probably unrelated. Also, the parenthetical qualification itself used the single symbol "P" to mention apparently distinct objects. Perhaps Lewis intends the same flexibility in symbols again. In other words, "P" in (8) might denote one particular probability function, while "P" in (9) and (10) might denote other, distinct functions. This would not be plausible in most proofs, but Lewis did mention "the same P" which suggest some P's might not be the same.

Such a reading would make the proof useless. The proof is intended to be a *reductio* for the existence of a universal probability conditional. Within the scope of the *reductio*, it is assumed that probability function P assigns non-null probability to events AC and A~C. If any additional probability function is discussed within the *reductio*, then it will not follow the original function must be trivial. Whether or not there is any complex use of symbols with varying or ambiguous denotation, what the proof purports to show is that simply assuming non-null probability for AC and A~C suffices to entail triviality. If the proof shows merely that *some other* probability function is trivial, then the proof has no interest. The most charitable reading of the proof is one that treats the symbol "P" as denoting the same function, whenever that symbol is used.

A more attractive possibility is that the parenthetical qualification at [3] is unnecessary, or at least irrelevant to the proof of the triviality theorem. This is an interesting suggestion, and it seems to have been the silent conclusion of most readers, since the qualification which has caught my attention here has not been a prominent topic in other published literature. The problem with this interpretation is that the qualification at [3] is critical to thesoundness of Lewis' argument.

If principle (7) makes an assertion about a hypothesized probabilistic conditional connective, then as such, it cannot be is false. Instead, we must ask who else shares the hypothesis. Much of the subsequent literature on Lewis' theorems treats a condition such as (7) as definitive of the conditional connective. (Hajek and Hall's (1994) is the encyclopedic technical survey. They consider (7) as well as weaker alternative assumptions which lead to variations on the triviality result.) But Stalnaker, in his "Probability and Conditionals," explicitly rejects the proposition Lewis labeled (7). When Stalnaker comes to discuss criteria for probability functions which assign values to propositions formed with his corner conditional, he accepts a proposition corresponding to Lewis' (6) when he writes:

> The absolute probability of a conditional proposition – a proposition of the form A > B – must be equal to the conditional probability of the consequent on the condition of the antecedent.
>
> (16) $Pr(A > B) = Pr(B \mid A)$

Stalnaker observes that his (16) is too weak to give a useful restriction on probability values assigned to the corner conditional. Then he writes:

> The following generalization of the proposed requirement suggests itself:
>
> (17) $Pr(A > B \mid C) = Pr(B \mid A\&C)$
>
> This condition, however is clearly too strong.

Of course, Stalnaker's (17) is exactly Lewis' (7), with only a trivial change in notation. I will refer to both as "the importation principle" unless I need to emphasize one or the other author's text. In his discussion of importation, Stalnaker had anticipated its negative significance, without recognizing Lewis triviality theorem itself as a consequence. Stalnaker goes on to specify a rather more complex criterion requiring a hierarchy of "sub-functions" which represent iteration of conditionalization. Within his hierarchy, a probability function which satisfies (16) might not also satisfy (17), but for every function which does satisfy (16), there must be some other function in the hierarchy which satisfies an iterated application of (16). It is perhaps in relation to Stalnaker's complex criterion for "C2-extended probability functions" that Lewis' parenthetical qualification at [3] can now be understood. A defender of Stalnaker's system might well object to Lewis' inference of his (9) and (10)from (8), on the grounds that Stalnaker's hierarchical criterion for iterated conditioning allows for distinct probability functions to satisfy (8) versus (9) and (10), while Lewis' notation (and subsequent reasoning) requires the three equations to denote a single probability function. Stalnaker's paper is unmistakable in recognizing a danger from the principle Lewis adopts in his (7), and unmistakable in constructing a route around the problem. I conclude that Lewis' proof, whatever its other merits, does not defeat Stalnaker's claim to have developed a logic of conditionals in which the conditional expressions can coherently have the conditional probability of their components. Lewis' proof assumes a relation between probability and the conditional which Stalnaker's system deliberately and explicitly avoids. (I will discuss later in this section whether Stalnaker's own proofs to the same effect have been correctly interpreted)

The importation principle creates a subtle disanalogy between conditionalization in probability theory, and any logical connective which hopes to satisfy Lewis' assumptions. The disanalogy occurs because logical conjunction and set intersection are commutative operations, while repeated conditioning in probability theory commutes only in elementary cases. Here are two instances of (7):

(7) $P(A \rightarrow C/B) = P(C/AB)$

(7') $P(B \rightarrow C/A) = P(C/BA)$

If we require a probability conditional $\rightarrow$ to satisfy (7) for all values of A, B, then it must also satisfy (7'), because each is a substitution-instance of the other. But the right-hand sides of (7) and (7') are identical in any language or logic which allows conjunctions to commute. Certainly this is the case in every language studied heretofore in relation to the triviality thesis, and in every familiar philosophical development of probability theory.

Now, if the right-hand sides are equal, then by transitivity so are the left-hand sides, giving us:

(7X) $P(A \rightarrow C/B) = P(B \rightarrow C/A)$,

which might be less plausible. Stalnaker's hierarchical criterion for iteration does not validate any of (7)-(7X).

Formulas similar to the importation principle can be derived in classical elementary probability theory, as Lewis observes. Probability theory itself has no explicit notation for repeated conditionalization, but it does employ the general result from measure theory that the set of probability functions on any measurable space is closed under conditionalization. So, for every probability function $P(\cdot)$ and event B we define the function symbol $P_B(\cdot)$ as a conditional probability by:

$$P_B(\cdot) =_{df} P(\cdot|B)$$

Then

$$\begin{aligned} P_B(C|A) &= P_B(AC)/ P_B(A) && \text{by definition of } P(\cdot|\cdot) \\ &= P(AC|B)/P(A|B) && \text{by definition of } P_B \\ &= [P(ABC)/P(B)] / [P(AB)/P(B)] && \text{by definition of } P(\cdot|\cdot) \\ &= P(ABC) / P(AB) && \text{canceling } P(B) \\ &= P(C|AB) && \text{definition of } P(\cdot|\cdot) \end{aligned}$$

In the Stalnaker C2 logic, the importation formula is not a theorem, among many other familiar principles of conditional reasoning which fail notoriously for subjunctive reasoning. Stalnaker catalogs several of these in "A Theory of Conditionals" (*IFS*, p. 48). The probabilisticanalog of importation fails outside the context of elementary probability theory, which includes only probability defined on Boolean combinations of events. We shall see in subsequent sections that modal assertions or predicate expressions (which in

Fregean terms are propositional functions rather than propositions) require more complex probabilistic constructions such as random variables and expectations[13]. We will address the meaning and use of the importation formula in relation to Stalnaker's system C2, where we can use it to show not only that Lewis' criticisms are mistaken, but also that while Stalnaker was correct to take them seriously, there was also no need to give up on conditional probability for C2 conditionals – only to interpret the probability more carefully.

