Similarity-Based Likelihood Judgment

by

Joshua J. Stern

B.A. Epistemic & Formal Systems, and B.S.E. Systems Science & Engineering,
University of Pennsylvania (1987)

Submitted to the Department of Brain and Cognitive Sciences in
partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

Massachusetts Institute of Technology
September, 1991

Signature of Author _____
Department of Brain and Cognitive Sciences
September 5, 1991

Certified by _____
Daniel Osherson
Professor, Brain and Cognitive Sciences
Thesis Supervisor

Accepted by _____
Emilio Bizzi
Chair, Department of Brain and Cognitive Sciences

Similarity-Based Likelihood Judgment

by
Joshua J. Stern

ABSTRACT

People readily quantify a wide variety of opinions about uncertain contingencies by making likelihood judgments. In this work the psychological nature of likelihood judgments based on information contained in similarity is examined. Specifically, similarity based likelihood judgment is interpreted as a judgment strategy that a person may use to produce an attribution of likelihood to the proposition that an individual, i, has some property, P, in situations in which 1) P is known to be a property of a certain type, 2) some other related individuals are known to have the property, 3) some other related individuals are known not to have the property, and 4) the person has beliefs about the similarity of i to the other related individuals. If no other information relevant to the likelihood that individual i has property P is available, then similarity-based strategies attribute likelihood in such a way that the likelihood that i has P is an increasing function of the similarity (in a P related way) of i to the related individuals that are known to have P and a decreasing function of the similarity of i to the related individuals that are known not to have P.

Similarity-based likelihood judgment is described here as a particular example of a default likelihood judgment strategy. The abstract question of what constitutes an adequate description for a default likelihood judgment strategy is considered at length in this work. It is proposed that, in general, successful default likelihood judgment strategies represent evaluations of particular kinds of conditional probabilities. These evaluations are typically based on some limited quantity of evidence, the form and type of which may vary from one strategy to another. It is argued that a complete description of similarity-based likelihood judgment as a specific strategy must have at least two parts. The first of these parts is a description of which conditional probability is evaluated by similarity-based likelihood judgment; the second part is a precise description of how this conditional probability is quantified.

The principal empirical finding reported in this work relates to the role of probability measures in the second part of the description of similarity based likelihood judgments.

In general, Psychological research supports the claim that the collection of all likelihood judgments produced by most individuals are not well described by a single coherent subjective probability measure. It is proposed however, that, under certain conditions, the similarity based likelihood judgments that a person makes are well described as judgments of conditional probability that are coherent with respect to a certain local probability distribution, and that the information content of this distribution is a function of the person's beliefs about the similarities of the related individuals involved.

The proposed theory of similarity based likelihood judgment is empirically evaluated according to its ability to predict quantitative attributions of probability produced by human subjects in an appropriate judgment context.

Table of Contents

I Introduction

I.1 Intro to the Psychology of Likelihood

Human beings are capable of producing opinions concerning the likelihood of an enormous variety and kind of possible events, and they do so within a wide range of informational contexts. Examples of some possible judgments - that the probability of a global oil shortage before 2010 is 0.1, that the odds of getting 13 or more heads in 20 tosses of a fair coin is near 1 in 3, and that the chances that it will rain on any given day in in Boston during April are about 50/50 - readily bear out this claim. Modern research on the psychology of likelihood judgment supports the view that the thought processes responsible for producing this range and variety of opinion are heterogenous in character - see Kahneman et al. '82 for an overview of modern psychological research on likelihood judgment. A partial list of reasoning patterns that are often thought to describe semi-distinct varieties of human likelihood judgment includes the following items.

1a) Relative frequency- if n individuals of type C are known and m of them are determined to have property P then it is common to judge the likelihood that some new individual of type C has the property P to be m/n. People's use of this reasoning strategy does not, in general, depend on whether the sample of C's and the new individual were "randomly" chosen. See Estes'76 for a review of related psychological findings and Pollock'90 for a normative theory and references.

1b) Availability - if m out of the first n individuals of type C that one can bring to mind have property P then it is common to judge the likelihood that a random new individual of type C has the property P to be m/n. See Tversky et al. '73 for a more extensive description and review of relevant psychological findings.

1c) Statistical adjustment - one may have learned as declarative knowledge that, for example, that the statistical incidence of heart attack among pre-menopausal women in the U.S. population is 1 in 17. Such knowledge may be used as a starting point for other judgments, such as the likelihood of

heart disease for a pre-menopausal woman in the U.S. who is 20-30 lbs. overweight.

2) Insufficient Reason - it is psychologically natural to attribute a probability of 1/n to each of n possible outcomes of an uncertain event in situations where the n outcomes are disjoint, exhaustive, and stand in identical evidential relationships to the unknown event. For instance, the probability that the next major earthquake (over 7 on the Richter scale) will occur on a Monday (E.S.T.) may be judged as 1/7, reasoning that the earthquake must occur on one of the seven days of the week and there is no available reason for thinking that its occurrence on any particular day of the week is more or less likely than any other day. This is an example of the application of insufficient reason, originally formulated by Laplace. For justification and extension of this doctrine see Jaynes '79.

3a) Causal Schemas - a person's judgments about the probability of an event of type A occurring given that an event of type B occurred can unquestionably be influenced by that individuals personal record of experience with regard to the statistical association of these types of events. There is also evidence that such judgments are influenced by a person's ability to perceive/intuit a mechanistic or other type of causal chain leading from events of type A to events of type B, although these chains apparently do not need to be very well elaborated. For example, one might believe that taking large dosages of Vitamin C somehow causes a decrease in the likelihood of contracting a viral infection, or that drinking a particular brand of beer causes an increase in the likelihood that a randomly chosen person will desire to engage one in conversation. It has been experimentally demonstrated that the presence of perceived causal connections between two events, A and B can have a distinct, extra-statistical effect on judgments of the likelihood of A given B. See Tversky et al. '82 and the volume edited by Hilton for experimental results. See also Pearl '88, the volume edited by Hilton and Salmon '84 for discussion of the role played by intuitions of causality in reasoning and Pearl '88 for a treatise on the computational efficiency of representing probabilistic information in terms of causal relationships.

3b) Explanation/attribution based schemas - in some situations it may not be intuitively acceptable to attribute a likelihood to a proposition without also providing an "explanation" of that attribution. Such an explanation might be a predictive rule that could be used to assign likelihood to other unconsidered cases as well. This type of reasoning seems to commonly occur in classification sorting tasks (see Medin et al. '87) and tasks related to the attribution of social properties (see Hilton & Slugoski). The volume edited by Hilton contains a group of papers focusing on different aspects of this type of strategy and its domain.

4a) Representativeness - the likelihood that an individual, c, is a member of a class, D, may often be judged according to the perceived similarity between c and a representative description of D. See Kahneman et al. '72 and the group of papers edited by Kahneman et al. '82.

4b) Similarity - the likelihood that an individual c has some property P may be judged according to the known presence or absence of property P among other individuals that are considered to be similar to c. The precise form of such judgment, at least for certain special cases, is the central focus of this paper - references in the next paragraph.

It would be misleading to speak as if there is a precise catalog of likelihood judgment algorithms that human beings, in general, possess. Many likelihood judgments are undoubtably the result of thought processes which are highly malleable and can be consciously and voluntarily manipulated - formally learned probability calculations are one clear-cut example (see e.g. Nisbett et al. '83 for details). Nevertheless, numerous experimental evidence suggests that similarity-like factors play an important role in a wide range of human inductive reasoning including stimulus generalization (Shephard, '57 Nofosky '84) and categorization (Brooks '78, Medin & Schaffer '78) in addition to likelihood judgment (Rips '75, Osherson et. al '91). Experiments reported in the latter two sources specifically establish that the strength of relevant similarity relations can be a statistically dominant factor influencing human probability attribution in certain informational contexts. This evidence gives reason to believe that an appropriate theory of the contribution of similarity to likelihood judgment may describe an important

fragment of human reasoning, albeit a fragment of somewhat unknown size and character.

Similarity based likelihood judgment is a variety of provisional or default reasoning - as indeed, are all of the other judgment schemas listed above. The present work will focus in detail on the form and nature of similarity based likelihood judgment as a default reasoning strategy. In the next section I describe what is meant by "default reasoning", summarize the recent intellectual history of the topic, and describe some important research issues related to the study and description of default reasoning strategies for likelihood judgment.

I.2 Theories of default reasoning and judgment - general remarks

The reasoning patterns 1a) - 4b) may be referred to as "default strategies" because the role that they play in reasoning is to assert that some kind of evidence, E, provides a default or provisional reason for attributing a likelihood value, V, to some event or proposition, A. This attribution is provisional because it can be, and often is, entirely overridden by the introduction of additional evidence. A classic example: the knowledge that the individual "Tweety" is a bird is evidence for attributing a high likelihood to the proposition 'Tweety flies', but the knowledge that Tweety is a bird taken together with the further knowledge that Tweety is a penguin is evidence for assigning a very low likelihood to 'Tweety flies'. The provisional nature of default reasoning stands in contrast to the truth-preserving property (validity) of deductive reasoning. The rules of deductive reasoning are such that a conclusion which is deductively inferred from true premises must be true. If all men are mortal and Socrates is a man then the truth of the proposition 'Socrates is mortal' is assured. If it were somehow ascertained that Socrates is immortal, then it could be confidently assumed that either not all men are mortal, or Socrates is not a man. In the Tweety example, the thesis that Tweety probably flies is (inductively) inferred from the premise that Tweety is a bird. The contrast with deductive inference is highlighted by noting that additional information casting doubt on the likelihood of Tweety flying does not in any way cast doubt on Tweety's status

as a bird. The activity of inductively leaping to default conclusions is also commonly referred to as non-monotonic reasoning because of the way in which default conclusions can later be altered or retracted, in contrast to the "monotonic" way in which deductively inferred conclusions are cumulative.

I.2.1 The nature and status of recent normative views

In recent history, the terms default reasoning and non-monotonic logic/reasoning have been used most commonly in the field of Artficial Intelligence. Until fairly recently however, researchers in AI have mostly been adamant in claiming that default reasoning has little to do with probabilities. In an influential paper, John McCarthy and Patrick Hayes '69 rejected the use of probabilities for non-monotonic inference in favor of some form of (as yet to be discovered) non-monotonic logic. Two of their complaints with probabilities were that "It is not clear how to attach probabilities to statements containing quantifiers in a way that corresponds to the amount of conviction people have," and that "The information necessary to assign numerical probabilities is not ordinarily available," and therefore probabilities are "epistemologically inadequate," as a representation of provisional belief. In the wake of this paper, researchers in AI have often gone out of their way to banish probability from their land. It has been traditionally argued that, for example, the conclusion that 'Tweety flys' is non-monotonically inferred from 'Tweety is a bird' on the basis of what is "typical" as opposed to what is "probable" and that the former proposition is provisionally "accepted" rather than assigned high probability - see Reiter '80 and Pearl '88 for conflicting viewpoints related to these distinctions. This reluctance to identify provisional conclusions with conditionalized probabilities is somewhat suprising given the neat fit between what I have just described as the defining characteristics of default reasoning and the semantics of conditional probability statements. The beliefs that the probability that Tweety flies given that tweety is a bird is high and that the probability that tweety flies given that tweety is a penguin is low can easily be accommodated within a single probability measure. In fact, if P stands for a proposition to be assigned a probability based on evidence E, then the

statement that 'the conditional probability of P given E is v,' in general, says almost nothing about the conditional probability of P given E and E'.

Recently, a number of researchers in AI have argued that probability theory is an essential tool in the study of reasoning - See Cheeseman 88', Pearl 88', Bacchus 90', Halpern 87', etc. Perhaps not coincidentally, this new acceptance of probability into AI has been accompanied by three kinds of theoretical innovations which address many of the traditional criticisms surrounding the application of probability to default reasoning. The first of these innovations has been the introduction of effective systems for combining deductive reasoning involving quantified statements in a first order language with probabilistic/statistical evidence (Halpern '87, Bacchus '90, Geffner & Pearl '90), the second, a recognition that the essential evidential structure of probabilistic relationships is not necessarily dependent on the use of a representation for gradations of likelihood that has the cardinality of the real line (Aleliunas 90')., and the third, has been the introduction of techniques for manipulating and drawing conclusions from incompletely specified probability measures (Jaynes '79, Cheeseman '83, Levi '80, Dempster '67, Shafer '76, Pearl '88). Each of these innovations, which generalize classical theories and applications of probability, have lent support to the idea that traditional probability theory and its descendants may play an important role in the description of a wide range of reasoning patterns. To this date however, it is an understatement to remark that there is still no consensus on this issue within AI as a whole.

Views on the nature of default judgment and reasoning in the recent history of Cognitive Psychology (since the late 1960s) have been radically different from those in AI. Psychologists have generally assumed, sometimes implicitly and sometimes explicitly, that any theoretical object which fit the description "a normatively justifiable strategy for default belief assignment" must be some straightforward translation of the basic principles of mathematical probability theory into a reasoning algorithm. For example, Daniel Kahneman and Amos Tversky write, "Although no systematic theory about the psychology of uncertainty has emerged...Perhaps the most general conclusion, obtained from numerous investigations, is that people do not follow *the principles of probability theory* in judging the likelihood of

uncertain events...Apparently, people replace *the laws of chance* by heuristics, which sometimes yield reasonable estimates and quite often do not," [p.32 Kahneman and Tversky. '82, my italics added]. This statement seems to indicate that these authors are thinking of "the laws of chance" as something which could be used to "yield reasonable estimates," - i.e. a procedural rule or reasoning algorithm, and not merely probability theory per se. It's not at all clear though, as the controversey surrounding the use of probabilities in AI indicates, what this estimating procedure or algorithm which is a straightforward application of the laws of chance to reasoning looks like, or if it exists.

The algorithm which comes closest to being a straightforward translation of mathematical probability theory into a reasoning strategy is the following well known one:

> A reasoner begins "life" with a prior probability . This is a representation pairing conjunctions of every hypothesis the reasoner might ever desire to evaluate and every experience that might evidentially bear on any hypothesis with a representation of quantities in such a manner that the set of quantities is isomorphic to a probability measure. Upon the arrival of new evidence, which by assumption the reasoner has anticipated the possibility of, the current probability measure of the reasoner, old(), is updated by conditionalization and replaced with the new probability measure, new(), according to the following procedure.
>
> For every event e_i such that there are no events which are proper subsets of e_i (or which imply e_i)
>
> $new(e_i) = old(e_i \ \& \ evidence) \ / old(evidence)$.
>
> When the reasoner wants to evaluate the likelihood of some event (which by assumption the reasoner has anticipated the possibility of) this evaluation proceeds by computing the sum of the probabilities assigned to any disjoint set of events $\{e_i\}$ whose union is equal to the event to be evaluated.

This algorithm, which I will refer to as the Orthodox Bayesian algorithm, does not really seem to be a serious possibility for employment by any human or mechanical reasoner that deals with a moderate number of

different hypothesis and evidence statements. One reason for this is the practical impossibility of anticipating all possible forms of evidence that could be received, all hypothesis that could be entertained, and what the precise effect of each combination of evidence on the various hypothesis should be. Another reason for is that for every set of n hypothesis and evidence statements which are statistically relevant to one another, the reasoner must represent on the order of 2^n numbers in his/her/its implementation of such a scheme. For even moderate n this number is impossibly large (and doubles in size for n+1).

I don't think that anyone studying the Psychology of Reasoning really wants to claim that the Orthodox Bayesian algorithm is the only rational way to produce likelihood judgments. It is important though for the Psychology of Reasoning to fully examine the consequences of the intractability of this algorithm. The reason for this is not primarily because it will change our perceptions that people commonly commit errors in their attributions of probability, but because there is still a lot at stake in where we place the blame for the errors that they do commit - because the nature of this blame has an important effect on how we study the algorithms which people use. Gilbert Harman [Harman '86, p.7] offers a convenient catalog of distinctions between different kinds of errors which could be committed while reasoning:

> i. One might start with false beliefs and by reasoning be led into further errors.
>
> ii. One might reach a conclusion that is perfectly "reasonable," even though it happens to be mistaken.
>
> iii. One can be careless or inattentive or make mistakes in calculation; one can forget about a relevant consideration or fail to give it sufficient weight; one can fail to remember some fact, etc.
>
> iv. One can arrive at one's view in accordance with an incorrect rule of reasoning, thereby violating the correct rules.

Several of the default judgment strategies I have listed as 1a) - 4b) seem to be thought of in the literature on the Psychology of Reasoning as involved in the production of errors of type iv. To the extent that Psychologists believe

that type iv. errors are intrinsic to the nature of the algorithms that human beings primarily use to produce likelihood judgments, one of the consequences alluded to above is that we will forgo attempts to understand these algorithms as rational strategies to be analyzed according to computational principles. As the above catalog of errors should make clear, knowing that an error in judgment occurred and that a certain reasoning procedure was employed in producing that judgment is not, by itself, sufficient information to view the error as one of type iv. If a reasoning strategy can "sometimes yield reasonable estimates," as has been claimed for most all of the strategies 1a) - 4b), it would seem necessary, in order to generally classify applications of such a strategy as type iv. errors, for one to presuppose the theoretical existence of some other strategy which produces reasonable estimates with far greater regularity. If these strategies (1a - 4b) are also thought of as sub-optimal in some meaningful sense, then this preferred strategy should be tractably realizable. It's not at all clear that the Orthodox Bayesian algorithm would necessarily be the preferred strategy if it were tractable [see e.g. Bacchus et al. '90 and Kyburg '83], but it is clear that it is not tractable for human beings or current artificial technologies that seek to move beyond the confines of small specialized domains.

Perhaps, what is, or should be, claimed is not that the "heuristic" strategies are uniformly incorrect as rules of reasoning but that they are incorrectly applied in some circumstances. At this point though, the (I think useful) distinctions between type iv. errors and type ii. and type iii. errors begins to blur. Are we saying that these heuristic rules are incorrectly applied on occaison in virtue of the fact that they happened to yield erroneous conclusions on those occasions? Or are we saying that they were incorrectly applied because important considerations were mistakenly overlooked that should have blocked or modified their application? This latter description seems to apply to many of the cases in the literature. For example, it has been pointed out that people commonly judge the proportion of words in English ending in "ing" to be greater than the proportion of words in English ending in "g", presumably because it is easier to explicitly think of words ending in "g" by concentrating on the retrieval of words ending in "ing" than by another technique that would be adopted if the prevalence of words ending in "ing" was not actively under consideration. As a consequence of this,

judgments of the probability that a word randomly selected from English text ends in "g" are often erroneously low because the fact that words ending in "ing" are common examples of words ending in "g" is frequently and deleteriously not taken into consideration.

It might be argued that such a failure is really a type iv. error after all, because any reasoning algorithm which systematically fails to take important relevant facts into consideration must be a sub-optimal rule of reasoning. And so, to continue this line of thought, it could be argued that the strategy of evaluating frequencies by sampling the cases that one can think of (strategy 1b) must be sub-optimal, because it will chronically suffer from failures to consider some facts of this type - i.e. those that are not readily thought of. Once again though, such a claim seems to presuppose the existence of tractable reasoning algorithms which do not systematically overlook some important and relevant facts.

Since the Orthodox Bayesian algorithm is not tractable, it is natural to inquire concerning what other strategies might count as reasoning according to "the laws of chance". One natural place to look for such strategies is in how people reason when they apply formally learned statistical principles - when they, for example, solve a problem appearing in a textbook on probability or mathematical statistics. Clearly people of moderate intelligence can readily acquire the ability to reliably solve such problems. Such problems are different however, in crucial respects, from the situation that a reasoner confronts when required to produce a judgment of likelihood in the real world. These textbook problems are essentially exercises in computing the immediate consequences of certain probabilistic/statistical assumptions. They take forms like "Given probability measure Pr, how likely is the compound event E?", "Given the compound event E occurred, how much more likely is probability measure Pr1 than probability measure Pr2 to have produced it?", and "Given that Pr1 is true, how likely is an observation of the compound random variable R to taken on a value that a priori was more than 20 times more likely given Pr1 than given Pr2?" Sometimes these type of problems are expressed as "word problems", which generally means that they are described in a dialect of English for which there are established conventions of translation (at least for "the initiated") allowing the unique recovery , from

the English description, of a well defined type of statistical question like the foregoing examples. The ability to solve these type of problems is not directly related to most of the skills that a person would/should employ in evaluating the likelihood of some real world proposition. For example, answering a question like "How likely do you think it is that the train will be less than 5 minutes late this (Friday) morning (given all of the beliefs and experiences that you either innately possessed or have acquired in your lifetime)?" Formally learned statistical skills are directly relevant to the solution of the superficially similar word problem "Given that the distribution of the lateness of this train (in minutes) recorded over a long run of days was well described by a Poisson distribution with parameter 6.72, and that there is nothing special about today, what is the likelihood that the train will be less than 5 minutes late?" Even if a person had direct experience with the statistical information referred to in the latter question, in order to translate the former question into the latter one, a person would need to have anticipated the question in order to form a precise belief about the distribution of the arrival times of that specific train, or have stored all of the data in their head and be able to retrieve it, and they would also need to believe that there was nothing special about today's train - i.e. the person would need to decide that among all of their beliefs, that particular belief about the general distribution of the train's arrival is the relevant one on which to base their judgment, rather than, say, the distribution of that train's arrival on Friday mornings. To summarize, the existence of algorithms for consistently reasoning to correct answers on textbook probability problems does not imply the existence of tractable related algorithms for reasoning "correctly" or straightforwardly applying the "laws of chance" to produce estimates in a complex environment.

Although researchers in neither Artificial Intelligence nor Cognitive Psychology have any concrete general models for describing how default likelihood judgments should be made, many people in each field have been drawn to the idea that probabilistic and statistical principles could play a role in such a theory. Such principles certainly do play an important (though, at present, not precisely describable) role in the work of the Applied Statistician, who, after all, is a notable example of someone who can use their intellect and intuition, extended by various artifactual tools, including statistical

theorems, to produce objectively successful likelihood judgments and predictions about a range of complex empirical phenomena (though not, of course, in "real time"). What can we precisely say about the ways in which the Applied Statistician makes likelihood judgments? For what it's worth, I conjecture that the following stepwise description partially and significantly describes most, if not all cases (although the order of steps is mostly unimportant and may well vary from case to case).

> First, some target proposition, T, such as 'the wheat harvest next year will be in excess of two billion bushels' is recognized as something to be assigned a likelihood. Second, T is represented as the conjunction of a "frame predicate" and a "specialization predicate". In this example, T might be represented as the conjunction of the frame predicate 'x is a (generic) wheat harvest from our country that is in excess of two billion bushels' and the specialization predicate 'x is next year's wheat harvest'. Third, a "reference predicate" is selected, with the property that every example of the frame predicate is also an example of the reference predicate. In this case, the reference predicate might be 'x is a wheat harvest from our country.' Fourth, a body of evidence is selected/obtained which can be used to evaluate the conditional probability of a "random" example satisfying the frame predicate given that it satisfies the reference predicate. In this case, this might be a record of the number of bushels of wheat produced by harvests from the past 15 years. Fifth, an actual number is assigned to this conditional probability, on the basis of the available evidence, using some type of estimation technique that is deemed appropriate. Maybe, in this case, a Gaussian distribution for number of bushels of wheat harvested in a given year is first estimated on the basis of the evidence, and then the probability mass that this distribution assigns to numbers greater than two billion is computed. Sixth, the number obtained in the fifth step is accepted as the probability of T.

Steps one through six are clearly, at best, an incomplete description of what the Applied Statistician does, because even if this description were generally correct, as far as it goes, we still don't have any precise criteria or basis for the different choices that are made for frame predicate, reference predicate, evidence, and estimation technique. Clearly intuition, past experience, theory, learning, pragmatic convenience, and a host of other factors play a role in these choices. Nevertheless the description above, vague as it is, may provide a useful starting point for both normative and descriptive investigations. One reason is that by focusing on the activity of likelihood judgment in this way, we may come to greater understanding of these other choices. Another reason is that the very fact that a variety of people are able to enjoy some measure of objective success working as Applied Statisticians suggests that there may, in general, be a relatively wide latitude in the ways that these choices can be made and still produce "reasonable" judgments. In section II.3 I relate the descriptive framework above to theories of "Direct Inference" that have been proposed in the Philosophical literature, provide a slightly more precise description of the steps described above, and make use of this description in the formal specification of a theory of similarity based likelihood judgment.

I.2.2 The nature of descriptive theories of default judgment

One approach to the complete specification of a default reasoning strategy for likelihood judgment is to decompose the descriptive task into three separate parts. The first part of the specification is a description of how the strategy works when it is concretely instantiated on particular occasions and applied to produce an attribution of likelihood. Such a description states which type of proposition the strategy can be used to attribute likelihood to, which type of beliefs/knowledge the strategy makes use of as evidence for this type of attribution, and, algorithmically, how particular tokens of this type of evidence are made use of to produce attributions of likelihood to particular propositions.

It is true but useless to say of such a strategy that the likelihood it attributes to a proposition is an evaluation of THE conditional probability of that

proposition given all of the current evidence. The use of the definite article could only be justified under conditions in which a probability distribution accomodating the proposition had already been established, in which case we wouldn't really need a reasoning strategy to evaluate it. What I will call Part I of the description of a default reasoning strategy provides some additional information - a specification of which type of conditional probabilities a given strategy evaluates and how this evaluation proceeds. Most of the theories that have been proposed as descriptions of human reasoning strategies 1a) - 4b) primarily address this part of the specification problem. At least to the extent that these proposals make any precise claims, they seem to make empirical claims of the following type.

Type of Claim: There is an algorithm, A, operating on evidence of type E, which people employ to produce (or affect other independently produced) attributions of probability to target propositions of type P according to reasoning pattern R in internally represented judgment contexts of type C. It is important to note that C is not a complete description of a person's epistemic state. It is rather a partial description which provides necessary but not sufficient (descriptive) conditions for the algorithm to be applied.

The capitalized letters above represent variables that would be concretely defined in a precise version of such a theoretical claim. The present work will propose a precise theory of this form for similarity based likelihood judgment. My description of what I take to be the remaining parts of a complete specification of a default reasoning strategy will hopefully make clear what other desirable knowledge about a default reasoning strategy is left out of such a description.

The second part of a complete specification is a description of the circumstances in which a default strategy will actually be instantiated (internally represented) in a particular way and used to produce a likelihood judgment. An example may help to clarify the contrast between the two parts of the description of a reasoning strategy for likelihood judgment that I am referring to. Consider the following reasoning scenario and "strategy " (the "reasoner" in this example is a life insurance company but the issues I am calling attention to are not affected by this).

Scenario:

Sam is a 49 year old man who has a history of heart disease in his family (his father had a heart attack). He himself has never had any heart problems and his serum cholesterol level is normal. However he was diagnosed as mildly diabetic in his early thirties, and had been a chronic smoker up until that time. What is the likelihood that he will live past the age of 75?

Strategy:

Life insurance companies routinely evaluate questions like this using some variety of a relative frequency strategy (1a). The likelihood that Sam will live past 75 would be evaluated by consulting an actuarial table. Cells in these tables are indexed by factors that are considered particularly relevant to life expectancy. Typically, Sam would be identified as belonging to some cell of such a table based on a yes/no listing of some particular set of his characteristics: history of heart disease in family - yes, history of heart disease personally - no, diabetic - yes, etc. Each cell contains a histogram recording the actual lifespan of the known cases which fit the criteria of membership for belonging to that cell - i.e. 5% of those cases lived less than 20 years, 25% lived between 20 and 30 years, etc.. The estimate that such a table provides for the likelihood of Sam living past 75 is equal to the relative frequency of people living past 75 (or the nearest approximation to that in the relevant histogram) among the tabulated cases with Sam's distinguished set of relevant characteristics (taking into account, of course, the fact that he has already lived to be 49).

According to 1a) above, a likelihood judgment strategy is a "relative frequency" strategy if the likelihood that a given individual who belongs to a class, C, has a given property is estimated as the fraction of the known set of individuals in C which have that property. The property in question in the "Sam" example is 'x will live past 75 years of age' and the class is determined by the description of the cell that Sam is assigned to. It is obvious that an actuarial table containing a given set of case descriptions can be built in an enormous variety of ways - choice of cells and choice of histograms can both vary. It is also clear that, in general, different tables

would provide different relative frequency estimates for Sam's longevity. In order to specify how a life insurance company will actually judge the likelihood that Sam will live past 75 years of age it is obviously not sufficient to specify that it uses a relative frequency strategy based on this particular set of cases. The descriptive label "relative frequency strategy" tells us, in this case, that the estimated likelihood that is produced by such a strategy ultimately results from evaluating the ratio m/n in some cell of some table. But which table? We do not know since many could have been built from the given set of cases. In order to specify a real reasoning strategy that would tell us how all of the information that went into making the table figured or did not figure in the given estimate, we need some more information. I am calling this other information a description of how the relative frequency strategy was instantiated to evaluate the target proposition on the basis of the known cases.

There is actually more that we would like to know even beyond the description of how a reasoning strategy is to be instantiated in some situation and how the instantiation will be used to produce an evaulation of likelihood. To see this note that if E stands for all currently available evidence, P stands for a proposition to be evaluated, and A stands for some new piece of evidence, then the constraints that probability theory by itself imposes on the relationship between Prob(P | E) and Prob(P | E&A) are very weak. Specifically, if Prob(P | E) is neither 0 nor 1, then Prob(P | E&A) can, in general, be any probability at all. The additional information that we would like, the third part of a complete specification, is a description of how a given judgment would affected by the introduction of new evidence. This third part, which is sometimes referred to as a description of epistemic commitment, may or may not be different from the second part in the case of different "reasoners". If the effect of new evidence is to cause a new judgment to be produced on the basis of the old evidence plus the new evidence considered equally, then the description of this third part is subsumed by the description of the first and second parts. However, there are often practical reasons why this latter strategy is too costly. Taken literally it might require one to save all of the old evidence and recompute all of one's inductively inferred beliefs after the receipt of new evidence [see Harman and Gaerdenfors for extensive discussion of this and related issues].

There is a substantial body of experimental data on topics such as biases of confirmation which indicates that the judgment strategies which human beings actually make use of often exhibit significant intransitivity relative to the order in which evidence is received [see Ross and Anderson '82 for an overview]. Such intransitivity implies that old evidence and new evidence are not treated identically in some cases. Sam and the life insurance company again provide a convenient example to illustrate some of the issues.

Consider three different types of evidence that might cause different types of adjustment to the estimate of Sam's longevity. The least significant effect would be produced by the introduction of a new case, fitting the same cell description as Sam, into the database of cases. If the individual described by the new case had lived past 75 years of age then the estimate for the likelihood that Sam will would be changed from m/n to (m+1)/(n+1). When n is large this is an insignificant adjustment, but is nevertheless an adjustment that depends on m and n together, and is not merely a function of the previously attributed probability m/n. To restate this point slightly differently, the significance of an assertion that 'the probability of Sam living past 75 years of age is p (= m/n)' is different than the significance of the assertion that 'the probability of Sam living past 75 years of age is p (= cm/cn for some large constant c)' even though the assertions appear identical. The ambiguity is only resolved upon the introduction of further modifying evidence. A more radical effect on the insurance company's estimate of Sam's longevity would be achieved by the discovery of relevant information about Sam that was not included in his case description, such as the fact that he has a tumor in his liver or that he works in a coal mine. In such a case Sam would be assigned to a different cell the company's actuarial table and the given estimate would be revised (as would be the cost of the policy they are willing issue). Relative to this type of information, description of the "epistemic commitment" of the insurance company to the estimate of Sam's longevity must include a list of factors that are relevant to the actuarial table and could potentially be discovered of Sam. Finally, there are undoubtably bodies of new evidence which would cause the insurance company to modify the actual cell structure of the table. Such information might be, for example, the new scientific discovery that some chemical used in a particular

kind of manufacturing process causes cancer. Whether or not someone has worked at that type of manufacturing plant now becomes an important factor in their case description. If the complete descriptions of all the cases that were used to build the old table are available then a new table with the additional factor cross tabulated with the old factors could be built from scratch. In such a case the resulting likelihood judgments produced by the reconstituted table would presumably be transitive relative to the ordering of the old and the new data. Otherwise, if all that is available are the old histograms, then some other technique for introducing the new factor must be adopted and the result will be intransitive relative to the ordering of the old and the new data.

A final general comment on descriptive theories of default reasoning is that these theories need not propose or endorse a "theory of probability". What I mean by a "theory of probability" is an attempt to directly specify precisely what it is that a person holds to be true of the world when they say "the probability of event A is p." Historically, a great deal of thought and argumentation has been devoted to such questions [see Fine '73, Gigerenzer '89 for reviews]. Despite these efforts, no consensus view of even a semi-formal referential semantics of probability attribution has emerged. The theoretical objects of study for theories of default likelihood judgment are relationships between probability attributions and other held beliefs; in this work, I focus on how judgments of likelihood are related to judgments of similarity, which are related in turn to featural and categorical knowledge. The study of these relationships is exemplary of a non-definitional approach to understanding the meaning of likelihood judgments. One non-definitional route to such understanding is through specification of the causal factors determining such judgments and the role such judgments play in reasoning and decision making. This priority of focus is common to most all psychological research on reasoning strategies 1a)-4b) cited above.

I.3 Similarity based likelihood judgment: plan of study

The strategy of similarity based likelihood judgment that is explored here is a default inference that is, roughly, instantiated and applied as follows.

Suppose that '*individual i has property P*' ' is an uncertain proposition of interest, that P is a property of type **P**, and that individuals $i_1,...,i_m$ are each saliently known to have property P, and that $i_{m+1},...,i_n$ are each saliently known not to have P. Then in the absence of other information, the probability that some new individual i_{n+1} has P is assumed to be some function of the similarity(**P**) of i_{n+1} to $i_1,...,i_n$ - where the notation similarity(**P**) is intended to express the idea that there are different kinds of similarity relations and that the appropriate similarity relations in this context are those appropriate to inferences related to **P**. This function of the appropriate similarity relationships is shown to vary positively with the appropriate similarity of i_{n+1} to $i_1,...,i_m$ and to vary inversely with the appropriate similarity of i_{n+1} to $i_{m+1},...,i_n$.

The greater part of the thesis will be devoted to advancing and experimentally evaluating a precise specification for the instantiated form and product of this reasoning pattern. A theoretical model for the nature of similarity relationships is also proposed. The proposed theory identifies similarities with subjective conditional probabilities and relates similarity based likelihood judgment to a variety of statistical estimation. Less precise relationships between similarities, features, and categories are also proposed and examined.

The point was made in section I.2 that this type of specification only speaks to one part of what we would ideally like to know about similarity based likelihood judgment as a default reasoning strategy. The examples provided there also helped to make clear why these other, missing parts of a complete description are elusive: in a situation in which the reasoner has even a moderate amount of background knowledge to potentially make use of as evidence for influencing a judgment of likelihood there will be a combinatorial explosion in the number of syntactically permissible instantiations for most reasoning strategies. In the relative frequency example above, an instantiation was determined by the combination of the set of known cases, a description of Sam, a choice of factors used to pick out a cell of the actuarial table - more generally referred to as a reference class (see section II.4), and the choice of histogram within a cell. It turns out that there

are also a number of different choices to be made in the instantiation of a similarity based strategy. The possibility that different reasoning strategies (1a-4b etc.) could be potentially be instantiated on the basis of a single moderately sized body of evidence only makes the situation more complicated, as does the consideration of how judgment would be affected by the introduction of further, unanticipated evidence into the knowledge base. Confronted with this complexity, it is self evident that any set principles which might provide a theoretical basis for predicting which specific judgment strategy and instantiation will be utilized in particular instances merits serious consideration. One possibility which would make things simpler would be if the actual choices among competing strategies which people commonly make in their reasoning are compatible or nearly compatible with the choices that are suggested by normative principles. We know or course that people sometimes commit elementary errors, but it may turn out to be the case that these errors can be isolated and set apart from normatively comprehensible principles of reasoning. But what are such principles? A natural place to look is at theories concerned with statistical model choice - i.e. the competitive selection of an optimal or near optimal statistical model from a set of alternative models types offering different and frequently incompatible descriptions of a given random environment. Unfortunately, this research topic has only been gradually taken up in any generality by the statistical community starting in the early 1970's, and so the theory that has so far been developed, while possessed of enormous potential, is still at a relatively early stage of development. What research there has been in this area has been pioneered by, and is most closely associated with, the related work of Akaike '74, Schwarz '78, and Rissanen '86, '87. The interested reader is referred to Sakamoto et al. '86 and Rissanen '90 as readable references and starting points for this literature.

II Description of similarity based likelihood judgment

The description of similarity based likelihood judgment as a statistical strategy will proceed as follows. Section II.1 will describe, in statistically meaningful terms, the nature of the evidence that is required to instantiate this pattern of judgment. Section II.2 will provide a qualitative theory of this

reasoning pattern. This qualitative theory will be useful for both understanding the theory and its experimental confirmation (to follow in Chapter III) as well as to help formulate rival descriptions to be competitively evaluated. I will often refer to any judgment pattern that satisfies these qualitative criteria as a model for similarity based likelihood judgment, though a unique quantitative model will ultimately be defended. Section II.3 will present a theory of the nature of similarity relations themselves. Section II.4 will introduce a formalism in which the instantiation of an instance of similarity based likelihood judgment, as well as the default assumptions of that instantiation, may be described syntactically. The formalism introduced there is closely related to (and capable of describing) theories of "direct inference" that have been proposed in the philosophical literature - primarily to describe normative theories of default strategies 1a and 1c. It is hoped this formalism will ultimately prove useful in providing a uniform descriptive language for talking about 1a - 4b and other default reasoning strategies. Section II.5 will complete the description of similarity based likelihood judgment by providing a quantitative specification of the reasoning algorithm that is used to produce numerical likelihood judgments from the (statistical) information contained in an instantiation of the strategy.

II.1 A Scenario for Inference

Before presenting an abstract definition for the type of reasoning scenario to which similarity based likelihood judgment applies, it will be helpful to consider a concrete example which has the same "informational skeleton" as the abstract version.

> Bill "from Indiana" is dining at a Roman trattoria. Before he left Indianapolis for his Mediterranean vacation, he told his "foodie" friends that he was looking forward to increasing his familiarity with Italian cuisine, and asked them for their menu recommendations. Each of them gave him a long list, but all that he can remember at the moment is that Sam recommended he try pasta e fagioli and Sally recommended he try pasta puttanesca. Bill is ignorant concerning the contents of both dishes and they are the same price on the menu. He decides to order the pasta

puttanesca because he feels that Sally's taste in pasta dishes tends to be more similar to his own than Sam's does.

The informational skeleton of the above scenario is intended to fit the following abstract description.

X needs to know if some individual, i_0, has some particular property, P. Not being able to test the matter directly, and lacking any certain knowledge, X is required to take his/her/its best guess - or alternatively, the situation requires X to choose to either engage in some activity involving i_0 or in the same activity with some other individual i_1, and so X must consider relative likelihoods for the truth of $P(i_0)$ and $P(i_1)$ - here I use the symbolic terminology P(i) to abbreviate the statement 'i has property P' and $\neg P(i)$ to abbreviate the statement 'i does not have property P. Assume that the relevant knowledge and beliefs available to X are described by the following four given assumptions:

1) that P is known to be a property from the class **P**,
2) that i_0 and i_1 are members of a finite set $\{i_k, 0 \leq k \leq m\}$ of related individuals*,
3) that X has beliefs about what proportion of the time each distinct pair of individuals in the set $\{i_k, 0 \leq k \leq m\}$, say i_j and i_l, had matching and non-matching values for properties in the class **P** - i.e. X has beliefs about the relative likelihoods of $P(i_j)\&P(i_l)$ vs. $\neg P(i_j)\&P(i_l)$ vs. $P(i_j)\&\neg P(i_l)$ vs. $\neg P(i_j)\&\neg P(i_l)$ for random P in **P**, and
4) that $i_2 \ldots i_n$ in $\{i_k, 0 \leq k \leq n\}$ are known by X to have P and $i_{n+1} \ldots i_m$ in $\{i_k, n+1 \leq k \leq m\}$ are known by X not to have P.

* This stipulation is intended to insure that it makes sense to talk about the similarities of each pair drawn from $\{i_k, 0 \leq k \leq m\}$ in the same sense of "similarity"- this idea will be formalized and elaborated in section II.3.

The type of information given by 1) -4), hereafter to be referred to individually as GI-1...GI-4 and collectively as the GIs, is available in many judgment situations. In the food example given above we could give the

label i_0 to Bill, i_2 to Sam, and i_3 to Sally. There were two properties: P1 was "x likes pasta e fagioli" and P2 was "x likes pasta puttanesca". The class **P**, containing P1 and P2, was something like the class of positive preferences for particular pasta dishes. The inference made was that Bill judged P2(Bill) to be somewhat more likely than P1(Bill). The basis for this judgment seemed to be primarily a function of his knowledge that P1(Sam) and P2(Sally), and his belief about the frequency with which his tastes relative to **P** match that of Sam and Sally respectively. In section II.2.2 I propose a theoretical account of similarity according to which it is proper to refer to such beliefs as estimated similarities.

II.2 Qualitative Description of Similarity Based Judgment

The following qualitative properties characterize a class of strategies for computing an estimate of likelihood in the reasoning scenario described in II.1. Some of these characteristics can be tested independently of particular quantitative formuli for producing estimates. All of them are useful for thinking about the nature of similarity based likelihood judgment.

> i. the probability of $P(i_0)$ will vary positively with the similarity of the pairs consisting of i_0 and each of the individuals known to have P and negatively with the similarity of the pairs consisting of i_0 and each of the individuals known not to have P.

> Principle i. implies the weaker principle i'.

> i'. the probability of $P(i_0)$ will vary non-negatively with the similarity of the pairs consisting of i_0 and each of the individuals known to have P and non-positively with the similarity of the pairs consisting of i_0 and each of the individuals known not to have P.

Although I will test Postulate i. directly, I include the definition of i.' here because at a later point in the thesis i.' will provide a natural point of commonality for a variety of different models for similarity based likelihood

judgment that I shall experimentally evaluate. What principle, i, (and its slightly different alternative version, i') says is that if i_0 is more similar to i_1 than to i_2 then no matter what other individuals are known to have P and not to have P, if we compare the estimate that is produced for the likelihood of $P(i_0)$ on the basis of the known set of cases plus the information that i_1 has P to the estimate for the likelihood of $P(i_0)$ formed on the basis of the known set of cases plus the information that i_2 has P then the former likelihood estimate will be found to be greater (not less) than the latter. The reverse would be true if we had instead formed the two estimates on the basis of the added information that i_1 does not have P in the former case and the information that i_2 does not have P in the latter. Note that this is simply a claim about similarity based likelihood judgment, not a universal property of human reasoning. Relative to the "Bill" example, principle i. makes predictions like the following: if Bill believes his taste in pasta dishes to be more like that of Fred than that of Sam, then Bill would think it more likely that he would like some unknown pasta dish when (the relevant information he has available is that) he knows Fred had liked it and Sally did not than he would be if the information had had was that Sam had liked it and Sally had not. Principle i' is slightly more conservative, saying only that Bill would not think it less likely that he would like some unknown pasta dish when (the relevant information he has available is that) he knows Fred had liked it and Sally did not than he would be if the information had had was that Sam had liked it and Sally had not. The truth of principle i implies the truth of principle i'. Here is a second principle.

> ii. in the absence of information other than the GIs given above, and if the unconditional probabilities of the individuals in the set $\{i_k, 0 \leq k \leq m\}$ are identical, the probability of $P(i_0)$ will be a function of only the structural information in GI-4 and the set of values given by the similarity function applied to each pair of individuals in the set $\{i_k, 0 \leq k \leq m\}$, holding the class **P** containing P constant.

Principle ii says that if all of the individuals about whom we surely know whether they have the property P or not are in the set $\{i_k, 1 \leq k \leq m\}$, and ab initio, we don't have any reason to think any of the individuals in the set $\{i_k, 0 \leq k \leq m\}$ are more likely than any others from the set to possess a

randomly chosen property from the class **P**, then the estimate for the likelihood that i_0 has P can be computed from the information contained in the combination of 1) an (m+1) x (m+1) matrix in which the j,k th entry is the value of the similarity of i_{j-1} to i_{k-1} relative to **P** and 2) a length m vector where the lth entry is a 1 if i_l is known to have P and a 0 if i_l is known not to have P. Principle ii. says that we should be able to predict Bill's judgment of the likelihood he will like the pasta dish using the information about who he knows liked the dish and how similar he believes their tastes to be to his own. Actually, principle ii. hedges this assertion by adding the proviso that Bill has no other relevant information. It would be nice to have a general theory that would allow us to say what other information is relevant, but none appears to be immediately forthcoming. In many cases, intuition allows us to declare specific facts as relevant or irrelevant. For the "Bill" example, the fact that pasta puttanesca is made with anchovies seems relevant. The fact that Frank Sinatra likes marinara sauce does not. It seems reasonable to assume that we derive these intuitions by simulating our own reasoning processes faced with such a situation and taking note of whether the information in question would alter that reasoning process. Clearly these intuitively derived conclusions are themselves provisional. The fact that Frank Sinatra likes marinara sauce could be deemed relevant if we were informed that Bill was a raving Frankophile.

Principles i. and ii., hereafter to be referred to as the Similarity Reasoning Postulates, are satisfied in the inductive models of Shephard '57, Nofosky '84, Medin & Schaffer '78, the similarity components in the models of Rips '75 and Osherson et al. '90, '91, and in many of the proposals in the area of artificial intelligence that has come to be called Case Based Reasoning. These relationships will be discussed further in section III.4. Postulates i. and ii. do conflict with some coherent ways in which reasoning might proceed. For instance, in general they will not be satisfied by abductive or explanation based inferences (strategy 3b) that seek to pin down the available pattern of data to some particular property or cause that may well deviate from the (statistically) general pattern described by similarity . The theoretical perspective I endorse does not deny that this type of leaping to specific conclusions is a common feature of human reasoning. I propose however,

that it is supported by some sort of information other than that contained in the GIs.

II.3 Similarity Itself

Most people feel that they can sensibly answer questions like "Which of the following is more similar to a raccoon: an opossum or a grizzly bear?" and "On a scale from 1 to 10 how similar are a raccoon and a grizzly bear?" Within Cognitive Psychology, what is generally meant by a "theory of similarity" is an account of how a person's answers to questions like these can be predicted on the basis of their other knowledge and beliefs. I refer to this type of an account as a theory of how similarity is estimated. The use of the term "estimated" is justified by the fact that it is possible to change a person's opinion about the relative similarities of two distinct pairs of individuals by providing that person with additional information about the individuals being considered. For example, someone who thought raccoons more biologically similar to opossums than to grizzly bears might revise this opinion if informed that evolutionary biologists believe raccoons to be members of the bear family (in the same sub-family as giant pandas) while opossums, which are marsupials, are evolutionarily distant from living mammals. That person might then revise their belief about the overall biological similarity of raccoons and opossums while continuing to believe that raccoons and opossums are the more perceptually similar than raccoons and grizzly bears.

A question logically distinct from the issue of how similarities are estimated is the question of what other beliefs are entailed by (estimated) similarities. The list of such beliefs should undoubtably include more than just similarity judgments and their logical consequences. Since the theory of similarity proposed here is expected to fulfil the role assigned by Similarity Reasoning Postulates i and ii, this second issue is the more central concern in the current context. Consequently, the "theory of similarity" proposed here is primarily a theory about the meaning and consequences of similarity as a belief.

I assume that what we commonly refer to as "similarity" is a mentally represented/computed function from ordered pairs of individuals and a domain to some partially ordered set of values. With this assumption in mind (but otherwise pre-theoretically), let Sim(i,j,D) stand for the amount of likeness of individual i to individual j with regard to their D-ness. Considerations of both similarity entailment and similarity estimation suggest that part of the notion of a domain, or way in which two individuals exhibit similarity, includes the specification of a specific class of properties among which those two individuals are to be compared. The belief that two things are similar in their D-ness entails some sort of expectation that they generally share D related properties. This view is completely consistent with the generally held idea that learning of some new D-related property that i and j share is cause to re-evaluate the estimate of Sim(i,j,D) upwards, at least by some tiny amount (depending, perhaps, on how much one already knows about i and j with respect to D) (see e.g. Tversky 77,78). Let us assume that a domain is a fixed (though perhaps uncountably infinite) set of properties, and that it is sensible to speak of a subjective probability distribution on the likelihood of encountering particular types of properties from a domain. Given these assumptions, the following definitions describe the belief entailments of similarity.

Let **D** be some fixed domain, i and j an ordered pair of individuals, P a property that is randomly chosen from **D**, and let the similarity of i to j with respect to **D** be abbreviated by Sim(i,j,D). For the present I equate a domain with an intuitively identifiable, but otherwise arbitrary, class of properties. In other words, a person knows a D-type property "when he or she sees one." A full explication of the notion of a natural domain would presumably provide explicit coherence restrictions on such property classes. The general view of similarity I shall propose is vaguely summarized by the statement that

a) Sim(i,j,D) = the subjective likelihood of the equivalence or substitutability of i for j relative to a randomly chosen P from **D**. In other words, given that some **D** related property holds of j, how likely is it that it would be just as "acceptable" to regard i as having that property. If we regard the notion of having a property as classical rather than potentially "fuzzy", then this

description can be slightly cleaned up as follows. Sim(i,j,**D**) = the subjective likelihood that P(i) has the same truth value as P(j) for a randomly chosen property in **D**, given that that truth value is "of interest". The rational for this last proviso will be made clear below.

One potentially confusing issue for shared-property approaches to similarity is that often one deals with the similarity of things that are not individuals in a logical sense, such as the similarity of soccer balls and basketballs. The main issue is the treatment of properties which hold of some basketballs and not others (e.g. "made of leather"). I will treat this issue by syntactically representing the "individuals" in the first two positions of the Sim function as sets, even when these sets contain only one member. I also assume that each set has a well defined subjective probability distribution over the selection of its members. The second version of line a) can now be re-expressed as

b) Sim(A,B,**D**) = the subjective likelihood for P randomly chosen from **D**, i randomly chosen from A, and j randomly chosen from B, that P(i) has the same truth value as P(j).

The reader should carefully take note of what has been smuggled in here during the course of the last two paragraphs. The function described by b) has an inherent ambiguity about it. Let us say that A is **homogenous** with respect to **D** if A and **D** are such that for every P in **D** and every i and j in A, either P(i) and P(j) or ¬P(i) and ¬P(j). Then if A and B are both homogenous with respect to **D** , Sim(A,B,**D**) will be something akin to the conditional probability of P(A) given P(B) for random P in **D**, where what is meant by P(A) is P(i) for <u>all</u> i in A. If B is homogenous with respect to **D** but A is not then Sim(A,B,**D**) may represent something like the expected proportion of A's with P given P(j) for j in B and random P in **D**. If both A and B are not homogenous with respect to D then Sim(A,B,**D**) is something like a covariance of A and B with respect to their **D**-ness. If one were designing an artificial language then it would be time to backtrack at this point and retrace one's steps because it is clear that this invitation to ambiguity can potentially get one in trouble when reasoning. In fact, a variety of observed "fallacies" in the literature on human reasoning with probabilities seem to arise as a

result of ignoring such ambiguities. Shafir et al. '90, Osherson et al. '90 and Osherson et al. '91 document a type of error which they term an "inclusion fallacy", and which is particularly relevant to similarity based likelihood judgment. An example of the inclusion fallacy which has been experimentally observed is the following: given the information that "all mice have sesamoid bones", human subjects will often assign a higher probability rating to the proposition that 'all mammals have sesamoid bones' than to the proposition that 'all hippopotamuses have sesamoid bones'. This pattern of reasoning is not a reflection of any confusion concerning whether hippopotamuses are mammals. These type of errors which are apparently quite common in reasoning about a variety of situations (see Shafir et al. '90), seem to reflect an undiscriminating application of similarity based likelihood judgment. A reasonable interpretation of why the proposition 'all mammals have sesamoid bones' is judged more likely than the proposition 'all hippopotamuses have sesamoid bones' because the similarity of mammals and mice as defined by b) above is greater than the similarity of hippopotamuses and mice. Harking back to the discussion of different types of reasoning errors that appeared in section I.2.1, it may be noted that if these judgments do indeed arise as a result of a similarity based judgment scheme, this does not mean that similarity based judgments are not reasonable in slightly different situations. For example, given that all mice have property P, it is not necessarily an error to produce a higher estimate of the proportion of all mammals which have property P than of the proportion of all hippopotamuses which have P. Nor is it an error to think that the average mouse shares more properties with the average mammal than with the average hippopotamus. In general, similarity based likelihood judgment will avoid inclusion errors when the property which appears in a target proposition and the properties used in the estimation of similarities are homogenous. If P is a property that is taken from a class of properties known to be homogenous with respect to mammals then we would not expect the above pattern of reasoning to be compelling for subjects. So for instance, the likelihood that all mammals contain "Wilson neutrino particles" (whatever these may be) given that all mice do would presumably be judged equal to the likelihood that all hippopotamuses contain Wilson neutrino particles given that all mice do. In the remainder of this section I will continue to discuss Sim(A,B,D) in terms of subjective probability, and A and B are to be

understood as homogenous with respect to D unless something else is indicated. The theory of similarity based likelihood judgment that will be proposed will have a statistical interpretation when these conditions hold. The reader is advised of the potential for ambiguity contained in this defintion of similarity (line b)), and the related definitions which follow.

The following four subjective probabilities provide the building blocks for a set of similarity formuli which are appropriate formal interpretations for b) under slightly different conditions.

$p_{11}(A,B,D)$ = probability of P(i) and P(j) for i in A, j in B, and P in D.
$p_{10}(A,B,D)$ = probability of P(i) and not P(j) for i in A, j in B, and P in D.
$p_{01}(A,B,D)$ = probability of not P(i) and P(j) for i in A, j in B, and P in D.

$p_{00}(A,B,D)$ = probability of not P(i) and not P(j) for i in A, j in B, and P in D
$= 1 - (p_{11} + p_{10} + p_{01})$

In the text below I omit the explicit notation of the dependence of these quantities on A,B, and D. One straightforward interpretation of b) is the conditional probability of P(i) given P(j). Algebraically, this quantity is expressed by line c).

c) $Sim(A,B,D) = p_{11} / (p_{11} + p_{01})$.

In general, the formula c) is asymmetric. Notice that since p_{00} and p_{10} do not appear in formula c), properties which entities in B never possess do not play a role in this similarity computation. Formula c) is the natural interpretation of formula b) under the stipulation that properties which entities in B do not possess are not "of interest". One reason why such properties would not be of interest is that we have some role for B already "staked out", and what formula c) expresses is the likelihood that an A could serve just as well.

Another reasonable interpretation of b) is simple agreement in truth value relative to properties in D (here, every property is "of interest"):

d) $Sim(A,B,D) = (p_{11} + p_{00})/ (p_{11} + p_{01} + p_{10} + p_{00}) = p_{11} + p_{00}$

This formula may be read as "the probability that either both i and j will have property P or both i and j will not have property P, for randomly chosen i from A, randomly chosen j from B, and randomly chosen P from D." In some situations, it is just the properties or features that neither member of a pair of individuals possess that are not of interest. For example, the property of having a smooth glossy surface, which is a feature of billiard balls, would typically have no role to play in evaluating the similarity of footballs and baseballs (presumably because neither type of ball is ever glossy), even though it is a property of some balls featured in games and sporting events. On the other hand, if a property represents a value along a dimension, like 'bigness' with respect to 'size', then the joint absence of 'bigness' may well be of interest and contribute to similarity. For example, the fact that neither killer whales nor walruses have legs may be a contributing factor to their perceived similarity as species of mammals. Of course, facts of this type could be coded 'positively'. Each mammal can be considered to possess the property no_leg(), where no_leg(x) is true just in case x has no legs. What these examples essentially show is that there are many degrees of freedom in the proposed similarity framework - which form of function is used, which type of properties are "of interest", which type of properties are represented, etc. Yet another degree of freedom is related to the known observation that the "contrast set", or the set of all individuals whose similarity of D-like properties is implicitly of current interest, plays a role in the "weighting" or the probability assigned to different types of properties. For example, if billiard balls were contained in the implicit constrast set involved in the aforementioned judgment of the similarity of footballs and baseballs, then not having a glossy surface could well play a role in judgment. Until most of these degrees of freedom are removed, questions like which similarity function is "the right one" will be indeterminate.

Relative to the probabilistic representational scheme introduced above, the removal of all properties not possessed by either individual of a pair from the domain under consideration, which may correspond psychologically to the choice of just that pair of individuals as the implicit contrast class, results in the equality $p_{00} = 0.0$. In this case formula d) reduces to

e) $Sim(A,B,D) = p_{11} / (p_{11} + p_{01} + p_{10})$

In each of the formuli b) - e), the similarity of individuals has been described as a function of the subjective beliefs called p_{11}, p_{01}, p_{10}, and p_{00} (with the latter value being arithmetically redundant). If similarity is indeed a function of the values p_{11}, p_{01}, and p_{10}, then the values of the similarity pairs mentioned in Similarity Reasoning Postulates i. and ii. above are functions of the information contained in GI-3 of section II.1. For clarity, I separately name this principle, which is intended to be paired with principles i. and ii. from above, Similarity Reasoning Postulate iii.:

> iii. holding the domain D constant, the similarity functions defined on pairs of individuals are themselves functions of the type of information contained in GI-3.

It is instructive to compare the framework introduced above with the familiar "Features of Similarity" model of A. Tversky. [Tversky '77] The latter theory applies to situations in which each individual is considered to possess a finite set of distinguished features. In such cases it is possible to define, relative to an ordered pair of individuals i and j, the following sets:

A = the set of features that both i and j possess,
B = the set of features that i has and j does not, and
C = the set of features that j has but i does not.

Tversky proposes that there are non-negative constants c_1, c_2, and c_3, with values dependent on context, such that the judged similarity of i and j is representable by the formula

f) $Sim_T(i,j) = c_1 f(A) - c_2 f(B) - c_1 f(C)$,

where f is some function of the overall salience of the features in a given set. In his '77 paper, Tversky also notes the possibility of an alternative functional form,

g) $Sim'_T(i,j) = f(A) / (f(A) + c_2 f(B) + c_1 f(C))$,

that would adequately capture his data (and theoretical perspective) in situations where similarities were normalized to assume values lying between 1.0 and 0.0. Formula f) is clearly quite similar to formula e) above. Both formuli can be described as a the ratio of the 'weight' of the matching features to the sum of the 'weight' of the matching features plus the 'weight' assigned to the features possessed by i and not j plus the 'weight' assigned to the features possessed by j and not i.

Their are two main differences between the theory proposed above and the "Features of Similarity" theory. The first of these is that the theory given above assigns a probabilistic interpretation to the "weights" of the different feature sets, and these interpretations make predictions about how similarities participate in inference and likelihood judgment, which is the central topic of this work and which will be described in what follows. The other major difference between the two frameworks is that the "Features" theory, at least formally, assumes that each individual really has some feature set which is determinate and is consulted when making similarity judgments. This feature set will be a function of the type of similarity being judged, but is otherwise fixed for an individual. As I understand the "Features" theory, if a contextually relevant feature of an object is, in general, saliently known to a person and yet does not play a role in a judgment made by that person concerning the similarity of that object and some other, that omission would be classified as a performance error.