Even supposing we have found a fatal objection to Lewis' use of his proof against Stalnaker, there are still many other arguments to the triviality conclusion. Perhaps most important, it is well known that Stalnaker accepted Lewis' argument, and extended Lewis' results (in the "Letter to Van Fraassen".) There is a potentially large body of literature to be examined here, but I will discuss only Stalnaker's consideration of his (17) which I quote below, since it contains what I take to be the essential issue.

To explain why (17) is too strong, Stalnaker makes an argument which anticipates Lewis' own device of conditioning on a component of the conditional proposition:

> In fact, the adoption of the strong requirement, (17) would trivially give all counterfactual propositions a probability of one, collapsing the distinction between knowledge and necessity, and reducing the probability system, P3 to one roughly equivalent to P1. This can be seen by the following argument: Suppose A > B represents a counterfactual – that is a conditional proposition whose antecedent is known to be false. Then
>
> $\Pr(A) = 0$, so $\Pr(\sim A) = 1$.
>
> But for all C such that $\Pr(C) = 1$ and for all D,
>
> $\Pr(D \mid C) = \Pr(D)$.
>
> Therefore
>
> $\Pr(A > B) = \Pr(A > B \mid \sim A)$.
>
> But by requirement (17),
>
> $\Pr(A > B \mid \sim A) = \Pr(B \mid A \,\&\, \sim A)$,
>
> Which always equals one. Therefore $\Pr(A > B) = 1$. But all we assumed was that A > B was counterfactual.

Stalnaker's last inference refers to a conditional probability whose hypothesis is the contradictory formula 'A & ~A'. This is a problem with his argument, even though his own (17) does not exclude conditionalization by absurdity. His system C2 would not have any significant limitation based on failure to accommodate conditionalization on absurdities. The real error in his argument is elsewhere.

Stalnaker seems to me to be mistaken (or at least to be working in a direction different from mine) in attempting to define the probability of a proposition without reference to the truth of the proposition. The problem created by this separation between the truth value (or values) and the probability of propositions appears in the first step of his argument, that the probability of the antecedent in a counterfactual conditional must be zero, on the grounds that it is known to be false. Where my approach differs from Stalnaker's is that I define the probability of the proposition in terms of the points where it is true. That is, I use the various truth values of a proposition to compute its probability. I depend on a proposition to be true in some instances and false in others to give the proposition a fractional probability, greater than zero but less than one. As Stalnaker is arguing above, he infers that the antecedent of a counterfactual must have probability zero simply from the fact that the proposition is known to be false in the actual world. I do not so much object to Stalnaker's view as wrong, but rather I am suggesting that he is making inadequate use of the resources available within probability theory, and within his C2 logic, to represent the complex relationship between truth and probability. There is, I will argue, no contradiction between a proposition being known true or known false in the actual world, but still having a fractional probability for an epistemic subject in that world. A modal logic such as C2 has exactly the formal resources to resolve the apparent *aporia*. The resolution follows Ramsey's footnote – for anyone who considers the antecedent of a conditional to be a settled falsehood, to evaluate either truth or the probability of the conditional is only to investigate which conclusions might follow from certain laws or hypotheses. That will be the subject of the following section.

VII. A Probabilistic Theory of Backward Subjunctives

The theory I will present in this section is intended to provide an example of how propositions may be known with certainty to be true or false in the actual world, but still assigned a probability strictly between 0 and 1, when that assignment is part of a semantic analysis. The theory is not intended as a serious candidate for a semantic analysis of English backward subjunctive conditionals. A thorough analysis would address both deterministic as well as statistical cases, bridging the two by means of quantifiers. The model here has just enough initial plausibility, I hope, to serve as an illustration of probabilistic techniques, and the reader is requested not to jump too quickly to seeking counterexamples. There will be plenty of those, but for now I simply want to show how a probabilistic theory gets past the problem of probability for known events.

In Chapter 17 of “A Philosophical Guide to Conditionals” Jonathan Bennett expresses a certain degree of dissatisfaction with existing possible-worlds based analyses of backward subjunctives. I will show an alternative theory of such conditionals based on probability which bypasses some well-known problems. Consider one of Bennett’s examples:

(1) If the plane had arrived at 2pm, it would have to have left at noon

Bennett observes that conditionals of this type can be identified by the feature of “inference to the best explanation” which connects the antecedent to the consequent. Inferences of this type are commonly studied in statistical theory, and a surprisingly effective account of them can be stated by appealing to one of the basic mathematical results in the field, the Radon-Nikodym theorem. The structure within which this theorem is stated is somewhat more complex than most philosophers expect in a probability theory, but the extra structure will prove to be ideal in understanding the role of context in the evaluation of these conditionals.

An inference to a best explanation is based, at some point, on a range of propositions which are potentially available as explanations. Specifying this range of possible explanations is critical to application of the Radon-Nikodym theorem. Typical philosophical examples of backward subjunctive conditionals identify only the one explanatory possibility which is asserted as the subjunctive consequent, but it is not uncommon to see discussion of other possibilities in debate over an example. If no other possibility is identified by context, then at least the complement of the antecedent is implicit as an alternative.

The Radon-Nikodym theorem operates (in the context of inference to the best explanation) by asking, first, for a probability distribution on propositions or expressions of a language which take values of true or false on possible worlds. Second, the range of potential explanations R must also be identified, and must be a subset of the expressions in the language. Given this information, the theorem guarantees that each of the various possible worlds will have its own coherent probability function for backward conditional expressions of the form:

(2) If ---consequence---, then ---explanation---

These probability values are written as a three-place function

(3) $P(C \mid R)(\omega)$

which expresses the dependence of this probability measure on all three factors – the consequence proposition C, the range R of explanation proposition, and the possible world ω which is, intuitively, the "point of view" from which the conditional is evaluated. This value is known as the *conditional expectation of* C *given R at* ω.