The theory which I propose is agnostic concerning the existence of a finite distinguished feature set. However there are, I believe, convincing arguments suggesting skepticism about theories viewing similarity as solely a function of such feature sets. For example, assuming that you believe a Chihuahua is more similar as an animal to a German Shepherd than to a Siamese Cat, it is suprisingly difficult to identify features which the Chihuahua and the German Shepherd have in common that the Chihuahua and the Siamese Cat do not. It may be easier to think of features shared by a Chihuahua and a Siamese that are not shared by a German Shepherd. The most important factor responsible for the perceived similarity of German

Shepherd and Chihuahua seems to be simply the knowledge that Chihuahuas and German Shepherds are both Canines. A follow up paper by Tversky and Gati partly acknowledge and document this phenomena [Tversky '78] . Tversky and Gati studied the effect that experimental manipulations inducing the salience of categories had on estimated similarities. They concluded that categories or "clusters" have an important effect on perceived similarities in that they "highlight those features on which the clusters are based." This analysis applies most directly to situations where it is clear that membership in a category is based on a particular set of features. Of course, it is always possible to claim that an animal known to be a Canine has a feature canine() which is strongly weighted when comparing that animal to other animals with the feature canine(). Something akin to this seems to be the view taken by Tversky and Gati who state that "A feature may acquire diagnostic value (and hence become more salient) in a particular context if it serves as a basis for classification in that particular context." So if the Chihuahua and German Shepherd are classified together because we know that they are Canines, then the feature canine() has strong diagnostic significance.

It is possible to construct scenarios in which this story becomes strained beyond the point of tenability. Suppose for example, that a zoologist told you that deer and rhinoceros have digestive tracts strongly resembling one another, but you know nothing else about the digestive system of either mammal. This information might well influence your judgment about the similarity of the two mammals. It does not however, seem representable as a particular feature that both mammals have. If the information is represented as a new feature which deer and rhinoceros share, does that mean it is also a feature distinguishing deer from horse? It seems more plausible to interpret the psychological effect of the information as simply causing one to believe that the digestive systems of the two mammals are more similar, and hence that the two mammals are more similar overall, independently of particular known features. In the language of the theory presented above, relative to the individuals deer and rhinoceros and the domain 'properties of digestion', p_{11} increases, while p_{01} and p_{10} decrease. These quantities would change relative to the domain 'biological properties' also, though the change would be smaller. One might plausibly infer

changes in perceived similarities of other mammals on the basis of this information - surely without bestowing new features upon them. Of course factual knowledge about particular properties can also have these effects as well. While agreeing with Tversky that judged similarity is partly a function of factual knowledge about common and unique features, I suggest that perceived similarities are functions of more than such knowledge alone, even when accounting for the effects that alternative judgment contexts may have on the weighting of features.

As stated above, the theory proposed here is not intended to be a theory of precisely how similarities are estimated. The "features of similarity" theory may be regarded as the best current model for such a theory. If estimated similarities are the type of beliefs described above however, it makes sense that the factors contributing to perceived similarities should primarily be factors that have a (rational) role to play in the formation of beliefs about the related probabilities. One prediction made by such a construal is that the "weight" assigned to categorical features like canine() in the computation of the similarity of two individuals with respect to a domain should be proportional to the probability that a property chosen from that domain will be one that is homogenous among all members of the category. I will discuss this idea in greater detail in section III.6, and point out how it can be used to make sense of a number of different results in the field. Assuming momentarily, for the sake of argument, that this point of view is correct, it is admittedly possible to describe such a phenomena by stating that the salience of categorical information will be increased as a function of the judgment context. But this seems to be only a way of redescribing the fact that some judgment contexts make categorical information more important than others, and a salience based account does not really offer a reason for why this should be so.

II.4 A formalism for describing a default judgment strategy

In this section I consider a particular framework for describing default likelihood judgment strategies. Part of this framework will be a formalization of what I proposed as a partial description of what an Applied

Statistician does at the conclusion of section I.2.1. I am immediately intent on using this framework to describe the instantiation of similarity based likelihood judgment but the framework is also intended to be general enough so that descriptions of other default strategies could easily be added to the basic structure.

An important part of the description I will introduce is closely related to a class of normative proposals for likelihood judgment that are known as theories of "direct inference." These theories may be loosely thought of as theories about how to set up "the right" statistical model to use in estimating the likelihood that some particular proposition is true. The main hurdle such theories face is referred to as "the problem of choosing the right reference class." The "Bill example" of section II.1 is already complicated enough to illustrate the main issues. In that example, the protagonist reasoned to the belief that he would probably like pasta puttanesca. I wrote the story in such a way as to suggest that this belief was the result of some process of deliberation and, in some sense, based on his belief "that Sally's taste in pasta dishes tends to be more like his own than does Sam's." The intuition driving theories of direct inference is that, on an abstract level, two major steps in Bill's reasoning process can be distinguished:

Step 1: Bill estimates the indefinite probability that he will enjoy some random pasta dish given that he knows that Sally did (and doesn't know whether Sam did or did not);
Step 2: Bill accepts the likelihood value given by the estimator of Step 1 as an estimate for the probability that he will actually enjoy the pasta puttanesca dish, should he order it.

Regardless of one's opinion about the descriptive plausibility of such a story (not to mention the legitimacy of calling it similarity based reasoning), a question which immediately invites rational scrutiny is why Bill *should* reason in this way and not some other. Why shouldn't he base his reasoning, for example, on an estimate of the indefinite probability that he will enjoy an Italian dish that Sally liked? Or an Italian dish served in Italy? One strategy for answering such questions is to seek a definition of what exactly constitutes a way of reasoning to such a conclusion about likelihood,

and adequate criteria for preferring one way of reasoning to another. These are essentially the motivations for the theories of direct inference which have been advanced historically by Hans Reichenbach, and more recently by Kyburg '83, Levi '80, Pollock '90, and Bacchus '90. It can be clearly seen that the first question, "what exactly constitutes a way of reasoning?" is the question that I (somewhat vaguely) addressed in section I.2. I do not believe that any of the theories of direct inference that have been proposed so far provide definitive criteria for preferences among different strategies for likelihood judgment. In section I.3 I mentioned what I take to be the most promising current candidates for such criteria, though in many respects these theories are also not yet up to the task. I do think though that the proposed theories of direct inference provide an excellent perspective from which to view similarity based likelihood judgment, and suggest an appropriate framework for describing the epistemic context from which these judgments originate.

In an often quoted paragraph that proved seminal for research in this area, Reichenbach stated that "If we are asked to find the weight (*probability assignment*) holding for an individual future event, we must first incorporate the case in a suitable reference class. An individual thing or event may be incorporated in many reference classes... We then proceed by considering the narrowest reference class for which suitable statistics can be compiled." [quoted in Pollock, p.110, my parenthetical synonyms inserted] This idea, which is related to the insurance company example of section I.2.2, can be seen to apply to the questions about Bill's inference that were posed above. If Bill had adequate information from which to form a reliable opinion about how probable it was that he would like an Italian pasta dish, served in Italy, that Sally had enjoyed in America, then this opinion should take precedence in influencing his actions over an opinion formed solely on the basis of the likelihood that he would enjoy any pasta dish that Sally had. The "narrower" reference class here is Italian pasta dishes served in Italy that Sally enjoyed. The wider superset is all the pasta dishes that Sally enjoyed. Pollock '90 offers an appealing description of the epistemic commitments involved in direct inference which I now paraphrase: being warranted in believing that c, an individual, belongs to a class X provides a prima facie or presumptive reason for thinking that the probability of F(x) for x in X is a

good estimate to accept as the likelihood of F(c) (where "the probability of F(x) for x in X" should be given the de dicto reading - i.e. the probability value picked out in that way, whatever that value may be); however if Y is a subset of X , then knowledge that prob(F(y) | y is a Y) is not equal to prob(F(x) | x is an X) "defeats" the reason for believing that the prob(F(x) | x is an X) provides an appropriate estimate of conditional probability to accept as the probability of F(c). Pollock also includes the requirement that the predicate F be "projectible" with respect to the class X in order for prob(F(x) | x is an X) to serve as an appropriate estimate. For a discussion of projectible predicates see Goodman '55.

If interpreted as a psychological theory, Pollock's view would suggest that the likelihood judgment made in Step 2 above is properly viewed as an estimate which Bill has a prima facie reason for accepting, but which he may withdraw for various reasons. Theories of direct inference usefully distinguish at least three distinct classes of beliefs supporting the assignment of a probability to a singular proposition (a proposition that need not involve quantifiers when represented in first order logic), and therefore three corresponding classes of reasons for modifying such an assignment. In the terminology of the preceding paragraph, reasons for modifying the estimate of F(c) made on the basis of prob(F(x) | x is an X) are: there may be reason to modify the estimate of prob(F(x) | x is an X) itself, there may be reason to doubt that c is really an X, and there may be reason to believe that prob(F(y) | y is a Y) (again, de dicto) is a more appropriate estimate to accept for prob(F(c)) than prob(F(x) | x is an X). One reason for thinking that prob(F(y) | y is a Y) is a more appropriate estimate than prob(F(x) | x is an X) for F(c) would be if Y is a narrower reference class than X and the estimate prob(F(y) | y is an Y) is at least as reliable as the estimate prob(F(x) | x is an X). If both of these conditions hold then we may view the former estimate as dominating the latter and it is to be preferred. Obviously more general criteria are needed for choosing between estimators in the general case.

The definition of an **inductive argument** that is given below is designed to provide a formal descriptive framework for explicitly representing those aspects of provisional reasoning about likelihoods that are relevant to direct inference in particular and default likelihood judgments in general. The

framework is to be capable of formally describing instantiations of similarity based likelihood judgment as well as instantiations of other default strategies - e.g. 1a, 1c. At a general level, the proposed framework incorporates the following type of principles, objects, and actions which play a role in many descriptions of reasoning.

a) There is a language of sentences which represent statements about the world. Some of these statements may have a statistical character, such as 'between 95% and 97% of American households own televisions.'

b) Some of these statements (described in 'a)') are accepted as true, some are only regarded as being likely to a certain degree. The framework will represent the acceptance of a statement as true by the assignment of a likelihood to that sentence that is over some threshold.

c) Beliefs about statements, represented by the assignment of likelihood to these statements, may play a role in the assignment of likelihood to other statements.

d) Different reasoning patterns by which statements are assigned likelihoods, as well as the evidence that these reasoning patterns draw on, may sometimes be distinguished and noted within the inductive argument.

e) If the reasoning process by which and/or evidence according to which a statement was assigned a likelihood is not to be distinguished within the boundaries of the argument then this statement is to be specially noted as a "premise" of the argument.

f) A "justification", consisting of the specification of a reasoning process and a body of evidence for that reasoning process, must be given to every statement assigned likelihood which is not to taken as a premise. This feature of a reasoning process has proposed for AI systems by Doyle '79. This feature has a descriptive utility in the current context. It's main computational utility is discussed in g).

g) A statement, S', once assigned a likelihood, may be re-assigned some different likelihood at a later point in time/in an argument. As noted in section I.2, this is a common feature of probabilistic and other types of non-monotonic reasoning. A question which naturally arises is "What of other assignments of likelihood to other statements for which the earlier assignment of likelihood to S' participated as a justification? Are these now to be retracted or recomputed?" In the case of an all-powerful reasoner the answer would appear to be "yes". However, it is easy to see that once initiated, such a process could start an explosive chain reaction in which not only beliefs that were immediately justified on the basis of the revised premise but also beliefs that were justified on the basis of those beliefs and so on would all have to be retracted and or recomputed. Another problem is that to recompute all of these beliefs one would like to have access to most all of the evidence that was used in computing them in the first place, and storing this evidence indefinitely would also be prohibitively costly. Related to this issue is empirical evidence that in some situations, such as experimental debriefing, people do not revise beliefs which appear to have been justified on the basis of evidence that was later retracted. See Ross & Anderson '82 for a review. For further discussion of these issues and some proposals for strategies of belief revision that might be adopted which fall short of the retraction of all unjustified beliefs, the reader is advised to consult Harman and Gaerdenfors.

By design, the inductive argument parallels, as closely as possible, the more familiar deductive argument or proof.

Definition:

Let **L** be a language containing, among other things, a set **R** of relational terms admitting universal quantification and statistical quantification. By this I mean that **L** should be able to express propositions like 'all f's are g's' and 'between 50% and 70% of the f's that one might happen to examine are g's'. Also let **V** be a totally ordered set of likelihood values (typically the real interval [0,1]). Ordered pairs (l,v), representing the assignment of a likelihood value v to a sentence l are to be be called **evaluations**. An **inductive argument** is a sequence of ordered pairs $((l_1, v_1), j_1)...((l_m, v_m), j_m)$,

where the first component of each pair is an evaluation, (l_i, v_i), and the second, j_i, is a prima facie reason or **justification** for this evaluation. A justification will often be a depend upon previously established evaluations which will be called **assumptions**. This is analogous to common specifications of deductive proofs. One feature of an inductive argument that is clearly distinct from deductive proofs is non-monotonicity - a sentence may appear in more than one evaluation with different likelihood values attached to it. An evaluation (l_i, v_i) will be called **current** for an argument if it appears somewhere in the argument and there is no evaluation (l_k, v_k) such that $l_i = l_k$ and $k > i$. A set of evaluations will be called **current** if all of its members are current. A **justification** will be called **current** if its assumptions are **current** - but see g) immediately above. Three types of justifications are particularly relevant to present purposes. As with deductive arguments, two sorts of justifications that should naturally be allowed are **premises** - special evaluations that are assumed without further justification - and deductive inferences based on earlier evaluations. It will be convenient to suppose that there is a single value **a** in **V** that plays the role of an acceptance threshold for sentences in **L**. Specifically, current evaluations (l,v) in which v is greater than or equal to **a** are candidates to serve as antecedents for deductive justification. If g is a set of evaluations and l is a sentence then the terminology "l is deductively inferable from g" will have the special meaning that there is a current subset of g, $\{(l_1, v_1),\ldots,(l_{i-1}, v_{i-1})\}$ such that $v_1,\ldots v_{i-1}$ are all greater than or equal to **a** and l is logically implied by $(l_1\&\ldots\&l_{i-1})$. If l is deductively inferable from g, this is a justification for the evaluation (l,**a**). It is desirable that the logic that is to be used for deductive inference be somewhat more general than ordinary first order logic so that it may make inferences which follow deductively from statistical statements as well as propositional ones. So for example, from 'the percentage of birds that fly is greater than 80%' one should be able to inductively infer things like 'the percentage of birds that fly is greater than 75%.' Halpern '87 and Bacchus '90 have developed proof theoretic logics that could serve this purpose - see also Geffner and Pearl '90 and Pearl '88 for a suprising extension of a proof-theoretic logic to a type of non-deductive (non-monotonic) probabilistic inference.

The formal syntactic specification of a justification is an ordered pair (*,j'), where * is a member a special class of symbols denoting types of justification, and j' has whatever further syntax is appropriate to *. For deductive justifications, j' is a list of assumptions from which the sentence in the current evaluation was deductively infered. The symbol *prem* will denote a premise type justification and the symbol *ded* will denote a deductive justification. Note that an inductive argument is only intended to represent a particular sequence of reasoning steps with a beginning and an end. For this reason, the appearance of an evaluation as a premise within a particular argument does not note a distinction between cases in which this premise arose from innate knowledge, as a result of some primitive perceptual experience, or was actually the conclusion (final evaluation) of some other involved inductive chain (if the latter two are indeed different).

What has been called an inductive argument up to this point is so general that it doesn't really say anything. One could tack on almost any set of reasoning principles that one wanted. For example, a purely Bayesian system defined on a finite field of propositions could be specified by allowing assignments of likelihood (priors) to all semantically distinct conjunctions of propositions as initial premises, evidence statements which are "accepted" as further premises, and the law of conditionalization as an additional justification for asserting new assignments of likelihood. Since the law of conditionalization is a theorem which follows from any standard axiomatization of probability, this updating rule could even be subsumed as part of *ded*. The proposed framework acquires content through the specification of which justifications are allowable and what priority is to be given to each of them.

The following definitions specify a genuinely inductive variety of justification intended to represent the type of default reasoning embodied in direct inference and the production of likelihood estimates by an Applied Statistician (and Step 1 and Step 2 above). I use the type symbol *dir* to denote this type of justification.

Let g stand for a set of evaluations, and let r and s be members of **R** with identical arity, n (i.e. relations which each take n arguments). Let

proposition that for every $x_1...x_n$, $r(x_1...x_n)$ implies $s(x_1...x_n)$ (i.e. all r combinations are also s combinations) be deductively inferable from g. A **direct estimate** is a function e from triples (r, s, g) to V. Semantically, the expression $e_L(r,s,g) = v$ (in V) indicates that e_L estimates, as a function of the evidence given by g, that the likelihood of a random s-combination also being an r-combination is v. In what follows, let f also be a member of R with arity n (the arity of r and s), and let g_1 and g_2 be sets of evaluations.

Given the type definitions above, (*dir*, (f, r, s, g_1,g_2,e_L)) is a justification for (l_i, v_i) iff conditions (i) - (iv) jointly hold:

(i) for every n-tuple $x_1...x_n$, the conjunction ($f(x_1...x_n)$ & $r(x_1...x_n)$) is semantically equivalent to l_i;
(ii) it is deductively inferable from g_1 that for every $x_1...x_n$, $r(x_1...x_n)$ implies $s(x_1...x_n)$;
(iii) g_1 and g_2 are subsets of $\{(l_1, v_1),...,(l_{i-1}, v_{i-1})\}$; and
(iv) $e_L(f,s,g_2) = v_i$.

The justification (*dir*, (f, r, s, g_1,g_2,e_L)) is considered current if g_1 and g_2 are current. Here is the idea. One would like to assign a likelihood from v to l_i, and, since neither l_i nor its negation are deductively inferable from what has gone before, the likelihood to be assigned must be estimated. This is done by relating l_i to some more general class of events, through the device of first representing l_i as the conjunction of a "frame predicate", f, and a "specialization predicate", r, and then relaxing the restriction imposed by r in order to make use of knowledge/cases from the more general reference class, s (it is accepted, according to g_1 that all r's are also s's). Based on the knowledge/cases contained in g_2, an estimate of 'the likelihood of an n-tuple being an f given that it is an s' is produced, and this estimate is accepted as 'the likelihood of an n-tuple being an f when it is an r (and so an s as well)', which is equivalent to satisfying the proposition l_i. In general, the acceptability of this estimate will depend upon the extent to which patterns which hold true in general (statistically) for s's hold true for the subclass r as well.

In the discussion of section I.2 of the thesis I described an instantiation of a default reasoning strategy for likelihood judgment as a decision to provisionally accept a certain judgment of a conditional probability as an estimate for the likelihood, or unconditional probability of a certain proposition to be evaluated. Given the framework outlined above, it is possible to be more specific. Let's take up the "Bill" example again.

Letting f(x) = Bill likes x, r(x) = x is pasta puttanesca, and s(x) = x is a pasta dish that Sally likes, the reasoning process of Step 1 and Step 2 above can be described in the framework of the inductive argument as follows.

1) ((Bill likes p % of the pasta dishes that Sally does, **a**), (*prem*))
2) ((pasta puttanesca is a pasta dish, **a**), (*prem*))
3) ((Sally likes pasta puttanesca, **a**), (*prem*))
4) ((Bill likes pasta puttanesca, (p/100)), (*dir*, (f, r, s, {2,3},{1},e_L)))

What I am calling the "instantiation" of the direct inference strategy in this example is the determination that the conditional probability to be evaluated is the conditional probability of f(x) given s(x), and the determination that this probability is to be evaluated on the basis of the evidence in the set {1}. This conditional probability is to be accepted as the likelihood of 'Bill likes pasta puttanesca'. What I am calling the inference strategy itself is determined by e_L. The estimator of choice is obvious for this simple case. A more complex estimator will be necessary in the case described by the GIs. Such an estimator will be described in section II.4. Using the framework provided above it is possible to give a formal description of the pattern of reasoning = (estimator + instantiation) that I am calling similarity based likelihood judgment. This description is complete except for the specification of the estimator (see II.4) - I reproduce the GIs here for convenience.

GI-1) that P is known to be a property from the class **P**,
GI-2) that o_0 and o_1 are members of a finite set $\{i_k,0...k..m\}$ of related individuals,
GI-3) that X has beliefs about what proportion of the time each distinct pair of individuals in the set $\{o_i\}$, say i_j and i_k, had matching and non-

matching values for properties in the class **P** - i.e. X has beliefs about the relative likelihoods of $P(i_j)\&P(i_k)$ vs. $\neg P(i_j)\&P(i_k)$ vs. $P(i_j)\&\neg P(i_k)$ vs. $\neg P(i_j)\&\neg P(i_k)$ for random P in **P**, and

GI-4) that $i_2 \ldots i_n$ in $\{i_{k,0\ldots k\ldots m}\}$ are known by X to have P and $i_{n+1} \ldots i_m$ in $\{i_{k,0\ldots k\ldots m}\}$ are known by X not to have P.

The default reasoning strategy known here as similarity based likelihood judgment can be formally expressed as the inductive argument appearing immediately below. There are actually two mildly different versions of basically the same estimator that I would like to consider. Suppose that the set $\{i_{k,0\ldots k\ldots m}\}$ of related individuals is a subset of $\{i_{k,0\ldots k\ldots m+q}\}$ of related individuals. In one of the two cases q will be equal to 0, in which case the former set will not be a proper subset of the latter. In the other case $q > 0$. The cardinality of the set $\{i_{k,0\ldots k\ldots m+q}\}$ will be a variable in the following argument.

1) $((\mathrm{Sim}(i_1,i_2, \mathbf{P}) = p_1, \mathbf{a}),$ (*prem*))

•
•
•

k) $((\mathrm{Sim}(i_{m+q-1},i_{m+q}, \mathbf{P}) = p_k, \mathbf{a}),$ (*prem*))

k+1) $((i_2$ has P, **a**), (*prem*))

•
•
•

k+n) $((i_n$ has P, **a**), (*prem*))

k+n+1) $((i_{n+1}$ doesn't have P, **a**), (*prem*))

•
•
•

k+m) $((i_m$ doesn't have P, **a**), (*prem*))

k+m+1) ((P is in **P**, **a**), (*prem*))

Conclusion:

k+m+1) $((i_1$ has P), v), (*dir*, j'))

where j' =

('i1 has x',
'x is P',
'x is in **P** & i_2 has x & ...& i_n has x & i_{n+1} does not have x & ...
& i_m does not have x',
{(k+m+1)},

$\{1, \ldots, k+m\}$,
e_L)
and for f = 'i_1 has x'
and s = 'x is in P & i_2 has x & ...& i_n has x & i_{n+1} does not have x &
...
& i_m does not have x'
$v = e_L(f,s,\{1, \ldots, k+m\})$.

This argument reaches a conclusion about the likelihood of $P(o_1)$, using the GI's in an essential way. Recall that by Similarity Reasoning Postulate iii, the appropriate similarities are functions of GI-3. So evaluations entered as premises on lines 1) through k) are information given by GI-3 (with the "OK" of GI-2). Lines k+1) through k+n) are information given by GI-4. Line k+m+1) is the proposition given by GI-1.

The first of the two different versions of the estimator corresponds to the scenario in which the only similarities which play a role in the estimate are the similarities between members of the set involving the unknown case of interest and the set of known cases. In the other version, similarities between members of a set involving the unknown case of interest, the set of known cases, and some other unknown cases will play a role in the estimate. It is not psychologically plausible to think that too many other unknown cases will play a role in this reasoning strategy. But it is not implausible that a few might and it is convenient to allow this freedom for two reasons. The first of these reasons is that this freedom is necessary to establish a connection between the estimator described in the next section and existing computational strategies for associative memory and perceptual inference in neural networks that I will discuss in section IV. The other reason is that is will be experimentally more convenient to test this alternative version of the estimator.

Notes:

The argument above idealistically treats similarities as estimates of arbitrary precision (assuming V is [0,1]). A more realistic treatment would replace statements like ($(Sim(i_1,i_2, P) = p_1$, a) with statements like ($(Sim(i_1,i_2, P)$ is between c1 and c2 , a). The inductive argument given above, supplemented

with the details of an estimating procedure that are provided in the next section (II.4) and perhaps modified as indicated, is intended to be a precise description of the form of default reasoning that I have called similarity based likelihood judgment. It is not intended to be a description of a non-monotonic logic or a system of reasoning. The main obstacle to turning this description into a suitable non-monotonic logic is, as indicated above, the lack of an adequate theory for which direct inference step should be performed when more than one is possible. The description of direct inference is close to the different ones given by Kyburg '83, Pollock '90, and Bacchus '90. Bacchus actually does situate a rule of direct inference within a more general non-monotonic logic. He handles the problem of choosing among those reference classes which stand in a subset/superset relationship by having a rule in his logic stating that it is permissible to non-monotonically suppose that the conditional expectation of the frame predicate with respect to a narrow reference class is equal to the conditional expectation with respect to a wider reference class unless there is evidence to the contrary, in which case the inference is blocked. While this assumption at first appears liberal, it is conservative enough to make the logic inadequate to sanction obvious judgments in some situations. For example, suppose I know that the Canadian Football league has 20 teams and one of them is the Toronto Argonauts. I also know that exactly one team wins the championship each year and that this is almost always a team with a good offense and/or a good defense. I do not know whether the Toronto Argonauts currently have (or have ever had) either. If I also do not know anything about any of the other teams in the league then it seems reasonable to attribute a probability of 1 in 20 to the proposition that the Argonauts will win the championship next year. However, I could not use Bacchus' non-monotonic assumption to license such an inference because the reference class of teams in the league with a good offense or a good defense is narrower than the reference class of all teams in the league and I have evidence to the effect that the conditional expectation of x winning the championship given that x is a team in the league with a good offense or a good defense is not equal to the conditional expectation of x winning given that x is a team in the league. To summarize, I am blocked from inferring that each of the twenty teams has an equal chance of winning relative to my knowledge base because I know that my knowledge base is incomplete. This seems

undesirable. Of course one could argue that it would be foolish of me to give out 20:1 odds on the Argonauts to all takers but that is a different issue.

In an important sense the lack of criteria for choosing between estimates also makes the framework above incomplete as a general descriptive theory because it does not say which inference will be performed when available knowledge allows several (once again the issues of what is relevant and which instantiation to pick rear their ugly heads). Even though it is partial however, the description can be sensibly confirmed experimentally by creating a judgment scenario in which the information contained in a single set of GIs adequately summarizes a typical subject's relevant knowledge - where "adequate" is given the operational criteria that we can lay our hands on a tangible instantiation of that typical subject's GIs which can then be used to predict that subject's likelihood judgments accurately. Experiments designed along these lines are described in section III.

II.5 Maximum Entropy Estimates Based on Similarity

The estimator e_L appearing in the argument above is an estimator of the probability that i_1 has the property P given that P is drawn from the class **D**, i_2 through i_n have P, i_{n+1} trough i_m do not have P, and we have statistics/beliefs concerning the percentage of the time each pair (i_j,i_k) "agree" for P in **D** - these beliefs are summarized by the estimated similarities of these pairs for the domain **D**. The statement that the two different versions of the estimator referred to at the end of the preceding section are only minor variations of one another will be supported by the fact that it will only be necessary to provide a single description of an estimator in this section. The reader may bear in mind the two interpretations though. One in which k ranges between 1 and m+q with q=0 and the other in which k ranges between 1 and m+q with $q > 0$. In section II.2.2, different models for the probabilistic implications of similarity were discussed, each having a slightly different interpretation for "agree". In this section I shall assume model d), restated below.

d) $Sim(A,B,D) = (p_{11} + p_{00})/ (p_{11} + p_{01} + p_{10} + p_{00}) = p_{11} + p_{00}$ where

$p_{11}(A,B,D)$ = probability of P(i) and P(j) for i in A, j in B, and P in D.
$p_{10}(A,B,D)$ = probability of P(i) and not P(j) for i in A, j in B, and P in D.
$p_{01}(A,B,D)$ = probability of not P(i) and P(j) for i in A, j in B, and P in D.
$p_{00}(A,B,D)$ = probability of not P(i) and not P(j) for i in A, j in B, and P in D
$= 1 - (p_{11} + p_{10} + p_{01})$

Model d) differs from the other similarity models that were considered in section II.2 principally in its inclusion of p_{00} in the formula. This inclusion is appropriate in the judgment context described by the GIs because were are explicitly concerned in some cases with the evidential impact of the knowledge that one individual does not have a property on the likelihood that another individual does not have that property. The use of similarity model d) will also make the specification of the maximum entropy estimator simpler, although it could be done with other models as well. One should keep in mind also that the idea of maximum entropy estimation has considerable generality in the type of probabilistic information it can make use of and its computational rational, and so could be potentially used to combine information from similarity with other types of probabilistic belief.

I will assume throughout this section that the likelihood values entailed by the estimated similarities and the likelihood values to be returned by the estimator really are standard probability values on the conventional [0,1] scale. The relationship between this arbitrary scale and other arbitrary scales which subjects might make use of when expressing judgments of likelihood and similarity will be discussed in section III.4.4.5..

The idea of the estimating procedure is to use the pairwise similarities to estimate a probability probability distribution, pr, representing the likelihoods of various patterns of the individuals $i_1 ... i_m$ having and not having a random property in D, and then the estimate for the likelihood of i_1 having this random property will follow from conditionalizing (using pr) on the event that i_2 through i_n have this random property and i_{n+1} through i_m do not.