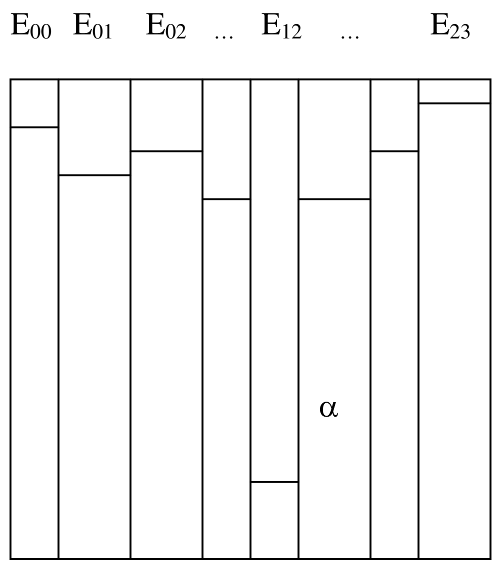
The semantic hypothesis I am offering based on this construction is, first, that a backward subjunctive conditional is always used, evaluated, or expressed in relation to (1) a probability function, (2) a range of potential explanations and (3) a specific world α, the "view-point." Second, I hypothesize that the backward subjunctive conditional is true if and only if the conditional expectation of the antecedent is maximized by the stated consequent, or,

(4) "If C given E" is true iff $P(C \mid E)(\alpha) \geq P(C \mid F)(\alpha)$

for every proposition F in the range of possible explanations. The "backward" nature of these conditionals is brought out in formula (4).

Any or all of the range of possible explanations, the view-point, etc. may be implicit, assumed, given by context, or otherwise partially obscure. It may be found in future work that uncertainty in evaluating particular backward subjunctive conditionals is explainable in terms of uncertainty in these parameters, but at present I intend to defend only the two theses already stated.

To get a first approximation of how this theory works, let us examine Bennett's example given above as (1). Here the view-point world is the actual world, and while the range of possible explanations is unclear, we get a plausible reading of the sentence by taking the range of propositions in question to be the various departure times for the plane. Finally, we suppose from Bennett's discussion that the plane is expected to take two hours to make the flight, which amounts to setting a high probability on propositions which assert flight times close to two hours, and dropping rapidly in probability if times different from two hours are asserted. This situation is conveniently represented by a square probability diagram:



Along the top side of the diagram are arrayed the various departure times for the flight, E_{00} through E_{23}, labeling the vertical rectangles. The vertical divisions which partition this diagram represent the events in the range of possible explanations. The vertical striping, labeled by the finite partition E_{00} – E_{23}, indicates that the diagram represents conditioning on the events from the given partition. On the left hand side are shown the two complementary events C, ~C to represent the consequence that the plane arrives at 2pm (or not). The horizontal divisions within each vertical column show the distinction within each column between the areas representing the success of the consequence C, versus its failure ~C. The letter α for the actual world is placed in the diagram in the region of ~C and ~E_{12} to show that both components of the conditional are false in the actual world – that is, the conditional is counterfactual.

An attractive feature of the analysis proposed here is that determining the truth value of the conditional does not depend on identifying any world other than α as similar, nearby, or related at all. All points in the probability space are distinguishable only by the regions where they are located. The only relations that are relevant in the conditional probability analysis are the relations between the events, which are depicted in the square diagram. A focus on relations among events and propositions, in an analysis of a subjunctive conditional, is what we might expect in view of conditionals role in reasoning. The probabilistic hypothesis here carries no burden of metaphysical assumptions. Here, as in deductive arguments, laws are indistinguishable from hypotheses.

The conditional expectation obtained from the Radon-Nikodym theorem cannot distinguish among possible worlds on any finer scale than the partition provided by the range R of possible explanations. The conditional expectation must take a single constant value within any cell of the generating partition. Therefore comparison of two competing explanations E_i, E_j can employ any worlds ω_i, ω_j for which the competing explanations are true. In the general case[14] (say with continuous variation in some relevant parameters) the conditional probabilities would be computed by taking the integral of the conditional expectation. Specific features of the worlds within each partition cell are irrelevant to these calculations. Variations which are not recognized by the partition are "averaged out" within each cell, which will reduce the role of so-called *ceteris paribus* considerations in the semantics of conditionals.

Finally, we should note that in a backwards subjunctive conditional, the consequence event C is typically not logicallydefinable in terms of the conditionalizing partition events $E_0 \ldots E_n$, or included among the range of possible explanations R. If it were, that would be the dialectical equivalent of considering the event one seeks to explain among the candidate explanations. Then the truth-conditions given here would identify the consequence itself as its own best explanation, which is absurd. Clearly some attention must be paid to the content of R. This consideration shows that the minimal two-cell R consisting of nothing other than the stated explanation and its converse might well be taken as determining the literal meaning of the conditional, with any more detailed range to be justified only when context, charity, etc. requires it.

Ranges of events which are strictly smaller than the fully detailed probability function allows for are called "sub-algebras". To define conditioning on these sub-algebras is one of the central technical advances in the modern theory of conditional expectation. In Bennett's example, we see the logical independence of the conditions from the consequences by noting that the arrival time of the plane is logically independent of the plane's departure time, however strong the statistical correlation may be. In terms of the square diagrams, logical independence is represented by non-empty joint events in both of the positive and the negative portions of every stripe. In the proof of Lewis' triviality theorem, the crucial step involved finding that certain joint events were necessarily empty while others necessarily had probability 1. But finding events with that property was a consequence of incorporating the consequence event C into the pattern of stripes which condition the probability diagram. We have thus found what might be considered a principled reason for excluding such constructions from the legitimate interpretation of counterfactuals, but we will find better reasons yet.

Summing up, modern probability theory does not require that a conditional probability function represent learning of the occurrence of a significant event by losing all information about other events. There is natural interpretation of the modal concept of "possible world" as a point in the sample space of a probability function. That interpretation allows a useful degree of flexibility for the semanticist who wishes to model the truth conditions of subjunctive or other hypothetical constructions in terms of probability calculations.

VIII. Lycan, Stalnaker, and Conditional Expectation

William Lycan's theory of conditionals[15], while not a probabilistic theory, shares the quantificational orientation of the power set construction described above. Both Lycan's theory and Stalnaker's, then, have important common elements with the present work. All three theories are subject to misunderstanding if familiar principles which apply to elementary probability are mistakenly applied to the more general concept of expectation of a propositional function.

To see the outline of the problems, consider the generalization whose probability was calculated in the case of the urn with nine blocks. By hypothesis, the urn contains many non-yellow blocks and many non-spherical blocks. Therefore the generalizations "All the blocks are yellow" and "All the blocks are spherical" are false, and known to be false. Given this information, how can there be any coherent assignment of non-zero probability to either assertion? Perhaps what we saw in the "non-triviality" proof was an intriguing exercise in abstract calculation, but are there grounds for calling it a calculation of the *probability*, literally so called, of any general proposition?