To be precise about the estimation of pr, it is necessary to first identify its domain (in the sense of functions, not similarities) - a set of mutually exclusive and exhaustive basic events to be assigned probabilities by pr. A **basic event** is a pattern of the m individuals having and not having a random property. Such a pattern is representable by a length m vector of 0's and 1's, where a 0 in the ith position stands for the condition that o_i does not have the property of interest and a 1 in the ith position stands for the condition that o_i does have the property. There are 2^{m+q} possible length m vectors of 0's and 1's, and hence a list of 2^{m+q} such basic events is exhaustive. From their definition it is obvious that they are also mutually exclusive. A distribution on this space of events may be thought to represent subjective beliefs about relative likelihoods of different patterns of the individuals under consideration having and not having a property randomly chosen from **D**. A standard probability distribution on this space assigns a number in the interval [0,1] to each basic event and the numbers sum to 1. The probability of a set that is a union of basic events is equal to the sum of the probabilities of the distinct basic events in the set. Since the estimated distribution, pr, is required to be consistent with the beliefs entailed by the similarities, the similarity values act as constraints on this distribution. To see the nature of these constraints, note that if $Sim(i_j,i_k,D) = .9$ then pr is required to be such that a total probability mass of .9 will be distributed to the set of basic events containing exactly those in which the ith and jth positions are either both 1 or both 0, and only a probability mass of .1 is left for the remaining basic events to share. If $Sim(i_j,i_k,D)$ was between .85 and .95, this would place an analogous type of constraint on pr.

Given the definitions above, it can be stated without any ambiguity that the chosen probability distribution pr is consistent with the constraints given by the similarity values of the $(m+q)(m+q-1)/2$ pairs (i_i,i_k) on the space of 2^{m+q} basic events, and that among the set of consistent distributions (assuming that this set is non-empty) it is the one that uniquely maximizes Shannon entropy. This is stated more formally below (with the aid of definitions of course).

$.be._k$	=	basic event #k (k between 1 and 2^{m+q})
val(i,k)	=	the value of the ith position of the kth basic event

$f_{ij}(.be._k)$ = 1.0, if val(i,k) = val(j,k) and 0.0, otherwise

The acceptable probability measures on the space of basic events are constrained to satisfy the (m+q)(m+q-1)/2 equations (or the (m+q)(m+q-1) inequalities)

$E(f_{ij}) = \Sigma_k f_{ij}(.be._k)pr(.be._k) = Sim(i,j,\mathbf{D})$

$(E(f_{ij}) \geq \Sigma_k f_{ij}(.be._k)pr(.be._k) =$ lower bound of Sim(i,j,**D**),
$E(f_{ij}) \leq \Sigma_k f_{ij}(.be._k)pr(.be._k) =$ upper bound of Sim(i,j,**D**)).

The entropy of a probability measure on a discrete space is defined to be

$Entropy(pr) = - \Sigma_k pr(.be._k)\log(pr(.be._k))$.

It is a well known fact that if the constraint equations (inequalities) are satisfiable by any probability measure then there is a unique measure, pr, which has greater entropy than any other probability measure satisfying the equations (inequalities), and the following descriptions will be true of pr [e.g. Jaynes '79, Kullback '59, Bishop et al. '75]:

There exist (m+q)(m+q-1)/2 constants c_{ij} and a special constant c_0 such that

a) $pr(.be._k) = c_0 \cdot EXP[\ \Sigma_{ij}\ c_{ij} \cdot f_{ij}(.be._k)]$ and

b) $1/c_0 = \Sigma_k\ EXP[\ \Sigma_{ij}\ c_{ij} \cdot f_{ij}(.be._k)]$

c) Upper bound of $Sim(i,j,\mathbf{D}) \geq E(f_{ij}) = \Sigma_k\ f_{ij}(.be._k) \cdot pr(.be._k)$

$= (d/dc_{ij})\ (\Sigma_k\ EXP[\ \Sigma_{ij}\ c_{ij} \cdot f_{ij}(.be._k)])$

Lower bound of $Sim(i,j,\mathbf{D}) \leq E(f_{ij}) = \Sigma_k\ f_{ij}(.be._k) \cdot pr(.be._k)$

$= (d/dc_{ij})\ (\Sigma_k\ EXP[\ \Sigma_{ij}\ c_{ij} \cdot f_{ij}(.be._k)])$

The definitions specify pr uniquely. Given a probability measure pr, an estimate for the probability of o_1 having a random property called P' from D given that i_2 through i_n have P' and i_{n+1} through i_m do not have P' is given by the conditionalization formula:

d) $pr(P'(i_1) \mid P'(i_2) \& \ldots \& P'(i_n) \& \neg P'(i_{n+1}) \& \ldots \& \neg P'(i_m)) =$

$$pr(P'(i_1) \& P'(i_2) \& \ldots \& P'(i_n) \& \neg P'(i_{n+1}) \& \ldots \& \neg P'(i_m)) \;/\; pr(P'(i_2) \& \ldots \& P'(i_n) \& \neg P'(i_{n+1}) \& \ldots \& \neg P'(i_m))$$

The wisdom of using this procedure to yield an estimate for the likelihood of $P(i_1)$ (where we have some specific P in mind) and possibilities for realizing such an estimate algorithmically and implementationally will be addressed in the discussion concluding the experimental evaluation of this estimator in section III.4.5.5 and in section IV of the thesis. The data analysis in section III.4.5.5 will, in effect, provide an example of how to handle the case in which the similarity constraints are not consistent. This case will also be discussed in section IV. The logic of using maximum entropy estimates calibrated on the similarity values will be discussed. For the time being, the proposed theory may come to seem less abstract if one takes note of the fact that this theory is simply one estimator which realizes, in detail, the general approach described by Similarity Reasoning Postulates i, ii, and iii. Postulates ii and iii taken together simply say that the likelihood estimate will be a function of the GIs alone. They are clearly satisfied by the proposed estimate. It is less immediately obvious that Postulate i is satisfied. This postulate stated that the probability of $P(i_1)$ will vary positively with the similarity of the pairs consisting of i_1 and each of the individuals known to have P (i_2, ... ,i_n) and negatively with the similarity of the pairs consisting of o_1 and each of the individuals known not to have P (i_{n+1}, ...,i_m). An argument to the effect that the proposed estimating procedure satisfies this condition is given below.

Lemma: The procedure above satisfies Similarity Postulate i.
Proof:

Let A be the set of .be.$_k$ such that
val(2,k) =1, ..., val(n,k)=1, val(n+1,k)=0, ..., val(m,k)=0

(i.e. A contains the events in which known cases given by GI-4 are correctly instantiated)

Let B be the set of $.be._k$ such that $.be._k$ is in A and val(1,k) = 1
(B contains the events in which the known cases are correctly instantiated and the conclusion is a 1 or "true")

Let C be the set of $.be._k$ such that $.be._k$ is in A and val(1,k) = 0
(C contains the events in which the known cases are correctly instantiated and the conclusion is a 0 or "false")

From d) above, the estimating procedure described returns

$$d)\ pr(P'(i_1) \mid P'(i_2) \,\&\, \ldots \,\&\, P'(i_n) \,\&\, \neg P'(i_{n+1}) \,\&\, \ldots \,\&\, \neg P'(i_m)) =$$
$$pr(P'(i_1) \,\&\, P'(i_2) \,\&\, \ldots \,\&\, P'(i_n) \,\&\, \neg P'(i_{n+1}) \,\&\, \ldots \,\&\, \neg P'(i_m)) \;/$$
$$pr(P'(i_2) \,\&\, \ldots \,\&\, P'(i_n) \,\&\, \neg P'(i_{n+1}) \,\&\, \ldots \,\&\, \neg P'(i_m))$$

which is equal to (pr(B) / pr(A)) = (pr(B) / (pr(B) + pr(C)).

From c), Sim(1,j,D) is increasing (decreasing) if and only if the expected value of the the corresponding statistic, $E(f_{1j})$, is increasing (decreasing). For a given j, let us call R(j) the set of basic events x such that $f_{1j}(x) = 1.0$ and S(j) the complement of R(j). If $E(f_{1j})$ is increasing then some probability mass assigned to basic events in S(j) must be reassigned to basic events in R(j). For j in the set {2, ... , n} this will cause an increase in the likelihood estimate d) because the basic events in B can only increase in probability and those in C can only decrease. To see this, it is sufficient to note that for j in {2, ... , n} the intersection of C and R(j) is the empty set. A symmetrical argument indicates that for j in {n+1, ... , m} an increase in Sim(1,j,D) will cause a decrease in the estimate d).

III Experimental Evaluation of Similarity Based Likelihood

III.1 Experimental Questions

A series of experiments were conducted to test the theory described above. These experiments focused on the following broad questions:

EQ1) Within an informational context that is well described by the GIs, is human likelihood judgment well described by Similarity Postulate i.?

EQ2) Within the same informational context as 1), is human likelihood judgment well described by Similarity Postulate ii.? How well?

EQ3) Is the relationship between similarities and likelihoods (in the same informational context as 1) and 2)) best described by the mathematical relationships given in sections II.2 and II.4?

EQ4) Are the factors which contribute to the estimation of similarities rationally commensurate with the beliefs about likelihood which the theory holds similarities to entail? This question evidentially bears on the related question of how descriptively accurate Similarity Postulate iii. is.

II.2.1 Likelihood Judgment Task - General Remarks

Common to a group of experiments examining questions EQ1) - EQ3) was a likelihood rating task designed to observe patterns of human judgment in an appropriate informational context. To be appropriate, an informational context is required to be structurally well described by the reasoning scenario formalized in section II.1. To some extent, this requirement must be balance against a second desideratum: that the judgment context is natural enough for the judgment schemas utilized to remain undistorted and the similarity representations that are engaged to be of a type easily computed or represented in long term memory rather than artifacts of the task at hand. For most subjects, both of the foregoing considerations are satisfactorily met by an experimental paradigm in which the following question is typical:

a) "P is a biological property of mammals having to do with bone structure. What is the probability that horses have P assuming that giraffes do but polar bears do not?"

It is helpful to observe that the type of information required by the GIs is available to someone evaluating question a). The likelihood to be judged is the likelihood of P(horses). The information of type **GI-1** is that P is a member of the class **P** of biological properties having to do with bone structure of mammals. It is also implicit in the presuppositions of the question that P is a type of property that is homogeneous within mammal species - i.e. for every species, either all normal members of that species have the property P or all normal members of the species do not have the property P. So **GI-2** is simply the recognition that horses, giraffes, and polar bears are each particular species of mammals. Assuming that the similarity relation of interest here is symmetric in its first two arguments, then by Similarity Postulate iii., the information required by **GI-3** is accounted for by a subject's opinions about the three quantities Sim(horse,giraffe,**P**), Sim(horse,polar bear, **P**), Sim(giraffe,polar bear, **P**). **GI-4** is accounted for by the information that giraffe has property P and polar bear does not have property P. Intuitively, it seems that most subjects will have little other information that is relevant to the likelihood judgment requested above. This intuition is confirmed experimentally by results presented below showing that a typical subject's likelihood judgments can be well predicted by functions mapping **GI-3** and **GI-4** to real numbers (on the same scale as the judgments). The estimator described in section II.3 is such a function. The quality of the accord between subject's judgments and the predictions of this estimator and others is examined below in detail. The experimental paradigm built around questions like a) is a minor variant of one that was used by Osherson '91. That study established the general feasibility of statistically predicting human likelihood judgments using information of the kind **GI-3** and **GI-4**. The estimator that was advanced in that study will be re-evaluated here along with other candidates.

Four experiments were conducted in which subjects were asked to respond to questions similar to a) above. Typically, answers to EQ1)-EQ3) were obtained by combining information from more than one of these experiments. In

each of the experiments involving likelihood judgment, subjects participated in two experimental sessions occurring on different days. On at least one of these days, a subject recorded his or her responses to a set of questions like a) presented in an individualized booklet. The booklets were randomly generated by a computer program using a process that will be described below. Because the nature and form of the likelihood judgment task and the booklets were similar among each of these experiments, I will describe the prototypical version of the task first, and indicate small modifications in the separate individual descriptions of each experiment to follow.

III.2.2 Likelihood Judgment Task - Subject Instructions

The instructions for this task were as follows:

> This experiment concerns your judgments about the probabilities that particular mammals possess particular biological properties. Before making a probability judgment, you will be told only that the property in question is related to some given aspect of mammalian biology and that some other mammals either do or do not possess that property. The aspects of mammalian biology that are dealt with in this experiment include the following: bone structure, digestion, dentition, thermal regulation, and fluid regulation.
>
> Examples of particular properties related to each of these aspects are given below. All of the properties to be considered in this experiment are possessed by some but not all kinds mammals. You will be asked to judge the probability that one kind of mammal has a property given examples of its occurrence and non-occurrence among other mammals. As you recall, probabilities are numbers between 0 and 1, though for convenience you should express these as percentages ranging from 0 to 100. Use numbers close to 100 to assign high probabilities, and numbers close to 0 to assign low probabilities.

Subjects were then shown the following list of exemplary properties, which they were told were unfamiliar, but chosen merely to convey of rough idea of the property aspect classes (which are named in the bold face type).

Properties involving the **bone structure** of mammals:

have skull orbit that is broadly continuous with the temporal fossa
have an extended humerus that is over twice the length of the clavicle
have about twice as many caudal as thoracic vertebrate

Properties involving the **dentition** of mammals:

have deeply hypsodent molars
have premaxillary gums
have transversally ridged pre-molars

Properties involving the **digestion** of mammals:

have their omasum and abdomasum seperately articulated
process their food caecotrophically
produce the enzyme ptyalin in their salivary glands

Properties involving the **thermal regulation** of mammals:

have their thermal neutrality point at about 20 C.
fail to initiate vasoconstriction via the smooth muscle of the peripheral arteries at temperatures below 5 C.

Properties involving the **fluid regulation** of mammals:

have a maximum urinary osmolality of about 2000 mOsm.
must consume a minimum of 5% of body weight in fluids daily to be at homeostasis..

All of the properties listed above are, in fact, properties related to the given aspects of mammalian biology, and they are all possessed by some but not all mammals. Subjects are told that these properties were chosen to be unfamiliar, that they are similar to, but not identical with, the actual properties to be considered in the rating task, and that the intended purpose of listing them was to be illustrate the nature of the different biological aspects referred to in the experiment.

Before beginning work on a particular booklet, subjects were required to assure the experimenter that they had some familiarity with the physical form, diet, habitat, body covering, and behavior of each of the seven

mammals appearing in that booklet. If they responded negatively, then they were presented with a second booklet and the question was repeated. This substitution process could be iterated until a satisfactory booklet was found. Only one potential subject required more than two booklet substitutions, and that candidate did not participate in the experiment.

III.2.3 Contents of the Likelihood Rating Booklets

Each subject receives a unique booklet containing 60 probability rating questions like the following:

P7 is a property related to mammals' digestion.

Given that:
lions have P7,
otters DO NOT have P7

What is the likelihood (0-100%) that tigers have P7?

Each question of this type will be referred to as an **argument**. Statements which appear in the argument above the solid line are referred to as **premises**. Premises which assert that a mammal species has a given property are referred to as **positive premises**. Premises which assert that a mammal species does not have a given property are referred to as **negative premises**. For example, the argument above has one positive premise, 'lions have P7', and one negative premise, 'otters DO NOT have P7'. The propositional form of the interrogative appearing below the solid line is refered to as the **conclusion** of the argument. The conclusion of the argument above is 'tigers have P7'. Only one property is mentioned in the premises and the conclusion of a given argument. Every argument had at least one positive premise and one negative premise. The mammal species mentioned in the conclusion is never mentioned in any of the premises. As a consequence, the truth of the conclusion or its negation is never deductively inferable from the premises. No mammal species is ever mentioned in more than one premise of a given argument. Sometimes the collection of the mammals

mentioned throughout the positive premises of a particular argument will be referred to as the **positive mammals** (of that argument). The **negative mammals** and the **conclusion mammal** are analogously defined.

All of the arguments in a particular booklet are formed using a set of seven mammals, randomly chosen for that booklet from a larger set of 47 familiar mammals. These are shown below in table I.

blue whale	gorilla	rhinoceros
bobcat	grizzly bear	seal
bison	hippopotamus	sheep
camel	horse	siamese cat
chihuahua	killer whale	skunk
chimpanzee	leopard	spider monkey
collie	lion	squirrel
deer	mole	tiger
desert rat	moose	walrus
dolphin	otter	weasel
elephant	ox	wolf
field mouse	persian cat	zebra
fox	pig	

table I. List of mammals available for experiment I, II, and III.

The limitation to seven different mammal species per booklet was motivated by the desire to permit the possibility of a within subject design in which each subject could comfortably give a set of similarity judgments for all the distinct pairs of mammals involved in the probability rating task - a similarity rating task was performed in the second experimental session of experiments II and III, and is described in a later section of the paper. Since at most seven mammal types are available to appear in a given argument, and exactly one mammal type appears in the conclusion of each argument and therefore not in any premise, every argument has at most six premises. The exact number of premises per argument ranges between two and six. Subjects were told that the particular property mentioned in each argument is unique, and they were instructed not to carry over information from one argument to the next. No two arguments appearing in a given booklet were identical.

If a pair of arguments can be made identical by taking one member of the pair and making all positive premises of that argument negative and all negative premises of that argument positive, then call that pair a **mirror pair**. The arguments P7 and P7' below exemplify a mirror pair.

P7 is a property related to mammals' digestion.

Given that:
lions have P7,
otters DO NOT have P7

What is the likelihood (0-100%) that tigers have P7?

P7' is a property related to mammals' digestion.

Given that:
otters have P7'
lions DO NOT have P7',

What is the likelihood (0-100%) that tigers have P7'?

No mirror pairs are allowed to appear together in a single booklet of arguments. A significant feature of the pseudo-random computer process that generates the argument booklets is that a priori, each member of any given mirror pair has an equal likelihood of appearing in a given argument booklet (though of course once one member of a mirror pair has been selected, the other cannot be). A consequence of this feature is that if the properties P1...P60 are to be interpreted as randomly chosen members from a class **P** of properties, then the a priori probability of any given mammal possessing a randomly chosen property from **P** is 1/2. This follows from the combination of the fact that any given mammal has an equal likelihood of appearing in either a postive or a negative premise, and the assumption/interpretation that the properties are randomly chosen. Neither this fact about the booklet generating process nor the possible sampling interpretation of the properties were discussed with any subject. The reader

is directed however, to observe the formal relevance of this feature to Similarity Postulate ii, restated below.

ii. in the absence of information other than the GIs given above -*if the unconditional probability of the* $\{o_i\}$ *are identical,* the probability of $P(o_0)$ will be a function of only the structural information in GI-4 and the set of values given by the similarity function applied to each pair of $\{o_i\}$, holding the class **P** containing P constant.

The experimental booklets given to each subject balance the conflicting demands that the arguments be random and varied on the one hand, and that they be reflective of natural experience on the other. The concern for the naturalness of the arguments reflected the belief that some arguments are actually unnatural in some sense. To see this, examine the following three arguments:

1.
Given that:
lions have P,
otters DO NOT have P,

What is the likelihood (0-100%) that tigers have P?

2.
Given that:
lions have P,
otters DO NOT have P,

What is the likelihood (0-100%) that sheep have P?

3.

Given that:

lions have P,

blue whales have P,

Siamese cats DO NOT have P,

humpback whales DO NOT have P,

What is the likelihood (0-100%) that grizzly bears have P?

Let us say that a belief is **opinionated** to the extent that it departs from a position of neutrality - a neutral belief often giving rise to a probability attribution near 50% on the (0-100) scale. For most people argument 1. gives rise to a fairly opinionated belief in its conclusion, and might typically result in a rating of, say, 75%. If the polarity of both the positive and negative premises were reversed the polarity of the belief would reverse as well, but it would still be opinionated. Argument 2. yields a relatively unopinionated belief in its conclusion, which might give rise to a rating of near 50%. Argument 3. seems confusing! It is hard to imagine any natural biological property that could fit this pattern. A reasonable strategy for a subject to adopt if faced with an argument like 3. would be to attribute a default rating of 50%. This story suggests that arguments 2. and 3. would be likely to yield relatively similar ratings (as compared with 1. and 2.) in spite of the differing psychological mechanisms that would be invoked by a typical subject considering them. To guard against such occurrences, precautions were taken in the otherwise random generation of the arguments. These precautions are described in Appendix A. A detailed description of the algorithms for the random but constrained choice of the sets of mammals and the sets of arguments is also given in Appendix A.

III.3 Experiment I - Stability of Likelihood Judgment

III.3.1 Motivation and procedure

Experiment I investigated the stability of subjects' judgments on the likelihood rating task being investigated here. This study was motivated by

the need to establish an absolute standard of performance by which models which predict subjects' judgments of likelihood could be evaluated. The plan for this experiment has subjects providing likelihood ratings for a booklet of 60 arguments during an initial experimental session, and then returning on another day and unknowingly performing essentially the same task. Twenty M.I.T. undergraduates participated in this experiment. They each attended two experimental sessions with the time interval between the sessions varying from a minimum of 1 day to a maximum of 2 weeks. Each session lasted about 40 minutes on average and 60 minutes maximum. The instructions and the likelihood booklet which a subject received in his/her initial session was exactly as described in section III.2. The instructions delivered in the second experimental session were the same as well. The likelihood booklet which a subject received in the second session differed from the specific booklet that he/she received in the first session in exactly three ways: the order of the arguments within the set of 60 was randomly permuted , the order of the positive premises (noticeably so if a given argument had more than one) was randomly permuted, and the order of the negative premises were randomly permuted as well. In all other ways, the booklet of arguments which a given subject evaluated in the second session was identical to the one which they evaluated in the first session. If a subject inquired of the relationship between the two booklets they were told that the booklets were "different".

III.3.2 Results: Part I

In this work, the primary standard used for assessing the quality of the correspondence between a set of likelihood judgments and a set of predictions for those judgments is the Pearson correlation, which was chosen primarily for its familiarity and scale-independent significance. In experiment I, the likelihood judgments given in the first session can be viewed as predictions for the judgments about the corresponding arguments given in the second session (or visa versa - equivalently). An argument in the second session is "corresponding" to one from the first if and only if it has the same premise set and conclusion. I will refer to corresponding arguments from session 1 and session two as **related pairs**. Session 1 and session 2 together provide 60

pairs of judgments about 60 related pairs. The median Pearson correlation for the 60 related pairs of judgments among the 20 subjects was 0.78, with a maximum correlation of 0.93 and a minimum of 0.50. The mean level of all the judgments was generally around 50 on the 0-100 scale for most subjects (mean across subjects: 48.8 s.d. 4.9). The mean value of a subject's judgments from the first session did not differ significantly from that of the second session for 19 of the 20 subjects (t-test, identical variances, 0.05 significance level), and the overall means, averaged across the 20 subjects, were 48.8 identically for both sessions. A histogram of the empirically observed pattern of "errors" corresponding to the difference between the rated likelihoods of session 1 and session 2, pooled together from the 20 subjects, is shown in figure 1. This distribution is, for practical purposes, symmetric around 0. Therefore attempts to model this distribution, with the ultimate goal of estimating the attainable level of performance of the optimal model, assume distributional forms that are symmetric around 0.

III.3.3 Analysis

Once the assumption of symmetry around 0 for the error distribution has been made, error distributions, both for the empirically observed data and proposed models, can be described by a one-sided profile distribution that assigns likelihood to deviations of absolute value between the judgments about related pairs in different sessions. A histogrammed version of the empirically observed error pattern, represented in this one-sided absolute value form, is shown in figure 2 (again, for the 20 subjects pooled together). Inspection of the likelihood ratings assigned by individual subjects in experiment I reveals that most subjects do not make full use of the 101 gradations of likelihood available on the 0-100 scale. Most subjects made judgments using mainly even multiples of 5 (0,5,10,15, etc.) or, even more commonly, using mainly even multiples of 10 (0,10,20, etc.) to convey their evaluations of likelihood. To accurately capture this phenomena, the observed error distribution for each subject may be histogrammed so that instead of being defined on the 201 integers between -100 and 100, the empirically observed error distribution is now defined on the 19 integers between -9 and 9, with an absolute value profile ranging between 0 and 9.

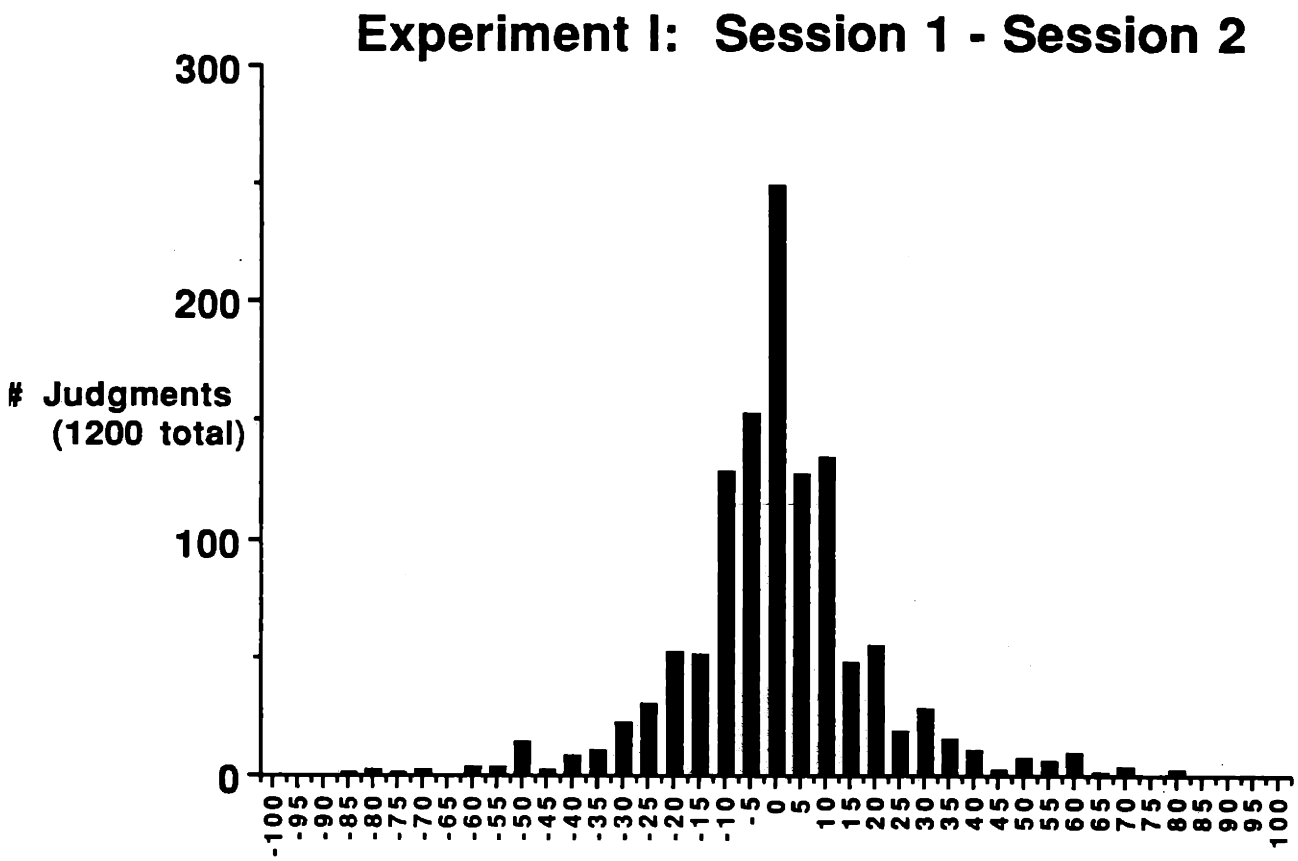

Figure 1. Histogram of observed difference between related pairs of judgments combined from the 20 subjects of experiment I

The sum of these newly defined error distributions is exactly what is shown in figure 2. Models for describing the distributional form of this error distribution were evaluated for individual subjects according to their ability to accurately describe this profile histogram of 10 values for each of the twenty subjects considered individually, after their appropriate parameters have been calibrated.

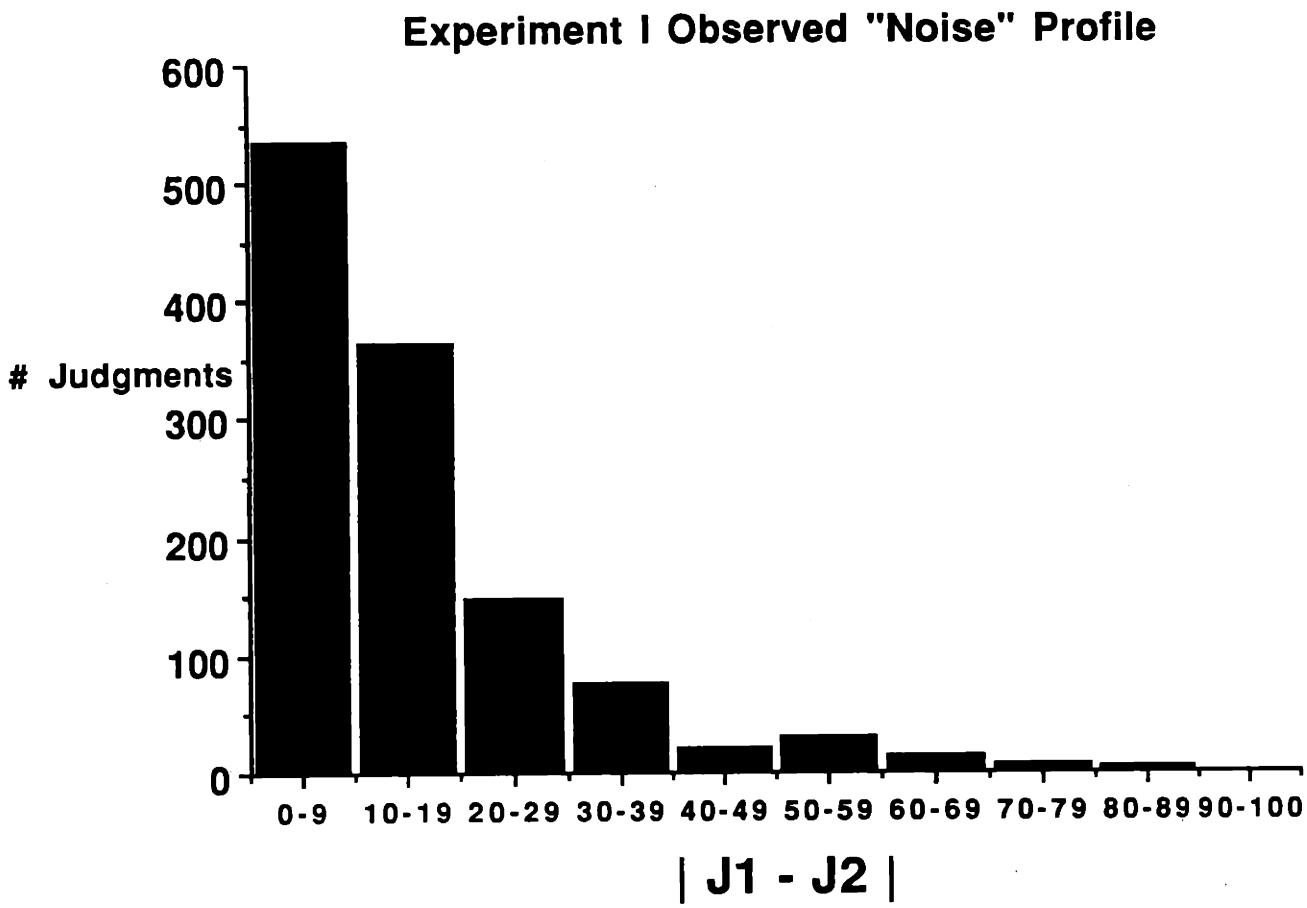

Figure 2. Observed absolute discrepancies between judgments of related pairs combined from the 20 subjects of experiment I.

The data from experiment I can be used to estimate how well an optimal descriptive model (actually optimal among the class of models that are insensitive to argument and premise order) can be expected to perform in predicting subjects' likelihood judgments for this task - the chosen standard of performance being Pearson correlation. The familiar formula for this correlation is given on line 1) (here J_i and P_i are random variables representing judgment and prediction for related pairs of arguments and J and P are random variables reflecting arbitrary judgments and predictions)

1) $E[(J_i - E[J_i])(P_i - E[P_i])] / (VAR(J) \cdot VAR(P))^{1/2}$

If we assume that optimal predictions are unbiased ($E[J_i] = E[P_i]$), then line 1) reduces to line 2).

2) $(E[(J_i P_i)] - E[J]^2) / (VAR(J) \cdot VAR(P))^{1/2}$
$= (E[J_i^2] - E[J]^2) / (VAR(J) \cdot VAR(P))^{1/2}$

One natural model for describing the "noise" in judgments for this task is to suppose that each judgment reflects the "true" or mean value rating for an argument plus some additive zero mean noise. Under such a model line 2) is equivalent to line 3) - where the variable N, independent of J and P, represents the additive noise.

3) $(VAR(J) - VAR(N)) / (VAR(J) \cdot (VAR(J) - VAR(N)))^{1/2}$
$= ((VAR(J) - VAR(N)) / VAR(J))^{1/2}$

The formula on the line immediately above represents the justification for the common rule of thumb that the square of the Pearson correlation is equal to the percent of the variance that is "explained" by a predictor. If the zero mean additive noise model were correct, then the variance of the additive noise would be equal to one half the expected square difference of two independent judgments of the same argument (by the same subject of course). This implication follows from the simple statistical facts, of line 4).

4) If N_1 and N_2 are independent and identically distributed r.v.s then
$E((N_1-N_2)^2) = (E(N_1^2) - 2E[N_1N_2] + E(N_2)^2)$
$= 2(E(N_1^2) - E(N_1)^2) = 2(E(N_2^2) - E(N_2)^2)$
$= 2Var(N_1) = 2Var(N_2)$

From this relationship it follows that for the zero mean additive noise model that VAR(N) is equal to $E[(JUDG-PRED)^2]$ which is equal to $(E[(JUDG1-JUDG2)^2])/2$. All of the above implies that under the assumption of zero mean additive noise, the best Pearson correlation that could be achieved between a typical set of judgments by a given subject (for that subjects particular set of arguments) and a deterministic model (not taking argument or premise order in to account) for making predictions about those

judgments is estimated by the formula of line 5) (where corr(J1,J2) is the correlation of the judgments of related pairs in session 1 and session 2).

5) $[(VAR(J) - VAR(N)) / VAR(J)]^{1/2}$
$= [1 - VAR(N)/VAR(J)]^{1/2}$
$= [1 - (1 - (corr(J1,J2))^2)/2)]^{1/2}$

So under this model, the median estimate of optimal model performance for this task is the result of applying this formula to the median performance estimate for the correlation between related pairs, 0.78:
$[1 - (1 - (0.78)^2)/2)]^{1/2} = 0.90$.
This would be our estimate of median optimal model performance if the assumption of zero mean additive noise held up. One way of testing this assumption in this context is described by the following procedure :

i. a particular (parametric) distributional form is chosen for the noise.
ii. the parameters of the chosen distribution are estimated independently for each subject according to the predictions that they make about that subject's judgment data; the distributions are constrained to predict the observed mean squared difference for related pairs from session 1 and session 2; subject to this constraint, they then attempt to accurately predict a histogram representation of the observed error distribution characteristics ;
iii. for each subject, the expected distribution of the difference between related pairs under the hypothesis of the fitted distribution is computed;
iv. for each subject, the fit between this expected distribution and the empirically observed distribution is tested (using a chi-sq test on histograms).
I give details of this procedure in appendix B.

Using this procedure, the null hypothesis that the noise is well described by a zero mean gaussian was rejected for 10 of the 20 subjects. Two alternatives that can be used singly or in combination were examined. One of these alternatives was a distribution that I refer to as "double-sided Poisson" that is defined on the integers and described by formula 6). Those readers familiar with the ordinary Poisson distribution will recognize the double sided Poisson as derived from an ordinary poisson distribution reflected around 0 (and re-normalized). Unlike the Gaussian distribution, the shape of the

Poisson distribution is not independent of the units used to describe the domain that it is defined on. The domain that this distribution will be defined on here is the integers between -9 and 9, corresponding to the indices of the histograms referred to above. This distributions has one free parameter, a positive real number which I will call L. Its density function is given by the formula on line 6).

$$6)\ f(k) = \begin{cases} (e^{-L} \bullet (L)^{|k|}) \,/\, (2 \bullet |k|!), & k \neq 0; \\ (e^{-L}), & k = 0; \end{cases}$$

An advantage of this distributional form (in terms of theoretical and computational simplicity) is that the distribution of its absolute value profile is Poisson. This distribution (6) is symmetric around 0 (hence 0 mean) and has a literal variance equal to $(L^2 + L)$. The interpretation of the variance relative to the 100-point scale is $100 \bullet (L^2 + L)$. The fit provided by this distributional form to the observed data proved adequate for 11 of the 20 subjects.

The second distributional innovation that was examined reflected the idea that while the noise in judgment is generally zero mean additive, there also may be some small percentage of time when, perhaps due to attentional lapses, the judgment given by a subject is entirely uncorrelated with the expected value of that subject's judgments for the particular argument being rated. Intuitively, such judgments may be thought of as corresponding to trials that would be thrown out for being too slow or inaccurate in a reaction time study. The noise distribution on these infrequent occasions will not be zero mean additive because while the judgment that will be produced is uncorrelated with the mean or mode of judgment for the given argument, the expected distance between that mean or mode and the judgment produced is not independent of the mean or mode. The expected value of this distance will depend on how "opinionated" judgments for that argument generally are. If they are generally far from the mean of judgments for other arguments, then the expected value of the distance on these occasions will be large. The mixture idea can be applied in tandem with either the gaussian or double sided poisson distributions. An additional parameter, ∂, is introduced into each model, representing the percentage of trials on which lapses occur. I

will call the gaussian, so augmented, the "gaussian mixture", and the augmented double sided poisson the "poisson mixture". To calibrate these mixture models so that their predictions about the observed error patterns may be evaluated it is necessary to calibrate two parameters in each case. As in the case without the mixture parameter however, one degree of freedom in the calibration is immediately eliminated by forcing the calibrated distribution to precisely "predict" the observed variance of the errors. The remaining degree of freedom then represents a ratio of how much of this variance is do to additive noise component, a quantity that is theoretically equal to $(1-\partial)^2$ times twice the variance of the inferred additive noise component, and how much is do to the contribution of the mixing component which is theoretically equal to $(1 - (1-\partial)^2)$ times the variance of the judgments themselves. It is the ratio of these contributions which has an impact on the estimated correlation of the optimal determinsitic model. The details of the fitting procedure itself are described in Appendix B.

The gaussian mixture was found to adequately describe the error distribution of 15 out of 20 subjects. The poisson mixture adequately described the error distribution of all 20 subjects. Of the four models considered, the poisson mixture provided the best description of the error distribution for all of the 20 subjects. The poisson mixture was therefore accepted as an appropriate description of the error data.

The assumptions that were used to derive the formulas given on lines 3) and 5) are no longer valid for the mixture models (when the mixture parameter is greater than 0.0). However, it can be shown that an augmented version of formula 3) provides a correct alternative for noise models that are mixtures of zero mean additive noise and uncorrelated random variables with the same variance as the judgments themselves. This formula is given on line 6) below, and a derivation is provided in Appendix C. Let N be here, as before, a random variable representing zero mean additive noise that is independent of judment, and let ∂ represent the probability of producing an uncorrelated judgment. Then the estimated correlation of a set of judgments and a set of optimal predictions is as follows.

7) $\mathrm{Corr}(J,P) = (1 - \partial)[\ (\mathrm{VAR}(J) - \mathrm{VAR}(N)) \ / \ \mathrm{VAR}(J)\ ]^{1/2}$

Since the variance of the double sided poisson distribution with parameter L is equal to $100 \cdot (L^2 + L)$, the estimate of line 6) relative to an inferred poisson mixture error distribution with parameters L and ∂ is given by 8).

8) $\text{Corr}(J,P) = (1 - \partial)[\ (\text{VAR}(J) - 100 \cdot (L^2 + L)) \ / \ \text{VAR}(J)\ ]^{1/2}$

III.3.4 Results: Part II

To summarize the preceding section, the predictability of each of the 20 subjects is captured by a formula involving three numbers, $(L,\partial,\text{VAR}(J))$. The estimation of these numbers is described in the previous section and in Appendix B. As mentioned above, the variance of the mixture model is constrained by the fitting procedure to exactly predict the observed variance of "error" between session 1 and 2. A formula for using the three numbers above to infer the expected correlation of judgments with the optimal model (relative to the particular set of arguments) is given by 7). The median estimated correlation of the optimal model, so inferred, was 0.88 with a maximum of 0.97 and a minimum of 0.67. The theoretically available advantage in predictive capability that a deterministic model has over these subjects' own stochastic performance can be perceived by comparing the above figures with the observed correlations of the related pairs from the two sessions - median 0.78, maximum 0.93, minimum 0.50. The median estimated mixing parameter was 7% with a minimum of 0% and a maximum of 20%. Figure 3 shows the observed profile distribution and the error distribution that was fitted to it using the double-sided Poisson model. Figure 4 compares the projected error distribution of the difference between a typical set of likelihood judgments made by each member of this population of 20 subjects and the expected values of each of those judgments with the observed discrepancies between judgments of related pairs of arguments made by those 20 subjects. The former distribution is to be interpreted as the projected error distribution of the theoretically optimal set of deterministic predictions for these arguments.

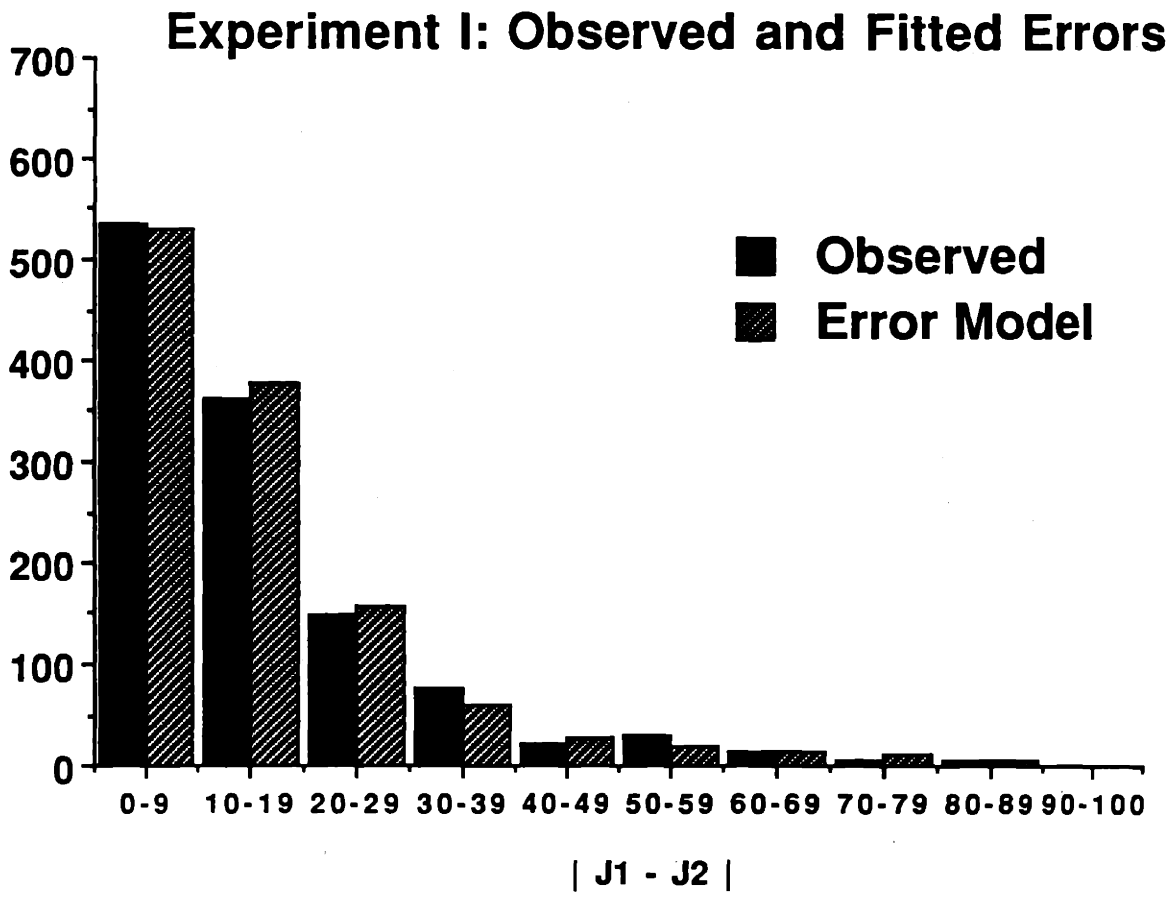

Figure 3. Comparison of observed discrepancies between judgments of related pairs for the 20 subjects of experiment I pooled together and the poisson mixture model fitted to each of these subjects individually and then pooled together.

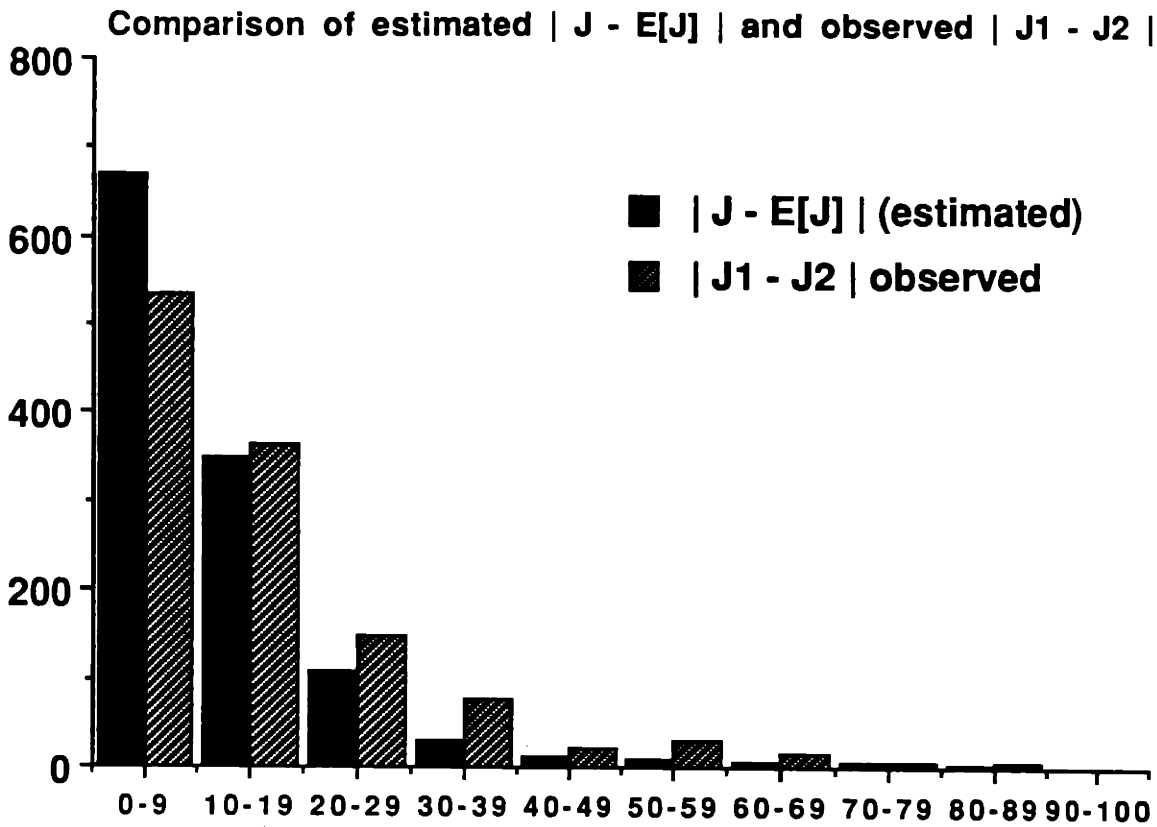

Figure 4. Comparison of the observed discrepancies and the predicted discrepancies between a set of likelihood ratings for 60 argument forms and the expected values of ratings for those argument forms.

Discussion

The decision to use the poisson mixture distribution as a model of the noise on this task should not be viewed as the advancement of a general claim about human likelihood judgments. As stated above, the primary purpose of this analysis was to acquire the ability to gauge how well a predictive model of judgment for this task (that does not take premise order into account) is performing on an absolute rather than a relative scale. When this standard of performance is measured by Pearson correlation, then the significant component of the fitted model above is the mixing parameter and not the form of the additive noise. The claim that judgment on this task does produce a percentage of judgments that are basically uncorrelated with the expected value of a given argument is a psychological claim, albeit a task

dependent one. However this hypothesis is not directly verifiable in the absence of a complete specification of a "noise" model. It will turn out however, that the evaluation of a predictive model in section III.4.5 will provide additional confirmation for the noise model presented above, because the distribution of the prediction errors of the model evaluated there will be qualitatively comparable to the form of the estimated distribution of the prediction errors of the optimal model that was estimated above.

III.4 Experiment II - Predicting Judgments Using Similarity Based Models

III.4.1 Motivation and general procedure

Experiment II provides the primary data for testing the ability to predict likelihood judgments on the task described in section III.2 using similarity based models. This data will be important to the evaluation of EQ1)-EQ3). The plan for this experiment involved two experimental sessions per subject. Subjects provide likelihood ratings for a booklet of 60 arguments during an initial experimental session, and then return on another day and provide a set of similarity ratings for each of the pairs of mammals which appeared in arguments evaluated during the first session. During the second session subjects also provided judgments about the informational relevance of knowledge of certain aspects to judgments concerning other aspects. In particular, subjects judged the relevance of the familiar aspects of mammalian biology which provided domains for similarity judgments to the unfamiliar aspects of mammalian biology that were typified by the list of properties read by subjects prior to the likelihood rating task. Twenty M.I.T. undergraduates participated in this experiment. They each attended two experimental sessions with the time interval between the sessions varying from a minimum of 1 day to a maximum of 2 weeks. Each session lasted about 40 minutes on average and 60 minutes maximum. The instructions and the likelihood booklet which a subject received in his/her initial session was exactly as described in section III.2.

III.4.2 Similarity and relevance rating - procedure

In the second session each subject was first asked to make judgments about similarity. A subject assigned a rating to all of the distinct pairs of the particular seven mammals that subject had dealt with in the probability rating task for each different type of similarity. When the order of a similarity pair is considered to be irrelevant there are 21 distinct pairs of seven mammals. Similarity pairs will be considered to be order independent throughout the experiment. An abbreviated version of the actual instructions given to subjects prior to similarity rating were as follows.

> This experiment concerns your judgments about the similarity of mammals. You will be asked to rate the relative similarity of all the distinct pairs of seven mammals with respect to a number of aspects of their biology. The aspects you will be rating include:
>
> overall biological similarity
> physical form
> body covering
> diet
> habitat
> behavior
> ancestral lineage
>
> In the booklet that you will be given there is a seperate table corresponding to each of the above aspects. In this table please record your ratings for the relative similarity of all mammal pairs for that aspect. The ratings should be given on a scale from 0 to 100, similar pairs receiving values close to 100 and dissimilar pairs receiving values close to 0.

Each subject provided a total of 147 (21 by 7) similarity judgments in all. Following the similarity rating, each subject rated the relative degree of relevance of each type of similarity, excepting overall biological similarity, to each of the unfamiliar aspects of mammalian biology dealt with in the probability rating task: bone structure, digestion, dentition, thermal regulation, fluid regulation. The instructions for this task read as follows.

> This experiment concerns your judgments about the degree to which information about some familiar aspects of mammalian

biology are relevant to making predictions about less familiar aspects of mammalian biology. The familiar aspects include:

physical form
body covering
diet
habitat
behavior
ancestral lineage

Examples of the unfamiliar aspects are given on the next page. In the attached booklet there is a separate table for entering judgments about each unfamiliar aspect. The rows of each table are labeled by the familiar aspects above. Please record in the appropriate rows your judgments about the degree to which complete knowledge about the familiar aspect of a mammal's biology would allow you to make predictions about the properties of that mammal that are related to the unfamiliar aspect. The ratings should be given on a scale from 0 to 100, with a rating of 100 standing for the judgment that knowledge of the familiar aspect allows perfect prediction, and a value of 0 standing for the judgment that information about a mammal's properties related to the familiar aspect is totally irrelevant to guessing about the properties of that mammal on the unfamiliar aspect.

A total of 30 (5 by 6) relevance judgments were required of each subject. The combination of the 147 judgments of similarity and the 30 judgments of relevance provided by each subject in the second session comprised a database to be used in modelling the probability judgments given by that subject in the first session.

III.4.3 Models for the domains of unfamiliar properties (instantiation strategies)

In section I.3 I first remarked upon the fact that, like relative frequency strategies for likelihood judgment, similarity based strategies would generally allow the possibility of being instantiated in more than one way in a given judgment context. Even though the likelihood task used in this experiment is only a semi-natural judgment context, it is already complex enough to

permit a multiplicity of potential instantiations of a similarity based strategy. One of the choices to be made by a particular instantiation is in the choice of a domain for each of the unfamiliar properties. Recall that the arguments that subjects evaluate appear as follows (italicization added).

P7 is *a property related to mammals' digestion.*

Given that:
lions have P7,
otters DO NOT have P7

What is the likelihood (0-100%) that tigers have P7?

The noun phrase that I have italicized specifies that P7 belongs to the domain of properties of mammals having to do with digestion. Of course P7 also belongs to the domain of all biological properties. The choice of similarity values will be functionally related to a subject's estimate of the likelihood that tigers have P7 in the argument above will depend on whether the domain of the similarity relations is taken to be the class of properties of mammals having to do with digestion or the class of all biological properties of mammals - or something else. The discussion in section II.3 about choosing a reference class is pertinent here. If the class of properties of mammals having to do with digestion is chosen as the similarity domain then the reference class used for inference will be the class of properties of mammals having to with digestion that lions have and otters do not have. Otherwise, if the class of all biological properties of mammals is chosen as the similarity domain then the reference class will be biological properties of mammals that lions have and otters do not. Clearly the former is the narrower reference class. If the estimates of conditional probability that can be produced relative to each reference class are of approximately the same order of accuracy, then it would be normatively preferable to use the narrower reference class. However, the subject performing this task has been pragmatically led to believe that he or she does not know very much about properties having to do with digestion. He or she has, in effect, been told that P7 is like the items on the following list of generally unfamiliar properties:

Properties involving the **digestion** of mammals:

have their omasum and abdomasum seperately articulated
process their food caecotrophically
produce the enzyme ptyalin in their salivary glands

From the point of view of descriptively modelling subject performance on this likelihood judgment task, a relevant question to be addressed is "How does a typical subject come to have subjective beliefs about the (conditional) likelihoods of mammals possessing and not possessing properties chosen from a domain about which they know, in detail, essentially nothing?"

Two different possibilities suggest themselves. One of these is that subjects simply fall back on a domain about which they do know something. In this case, this strategy would probably be realized by the choice of the class of all biological properties of mammals as the domain. If this were so, then we expect that the similarity values which a given subject would make use of when reasoning about the argument above would be approximately commensurate with that subject's judgments about the overall biological similarity of the pairs of mammals that are involved. One way of describing this first strategy is to say that subjects provisionally assume that "things" in the unfamiliar domain will stand more or less as "things" do in more familiar domains - i.e. "if lion and tiger are alike in most biological ways that I know of, it's a good bet that they will be alike in their digestion."

A second possibility is that subjects would attempt a more refined extrapolation of the knowledge that they do have to the unfamiliar domain. A subject might reason as follows, "I don't know much about digestion, but how a mammal digests seems like it would be related to what it eats, or at least more related to what it eats than to what kind of hair it has," for example. A description of this second strategy, comparable to the earlier description of the first, is that subjects provisionally assume that "things" in the unfamiliar domain will stand as "things" in the relevant familiar domains do. I attempted to capture this pattern of belief by positing that beliefs about probabilities, and hence similarities, in the unfamiliar domain could be modeled as a weighted mixture of beliefs about the relevant familiar domains. One way of thinking about this idea is as follows. It is a

mathematical fact that if one has k probability measures $pr_1, \ldots pr_k$ that are defined on the same algebra of sets, and a set of k real numbers $v_1, \ldots v_k$ that sum to 1 (i.e. $\Sigma_i v_i = 1.0, 1 \leq i \leq k$), then the "mixture" defined by $\Sigma_i v_i \bullet pr_i, 1 \leq i \leq k$ is also a probability measure on this same algebra of sets. The algebra of sets that we are interested in here, as in section II.4, is the algebra which has basic events that are patterns of of some given set of individuals, in this case species of mammals, having and not having a property. Given such an algebra and a probability distribution on the algebra, the similarity functions that were introduced in section II.2 are well defined. So if we define a probability measure on the unfamiliar domain, then these similarity functions will all be defined relative to that domain. The probability measures $pr_1, \ldots pr_k$ to be used are based on the subjective beliefs that subjects do have about the familiar domains of physical form, body covering, diet, habitat, behavior, and ancestral lineage. What this work proposes is that people do not generally represent beliefs about these probability measures that are are sufficient to totally specify them - or similarity based reasoning as it is being defined in this work would be irrelevant. What this work does propose is that people represent beliefs about conditional probabilities defined relative to these probability measures - similarities, and that these beliefs place constraints on the incompletely defined measures. Here is how those beliefs that are sufficient to define similarities can be constituted on the unfamiliar domain.

Let $\mathbf{REL(FA_k,UF_l)}$ stand for a given subject's judgment about the relevance of information about mammalian properties for familiar domain k to knowledge of mammalian properties for unfamiliar domain l, where FA_k ranges over {physical form, body covering, diet, habitat, behavior, ancestral lineage} and UF_l ranges over {bone structure, digestion, dentition, thermal regulation, fluid regulation}. The real numbers $v_1, \ldots v_k$ to be used in defining a mixture are given by $REL(FA_k,UF_l)) / (\Sigma_j REL(FA_j,UF_l))$. The theory of similarity presented in section II.2 held that the following primitives were the basis for constructing similarities:

$p_{11}(A,B,D)$ = probability of P(i) and P(j) for i in A, j in B, and P in **D**.
$p_{10}(A,B,D)$ = probability of P(i) and not P(j) for i in A, j in B, and P in **D**.
$p_{01}(A,B,D)$ = probability of not P(i) and P(j) for i in A, j in B, and P in **D**.

$p_{00}(A,B,D)$ = probability of not P(i) and not P(j) for i in A, j in B, and P in D
$= 1 - (p_{11} + p_{10} + p_{01})$

If the the probability measure on the unfamiliar domain that these quantities reference is constructed, as suggested above, from a mixture of the probability measures on the familiar domains, then these probabilities will themselves be equal in value to a mixture of the same proportions of their counterparts defined on the familiar domains.

In principle, the required quantities, p_{11}, p_{10}, and p_{00} for each pair of individuals relative to the unknown domain can be constructed by mixing, and the corresponding similarities on the unknown domain will be well defined. Our ability to determine the values of these new similarities, so defined, is more problematic because the data collected in the experiment do not give us direct access to the various p_{11}, p_{10}, and p_{00}, but only to the ratios of these quantities, and it is not true in general that the mixture of the ratios will be the ratio of the mixtures. This will be true however for similarity function d) because the function simplifies to a non-ratio form:

d) $Sim(A,B,D) = (p_{11} + p_{00}) / (p_{11} + p_{01} + p_{10} + p_{00}) = p_{11} + p_{00}$.