Dorothy Edgington makes a related criticism in unpublished[16] comments on Lycan's theory. Her example involves a conversation between a doctor and a patient who are considering an imminent operation. Suppose the patient asks the doctor, "Will I be cured if I have the operation?" Edgington's doctor replies, "I think so. I think, but I am not certain, that you will be cured if you have the operation. In my view it's about

80% certain that you will be cured if you have the operation." Edgington observes, intending to refute Lycan:

> The doctor clearly envisages a real possibility in which I have the operation and am not cured. That is why she is less than certain that if operated [upon] (O), I will be cured (C). The doctor knows that not all real and relevant O-events are C-events. Lycan's truth condition is not satisfied and [is] known to be not satisfied. But what is known to be false is not something the doctor should be 80% certain of.

Let us take Edgington's comment at face value, and suppose with her both that the general conditional proposition

(1) $(x)(Ox \supset Cx)$

is false and known false, and that according to Lycan's semantics for conditionals, this gives (1) a probability of zero. There is a problem here and we have collected most, but not all, of what we need to resolve it. We will address this problem in three steps. First we will discuss the practice of modeling propositions by sets of possible worlds. Next, we will discuss conditional probability in contrast to conditional expectation. Finally we will re-open the question of conditional probability in the C2 modal language. Then we will reply.

In probability theory there are two likely candidates which might serve as formal representatives for propositions. While events – sets of possible outcomes – are the most likely choice, we will see that the indicator functions of events have distinct advantages. (This comparison will bring out a particular advantage of using either a modal semantics or a first-order model for representation of probability theory. The classical, elementary theory of probability based exclusively on Boolean algebra or propositional calculus does not allow for the indicator function treatment I will outline here.)

Indicator functions are familiar to most philosophers from set theory, as the membership function or characteristic function of a set. For any set S in some universe Ω, the indicator function 1_S is the {0, 1}-valued function on Ω which equals 1 for members of set S, but 0 for other elements of the domain:

$$1_S : \Omega \to \{0,1\} : 1_S(\omega) = 1 \text{ iff } \omega \in S, \text{ and } 1_S(\omega) = 0 \text{ otherwise.}$$

Clearly there is a one-one correspondence between sets and their indicator functions, and for many purposes one would simply identify the two objects with each other, ignoring any distinction. But in probability theory we become interested in both the logical relations between propositions, and in the numerical relations between their probabilities. In particular, we become interested in both additive and also fractional or proportional relations among probabilities. Therefore the numerical character of the indicator function becomes convenient. In particular, we can take the average value of an indicator function in the same way we can average any collection of values. In a finite collection, we add all the values (perhaps weighted according to their probability) and divide by the number of values (or by the total of the weights). In an infinite collection we would integrate the function, and divide by the measure of the domain, but infinite collections will not be discussed further. The average value of a function, including an indicator function, is its ***expectation***. The expectation of an indicator function is exactly the probability of its indicated set.

For the purpose of explaining how a proposition can be both determined for certain and yet also have a probability different from 0 or 1, the indicator function is perfect. First, every indicator function is always both defined and either 0 or 1 at every point. Next, the indicator function has an average over all the points in the domain. The second fact depends on the first – computing an average value requires a collection of values to be averaged. Here is the critical point, both for Lycan's theory and for mine: If any proposition is to have a probability in the sense of an expectation calculated as an average of its indicator function, then that proposition must have an indicator function. Of course an indicator function must be defined at all the points in the probability space, and this is just what Lycan's theory and mine both do for the quantified conditional propositions. Lycan does not himself take up the problem of defining a probability space or a probability measure, or any other issue regarding the probability of conditionals. Lycan's theory contains no impediment whatsoever to embedding his account of conditionals within an account of probability for quantified expressions. On the contrary, if the points in a probability space are possible worlds, then specifying truth conditions for the quantified conditional in each possible world is a necessary precondition for calculating an expectation for the quantified conditional.

At this point it may seem that I am inviting Lycan out of his own embattled position into my own even flimsier one. This is merely a diversionary tactic. Instead, I will attempt to ally both Lycan's theory and my own with Stalnaker's C2. Together we may yet repulse the axis of the three theses, "No Truth Values", "Conditional Probability is Trivial" and the anti-modality, "The World is Everything That Is The Case."

Conditional expectations are the key to this strategy, and a modal logic is a natural setting to observe the advantages of comparing propositions to indicator functions. Where a first-order expression has a satisfaction value for every element of the domain, a modal expression has a truth value for each possible world. We will see that the maneuver of identifying the probability of a proposition with the expectation of its indicator function, when combined with a conditional-expectation account of conditional probability, will establish a formal context in which the importation assumption fails, and the triviality results do not hold.

We can see the outline of this solution by looking at the truth-conditions for Stalnaker's conditional, in his system C2. A Stalnaker conditional A>C has a truth value relative to a an *actual world* α and a *selection function* $f(A,\alpha)$. Intuitively, the selection function picks out the unique closest possible world to α in which the proposition A is true. The semantics of the corner-conditional states that A>C is true in the actual world α, iff C is true in world β whenever $f(A,\alpha) = \beta$. The conditions which Stalnaker sets for his selected-world function are met in the context of the power set of a first-order domain. Let's see how this works intuitively first, and then technically.

Consider a simple quantified sentence, (x)(Fx), interpreted in a domain Ω. Then sentence is true in Ω iff every x in Ω is F. In the probability space of samples whose domain is 2^{Ω}, any universally quantified sentence is satisfied by some sample from Ω, perhaps the empty set. Now take the union of all the samples which satisfy (x)(Fx) – that is, the largest sample from Ω which will satisfy (x)(Fx). If we interpret all the samples from 2^{Ω} as possible worlds, then this largest sample which satisfies (x)(Fx) satisfies the first condition set by Stalnaker for his selection function – it is unique, and (x)(Fx) is true there.

Adapting the notation of Nute (1984) we let *L* stand for the set of all quantified sentences and write $[\phi]$ for the collection of all samples which satisfy ϕ; that is, when ϕ is in L then $[\phi] \subseteq 2^{\Omega}$. We define a selection function that picks out from the actual world all those domain objects which satisfy the hypothesis ϕ relative to the base world i:

(S0) $s(\phi, i) : L \text{ x } 2^{\Omega} \rightarrow 2^{\Omega} : s(\phi,i) = U[\phi] \cap i$

where $U[\phi]$ is the union of all the samples in $[\phi]$. For the accessibility relation in 2^{Ω} we take the subset relation:

(R0) $\forall\, i, j \in 2^{\Omega},\ i\ R\ j \Leftrightarrow j \subseteq i$

Verification is trivial that s and R as thus defined satisfy the following properties:

(S1) $j = s(\phi, i) \rightarrow j \in [\phi]$

"a selected world satisfies its hypothesis"

(S2) $j = s(\phi, i) \rightarrow i\ R\ j$

"a selected world is accessible to the base world"

(S3) $s(\phi, i)$ undefined $\rightarrow.\ (\forall j)\ i\ R\ j \rightarrow j \notin [\phi]$

"some world is selected, if any accessible world satisfies the hypothesis"

(S4) $i \in [\phi] \rightarrow s(\phi, i) = i$

"the base world is selected, if it satisfies the hypothesis"

(S5) For any ϕ, ψ in L: $s(\phi, i) = j \in [\psi]$ and $s(\psi, i) = k \in [\phi] \rightarrow j = k$

(an analog of the anti-symmetric property for weak partial orders)

(S6) $i \in [\phi > \psi]$ iff. $s(\phi, i) \in [\psi]$ or $s(\phi, i)$ undefined

(the semantic definition of the corner-conditional)

To illustrate, we can verify the accessibility and the "anti-symmetric" properties, (S2) and (S5) as follows.

(S2) Let $j = s(\phi, i)$. Then by (S0), $j = U[\phi] \cap i$ which entails $j \subseteq i$. Therefore, $i \, R \, j$, by (R0).

(S5) By (S0), $j = U[\phi] \cap i$; by hyp., $j \in [\psi]$ and $j \subseteq U[\psi]$. So $j \subseteq U[\phi] \cap U[\psi] \cap i$. Then $j = U[\phi] \cap U[\psi] \cap i$ because any $x \in U[\phi] \cap U[\psi] \cap i$ is obviously also in $U[\phi] \cap i$, so $j = U[\phi] \cap U[\psi] \cap i$. Similarly, $k = U[\phi] \cap U[\psi] \cap i$ so $j = k$.

The semantics of the corner conditional stated in (S6) deserve special attention. There are both clear similarities and clear differences to the standard first-order semantics of the quantified conditional. The most important difference is the relation between the base world and the selected world. Stalnaker's conditional is given a truth-value at the base world, by means of the selection function, which identifies the one nearest world among all the accessible worlds. From the point of view of the selected world, the truth-value of the corner conditional is given directly by the semantics of the first-order quantified conditional. From the point of view of the base world, the selected world is perhaps the most relevant, or even the uniquely relevant vantage point for such an evaluation. Still, the selected world is only one among a potentially large collection of possible worlds.

With this perspective on Stalnaker's system, it is perhaps easier to see the relation between ordinary logic which assigns a unique truth value to each sentence, modal systems in which truth at one world is defined in relation to other worlds, and a probability space in which not only are truth values selected from {0,1} but also fractional values from the continuous unit interval. In Stalnaker's system, the Lycan truth condition is employed but "at arm's length" – in the selected world, instead of the base world. When evaluating the truth of a conditional, Stalnaker expects the same absolute guarantee of consequence-satisfaction following upon hypothesis-satisfaction as Lycan requires, but only at the selected world. Stalnaker's use of a modal syntax and semantics appears (and in some respects, is) fundamentally different from Lycan's first-order quantificational framework. The formal construction in this section shows that this appearance is rather superficial – a first-order model can provide all the structure needed to define both Stalnaker's accessibility relation and his selection function. Interesting differences remain among Lycan's, Stalnaker's, and the probability spaces I have defined, that much is clear. It should also be clear that some of the strengths of the three proposals can be seen as shared strengths, rather than as countervailing advantages.

Before leaving the formal discussion of C2 I want to emphasize the role of the individual possible worlds, as the loci of truth-evaluation for propositions. It is characteristic of modal systems (especially when interpreted in the style of Kripke) that every proposition has not one unique truth value, but instead a pattern of truth values distributed across all possible worlds. Each proposition is then a function, if we consider the set of all the possible worlds to be the domain and the set of truth values to be the range. The function defined in this manner is the indicator function of a set of possible worlds, and the set of possible worlds is more commonly identified with a proposition. With Stalnaker's C2, we can see an important analogy between Kolmogorov's theory of conditional expectation and the logic of conditionals. In Stalnaker's system the conditional is a proposition with a truth value at each possible world. The conditionals are in no way distinguished from the unconditioned propositions – we may identify the conditional propositions either with a pattern of truth values over the possible worlds, or with a set of possible worlds. The central, critical difference between the Kolmogorov theory of conditional expectation and the classical heory of conditional probability is that a conditional expectation is a function, while a classical conditional probability is simply a number. That is to say, in the Kolmogorov theory the unconditioned propositions have the same characteristics as the conditionals, in that both are representable by functions.

One of the lessons to be learned from Lewis' triviality result is that classical probability theory has limitations which are surpassed in the modern theory. When the bearers of probability may only be sets (or events, propositions, etc) in a field of Boolean operations, then only in the case of universal independence can there be events which have conditional probability. In the Kolmogorov theory, the bearers of probability are not only sets or events, but in addition there is a concept of probability (more properly called "expectation") for functions defined on the members of the sets. The classical theory is mirrored in the new theory by the one-one relationship between sets and indicator functions. The modern theory incorporates the old by identifying the probability of an event with the expectation of its indicator function. The modern theory extends the old by defining conditional expectation as, again, a function defined on the members of the sets.

To complete the rehabilitation of C2, we return to Stalnaker's discussion of the relation between conditionals and conditional probability in the closing pages of "Probability and Conditionals." Stalnaker considers the importation principle:

(17) $\Pr(A > B \mid C) = \Pr(B \mid A \,\&\, C)$

which he rejects for the inappropriate reason that "[t]he antecedent A may be a *counterfactual* assumption with respect to the condition C. That is, the antecedent A may be incompatible with the state of knowledge selected by the condition C." It is true that A and C might be incompatible, but the counterfactual status of a conditional has already been handled by Stalnaker and Thomasson in the truth-conditions of the corner conditional. . What makes a conditionalcounter factual is an inconsistency between its antecedent and the base world i where the conditional is to be evaluated. A means of reconciling these inconsistencies is already built into Stalnaker's logic, before any considerations of probability are raised. That is the selection function $s(\phi,i)$ which sometimes gives us a new possible world, distinct from i, in which the hypothesis ϕ is realized. Sometimes the selection function is undefined, which is what happens (by design) when a hypothesis is unsatisfiable in any accessible world. I would say, Stalnaker has already bitten the bullet which gives him pause at (17). Conditionals whose antecedents are unsatisfiable, or even absurd, already have their appropriate place in C2. On the other hand, (17) would generate the problem of triviality which Stalnaker did not foresee. I believe that the complexity which Stalnaker introduces after rejecting his (17) is what led Lewis to insert his "parenthetical qualification" of his (7), creating an ambiguity in his first result which is not resolved until the work of later authors. Lewis' second result seems to pay more attention to Stalnaker's complex system of subfunctions and levels of conditionalization, but in my view both Lewis' attack and Stalnaker's capitulation were misplaced.