Recall that this function, which represents subjective beliefs about the likelihood of "matching", is pragmatically justified in informational contexts where the joint absence of a property in two individuals is of equal interest to the joint possession of a property. There are two reasons for thinking that this is an appropriate form of similarity for use in the judgment context of the likelihood judgment experiment. The first of these is that it is known that the similarities of the conclusion mammal to both positive and negative mammals play a role in judgment. This is confirmed by both the analysis below and by the related experiments in Osherson '91. If this role is in accord with Similarity Postulate i. it follows that an increase in this similarity implies an increase in the probability of the two mammals both failing to possess a particular property. This thesis is, in part, an argument to the effect that this belief should be interpreted as arising out of a more general belief in the probability of the two individuals failing to possess a generic kind of

property (generic relative to some domain, at least). Under the given interpretation of similarity as a belief about this probability, this implication and its converse are both true since they are judgments about the same thing. The second reason for thinking that d) is an appropriate form of similarity is that it is symmetric in both the order of its arguments and the polarity of properties. It was pointed out above in section III.2.3 that the random construction of the likelihood booklets by computer was done in such a way as to make the empirical distribution of mammals, as they appear in the positive and negative premises of the arguments, statistically consistent with these symmetries.

The ideas just expressed argue for why it would be appropriate to use similarity function d) in the likelihood judgment task. They do not argue that a subject's similarity judgments about pairs of mammals, delivered on a later day, should correspond to formula d), rather than c) or e) for instance. Nevertheless, I shall make use of the following model for describing the second similarity instantiation strategy.

Let $\mathbf{SIM(i,j,UF_l)}$ stand for that subject's estimate of the similarity of mammal type i to mammal type j relative to unfamiliar domain UF_l and $\mathbf{SIM(i,j,FA_k)}$ stand for that subject's estimate of the similarity of mammal type i to mammal type j relative to the familiar domain FA_k. Assume that the range of each of the boldface variables is normalized to lie between 0.0 and 1.0. A formula for expressing similarity under the weighted mixture scheme and the assumption of similarity model d) is the following:

$$SIM(i,j,UF_l) = (\Sigma_k SIM(i,j,FA_k) \bullet REL(FA_k,UF_l)) \; / \; (\Sigma_k \, REL(FA_k,UF_l))$$

In the future these will be referred to as **mixture similarities**.

Remarks on choices of similarity:
In the immediately preceding paragraphs I have described the motivation for the two models which each attempt to describe the values of subjects' estimated similarities for the unfamiliar domains. One of these models uses the values of overall biological similarity. The other uses mixture similarities. Other formulas for computing similarities that attempted to

capture the logic of the second domain instantiation strategy were tried, but none were generally as useful in the describtion of observed subject performance as mixture similarities, and so I will not bore the reader with their description and more tenuous motivation for their precise algebraic form. As things turned out, the actual amount of contrast among the two similarity functions that were used was disappointing. This was due to the fact that there were generally strong correlations among the judgments of relevance for the different unfamiliar domains as well as strong correlations in values of the similarities of two individual mammals rated across the different familiar domains. As a consequence of this, differences in the descriptive performance of the two choices of similarity strategy were never large. Some statistically significant patterns of interest did emerge though which will be described below.

III.4.4 Results Part I: Evaluation of EQ1 (and a domain)

EQ 1) is concerned the descriptive validity of Similarity Postulate i. Similarity Postulate i., or its minor variant Similarity Postulate i.', seem to be common to almost all similarity based strategies for inductive reasoning. The disconfirmation of i. as an accurate description of human performance on the likelihood rating task would show that these reasoning strategies, as a group, are on the wrong track. What this postulate predicts, applied to this context, is that the perceived likelihood of the conclusion statement of a given argument will be an increasing function of the each of the positive similarities and a decreasing function of each of the negative similarities. So for example, if we assume that the similarity of lions to tigers (for the relevant domain) is greater than the similarity of siamese cats to tigers (for this domain) then the judged likelihood of argument 1) below should be greater than the judged likelihood of argument 2).

1)
Given that:
lions have P*,
otters DO NOT have P*,

What is the likelihood (0-100%) that tigers have P*?

2)
Given that:
siamese cats have P*,
otters DO NOT have P*,

What is the likelihood (0-100%) that tigers have P*?

Although we might expect that the judged likelihood of 1) would also be greater than the judged likelihood of 3), we need more information to determine whether Similarity Postulate i. makes a prediction about this.

3)
Given that:
siamese cats have P*,
raccoons DO NOT have P*,

What is the likelihood (0-100%) that tigers have P*?

Similarity Postulate i only makes direct predictions about pairs of arguments where one member of the pair "dominates" the other. To be precise, $argument_1$ is said to **dominate** $argument_2$ relative to a domain, or kind of similarity, if and only if:

i. both arguments have the same number of postive and negative premises;
ii. if we rank ordered the positive similarities (for the relevant domain) for each argument from largest to smallest, the similarity of the kth postive similarity of $argument_1$ is greater than or equal to the kth positive similarity of $argument_2$ for every k between 1 and the number of postive premises;
ii. if we rank ordered the negative similarities (for the relevant domain) for each argument from largest to smallest, the similarity of the kth negative similarity of $argument_1$ is smaller than or equal to the kth negative similarity of $argument_2$ for every k between 1 and the number of negative premises.

We can say that argument$_1$ **strictly dominates** argument$_2$ if argument$_1$ dominates argument$_2$ and at least one of the comparisons of similarity mentioned in ii. and iii. above is a strict inequality. Relative to a fixed domain, **D**, it will also be convenient to refer to the set of values S = {Sim(m,c,D) | m is a postive (negative) mammal and c is the conclusion mammal} as the set of **positive (negative) similarities.**

The data collected for experiment II allow the evaluation of Similarity Postulate i. relative to a set of similarity functions. An evaluation of Postulate i relative to overall biological similarities and mixture similarities was conducted as follows: for each of the two sets of similarities in turn, each subject's argument booklet was searched (by computer) for pairs arguments such that one argument strictly dominated the other, according to that similarity set; all pairs of arguments so identified were then labeled as confirming or disconfirming Postulate i, relative to the chosen similarity, according to whether the member of the pair which strictly dominates was accorded a higher likelihood rating; finally, the percentage of confirming pairs among all pairs satisfying the strictly dominating relation relative to the domain was computed.

Since twenty subjects each evaluated 60 randomly generated arguments for this experiment, there were 20•("60 choose 2") = 35,400 distinct pairs of arguments involved in the experiment. Contingent on the judgments of overall biological similarity given by the individual subjects, 1,808 of these pairs turned out to satisfy the strictly dominating relation relative to this function. Of these, 1,415 (= 78.3%) were confirming instances for Similarity Postulate i. Contingent on the judgments of aspectual similarity and relevance given by the subjects that were converted into mixture similarities, 254 pairs turned out to satisfy the strictly dominating relation. Of these, 181 (= 71.3%) were confirming instances for Similarity Postulate i. Both ratios (1415/1808 and 181/254) are highly significant in their departure from chance ($p < .0001$). The difference between these percentages was also significant by a chi-square analysis ($p < .02$).

Clearly both of the above results indicate that Similarity Postulate i. is correct most of the time. Since subjects' judgments on the likelihood task are noisy,

it is not immediately clear what kind of results we should interpret as supporting the view that Similarity Postulate i is correct in general. In fact, we must define what it would mean for it to be correct in general. The most natural definition would seem to be that Similarity Postulate i. is correct in general, relative to a similarity model, if the percentage of confirming instances for the postulate, determined by the same algorithm as above, would be 100% when the likelihood ratings for each argument were replaced by the expected value of those ratings, and the same for similarities.

In an effort to partially gauge the effect of "noise" on the confirmation of SImilarity Postulate i., I attempted to evaluate the expected frequency with which the likelihood ratings given to a pair of arguments will be reversed in magnitude from the mean values of the likelihoods accorded to those arguments. Since there is noise in the judgments, sometimes the conclusion of one argument will be judged more likely than the conclusion of another argument even though, on average, the conclusion of the second is judged more likely than the first. This percentage of reversals was estimated using the data from experiment I as follows.

Recall that in the terminology of experiment I, a pair of arguments, one from session I and one from session II, were called **related** if they only differed in the order of their premises. Now I will call a pair of arguments from session I of experiment I together with a pair of arguments from session II of experiment I **corresponding** just in case the pair consisting of the first members of each of these new pairs are a related pair and the pair consisting of the second members of each of these new pairs are a related pair. Also let A_i equal the set of distinct pairs of arguments from experiment I, session I of subject i for which the likelihood rating given to the first member of the pair is greater than the likelihood rating given to the second member of the pair. Now let B_i equal the set of distinct pairs of arguments from session II of subject i such that the corresponding pair of arguments from session I were members of A_i and the likelihood ratings given to the arguments in session II are oppositely ordered in magnitude from the order of the corresponding pairs in session I. Let the variable p stand for the ratio of the sum of the cardinality of the sets B_i, where i ranges over the 20 subjects participating in experiment I, divided by the sum of the cardinality of the sets A_i. In other

words, p is the percentage of corresponding pairs which reversed their order from the first to the second session. The value of p for experiment I was 16%, and so (1-p) = 84% of the corresponding pairs retained their ordering.

In the analysis of experiment I it was pointed out that the comparison of judgments from session I and session II essentially measures the results of two independent occasions on which corrupting noise was added: the first time judgments were given and the second time judgments were given. Looked upon from this point of view, we expect that the figure of 84% measures the percentage of pairs which "refused" two chances to switch plus the percentage of pairs which (ala Cosi Fan Tutti) switched twice. We can solve for the percentage of pairs which switch each time, call this p*, according to the equation

1) $(1 - p^*)^2 + (p^*)^2 = .84$

This equation has two possible solutions, $p^* = .913$, and $p^* = (1-.913) = .087$. From context however it is clear that the former estimate is the desired one. The figure of 91.3% is the estimated percentage of time that the rank order of a subject's ratings of a pair of arguments reflect the rank order of the expected value of those ratings.

Returning to the consideration of the statistical correctness of Similarity Postulate i, the expected number of confirming instances that would be observed is equal to the percentage of time that the postulate is really correct, call this pc, times the percentage of time that a given subject's judgments relative to the potentially confirming pair of arguments did not "switch", plus the percentage of time that the postulate was really incorrect, (1-pc), times the percentage of time that the subject's judgments did switch. The percentage pc can therefore be solved for by the following equation.

2) $(1-p^*)(pc) + p^*(1-pc)$ = the observed percentage of confirming pairs.

Substituting in the estimate of .913 for p* and .783 (the percentage of confirming pairs for overall similarity) for the observed percentage of confirming pairs, we arrive at an estimate of 84.3% for the percentage of time

that Similarity Postulate i is correct. This figure is based on the assumptions that overall biological similarity is the correct choice of similarity function to pick out the dominating pairs and that the subject's judgments about overall biological similarity are "noiseless" in the sense of picking out the dominating pairs with perfect accuracy. So the figure of 84.3% correctness is a conservative estimate for the true performance of Similarity Postulate i, although from the currently available data we are unable to estimate how conservative this estimate is.

III.4.5 Results Part II: Evaluation of EQ2 and EQ3

EQ2 and EQ3 essentially ask how well the subjects' likelihood judgments for these arguments can be predicted using only information describing the form of the arguments and the similarity values of the mammals involved, and whether the form of the model generating the best predictions is that described in section II.4. and II.5. The reader will forgive me for ruining the suspense if I announce at this juncture that the model described in section II.4 and II.5 does admirably describe subject performance both in the comparative and the absolute senses. The order I will take things in will be to work towards describing the data analysis which establishes the predictive capability of the model in II.4 by first describing the evaluation of a class of simple models. The relative predictive capabilities of these simple models are interesting in their own right, in part because they are so simple, and in part because they touch base with a number of previous models proposed in the field.

III.4.5.1 Simple Models

I will now consider a variety of simple models for predicting the probabilities assigned by individual subjects. It will be helpful in proceeding to define a set of data structures that models may make use of. Recall that probability ratings are assigned to the conclusion statements of particular arguments.

An **argument form** is a data structure capable of summarizing all of the information about any individual argument pertinent to the class of models to be considered here - other than the actual rating given to the argument. An argument form consists of the following information:

a property aspect - one of bone structure, digestion, dentition, thermal regulation, fluid regulation;
a conclusion mammal - the mammal mentioned in the conclusion;
a set of positive mammals - those mammals known to possess the property;
and a set of negative mammals - those mammals known not to possess the property.

A **similarity table set** is a set of five tables, one for each property aspect, specifying what a particular model takes to be the similarity of a particular pair of mammals relative to a particular property aspect. For instance, a similarity table set might have a value of '90' for the similarity of lion and bobcat with respect to dentition and a similarity of '50' for that pair of mammals in the respect of their thermal regulation. Note that these values cannot be literally identical with the ratings given by a subject, but are in some sense inferred by a model, though the extent of the inference may be something trivial like merely equating the value of similarity of digestion with the value of overall biological similarity - which was rated. I will use notation of the form sim(otter,grizzly,digestion) to represent the similarity of otter and grizzly for digestion.

A calibrated model for this task is a function from an argument form and a similarity table set to real numbers between 0.0 and 1.0. The probability ratings themselves are normalized, after dividing them by 100.0, to lie between 0.0 and 1.0 as well. Models are evaluated by their success in predicting the probability ratings. An important feature of the probability rating task is that subjects are essentially judging the relative likelihoods of exactly two mutually exclusive alternatives: the conclusion mammal either has the property or it does not. The resulting likelihood judgments may be viewed as the result of weighing the balance of evidence supporting each of the two alternatives. A natural class of simple similarity models is obtained by adding to this notion the idea that support in favor of the conclusion

mammal having the unknown property comes from the set of similarities of the conclusion mammal to the positive premise mammals, and support in opposition to having the property comes from the set of similarities of the conclusion mammal to the negative premise mammals. This idea forms a common thread running through a long history of psychological models describing choice and categorization. See Nofosky 1990 for a review and analysis of many aspects of this line of research. The following definitions will also be useful in specifying models inspired by this idea:

Call **POS**(arg) the set of pairwise similarities generated from a particular argument form, arg, and a similarity table set by taking all available instances of sim(conclusion,mam_i,property) where property is the property aspect of arg, conclusion is the conclusion mammal of arg, and mam_i ranges over the positive mammals of arg .

Let **NEG**(arg) be the set of pairwise similarities generated symmetrically from a particular argument form by the considering the set of negative premises.

I will abbreviate these sets by POS and NEG where unambiguous context allows. The following abbreviations will stand for familiar functions applying to sets of real numbers:

CARD() - the function that returns the cardinality of a finite set of numbers;
SUM() - the function that returns the sum of a finite set of numbers;
MAX() - the function that returns the maximum value in a finite set;
MIN() - the function that returns the minimum value in a finite set;
AVG() - the function that returns the arithmetic mean of a finite set.

Here are some illustrative examples of calibrated similarity models.

Model example1 is an obtuse model that considers only the number of positive and negative premises, ignoring similarity values - this might be called an "urn model".

model 1: CARD(POS) / (CARD(POS) + CARD(NEG))

Example 2 is more complicated. This model forms a linear sum from the positive and negative similarities, and then passes this sum through a sigmoidal function of the type commonly used in neural networks and other applications. Note that as the linear sum in brackets that is being exponentiated grows more negative, the whole expression approaches 1.0, and as the linear sum grows more positive, the whole expression approaches 0.0. Here EXP[X] stands for e^X.

model 2: 1.0 / (1.0 + EXP[$\underline{-0.05}$ + $\underline{-0.5}$•SUM(POS) + $\underline{0.6}$•SUM(NEG)])

The expression given as model 2 is well defined, but the choice of the underlined numbers, (-0.05, -0.5, 0.6), is not demonstrably motivated. It is clear that these numbers may need to vary in magnitude in order to accurately describe the patterns of judgment which reflect quantitative relations between judgments of similarity and of probability that are given on arbitrary scales (why should a similarity value range between 0 and 100 for example?). This problem is treated in the modelling procedure by letting these numbers start out as free parameters to be calibrated by some computational procedure in accord with a well defined mathematical specification.

III.4.5.2 Calibration and Assessment of Models

For a statistically valid evaluation of models containing calibrated parameters it is necessary that such models be evaluated with respect to the quality of their predictions rather than the quality of the accord between the data and the parameterized fits that are obtainable. In keeping with this principle, models were evaluated using a statistical technique known as jacknifing or cross-validation. See Efron '82 for theoretical details of these type of analysis. The procedure for assessing model performance was as follows.

One argument is selected from the 60 rated by a given subject. Call this chosen argument the "target". The model under consideration is then calibrated by a computational procedure that has access to the similarity and relevance judgments given by that subject as well as the other 59 arguments and the ratings assigned by the subject to those arguments. Once calibrated, the model is used to make a prediction for the likelihood rating given to the conclusion of the target argument. This prediction is then stored for later use and a different target is selected. Repeating this procedure 60 times, cycling through the complete set of rated arguments, eventually results in the generation of a complete vector of 60 predictions - one for each argument. The model is then assessed a score for that subject as a function of the match between this vector of 60 predictions and the actual ratings given by the subject. A few different functions were considered for scoring this match including the sum of the absolute differences between the predictions and actual judgments, the sum of the squared differences between the predictions and the actual judgments, and the Pearson correlation coefficient between the prediction vector and the judgment vector. It was found that these measures only rarely disagreed concerning which of two models provided a better fit for a given subject, and then only when both discrepancies were tiny. For reasons of familiarity the correlation coefficient was finally adopted as the measure of choice.

Successful models are those that achieve relatively high correlations for most subjects. Specifically, the median correlation across the group of twenty subjects was taken to be an overall summary of a model's performance. One model is judged to be significantly better than another is it achieves higher correlations for 15 or more of the twenty subjects. This corresponds to significance at the .05 level for a sign test with N = 20.

There are two important components to the procedure that fixes the free parameters in a model so that the model can be used to make a prediction. These components are the fitting criteria and the computational procedure used to find the "pseudo-optimal" values of the free parameters best satisfying the criteria. The term pseudo-optimal acknowledges the reality that for models in which the predicted values generated by the model depend on the parameters being fitted in a non-linear way, such as in model 2 above,

there is no sure-fire procedure for finding values which provide globally optimal fits for any reasonable fitting criteria. The computational procedure must therefore be thought of as a technique for approximating a given fitting criteria. The fitting criteria used in the modelling discussed in this section was the minimization of the sum of the absolute values of the differences between the 59 available subject ratings and current model predictions. Given the correctness of the noise model established in experiment I, minimizing the sum of the absolute values is a more efficient form of parameter estimation than the more familiar procedure of minimizing the sum squared discrepancy. The computational procedure used was the non-linear simplex minimization procedure found contained in Press et al. 88'.

II.4.5.3 Taxonomy of Simple Models

The models dealt with in this section vary in three different ways, the choice of a similarity function used to generate the similarity table set, the choice of a support function used to generate positive and negative evidence from the positive and negative exemplars, and the choice of scaling function that transforms the positive and negative evidence into a "probability" value. The range of choices that were used to construct simple models will be described below.

Similarity Function

The two choices of similarity function used were essentially those described in the evaluation of Similarity Postulate i:

1) a given subject's ratings of overall biological similarity; the abbreviation OVER(m_1,m_2) will stand for a given subject's rating of the overall biological similarity of m_1 and m_2;

2) the "mixture" similarities described above*; MIX(m_1,m_2,uf_i) will stand for a given subject's "mixture" similarity for m_1, m_2 and unfamiliar aspect i.

* The only additional modification to these similarity functions was that the "mixture" similarities were modified so that the mean and standard deviations of each set of 21 similarities (21 = 7 "choose" 2, the number of distinct pairs of mammals) corresponding to the similarities for unfamiliar property i (i.e. bone structure, digestion, etc.) were identical.

Support Function

These functions are applied separately to POS and NEG to generate weights of positive and negative evidence, abbreviated as EVPOS and EVNEG. The functions considered here, abbreviated as above, are MAX, MIN, AVG, SUM, and CARD.

EVPOS = MAX(POS), or
MIN(POS), or
AVG(POS), or
SUM(POS), or
CARD(POS).

EVNEG = MAX(NEG), or
MIN(NEG), or
AVG(NEG), or
SUM(NEG), or
CARD(NEG),

Scaling Function

These functions combine the two numbers, EVPOS, and EVNEG, into a probability between 0.0 and 1.0. The following functions are considered:

linear threshold - if (c1 + c2•EVPOS - c3•EVNEG) > 1.0, return 1.0
else if (c1 + c2•EVPOS - c3•EVNEG) < 0.0, return 0.0
else return (c1 + c2•EVPOS - c3•EVNEG)

sigmoidal - $1.0 / (1.0 + EXP[(c1 - c2 \bullet EVPOS + c3 \bullet EVNEG)])$

quotient - $(c1 + EVPOS) / (c2 + EVPOS + c3 \bullet EVNEG)$

c1,c2, and c3 are always positive constants.

Any combination of support function and scaling function gives a model that obeys Similarity Postulate i'. To see this, first note that for every choice of support function, an increase in the similarity of a mammal mentioned in a positive premise and the conclusion mammal can sometimes increase and never decrease EVPOS, and an increase in the similarity of a mammal mentioned in a negative premise to the conclusion can sometimes increase and never decrease EVNEG. Second, note that every scaling function is increasing in EVPOS and decreasing in EVNEG. Agreement with Similarity Postulate i' follows from these facts.

One reason these simple models are of interest is because they describe or generalize a number of well known models of categorization, choice and judgment. For example the models described by the use of the SUM support function and the Ratio scaling function (for some similarity function) are a generalization of Medin & Schaffer's Context Model [Medin78]. Nofosky90 contains an extensive discussion of the relationship between this model, models of choice proposed by Luce, and models for stimulus generalization and categorization studied classically by Shephard and more recently by Nofosky. The model described by the MAX support function and the Linear scaling function was proposed and tested on a closely related experimental task in Osherson91. Models employing the MIN support function and the MIX similarity function are related to explanation based and causal attribution models in the following sense. The MIX similarity function focuses the weight of similarity on the known aspects that are considered "relevant" to the unknown property class. So for example, if the unknown property is known to deal with thermal regulation then the familiar aspect body covering might be judged to be particularly relevant. Explanation based strategies, if they apply to this task at all, make predictions about the likelihood of a mammal having the unknown property according to the "explanation" of the observed pattern of mammals having and not having a

property. A common idea about what constitutes an explanation is a factor that is true of all of the positive cases and false of all of the negative cases. If such an explanation must focus on a single property, then the questions asked of the subject in the current task would seem to permit too many possibilities for this to be a tenable strategy. However, if the factor sought by an explanation could be something like "alike in body covering" then we would expect MIN MIX to be able to capture this because for the explanation to predict that the conclusion of the argument is likely it would be required that the conclusion mammal share this factor, which in this case would be the relevant similarity, with all of the positive premises and none of the negative premises. If this were true then the MIN MIX model would make EVPOS large and EVNEG small, and the prediction would follow. The procedure by which the arguments in the likelihood booklets were generated, as described in Appendix A, was designed to produce a significant number of arguments in which such an "explaining" factor could be found. As noted in section II.2, the main desideratum in the generation of the likelihood booklets was to avoid "unnatural" arguments. In essence, the procedure that was adopted was motivated by the idea that the availability of an "explaining" factor in a given argument (according to my intuitions about similarities and unfamiliar property domains) was a sufficient condition for guarding against the unwanted arguments.

III.4.5.4 Results: Simple Model Performance

The specification of a primitive similarity function, a support function, and a scaling function taken together (along with a calibration procedure) define a simple model. The table below summarizes the empirically observed predictive capabilities of the simple models. The table also includes for comparison a restatement of the observed median correlation of the two sessions from experiment I and the predicted optimal model performance, determined as described in section II.3.1. Significant differences between models listed in the table are indicated by a gap of a line.

Those models that used similarity in some intrinsic way (either OVER or MIX) did significantly better than those that did not. Those models are

represented by the choice of "CARD" as support function, and of course the constant function which achieves 0.0 correlation. Models using "MIX" for the primitive similarity function did slightly better than those using "OVER", although this difference was rarely significant. As noted above, comparison of the similarity table sets generated by using OVER and MIX confirmed that these functions were highly correlated. The support functions MAX, SUM, and AVG all did well, but no one function stood out above the others. Both sigmoid and linear threshold scaling functions did well. Again neither stood out. All models using ratio scaling were significantly worse than the most successful strata of models.

Support	Similarity	Scaling	Median Correlation
Estimate of the optimal model			.88
Observed correlation of Experiment I			.78
MAX	MIX	Linear	.76
MAX	MIX	Sigmoid	.76
MAX	OVER	Linear	.75
SUM/AVG	MIX	Linear	.75
AVG	OVER	Linear	.74
MAX	OVER	Sigmoid	.73
SUM	OVER	Linear	.72
MAX	OVER	Ratio	.69
SUM	OVER	Ratio	.67
MIN	OVER	Linear	.65
MIN	MIX	Linear	.34
Card	-----	Linear	.24
Card	-----	Ratio	.17
Constant	-----	------	.00

Table II. Descriptive Performance of Simple Models

II.4.5.5 Evaluating the Maximum Entropy Estimator

In this section I describe the evaluation of the maximum entropy estimator proposed in section II.4 for estimating likelihoods on the basis of information contained in similarities. First I will describe how the estimator that was described in generality, so necessarily abstractly, in section II.4 is applied as a model for predicting judgments on the likelihood rating task of experiment I and II.

The estimator of section II.4 is a function from an instantiation of a similarity based likelihood scenario to a judgment of probability. A precise syntactic description of the instantiation process was given in section II.4 and was exemplified using the story of "Bill" and the pasta. An example using one of the likelihood rating task arguments will help to clarify how the choices to be made in the instantiation process apply in the current context. Consider the following argument.

P7 is a property related to mammals' digestion.

Given that:
lions have P7
grizzly bears have P7
otters DO NOT have P7

What is the likelihood (0-100%) that tigers have P7?

The proposition that a subject is asked to assign a likelihood to is the proposition 'tigers have P7'. The description of similarity based likelihood offered in section II.3 holds that this is to be accomplished by the process of 1) choosing an appropriate conditional conditional probability to evaluate and then 2) evaluating this conditional probability on the basis of the available evidence. The choice of a conditional probability in step 1) is called an instantiation. For a similarity based strategy, the choice of this conditional

probability is equivalent to the choice of particular facts with which to bind the variables used in the description of the GIs. I will now describe possible choices for the GIs with reference to the argument given above.

GI-1: P is known to be a property from the class **P** - in this case P refers to P7. Possible choices of property domains that P7 is known to belong to include the class of properties having to do with the digestion of mammals (which some mammals possess and some do not) and the class of properties having to do with the biology of mammals (which some mammals possess and some do not).

GI-2: i_0 (the conclusion mammal) is a member of a finite set $\{i_k, 0 \leq k \leq m\}$ of related individuals (other mammals)- this fact relates to the choice of contrast class of similarity. In this case the contrast class could be considered to be either the seven mammals involved in the likelihood arguments for a given subject or all the mammals that the reasoner (the subject) is familiar with.

GI-3: the reasoner has beliefs about what proportion of the time each distinct pair of individuals in the set $\{i_k, 0 \leq k \leq m\}$, say i_j and i_l, had matching and non-matching values for properties in the class **P** - i.e. the reasoner has beliefs about the relative likelihoods of $P(i_j)\&P(i_l)$ vs. $\neg P(i_j)\&P(i_l)$ vs. $P(i_j)\&\neg P(i_l)$ vs. $\neg P(i_j)\&\neg P(i_l)$ for random P in **P** - in section II.2 a theory of similarity was presented which identified estimates of similarity with beliefs of this type.

GI-4: $i_2 \ldots i_n$ in $\{i_k, 0 \leq k \leq n\}$ are known by the reasoner to have P and $i_{n+1} \ldots i_m$ in $\{i_k, n+1 \leq k \leq m\}$ are known by the reasoner not to have P - which individuals are chosen to make up the set $\{i_k, 0 \leq k \leq n\}$ of "positive" cases and the set $\{i_k, 0 \leq k \leq n\}$ of "negative" cases is obviously constrained by the reasoner's knowledge. Beyond this however, the choice to be made is part of the instantiation strategy. For the properties appearing in the likelihood rating arguments, the only individuals known to the subject to be positive cases are those mammals appearing in the positive premises. However logically, there are countless individuals that could be chosen to appear as negative premises: basketballs, lasagna recipes, etc. - none of these individuals possess any biological property having to do with mammalian digestion. Clearly, in some sense, they are not chosen because they are not informationally relevant.

Ideally, we would like to have a precise syntactic definition of how "informationally relevant" is determined. One of the nice properties of the maximum entropy estimator however, is that even if such a totally irrelevant premise was chosen as a negative case, in theory it would only cause a waste of computational resources, rather than a distorted judgment. The reason for this will be described shortly.

The two principal choices to be made regarding the instantiation of the argument above are the property domain and the set of positive and negative cases. The maximum entropy estimator was tested in its ability to predict likelihood judgments for these arguments using an instantiation strategy in which the choice of property domain was the class of properties having to do with the biology of mammals (which some mammals possess and some do not) and the choice of positive and negative cases were exactly the positive and negative premises of each argument. These choices of instantiation can be partly justified on independent grounds and were partly required by pragmatic necessities related to issues of parameter fitting. The justification for the choice of the class of properties having to do with the biology of mammals as the property domain was the success which this choice enjoyed relative to confirming Similarity Postulate i. and in the evaluation of the simple models above. The justification for using some of the positive and negative cases contained in the premises of the argument is that this is obviously the most relevant information the subject has available, and in the case of the positive cases, the only information the subject has available of the required type. One justification for using all of these cases is that they are obviously salient in the subject's mind.