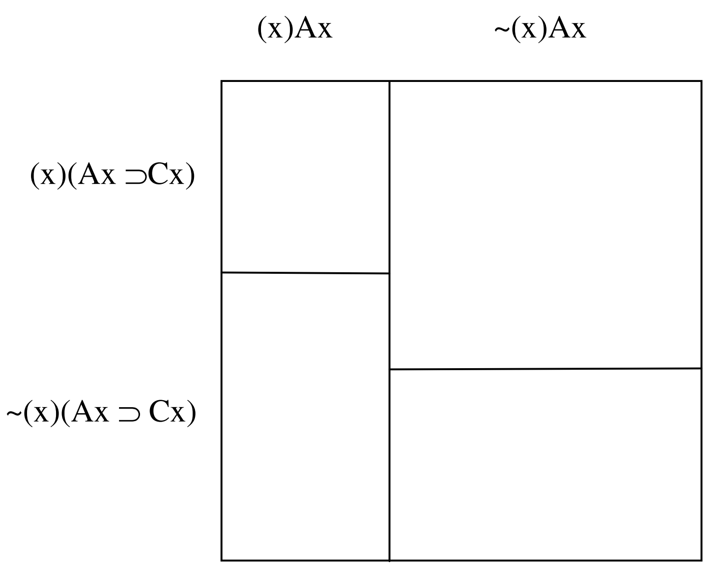
Now, with the Allied forces operating at full strength, we may return to the counter-example offered by Edgington against Lycan's analysis. The first observation to make *contra* Edgington is from the ground-level logical point of view which Lycan's system occupies. When the patient asks the doctor, "Will I be cured, if I have the operation?" the doctor is perfectly correct to say simply, "No. I can't say flatly that if you have the operation, you will be cured. That would not be true." This manner of speaking is perhaps unsophisticated, and perhaps even misleading, if not amplified with probabilistic information. But the patient asked a flat, factual question – not a probabilistic question. A flat factual answer is within the logical proprieties of the conversation. Lycan's system insists that the flat factual interpretation of the conditional is primary. In the muddy trenches of concrete communication, where literal meaning wrestles flat-footed with suppositions and implicatures, the outcome may not be nice or pretty, but it must be *compositional* if it is to enable further advances. Lycan's reading of the conditional serves uniquely well as the foundation of other approaches which only appear more sophisticated. Only the extensional reply specified by Lycan's semantics is literally on the topic raised by the patient's query. Evaluated in the actual world[17], the truth-value of the singular conditional (if Lycan is right) is determined by the general conditional, which all agree false.

The argument is not yet won; a flanking maneuver is next. Any flat factual answer is not only conversationally compatible, but also logically compatible, with a modal answer to the same question. At the same time, in the same breath, as he tells the patient "No," the doctor would again be correct to continue, "It is possible that if you have the operation, you will be cured. Certain cases respond well to this operation. You might be one of them." It is not just common; it is practically universal among modal systems to distinguish between the truth value of a proposition and its modally quantified version. There is no contradiction in asserting both "$\sim A$" and "$\Diamond A$" (or in Stalnaker's symbolism, "$\sim(O > C)$" and "$\Diamond(O > C)$"). The same point can be madein quantified logic: "$\sim(\forall x)(Ox \supset Cx)$" is consistent with both "$(\exists x)(Ox \supset Cx)$" and $Op \supset Cp$. Neither party in this conversation has raised any issue of mere possibilities. But that issue is clearly present: the patient herself will choose whether to have the operation. We must ask, which possibilities have which outcomes, and in what proportion? Edgington's example correctly situates the evaluation of a conditional in the human context of evaluating possible actions – propositions which the doctor and patient can choose to make true. We know that modal possibilities map onto quantifiers and predicates, and we know how to calculate those proportions.

The heavy artillery behind Edgington's position is the doctor's knowledge of the true rate of cures after the operation. We must explain how the doctor can, with logical consistency, communicate the probabilistic information which he does possess, by means of an expression he knows to be false. We know this answer – a propositional function may take the value "false" in any particular world, yet have an expectation greater than zero, or even close to one. Equally, the propositional function can take the value "true" in (say) the actual world, while having a low expectation across the worlds accessible therefrom. There is much more to probability than to attach a single "credence" number to each proposition, as a substitute for a truth value. Truth and probability coexist naturally, when language is employed for the purpose of anticipating future events, and planning how to meet them.

IX. Epilogue – Diagrams for the Space of Samples

The square probability diagrams used illustrate the triviality theorem can also help resolve some questions which may linger around the non-triviality of the space of samples. Following the basic principle of every square diagram, we may write an event, say (x)Ax and its complement ~(x)Ax along the top segment. Down the side segment we can draw any other event and its complement. We will diagram the conditional event (x)(Ax ⊃ Cx). Using the general drawing procedure which is non-committal as to the relative positions of the two horizontal divisions – the conditional probabilities – we obtain the following diagram:



Of course, we do have specific information about the relative probability of the events in this picture. First, for any two quantified expressions (x)Ax and (x)Cx, we know that the quantified conditional $(x)(Ax \supset Cx)$ has the conditional probability:

(1) $P[(x)(Ax \supset Cx)] = P[(x)Cx \mid (x)Ax] = P[(x)(Ax \,\&\, Cx)]/P[(x)Ax]$

And

(2) $P[(x)(Ax \supset Cx)] \cdot P[(x)Ax] = P[(x)(Ax \,\&\, Cx)]$

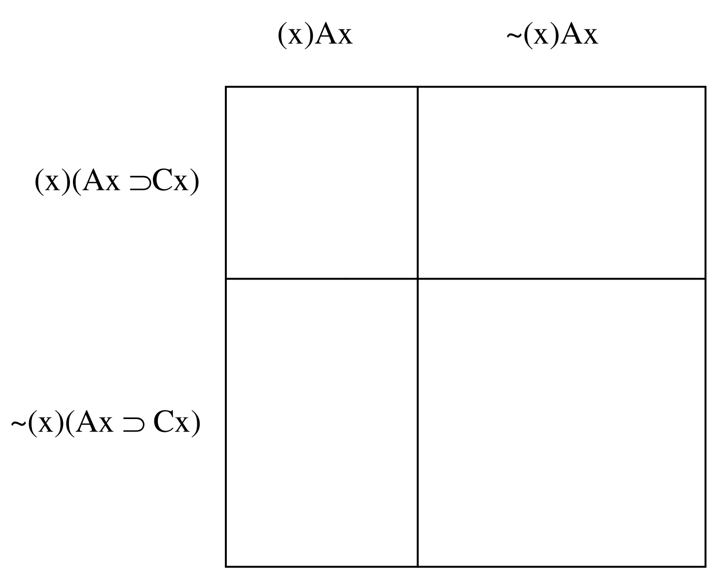
We also know from logic that:

(3) $(x)(Ax \supset Cx) \,\&\, (x)(Ax) \Leftrightarrow (x)(Ax \,\&\, Cx)$

so

(4) $P[(x)(Ax \supset Cx)] \cdot P[(x)Ax] = P[(x)(Ax \supset Cx) \,\&\, (x)(Ax)]$

Which says the conditional is probabilistically independent of is antecedent. Therefore we know that the two horizontal divisions in the diagram above must be collinear, and we can redraw the diagram as:



At this point we may ask a question along the lines of, "What happened to the triviality theorem?" The triviality theorem has (or has been claimed to have) been proved within probability theory itself, and to have no dependence on any logical assumption. If that is so, then we must observe that here in the space of samples, we have a new and different probability space, but it is no less a probability space than any other, and probability theorems must be true here as elsewhere. Whatever logical tactic has been used to construct a conditional event – or at least, an event which has a conditional probability – we have been led to believe that certain consequences will follow, merely from the existence of a conditional event. We have the conditional event – where is the triviality?

First, we must look at the hypotheses of the triviality theorem. In the probability space of samples, we have constructed an event with a conditional probability, and we can in some sense justify calling it a conditional event just because it is constructed with a conditional connective. But is this so-called conditional event – the quantified conditional event – also a conditional event in the sense of Lewis' argument, the many other arguments in the literature, and in the proof offered above?

What makes an event B the conditional event of C given A, in the Lewis sense are the two properties:

(CE1) $P(B) = P(C|A)$ and

(CE2) $P(B|E) = P(C|AE)$ for every event E where $P(AE) > 0$

the second property being the "importation" principle. It is interesting to note that a quantified analog of the importation principle holds in predicate calculus just as it holds in propositional logic, i.e., the follow is a theorem:

(5) $\vdash (x)(Ex \supset (Ax \supset Cx)) \supset (x)((Ex \,\&\, Ax) \supset Cx)$

which is provable by the same DeMorgan operations as are used to demonstrate the importation principle in propositional logic. But the context of probability theory is yet again different from either propositional logic or predicate calculus. In the triviality proof, we assumed as a hypothesis that an event is not conditional unless it satisfies the probabilistic importation principle. Now we are asking whether the quantified conditional satisfies the probabilistic importation principle, assuming only the probability structure which has already been defined for the probability of a quantified conditional. We are asking whether the following equation is an identity:

(6) $P[(x)(Ax \supset Cx) \mid E] = P[(x)Cx \mid E \cap [(x)Ax)]]$ for all E, $P(E)>0$

If it were an identity, then it would hold for $E = {\sim}(x)Cx$, if that is non-null, which would give us:

(7) $P[(x)(Ax \supset Cx) \mid {\sim}(x)Cx\ ] = P[(x)Cx \mid [{\sim}(x)Cx] \cap [(x)Ax)]]$

Which reduces to zero on the right, giving:

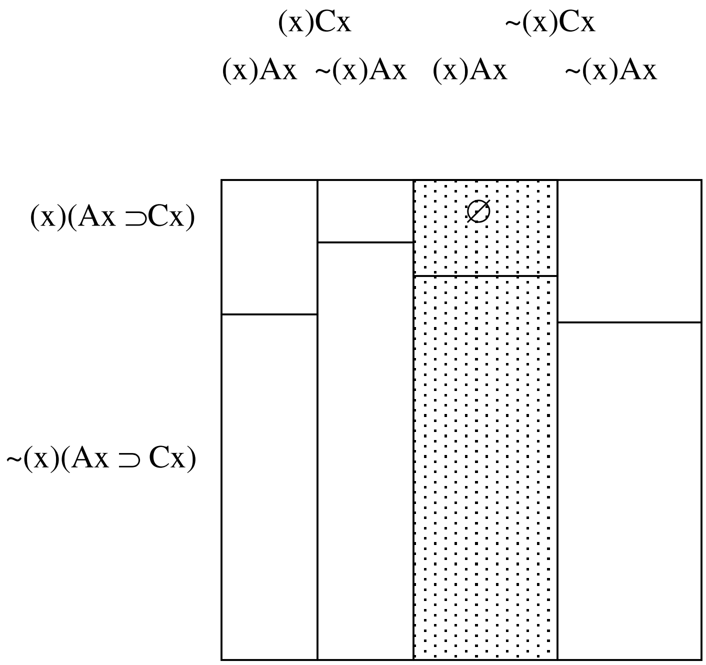
(8) $P[(x)(Ax \supset Cx) \mid {\sim}(x)Cx\ ] = 0$

which is one of the two facts required to establish the triviality result. This conditional probability can be zero only if

(9) $P[(x)(Ax \,\&\, {\sim}Cx)] = 0.$

We will analyze this possibility with a diagram. If we can show this equation is not an identity, then we will have shown the triviality result does not apply.

We draw the same diagram as was used for the triviality proof, with new labels for the quantified propositions. As before, because we are not asserting independence between events depicted down the side with respect to events depicted across the top, we draw the horizontal divisions as disconnected:



We draw a ∅ in the region for the intersection of the three expressions, (x)(Ax ⊃ Cx), (x)Ax, and ~(x)Cx, not because the equation shows a zero, but because that conjunction is logically impossible. This region must be empty because in any sample where "all the A's are C's" and in addition, "everything is A" then it must also obtain that "everything is C". The conditional is incompatible with the joint event formed by the conjunction between its antecedent and negated consequent. This relation was sufficient to establish the triviality result in the space generated by propositional calculus. But it is not determinative now, because the region for equation (9) is the entire shaded stripe, which is not necessarily empty. Intuitively, there is no contradiction between asserting that (e.g.) all the spheres in a sample are yellow, while not all the blocks in the sample are yellow. We have a concrete example in the case of the urn with nine blocks. Therefore the quantified material conditional does not satisfy the probabilistic importation property, and the triviality theorem is not contradicted, but rather, does not apply. Neither Stalnaker's corner conditional nor the quantified material conditional satisfy the importation property. That property is not relevant to the possibility of conditional events in a probability space.