Once the choice of an instantiation of the GIs has been made, the "job" of the estimator is determined. For the sample argument above, that job is to provide an estimate for the conditional probability that tigers have a randomly chosen biological property given that lions have this property, grizzly bears have this property, and otters do not have this property. The estimator, in accord with Similarity Postulate ii., is to produce this estimate using the information described by GI-3 and GI-4, which in this case amounts to using the pairwise overall biological similarities and the description of the argument form. Because of pragmatic consideration, which I shall describe

momentarily, the version of the estimator that will be tested here is the second, alternative version, in which some other similarities are used. The form of the maximum entropy estimator applied here produces an estimate of this conditional probability based on the description of the argument form and the pairwise similarities of all seven mammals contained in a given subject's likelihood booklet, whether or not all seven mammals appear in the particular argument being evaluated. This variation, while motivated by expediency, is also of considerable interest in its own right. The variation is a natural one in the sense that it is reasonable to suppose that the set of seven mammals which the subject has been told of and which appear repeatedly throughout the different arguments are at the forefront of the subject's mind. One of the attractive properties of the maximum entropy estimator is that it can make use of any information that constrains the probability distribution being estimated. The actual numerical differences between the two ways of producing estimates (use of just the similarities mentioned in each argument vs. use of similarities between all 7 mammals) would usually be slight in any case.

It is convenient to describe the maximum entropy estimator as an algorithm which proceeds in the following two steps:

a) A probability measure is estimated which assigns a likelihood to each of the 2^7 basic events corresponding to all the possible conjunctions of the seven mammals in the argument set having and not having a randomly chosen biological property. This probability measure is to be the maximum entropy distribution compatible with the estimated overall biological similarities of 21 distinct pairs of the seven mammals - but see below.

b) A likelihood judgment for each argument is produced by interpreting the argument as a request for the value of the conditional probability that mammal mentioned in the conclusion has a randomly chosen biological property given the pattern of mammals having and not having this property that is described in the premises of the argument.

These steps should not be taken as a description of a processing algorithm which people use. A discussion of plausible processing algorithms is given in

section IV of the thesis. Using steps a) and b) to generate a prediction of likelihood for each argument would be straightforward were it not for the fact that the similarity values that are immediately available, the judgments of overall biological similarity given by the subjects in session II of experiment II, are at least three steps away from the similarity values that are required by the model. These "three steps" reflect the following considerations.

i) Both the judgments of likelihood and the judgments of similarity given by the subjects are placed on arbitrary scales of magnitude. At best these scales are related to the conventional [0,1] probability scale by a constant multiplicative factor. It is a reasonable possibility however that they may be translated in non-linear ways. There is actually a precedent for such translations: in D. Kahneman and A. Tversky's Prospect Theory model of choice in lotteries Kahneman et al. '79. In this model, verbally communicated probabilities are scaled by a non-linear transformation in order to be commensurate with the role assigned to them by a theory of utility.

ii) The estimates of similarity may themselves be "noisy" - i.e. a given judgment may be different from the expected value of an independent sequence of productions of that judgment.

iii) In section II.2, a number of slightly different models for relating similarities to probabilities were proposed. It was argued that the exact form of this relationship could be determined by pragmatic factors. In the discussion of similarity instantiations in section III.4.3, some reasons for thinking that similarity model d) would be pragmatically appropriate to the likelihood judgment task were discussed. It was also pointed out that there would be no particular reason to think that subject's would be pragmatically motivated to use similarity model d) when producing judgments of similarity on a different day, in experiment II session 2. Fortunately, model d), the symmetric form of model c), and model e) are monotonically related to one another, so problem 3) is, in some sense, subsumed by problem i) from the point of view of data analysis.

The computational strategy that was adopted for surmounting these problems was as follows. The model for generating likelihood predictions was allowed to have 25 "free" parameters and one meta-parameter. As in the evaluation of the simple models, before making a prediction about the likelihood rating of a given argument, the free parameters were to be fixed by calibrating them according to which combination of parameters best fit the remaining data not including the rating for the argument to be predicted. This was done in turn for each of the 60 arguments. It is heuristically well known that one cannot generally calibrate 25 truly "free" parameters using only 59 data points and still generate a successful prediction about a 60th data point. In this case however, the way in which most of the "free" parameters were used prevented them from being really free in the conventional sense. In order to describe these parameters and their use it will be necessary to recall the specification of the maximum entropy estimator from section II.4. The estimator is specified by the following definitional equations:

There are 2^7 basic events to the space that the probability measure pr assigns probability. These are the possible conjunctive patterns of the seven mammals having and not having a given property. Let $be._k$ stand for the kth basic event (k between 1 and 2^7) and let the function val(i,k) be equal to 1 if the ith mammal has the property in the kth basic event and 0 if it does not. Twenty one special functions f_{ij} that map the space of basic events in the set {0,1} are defined as follows:

1) $f_{ij}(.be._k)$ = 1.0, if val(i,k) = val(j,k) and 0.0, otherwise.

There is an intrinsic relationship between these functions and the similarity formula d) shown immediately below in abbreviated form.

d) $Sim(i,j) = (p_{11} + p_{00}) / (p_{11} + p_{01} + p_{10} + p_{00}) = p_{11} + p_{00}$ where

$p_{11}(i,j)$ = probability of (the event that) P(i) and P(j).
$p_{10}(i,j)$ = probability of (the event that) P(i) and not P(j).
$p_{01}(i,j)$ = probability of (the event that) not P(i) and P(j).
$p_{00}(i,j)$ = probability of (the event that) not P(i) and not P(j)
= $1 - (p_{11} + p_{10} + p_{01})$

Specifically, the relationship is that $f_{ij}(.be._k) = 1.0$ just in case the event $.be._k$ is a subset of either the event that P(i) and P(j) or the event that not P(i) and not P(j) - i.e. $f_{ij}(.be._k) = 1.0$ if $.be._k$ is an event in which mammal i and mammal j "match" in terms of having or not having the property. The sum of the probability assigned (by pr) to the events $.be._k$ is such that $f_{ij}(.be._k) = 1.0$ is equal to $p_{11}(i,j) + p_{00}(i,j)$ which is in turn equal to Sim(i,j).

There exist 21 constants c_{ij} and a special constant c_0 such that

2) $pr(.be._k) = c_0 \bullet EXP[\ \Sigma_{ij}\ c_{ij} \bullet f_{ij}(.be._k)]$ and

3) $1/c_0 = \Sigma_k\ EXP[\ \Sigma_{ij}\ c_{ij} \bullet f_{ij}(.be._k)]$

The constants c_{ij} correspond to 21 of the 25 "free" parameters. The constant c_0 is actually determined by these parameters as is shown by equation 3). There are 21 similarity functions corresponding to the distinct pairs of the 7 mammals that are also determined by these parameters as shown by equation 4) - the parameters determine pr, which in turn determines the similarity functions.

4) $Sim(i,j) = E(f_{ij}) = \Sigma_k\ f_{ij}(.be._k) \bullet pr(.be._k)$

These similarity functions can be thought of as the "true" similarities. For the model, they are to be related to the subjects' judgments of overall biological similarity by the following equation.

5) $a \bullet Sim(i,j) + b = SIM(i,j,OVER) +$ "noise1"

What this equation expresses is that there are two more free parameters, a and b, defining a linear transformation from the current "true" similarities of the model to the subject's judgments about overall biological similarity. Any left over discrepancy contributes to the current estimate of the fitting error of the model. Given the probability measure pr, the conditional likelihood that the conclusion mammal, i_1 , has P, given that the mammals

i_2 ... i_n have P and the mammals i_{n+1}... i_m do not have P, is given by the familiar conditionalization formula expressed on line 6).

6) $pr(P'(i_1) \mid P'(i_2) \& \ldots \& P'(i_n) \& \neg P'(i_{n+1}) \& \ldots \& \neg P'(i_m)) =$

$$\frac{pr(P'(i_1) \& P'(i_2) \& \ldots \& P'(i_n) \& \neg P'(i_{n+1}) \& \ldots \& \neg P'(i_m))}{pr(P'(i_2) \& \ldots \& P'(i_n) \& \neg P'(i_{n+1}) \& \ldots \& \neg P'(i_m))}$$

Finally, these conditional probabilities are related to the subject's judgments of likelihood by another two parameter linear transformation.

7) $c \cdot pr(P'(i_1) \mid P'(i_2) \& \ldots \& P'(i_n) \& \neg P'(i_{n+1}) \& \ldots \& \neg P'(i_m)) + d =$
rating for argument 1 + "noise2"

where argument 1 has conclusion mammal, i_1, positive mammals i_2 ... i_n, and negative mammals i_{n+1}... i_m.

To summarize, Of the total of 25 free parameters, 21 were used to parameterize the unknown probability distribution, 2 were used to parameterize a linear transformation relating the unknown probability distribution to the likelihood ratings, and 2 were used to parameterize a linear transformation between the expected values of the probability distribution that correspond to the similarities of model d) and the subject's ratings of overall biological similarity. There are two sources of bad model fit that the free parameters are adjusted to minimize during the calibration procedure. These are represented above by "noise1", the discrepancy in the fit of the similarities of the model, after they are linearly translated, to the judged similarities, and "noise2", the discrepancy in the fit of the conditional probabilities of the model, after being linearly translated, to the subject's probability ratings. During the calibration process, the fit of the model to the data is described by the following equations.

8) Error of similarity fit = Σ_{ij} D((a•Sim(i,j) + b) - SIM(i,j,OVER))
where D is an error norm, such as squaring, or taking absolute value, a and b are two of the free parameters, Sim(i,j) is the "true" similarities which are a function of the current state of the model, and

SIM(i,j,OVER) is the current subject's rating of the overall biological similarity of mammal pair (i,j).

In other words, the contribution to the error of the fit from the similarities is equal to the sum of the contributions of the error of fit to each of the 21 similarity judgments, and each of these errors is some function of the discrepancy between the model and the given judgment.

9) Error of likelihood fit = Σ_k D((c•Condit(arg_k) +d) - Rating(arg_k))
where D is as above, c and d are two of the free parameters, and Condit(arg_k) is the condtional probability for argument k derived from the current model's version of pr.

Here the variable k ranges over all of the 60 arguments except the one that is currently left to one side and is to be predicted. The total error of the current model fit is described by 9).

10) Error of total fit = sw•"Error of similarity it" + "Error of likelihood fit".
where sw is the meta-parameter mentioned above.

Before making each prediction of the likelihood rating assigned to an argument, the parameters of the model are adjusted in an attempt to minimize the error of total fit. Note that if the meta-parameter sw is large, then the adjustment process will tend to ignore the error of the fit to the argument ratings, focusing instead on minimizing the error of the fit to the similarities, as the large value of sw will cause the error of similarity fit to be the source of most of the current fitting error as measured by 9). As the meta-parameter sw goes to infinity then there are actually no free parameters in the model that are being adjusted to fit the likelihood ratings since the values of the parameters are determined by those which allow the model form to best fit the similarity judgments. If sw is very large though, the model cannot adjust to compensate for problems i)-iii). If sw is very small we would not expect the calibrated model to be particularly successful in predicting because too much freedom would then be given to the parameters to fit the 59 calibrating likelihood judgments. The poor predictions such a procedure would make are analogous to what would happen if we had fit a noisy curve

by simply drawing connecting lines between adjacent points and were then expecting the direction of the tangent to the last line segment drawn to predict the next data point. An intermediate value of sw is optimal then. Instead of arbitrarily picking such a value, it was searched for by a computer algorithm.

The entire process of evaluating the maximum entropy estimator is described by the following hierarchical algorithm.

I For each subject, a one dimensional search is performed for value of sw that causes the function described by level II to return the optimal value of the correlation between model predictions and the likelihood ratings of that subject. The value of the optimal correlation returned by II is taken to be the performance of the maximum entropy estimator for the current subject.

II Each of the 60 arguments rated by a given subject is selected in turn and held aside. The other 59 arguments, the ratings for those arguments, the 21 judgments of overall biological similarity, and the current value of the parameter sw passed along by I, are passed to the function described by level III. This function returns calibrated values for the 25 "free" parameters and these are then used to make a prediction for the likelihood rating given by the subject to the argument form of the argument that was held aside. Note that fixing these values in effect fixes a single probability distribution such that each of the 60 arguments is interpreted as a specific conditional probability relative to this distribution. When this process is repeated 60 times, there is a vector of 60 predictions. The correlation between this vector and the actual likelihood ratings given by the subject is returned to I.

III The 25 parameters are adjusted to provide the best fit of the specified form to the 59 rated arguments and the 21 similarity judgments. The value of the fit for a given set of parameters is described by equation 9) above. The parameter sw used in this equation is the parameter passed on by II. The values of the fitted parameters are returned to II. Details of the fitting procedure are described in Appendix D.

Results and Discussion:

The median correlation of the predictions generated in the manner described above and the likelihood ratings of the 20 subjects of experiment II was .84, with a minimum of .72 and a maximum of .95. This performance compares favorably with the estimated performance achievable by the optimal model for the 20 subjects of experiment I: .88. A notable achievement of the maximum entropy estimator and the calibration process described above was an improvement in performance over the observed median intersession correlation of the 20 subjects from experiment I: .78. These figures, together with the performance statistics of some of the better simple models are shown in the table below.

The maximum entropy estimate clearly performed "well" in an absolute sense. The estimate of optimal model performance was taken from a different group of subjects and it is possible that those subjects participating in experiment I were more "noisy" than the subjects participating in experiment II. However the maximum entropy estimator, besides being handicapped by problems i), ii) and iii) mentioned above, was also handicapped by the

Support	Similarity	Scaling	Median Correlation
Estimate of the optimal model performance			.88
Performance of the maximum entropy estimate			.84
Observed correlation of Experiment I			.78
MAX	MIX	Linear	.76
MAX	MIX	Sigmoid	.76
MAX	OVER	Linear	.75
SUM/AVG	MIX	Linear	.75
AVG	OVER	Linear	.74
MAX	OVER	Sigmoid	.73
SUM	OVER	Linear	.72

Table III. Relative Performance of Maximum Entropy Estimator

relatively small number of data points on which to do calibration. Figure 5 shows a histogram of the absolute values of the "errors" between the predictions of the estimator and the actual subject ratings, pooled together from the 20 subjects of experiment II along with the histogram representing the estimated optimal attainable performance for the 20 subjects of experiment I. Figure 6 shows the two histograms of figure 5. along with the observed discrepancies between ratings of related pairs from experiment I. Comparison of these histograms reveals that the performance of the maximum entropy estimator is indeed "closer" to the estimated optimal attainable performance standard than to the observed errors.

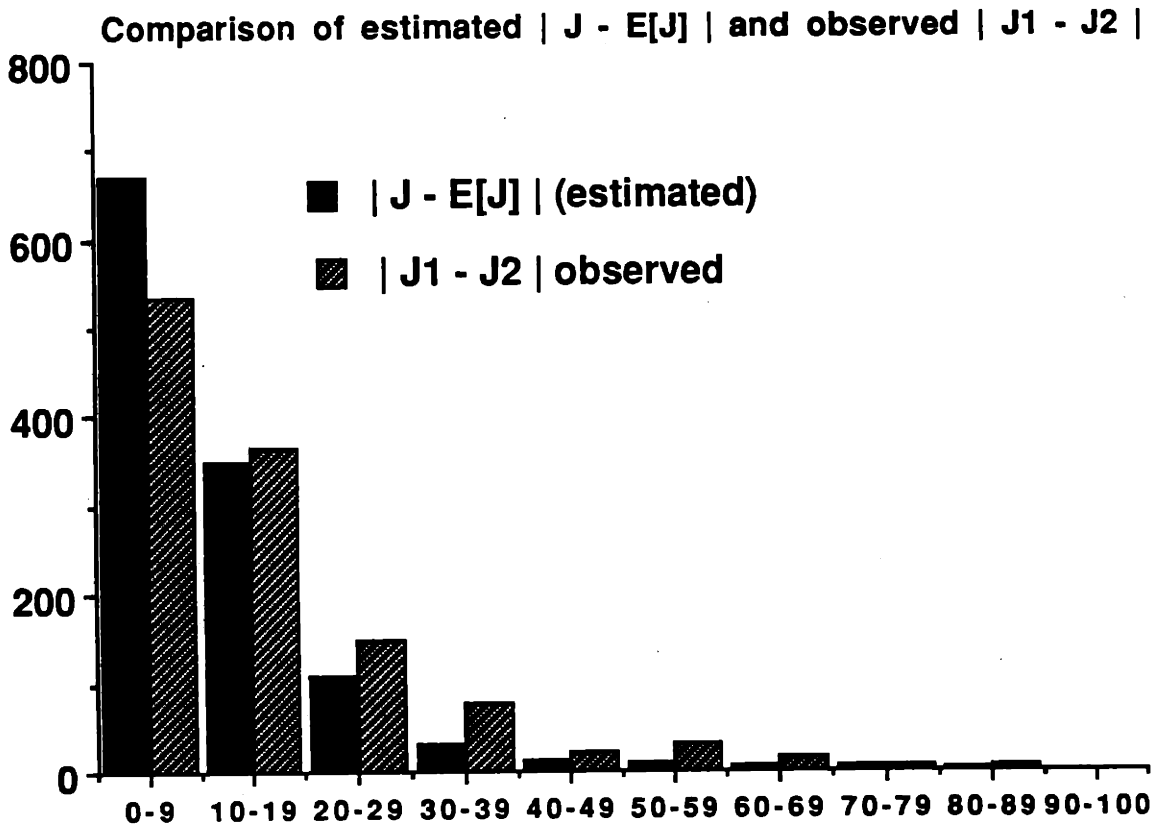

Figure 5. Comparison of the inferred optimal performance of a deterministic model for the subjects of experiment I with the performance of the maximum entropy estimator in predicting the judgments of the subjects from experiment II.

entropy estimator in predicting the judgments of the subjects from experiment II.

The comparison of the performance of the maximum entropy estimate with the performance of the various simple models is not a direct one, even though the test data was the same. The calibration procedure that was used with the maximum entropy estimate was significantly more complex. It seems likely that the performance of some of the simple models could be improved using the more sophisticated calibration procedure which, in effect, allowed "errors" in the similarity values to be inferred and adjusted. There are however, some non-statistical reasons for thinking that the maximum entropy estimator is more "natural" than any of the simple models that were examined proposed. I examine some of these below.

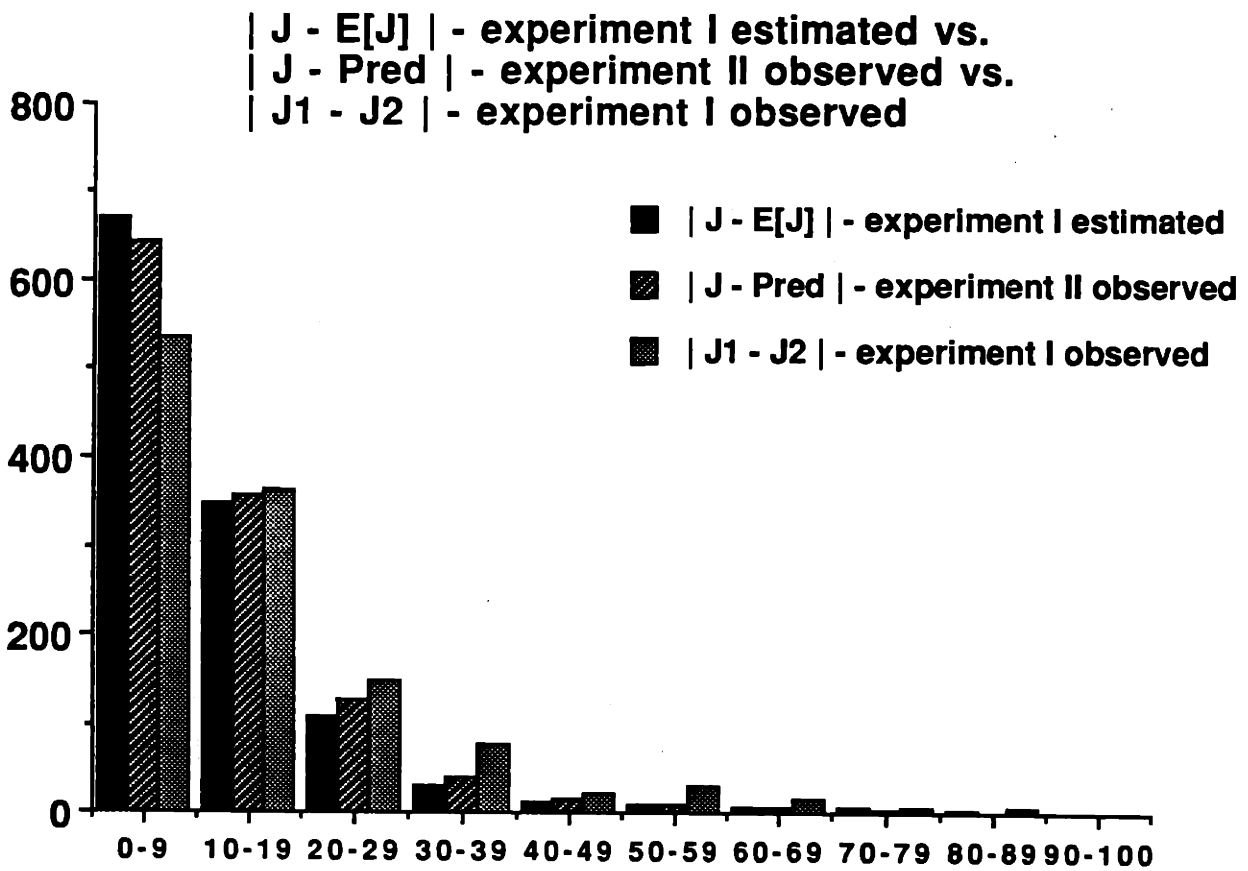

Figure 6. Comparison of the items from figure 5. with the addition of the observed discrepancies between ratings of related pairs from experiment I.

a) Continuity in the handling of deductive cases - All of the arguments that were presented to subjects in experiment I and II called for inductive judgments to be made - in the sense that a subject could not know for sure whether or not the conclusion mammal had the vaguely specified property. If the arguments were modified so that, while they retained their basic form, the conclusion mammal of a given argument was allowed to appear in the premises of that argument, then the arguments in which the conclusion mammal did appear in the premises would no longer be inductive. For example, given that bobcats have P12 and otters do not have P12 it is 100% likely that bobcats have P12, whatever P12 is. What I mean by "continuity in the handling of deductive cases" for a particular model is that the definition of the model does not have to be changed or elaborated to make it necessarily true that the model would assign a probability of 1.0 to the foregoing argument. This is true of the maximum entropy estimator for two reasons, one trivial and one substantive. The trivial reason is that the maximum entropy estimator proceeds by first computing a probability measure Pr, and then evaluating the argument as a conditional probability governed by that measure. No matter what probability measure is computed the probability of the event that bobcats have P12 given that bobcats have P12 and otters have P12 will always be 1.0 because the event being conditionalized on is a subset of the event being evaluated and conditionalization essentially means setting the probability of the event conditionalized on to 1.0, and any superset of a set having a probability of 1.0 must have probability 1.0 as well. The more substantive reason for the continuity of deductive cases is that the probability which the maximum entropy estimator assigns to a conclusion is a continuously differentiable function of the similarity of the conclusion to each of its premises, and for any positive premise individual, whether that individual is actually identical to the conclusion individual or not, the limit of the probability assigned to the conclusion as the similarity of the premise individual and the conclusion individual approach 1.0 (for any of the similarity models) will be 1.0. The reader can verify this property straightforwardly for similarity model d) through consideration of the proof given in the conclusion of section II.5 to show that the maximum entropy

estimator defined there agrees with Similarity Postulate i. Continuity in the handling of deductive cases will not generally hold for any of the simple models.

b) The maximum entropy estimator, which essentially uses the similarity values as a type of covariance statistic, is a form of computation that has been independently proposed several times in the connectionist and neural network literature. I shall explore some of these proposals in section IV. The application of a connectionist or any other simple model to the description of a "higher level" (i.e. involving conscious choice and strategies) judgment process is, of course, not intended to be an isomorphic description of processing. Nevertheless, the fact that this type of computation falls so naturally out of this paradigm is intriguing.

c) The maximum entropy estimator has an independent computational motivation for its form. In certain circumstances it is a type of maximum likelihood estimate for an unknown probability distribution. The following fact conveys a strong flavor of the idea. Suppose that the estimated similarities were actually based on empirical statistics recorded from a given set of data. The data set might be generated by first choosing n actual properties of mammals that are homogenous within a given species of mammal, and then recording for each of those chosen properties whether or not a given species of mammal has the property for each species of mammal under consideration - take for example the seven mammals from some argument booklet. Call a record recording for each of n properties, whether each of m mammals has the property or not a "data matrix". The four building blocks of the similarities, $p_{11}(i,j)$, $p_{10}(i,j)$, $p_{01}(i,j)$, and $p_{00}(i,j)$ can be determined for each pair i,j by counting up the number of the n properties that fall into the associated set and then dividing by n. Once these the values of these building blocks are fixed, the values of the similarities are fixed as well. Now given a certain set of similarity values derived in this way, it will generally be the case that more than one possible data matrix of n X m 0's and 1's could have given rise to exactly this set of values. If all we know about the data matrix that was actually recorded is the set of similarity values that were derived from it, then it makes sense to say that every data matrix which would give rise to exactly this set of values is equally likely. Suppose then

that we define a probability distribution on the possible patterns of the m mammals having and not having a property by counting the number of times that this pattern occurred among all of the possible data matrices that are compatible with the observed similarities. Call this distribution pr', and call the maximum entropy distribution compatible with the similarities pr. For finite size n, pr and pr' will not be exactly equal but they will be very close to one another, and the distance between them shrinks "exponentially" fast as n grows large (given any of a number of reasonable definitions of "distance" between probability measures). So pr is a good approximation to pr' for any n, and becomes essentially equal to pr' as n grows large. So one way of thinking about the maximum entropy distribution compatible with the observed similarities is that it is a very close approximation to the average or expected value of all the ways in which the similarities could actually have been generated. It will also be a close approximation to the mode of this distribution of "ways of generating the similarities." For a more formal discussion of these theoretical facts see Jaynes '79.

d) The maximum entropy estimator provides an independent interpretation for the notion of "unnatural" arguments that were discussed in III.2. This independent interpretation is essentially that the probability which the estimator assigns to the event which corresponds to the probability of the premises is so low in these cases that it falls below some threshold, indicating that the computations to be performed in estimating the ratio of the probability of this event to the probability of the conjunction of this event and the conclusion event would be unreliable or meaningless.

II.4.5.6 Discussion of EQ2 and EQ3

Recall that EQ2 was essentially the question of whether the likelihood judgments on this task could be predicted as functions of argument form and similarity. EQ3 asked whether the the maximum entropy estimator was the correct form for such a model. Evaluation of EQ2 and EQ3 on the basis of the data analysis from experiment I and experiment II is not a clear cut, yes or no. The following conclusions seem immediately defensible.

1) A variety of simple models perform well in predicting likelihood judgments as a function of rated similarities and argument form. The performance of the better versions of these simple models is only marginally inferior to the predictions of the same type of judgments generated by the having the same subject rate arguments that are identical except for the order of their premises on a different day. These simple models do not however capture all of the statistical regularity that is present in the judgments.

2) The maximum entropy model was able to capture most of the statistical regularity in the judgments as a function of, depending on how one looks at it, i. rated similarities, other rated arguments, and argument form, or
ii. "true" similarities and argument form.
Although both the simple models and the maximum entropy estimator calibrated parameters using other rated judgments, the simple models produced from this calibration process a function from rated similarities and argument form to likelihood judgments. The maximum entropy model used the calibration process to change the value of the similarities themselves. There are two possible ways of looking at this. One way, viewpoint ii. above, is to see the model as inferring true similarities from similarity judgments that are noisy and perhaps reflect a different similarity or choice of judgments about what type of properties are of interest. The other viewpoint is to see the form of the model as a useful one for estimating an unknown probability distribution, assisted by the similarity judgments, but perhaps being only approximate in its interpretation of similarities. Under either interpretation, an interesting consequence of the way in which the maximum entropy estimator was evaluated is that it implies that the statistical regularity in a set of likelihood judgments of the type appearing in the likelihood booklets of experiment I and II are well described as a set of conditional probability judgments from a single coherent probability distribution. This finding is particularly interesting in light of claims in the Psychological literature which argue for a disassociation between human likelihood judgment and probability. In section II.3 I briefly discussed some relationships between the ambiguity of an estimated similarity as a statistic and the "inclusion fallacy". I suggested there that errors like "inclusion fallacy" would be avoided if similarities were not being used inappropriately and that one form of appropriate usage which would avoid these errors would be if the individuals

A and B appearing in the expression Sim(A,B,D) were homogenous with respect to D, which is to say that for every i,j in A and k,l in B and for every P in D, P(i) if and only if P(j) and P(k) if and only if P(l). Experiments I and II pragmatically presupposed the homogeneity of the different mammal species with respect to the type of properties being reasoned about. The fact that probabilistic coherence was locally obtained in a typical subject's reasoning about a set of arguments may be interpreted as strong support for this view of the rational utility of similarity based likelihood judgment.

III.5 Experiment III: Does the property aspect matter to this task?