This explains at least that there is no logical contradiction between the triviality and the non-triviality results. But there remains a question about the status of the importation property. It has already been noted that Stalnaker's C2 does not satisfy importation, but one might have expected importation to hold in predicate calculus. Certainly at least one form of importation is true, as has been noted. Resolving this issue will illustrate a deep point about the roles of variables and quantifiers in logical expressions.

Consider the event whose probability we were evaluating just above, viz:

(10) $[(x)(Ax \supset Cx)] \cap [\sim(x)Cx]$

This joint event is represented by a conjunction of quantifier expressions:

(11) $(x)(Ax \supset Cx)\ \&\ \sim(x)Cx$

But the second conjunct is not a universally quantified expression. Rather, it is the negation of a generalization, and therefore an existential. We can now re-write the above as:

(12) $(x)(Ax \supset Cx)\ \&\ (\exists x)\sim Cx$

What might be done to effect a collection of quantifiers to the left – that is, to express the formula in prenex normal form? Given that all the atomic subformulas are monadic, we need not concern ourselves about the order of the quantifiers. Once the expression is in prenex normal form, the quantifiers may be permuted at will. But we know from the rules of predicate calculus that, because the two conjuncts have quantifiers with distinct scope, any equivalent formula which gives the quantifiers overlapping scope within the sentence, must employ distinct variables to respect the original scope distinctions. We rewrite a final time as:

(13) $(x)(\exists y)((Ax \supset Cx) \& {\sim}Cy)$

Propositional compounds of quantified expressions lead to expressions which are still in the broad category of generalizations, but we have left behind the simple cases of universally quantified formulas of a single variable. The failure of the importation principle in the context of quantified expressions is not due to any multiple valued, modal, or other non-classical feature of the logic, but instead is a consequence of the interaction between quantifiers, variables, and scope. Even in the simplified context of monadic predicates, variables and their scope introduce significant complexity which cannot be found in propositional calculus. The extensions of monadic predicates are only simple sets, but the extensions of molecular monadic expressions are relations. We are not forced by the desire to find a logical expression with conditional probability to look beyond classical logic, as long as we recognize classical logic to include predicate calculus. This should not be surprising, in view of Church's Thesis.

BIBLIOGRAPHY

Adams, E. 1975. *The Logic of Conditionals*. Dordrecht: Reidel.

Billingsley, P. 1995. *Probability and Measure, Third Edition.* New York: J. Wiley and Sons.

Eells, E. and Skyrms, B. eds. 1994. *Probability and Conditionals*. Cambridge: Cambridge University Press.

Gaifman, H. 1964. Concerning Measures on First-Order Calculi. *Israel Journal of Mathematics* 2:1-17.

Halmos, P. R. 1950. *Measure Theory.* New York: Springer-Verlag,

Lewis, D. 1976. Probabilities of conditionals and conditional probabilities. *Philosophical Review* 85: 297-315.

Milne, P. 2003. The simplest Lewis-style triviality proof yet? *Analysis* 63: 300-304.

Monk, J. D. 1989. Handbook of Boolean algebras. Amsterdam ; New York : North-Holland ; New York, N.Y., U.S.A

Ramsey, F. P. 1990. General Propositions and Causality reprinted in *F. P. Ramsey's Philosophical Papers*, 145-63. Cambridge: Cambridge University Press.

NOTES

[1] The standard texts in this field are by Hopcroft & Ullman, which are available at different levels of mathematical expertise.

[2] Personal communication, March 2006

[3] (I will rigorously keep my promise not to solve the problem of "What the Tortoise said to Achilles," but I would be derelict not to mention it, keeping a discreet distance.)

[4] Mathematically, the segment diagrams can represent any probability space defined on the unit interval of real numbers, [0,1]

[5] The proportional divisions in the square diagram can be constructed geometrically from he divisions in the segment diagram, using similar triangles. The construction is sometimes called the "fourth proportional."

[6] For example, we might number our blocks '1' through '9', print out the nine-place binary numerals from '000000000' to '111111111' on 512 slips of paper, draw a

slip at random, and use the pattern of digits to specify, for each block, whether it has been selected.

[7] These definitions entail the peculiar (but not unheard of) consequence that an empty collection is taken to satisfy any universal generalization. This feature of our definitions is essential to the calculations and proof we present below, so we accept it without further discussion. We also omit discussion of predicates with multiple variables, although an extension to such cases is not difficult.

[8] More generally, the second space will be the algebra of principal ideals generated from the extensions of open formulas with one free variable. See (Monk 1989)

[9] For concreteness, suppose that if the samples are drawn repeatedly, all objects are replaced before the next sample, and that each new trial is independent of all the previous trials.

[10] More general forms of sampling could be considered, as long as independence among domain elements is preserved. Nevertheless, for this elementary result the counting measure reveals the role of exponential function as parallel to the power set operation, which an interesting feature of our construction.

[11] Here a rich and important issue is reserved for future work.

[12] It is not necessary to suppose that the extension of F is not empty, because we are not conditioning on [Fx] itself. Even when [Fx] is empty, the extension of [$\forall$x(Fx)] is not empty – it will contain one element, the empty set.

[13] The importation formula for conditional expectation of random variables is:

$$E [E [X \mid G_2] \mid G_1] = E [X \mid G_1]$$

but this formula is defined not for events, but for G_2, G_1 as Boolean algebras of events. Moreover, it holds only under the condition that $G_1 \subseteq G_2$. I will discuss the advantages of conditional expectation versus conditional probability at various points through the essay, but most extensively in the section on backward subjunctives.

[14] Any account of backward subjunctives which hopes to account for both the linguistic data and the practical inference behavior would have to compare relations among events which correspond to statistical correlation, not must the maximum of conditional probability. A deterministic version of the same project might look at the gradients established by vector fields (such as force) and partition the event space into integral curves.

[15] Lycan, *Real Conditionals*

[16] From an "Author Meets Critics" session at a meeting of the APA

[17] An interesting question may be raised here, whether it is oversimplifying the question to suppose that the actual world is the only relevant variable. This issue will be touched upon in the next section.