III.5.1 Motivation and Procedure

One of the recurrent nagging issues from experiment I and experiment II was whether the particular choice of unfamiliar aspect which was featured in a given argument really had any effect on judgment. In experiment I it was found that Similarity Postulate i. held up better when a similarity function which did not vary with the different unfamiliar property aspects was used. In experiment II, some of the models which did vary the similarity table that they used as a function of the unfamiliar property aspect seemed to gain a slight edge in performance, although a statistically insignificant one. The following experiment, which is a slight variation on experiment I, was conducted in an attempt to resolve the issue. Ten subjects participated in two experimental sessions which had an elapsed time of from one to seven days between them. In the first session, subjects rated an argument booklet exactly as in the first session of experiments I and II. In the second session, as in experiment I, subjects rated another booklet of 60 arguments that for each subject was a variation on the booklet that he or she received in session I. Of the 60 arguments in the booklet which the subjects received in the second session of experiment III, 30 formed a related pair with a corresponding argument from the first booklet in exactly the same way as in experiment I. The property aspect, the set of positive premises, the set of negative premises, and the conclusion were all identical. For the other 30 arguments in this second booklet, the set of positive premises, the set of negative premises, and the conclusion were all identical in in content to a corresponding argument

from session I. However, what distinguished this second set of thirty arguments was that the unfamiliar aspect of the property involved was randomly reassigned. So no matter what the property aspect of the corresponding argument had been in the first booklet, in the second booklet it had an even chance (one in five) of being either bone structure, digestion, dentition, thermal regulation, or fluid regulation. The thirty related pairs of arguments which corresponded in aspect between the first session and the second session were randomly interleaved throughout the booklet of 60 arguments in each session with the thirty related pairs of arguments that did not necessarily correspond in aspect.

III.5.2 Results

The analysis of experiment III proceeded by computing separately, for each subject, the correlation between the first booklet and second booklet ratings of the 30 arguments in the set with unchanged property aspects, and then comparing this to the similar correlation between the 30 pairs of arguments with randomly reassigned property aspects. The median among the 10 subjects for the unchanged set was .62, the median for the randomly reassigned set was .59. The small discrepancy between these numbers was interpreted as providing evidence for small differences in most subjects perception of the different unfamiliar aspects, at least as they effect this type of judgment. I do not have a strong story to tell about the discrepancy between the .62 for the unchanged set here and the median of .78 for the nearly identical statistic in experiment I - the fact that the 10 subjects in this experiment were run near the end of an M.I.T. semester may have been the important factor. Experiment III was interpreted as providing further confirmation of the fact that differences in unfamiliar property aspect did not have a strong effect on judgment, though they did probably have some small effect.

III.6 Experiment IV - The factors influencing similarity

III.6.1 Motivation and procedure

Experiment IV investigates the factors which influence judgments of estimated similarity. It was pointed out in section II.2 that the question of which factors influence the estimation of similarities is logically distinct from the question of the meaning and consequences of similarities as beliefs. Nevertheless, if similarity based likelihood judgment is to have a utility as a reasoning strategy we would expect the factors which contribute to estimated similarities to be commensurate with that which is believed as their consequence. From the results of Tversky '77 and '78 we expect that knowledge of domain related features and categories will play an important role in the estimation of similarities, although Tversky does not provide any concrete model for integrating featural and categorical information - unless the model he provides is summarized by the statement that a category is a feature with high diagnosticity.

In addition to expecting a relationship between featural and categorical knowledge and estimated similarities, we also expect judgments of similarity to be reliable indicators of estimated similarities. The reasons why these two identities may not be necessarily identical include the following:

a) A particular judgment of similarity may be the result of a computation which, relative to the variables measurable by the Cognitive Psychologist, is a stochastic process.

b) Judgments of estimated similarity involve translating mentally derived quantities to numerical values on some essentially arbitrary scale and reporting them as such. The number of gradations in the scale to be used for reporting may be far different from the number of effective value gradations of the internally represented quantity and decisions about the translation between the two may be somewhat ephemeral and based on factors not available to the Cognitive Psychologist.

The model of similarity that was used in the maximum entropy modelling in the analysis of the data from experiment II was

We are, of course, uncertain about which featural and categorical knowledge might play a role in the estimation of any particular similarity for any particular subject. However, in spite of these uncertainties, as well as the uncertainties in the relationship between estimated similarities and any particular set of similarity judgments, it seems like a worthwhile project to see whether a computation which uses the featural and categorical knowledge that we might expect to be available to a subject, and which uses this knowledge in a manner commensurate with the interpretation of similarity given above, can successfully predict judgments of similarity. This was the goal of experiment IV.

A body of featural and categorical knowledge/beliefs about mammals that was collected in the course of the experiments described in Osherson et al. '91 was conveniently available. I will now describe how this data was collected and how it is used to construct a model of the beliefs that a typical subject has about the probabilities of pairs of mammals having and not having "biological" properties. The data I will now discuss was obtained in two different experiments with two different groups of subjects. The first of these experiments I will refer to as the "feature rating task" and the second as the "categorization task".

Feature Rating Task

Subjects first reviewed a list of 48 mammals and 85 features. Some 42 of the 48 mammals appear on the list of 47 mammals that figure in experiments I, II, and III. Data pertaining to mammals which either appear on the list of 48 mammals used in the feature rating task but not on the list of 47 mammals used in experiments I, II, and III, or which appear in experiments I,II,and III but not on the list of 48 mammals used in the feature rating task did not figure in the final data analysis of this experiment - though the former figure as a small percentage of the contrast class for the feature rating and categorization tasks and the latter as a small percentage of the contrast class for similarity rating. Subjects also reviewed a list of 85 familiar features of mammals. Abbreviations for this list of properties are listed in table IV, and

some sample properties are given in unabbreviated form in table V. Subjects always worked with unabbreviated properties; the abbreviations are for expositional ease. With the exception of animal noises (bleeting, roaring, etc., essentially unique to each animal), no other feature was listed by more than a single subject from a group of 10 M.I.T. students asked to supply features of mammals. Moreover, none of the 85 properties was judged to be inappropriate by more than one student in the same group. These pilot studies suggest that the 85 properties capture much of the common knowledge about familiar mammals.

black	white	blue	brown	gray
orange	red	yellow	patches	spots
stripes	furry	hairless	toughskin	big
small	bulbous	lean	flippers	hands
hooves	pads	paws	longleg	longneck
tail	chewteeth	meatteeth	buckteeth	strainteeth
horns	claws	tusks	smelly	flys
hops	swims	tunnels	walks	fast
slow	strong	weak	muscle	biped
quadraped	active	inactive	nocturnal	hibernate
agility	eats fish	eats meat	plankton	eats vegetation
insects	forager	grazer	hunter	scavenger
skimmer	stalker	newworld	oldworld	arctic
coastal	desert	bush	plains	forest
fields	jungle	mountains	ocean	ground
water	tree	cave	fierce	timid
smart	group	solitary	nestspot	domestic

Table IV: Abbreviations for properties figuring in the study

bulbous:	having a roundish or bulky body shape
longleg:	having long legs
chewteeth:	having molars that are good for chewing
agility:	having a high degree of physical coordination
ocean:	living in the ocean
bipedal:	having the ability to walk erect on their hind legs
coastal:	living near the edge of an ocean or sea

Table V: Sample unabbreviated properties

Prior to the performance of the feature rating, it was explained to subjects that a non-negative number was to be assigned to each mammal-feature pair, and that the number assigned should reflect "the relative strength of association between the property and the mammal". No upper bound was imposed on these ratings. Subjects were also told to expect that many of the properties would be negligibly associated with many of the mammals. A rating of 0 was

to be used for these cases. Each subject worked for one hour, evaluating 10 to 15 randomly chosen mammals on all 85 properties (faster subjects evaluated more mammals). Subjects worked individually, at a computer terminal, and had the opportunity to revise their prior ratings at any time. Random sampling of the mammals was constrained so that each mammal was evaluated by 12 to 13 subjects* (* this number is somewhat in excess of the figure reported in Osherson et al. '91 as more ratings were conducted subsequent to the submission of that paper).

Construction of mammal-feature database

Every subject's ratings were individually normalized by a linear transformation to range from a lowest score of 0 to a highest score of 1. Following this operation, the median value of the 12 to 13 ratings of each mammal-feature pair was computed and stored in database1. The numbers ranging from 0 to 1 in database 1 were then chosen for conversion to a value of either 0, 1, or * according to the following procedure. - where * signifies that they were removed from the database. A rating from database1 for $mammal_j$ and $feature_k$ was converted to a 1 in database2 if its value was greater than or equal to the median of all the 85 feature ratings for $mammal_j$ in database1 and it had a value of .1 or greater. A rating was converted to a 0 in database2 otherwise. If the ratings for a feature were such that it wound up with at most a single non-zero entry among all the 48 mammals in database2 then it was removed as a column from the database. There were 72 remaining features after the removal process was concluded. The general motivation for this conversion procedure was to assign a 1 to all mammal feature pairs such that "the amount of association that the typical subject has with this mammal possessing this feature is non-negligible, it is significant that the mammal possesses this feature relative to the contrast class of other mammals, and the feature is a basis for comparison among mammals". The specific choice of the threshold value .1 was arbitrary. The database of the converted numbers will hereafter be referred to as the mammal-feature database.

Categorization Task

An independent group of 30 subjects participated in a categorization task. The subjects performing this task first read the following instructions.

> This part of the experiment concerns your judgment about how to distribute mammals into natural categories. Your task will be to create biologically meaningful groups, and then for each group to indicate which of the 48 mammals belongs to it. It is permitted to leave a mammal uncategorized if there are no other mammals in the list with which it forms a biologically natural group. Groups can be of any size, and it is permissible to have overlap of members.

Categorization was carried out on a computer terminal. Subjects devised category names and indicated which mammals among the 48 were included in it. Review and revision of previous choices of category name and membership was possible at any time. The superordinate "mammal" was not allowed. The categorization procedure lasted roughly 30 minutes.

Mammal X Mammal Featural Categorization and Featural Similarity Databases

One 48 X 48 database, indexed by pairs of mammals, was constructed by recording in the (i,j) position a decimal number indicating the fraction of the 30 subjects in the categorization task which had formed some category that contained both the ith and the jth mammals. This database will be referred to as the categorization database, and abbreviated as **CP**.

A second 48 X 48 database, indexed by pairs of mammals, was constructed by computing for each pair (i,j) of mammals the fraction of the 72 features in the mammal-feature database for which either both mammal i and mammal j had a value of 1 or both mammal i and mammal j had a value of 0. This database will be referred to as the featural similarity database, and abbreviated as **FS**

The formula for estimating similarity that I will examine is given by the following equation.

1) $Sim(i,j,\{overall\}) = \partial \cdot CP[i][j] + (1-\partial) \cdot FS[i][j]$,

where {overall} stands for the set of properties related to overall biological similarity which some kinds of mammals have and some do not, and ∂ is a positive constant between 0.0 and 1.0. This model reflects the following set of ideas. Assume for sake of argument that the mammal-feature database represents the knowledge of some typical subject about occurrence of familiar properties among familiar mammals, and also assume that we fix a given set of intuitive sub-categories of mammals -e.g. canines, felines, bears, whales, etc. which are real biological categories for this proto-typical subject. Then of the features figuring in the mammal-feature database, there is some subset of these features, which I will call H, such that for every sub-category and for every feature f in H, either all members of the category have f (i.e. have a 1 in the f position) or all members of the category do not have f (i.e. have a 0 in the f position). Then p1 = card(H)/72 is the fraction of the familiar properties of mammals which are homogenous among sub-categories of mammals. Now suppose we want to estimate the likelihood that some mammal i matches some mammal j relative to some new feature f'. One possible estimate would be to simply use the proportion of matches among the, in this case 72, properties of the "observed" feature set (or some approximation of that quantity). This would seem to be our best guess if we believe that f' is "drawn" from a set of features that is more or less equivalent to the observed feature set. However, if f' is conceivably drawn from a set of unfamiliar features/properties of the type which figured in experiments I and II, then we might desire to modify our estimate for this unfamiliar domain. Specifically, if we believe that the proportion of properties in this unfamiliar domain which are homogenous with respect to sub-categories of mammals - call this p2 - is greater than p1 then we would want to modify our estimate accordingly. One way to do this is to pick a ∂' between 0 and 1 to use in the following formula.

2) probability of i and j matching in new domain

= Sim'(i,j,new domain)

= $\partial' \cdot CAT[i,j] + (1-\partial') \cdot FS[i][j]$,

where the function CAT[i,j] = 1 if i and j belong to the same sub-category and 0 otherwise. A choice of a particular ∂' between 0 and 1 is appropriate if one, in effect, believes that $p2 = \partial' + (1-\partial') \cdot p1$. It is a simple exercise to check relative to this latter formula that p2 ranges monotonically between p1 and 1.0 as ∂' ranges monotonically between 0.0 and 1.0. The idealistic justification for these equations is the following. Features or properties logically come in two types: those which are homogenous with respect to sub-categories and those which are not. Relative to the domain of familiar properties in the mammal-feature database, some fraction of p1 of the properties are homogenous and some fraction (1-p1) are not. It is not convenient to actually compute p1 however. In an idealized model of the new domain of unfamiliar properties, some fraction p2 are homogenous and some fraction 1-p2 are not. When confronted with a property in the new domain, we would ideally want to predict the probability of two mammals matching in the new domain given that a property belongs to the homogenous fraction according to whether the mammals belong to the same subcategory, and ideally we want to predict the probability of two mammals matching relative to properties in the inhomogeneous part of the new domain according to their tendency to match in the inhomogeneous part of the old, familiar domain. Since we wouldn't generally have any information about whether a property in the new domain was homogenous or not, this is all an "as if" story. But we can get an estimate of the probability that j matches i relative to a random property in the new domain that makes a prediction "as if" there is a p2 chance the property is homogenous and a (1-p2) chance that it is not by using formula 2) above.

Since different subjects have different opinions about the categories of mammals, the function CAT[i,j] in formula 2) is replaced by CP[i][j] in formula 1) above, where CP[i][j] corresponds to the percentage of subjects which place mammal i and mammal j together in a category. Since the story which says that every pair of mammals either belong to a single subcategory or do not is idealized anyway it is possible that using CP brings some additional advantage to the descriptive validity of the simple account given

above by taking some account of different kinds of homogeneity through the stochastic mechanism of having different subjects select different categories at different times and then averaging them together. Realistically, the fraction ∂' will vary across subjects, but for the purposes of the current analysis, in which I am only using some composite of a typical subject's knowledge base anyway, I will idealistically use a single global estimate of overall biological similarity with a single ∂ parameter.

The similarity judgments that were to be predicted were those provided by the 20 subjects of experiment II in their second experimental session plus the judgments provided by an additional group of subjects that participated in another experiment not described here. The first experimental session of this other experiment was a minor variation on the first session of experiment II and the second sessions of the two experiments were identical. Each subject in these experiments provided judgments for the similarity of each distinct pair of seven mammals relative to their overall biological similarity (among other similarity aspects). There are 21 distinct pairs of 7 mammals, and so 21 times 40 = 840 pairs in all. However, because some of the mammals appearing in these experiments did not appear in the feature rating and categorization tasks it was necessary to remove from consideration all of the pairs containing these now extraneous mammals. After doing this there was a remaining total of 647 rated pairs distributed among the 40 subjects. There were never more than 2 extraneous mammals in the mammal set of any subject, and so after the removal of the extraneous mammals there was either "7 choose 2" = 21 pairs, "6 choose 2" = 15 pairs, or "5 choose 2" = 10 pairs of mammals rated and potentially predictable for each subject. Of the 647 different instances of similarity judgments under consideration, 347 were actually distinct pairings of mammals, each appearing in the ratings only a single subject.

III.6.2 Data Analysis and Results

Because of the uncertainty concerning the scale of the similarity ratings, the predictions generated by formula 1) above were allowed to try and match the judgment data as best they could by adjusting the parameters of a two

parameter linear transformation as well as the value of the parameter ∂. These three parameters were adjusted to find the best match (in terms of the sum of the absolute value of the differences) between the adjusted predictions of formula 1) and the 647 judgments of the overall biological similarity of mammals. The technique for doing this was essentially the same as the technique for calibrating the simple models of section III.4.5. In this way, a value of ∂ was globally fixed at approximately .4. Subsequent to the calibration of ∂, the Spearman rank order correlation between formula 1) and each subject's 10,15, or 21 judgments was obtained, and finally the median of these rank order correlations was computed. A value of .74 for the median rank order correlation between the predictions of the model, representing a global knowledge base, and the judgments of overall biological similarity of the 40 subjects was obtained.

To further examine the nature of the information provided by the categorization data, the same analysis as above was performed again with the substitution of subjects' judgments of similarity of ancestral lineage used instead of their judgments of overall biological similarity. The value of ∂ that was now obtained was .81 with a median correlation of .66.

III.6.3 Discussion of EQ4

The rank order correlation of .74 for this similarity model relative to overall biological similarity provides support for the interpretation of similarity applied in the modelling of section III.4.5. The fact that the value of the parameter ∂ was .40, a value significantly greater than 0, supports the idea that the information contained in the subjects' categorizations of the mammals was statistically useful. There are at least two different kinds of reasons why this might be true though. One reason is of the type outlined above. The other hypothesis is the categorization data is simply another measurement of overall similarity itself, like unto the featural similarity, and the fact that information from the two is averaged by the best predicting function is simply a statistical mechanism for removing unwanted some uncorrelated variance from the estimate. Arguing against the complete truth of this second hypothesis though is the fact that the value of the ∂ parameter obtained when ancestral lineage was substituted for overall biological

similarity was increased to .81 (resulting in a model with a still respectable correlation of .66 with the ratings of similarity of ancestral lineage) argues that the categorization data is not simply a reflection of similarity itself. The formula that was used for combining the similarity and categorization data reflected the use of featural and categorical information in a way that was commensurate with the beliefs of the similarity model. Taken alone, this data provides only minimal support for the view that the statistical importance of categorization information to similarity with respect to a domain is related to the expected percentage of homogeneous properties in that domain. I argue that an analysis of the role of categories in inference in terms of a conceptual division of properties into the homogeneous and the inhomogeneous can provide a coherent interpretation for some previous results in the inductive reasoning literature. In order to do this it will be necessary to generalize the notion of "homogeneous" used above. This generalization will retain the idea that the statistical importance of categorization information can be expressed in terms of beliefs about patterns of mammals having and not having properties.

I will use the notation **(H,{C1,...,Cn})** to stand for the set of properties for which the following is true: i. they are "homogeneous" relative to the set of categories {C1,...,Cn}, meaning that given any property in H and any pair of individuals i and j belonging to a category Ck, either both i and j have the property or both do not and also ii. if animal i is in Ck and has the property and animal j is in Cl and $k \neq l$ then animal j does not have the property. The notation **(I,{C1,...,Cn})** = $(H,\{C1,...,Cn\})^C$ will in general be used to stand for the "inhomogeneous" properties which are the complement of (H,{C1,...,Cn}).

III..7 Probabilistic interpretations of some effects of categories on inductive reasoning

III.7.1 Another Interpretation of Rips '75

In his 1975 paper entitled "Inductive judgments about natural categories" L. Rips investigated the joint effects of similarity and typicality within a category on inductive judgments. He performed parallel versions of the same

inductive task involving two groups of items (or "individuals" in the current terminology), a set of birds and a set of mammals. In an earlier study, Rips et al. '73, he had obtained ratings from subjects for the similarities of all pairs of items in each each group as well as ratings for the similarity between the items of each group and their superordinate ("bird" or "mammal"). This similarity data was then used to infer a representation for each group in which the individual members and the superordinate were mapped onto points in a two dimensional Euclidean plain (details in Rips et al. '73). The interpretation given to to these planar representations is that the distance between the points in the plain representing two items is inversely related to the similarity of the items. This is true of the distance between basic items and the superordinates as well. The induction task that Rips had subjects perform was as follows.

> "Subjects read a problem concerning animal species inhabiting a small island. The problem listed the names of the species (e.g. robins, geese, and hawks) together with the fact that the number of animals in each was approximately the same. The problem then stated that all of the animals in one of the species (e.g. all of the robins) had a new type of communicable disease. Subjects were then asked to estimate, for each of the other species, the proportion of animals that also had the disease. We can let the *Given Instance* denote that species said to have the disease, and the *Target Instances* those species about which estimates must be made."

The first of Rips' two principle findings was that the estimate of the proportion of target instances having the disease given the uniformity of the disease among the given instances was well predicted by the following equation.

1) $prop.(T \mid G) = c1 - c2 \cdot d(G,T) - c2 \cdot d(G,C)$

where c1,c2, and c3 are positive constants, the function d(A,B) returns the distance between A and B as measured by the planar representation of the similarity data for the category, G stands for the given instance, T stands for the target instance, and C stands for the superordinate (e.g. mammals or

birds). The constants c1,c2, and c3 were determined by the best fit of a linear regression. The candidate terms to be included in the fitted equation included d(T,C) in addition to the terms appearing in the equation above. It was found that, after the constant term, the statistical importance of the various predictors corresponded to the following ordering: d(G,T), d(G,C), d(T,C). The significance of the latter term was not great enough to warrant its inclusion in the formula given above. Because d(A,C) and d(B,C) will not in general be equal for particular choices of A and B, the equation above implies the observed fact that judgments of prop.(A | B) were greater than judgments of prop.(B | A) when B was closer to the superordinate, or more "representative" of the category, than was A. Rips describes this phenomena in the following terms, "A representative instance, by definition, is one that shares many important properties with other instances in its class. If we learn that such an instance possesses some new property, then we assume that this property, too, is shared by other instances...The new property may be assumed to be based on other well-known properties of the typical instance which are, in fact, widely shared. Therefore, the new property itself, or at least susceptibility to it, should be common to many instances of the set."

The second important finding that Rips describes in his paper is that the effect of typicality can be eliminated by providing subjects with additional information. Rips performed a parallel version of the experimental task described above in which subjects are also told that "scientists know that any of the other animals could also contract the disease." It turns out that this information alone is sufficient to render the term d(G,C) in the equation above statistically insignificant, or as we are told, to eliminate the effect of representativeness. The group of subjects which were given this additional information were termed the "informed group" while the other subjects were termed the "uninformed group".

From the comments above, and others in the article, it is clear Rips assumes that, by and large for most subjects, questions of the form "Given that all mammals of type 1 have a disease, what is the proportion of mammals of type II that have the disease?" will elicit essentially the same judgment strategies and responses as questions of the form "Given that (all) mammals of type 1 have a certain property, what is the likelihood that mammals of type

II will (all) have the property?" If we accept this assumption, which seems to be supported by the data, then Rips' results are directly related to the experimental results reported in this paper concerning the factors which influence similarity and likelihood judgments about arguments.

What I would like to argue is that Rips' results are agreeably consistent with the story of homogeneous and nonhomogeneous sets of properties that I laid out in section III.6.1. The only superordinate categories which are mentioned in Rips' experiment are "bird" and "mammal". It is difficult to know what other sub-categories might implicitly have played a role in subjects' judgments on Rips task. My experience with the categorization task described in experiment IV suggests that ideas about "biologically meaningful groups" vary widely among different subjects and may reflect different criteria. Let us assume though that at any given time, a typical subjects will recognize some set {C1,...Cn} of sub-categories of birds or mammals, and when I speak about (H,{C1,...,Cn} and (I,{C1,...,Cn} it will be to these subcategories that I refer.

I will now consider a possible probabilistic analysis of the observed patterns of judgment on Rips' task. The conditional probability of interest, the probability that 'the target' has a property given that 'the given' does can be broken into cases according to whether P is in (H,{C1,...Cn}) or not. In other words, cases are determined by the question is 'the property' of a type that is particular to a given type or sub-category of mammal? If so, then knowledge that a member of a given sub-category has the property will imply that other animals that are not members of that sub-category do not. If this is not the case, then the sub-categories are essentially irrelevant. I realize that reference to 'the property' is somewhat bizarre in this scenario but as I stated above, I am assuming that the reasoning patterns take on this form because Rips himself describes them in this way and it is consistent with the data; 'the property' can be though of as 'a propensity of magnitude m to contract the disease'. I use the notation P(A) to mean A has 'the property'. The analysis of the conditional probability broken into cases is a follows.

2) pr(P(T) | P(G)) =
pr(P(T) | P(G) & (P in (H,{C1,...Cn}))•pr(P in (H,{C1,...Cn}) | P(G)) +
pr(P(T) | P(G) & P in (I,{C1,...Cn}))•pr(P in (I,{C1,...Cn}) | P(G)).

One reasonable interpretation of the Rips' instructions and the performance of subjects on the "informed" condition of the experiment is that the information that "any of the other animals could also contract the disease" is implies that P is in (I,{C1,...Cn}), because this information says that it is not the case that only a certain type of mammal has the property. If P is known to be in (I,{C1,...Cn}) then the formula above reduces to

3) pr(P(T) | P(G)) = pr(P(T) | P(G) & (P in (I,{C1,...Cn}))

To understand the effect of typicality on this formula, let's consider two animals, A and B, taking A to be a typical mammal and B to be an atypical mammal. Since the judgments were symmetric under the exchange of the given and target animals in the informed case, we assume that

4) pr(P(B) | P(A) & (P in (I,{C1,...Cn})) = pr(P(B) | P(A) & (P in (I,{C1,...Cn}))

Note that if one assumes that the unconditional likelihood of A and B are equal then for a fixed type of P, pr(P(A) | P(B)) = pr (P(B) | P(A) - I discuss this point further below, but it seems intuitively plausible that if P is known to be a type of property which any sub-categorical type animal might have, that the unconditional probability of its occurrence among any individual animal should be equal to any other, knowing nothing else. I will give the label c1 to the quantity on either side of the equality in 4). Plugging this back into the original formula 2), which still applies to the "uninformed" case, we get the following.

5) pr(P(T) | P(G)) =
pr(P(T) | P(G) & P in (H,{C1,...Cn}))•pr(P in (H,{C1,...Cn}) | P(G)) +
c1•pr(P in (I,{C1,...Cn}) | P(G)).

If we assume that a typical animal A and an atypical animal B are in different sub-categories, then the probability of one having the property given that the other has the property and the property is homogeneous among sub-categories will be 0 regardless of which animal is the given and which is the target. If we replace this quantity by 0 then the formula above reduces to

6) pr(P(T) | P(G)) = c1•pr(P in (I,{C1,...,Cn}) | P(G))

Therefore, we expect that pr(P(B) | P(A)) > pr(P(A) | P(B)) in the uninformed case if and only if pr(P in (I,{C1,...,Cn}) | P(A)) is greater than pr(P in (I,{C1,...,Cn}) | P(B)). Rips' discussion of the psychological origin of the asymmetry due to "representativeness" which I quoted above is more directly an argument to the effect that this inequality is expected than it is a description of his model. As he says, "A representative instance, by definition, is one that shares many important properties with other instances in its class. If we learn that such an instance possesses some new property, then we assume that this property, too, is shared by other instances...The new property may be assumed to be based on other well-known properties of the typical instance which are, in fact, widely shared." Putting aside questions about the factual content of this matter, either it is a valid psychological description, in which case we may except the prediction of the typicality phenomena from the formula above, or there is some other explanation for the descriptive success of Rips' model that depends on some other psychological belief or process. I will assume that the former is the case.

I have just given a probabilistic analysis of Rips' results in which everything seemed to come out reasonably in a way that is compatible with probability. Rips himself considers and dismisses the following simpler probabilistic analysis which superficially seems to contradict the argument made above.

7) pr(A | B) = pr(B | A)•pr(A) / pr(B).

This is Bayes' rule. It suggests that if the unconditional probability of two animals having the disease is equal, then the conditional probability of one given the other should be symmetric in the choice of given instance and target. If we want the typicality result to come out as observed - i.e. P(B | A) > P(A | B) where A is the more typical animal - then we must have pr(B) > pr(A). Rips' asked subjects about the unconditional probabilities of animals having the disease. Exactly what pattern of responses he obtained is not clear to me from reading his article, but they were evidently not the appropriate ones to make the formula compatible with the observations. It is hard to

definitively pinpoint where the contradiction in the two probabilistic analysis lies, but I suggest the following. The instance of Bayes' rule given above should actually be written out as

8) $pr(P(A) \mid P(B)) = pr(P(B) \mid P(A)) \bullet pr(P(A)) \,/\, pr(P(B))$.

The question that must be answered to apply such an analysis is whether the 'P' that one has in mind is when evaluating $pr(P(A) \mid P(B))$ is the same 'P' that one has in mind when evaluating $pr(P(B) \mid P(A))$ (where by same I mean drawn from the same reference class), or whether what actually takes place is an evaluation of $pr(P_1(A) \mid P_1(B))$ and of $pr(P_2(B) \mid P_2(A))$ where $P_1 \neq P_2$, in which case formula 8) is irrelevant. The analysis given above suggests that there are at root two different types of 'P' because they reflect different mixtures of the homogeneous and the inhomogeneous. To put it another way, on the basis of the knowledge that a typical or an atypical member of a superordinate category has a property we form different expectations about the likelihood of that property being one of the properties that pattern along the lines of biological sub-categories, and hence are rarely jointly observed in dissimilar animals, or one of the properties which are observed in many different kinds of animals. In Rips' "informed" condition, subjects are essentially told that 'the property' under consideration is of the inhomogeneous variety. In the uninformed condition, the property is effectively a statistical mixture of the two. What is shown to be required for the observed typicality effect, according to the analysis given above, is that there is a difference in the mixing proportion which is inferred on the basis of the knowledge that a typical vs. an atypical animal has the property. There may very well be a good reason for such an inference.

III.7.2 Representing Categories as Higher Order Statistics

The interpretations of similarity that have been proposed in this work have all focused on similarities as slightly different types of second order statistics. Relative to a distribution defined on a set {i1,i2,...,in} of individuals, a first-order statistic is a constraint on the expected value of some ik. A second order

statistic relates to the joint distribution of some ik and il. Typically, a category may have several i's as members. One way of representing the inductive significance of a category relative to a certain property domain is as an expectation about the proportion of properties from that domain for which it will be true that all of the members of the category will "agree" on the property homogeneously. In section IV.1 I introduce the notion of an exponential family and explain how the maximum entropy distribution compatible with a set of similarities is one example of an family. In general, such families describe distributions in terms of "factors" that influence the assignment of likelihood to events. For the distributions involved in the pure form of similarity based likelihood judgment that has been studied in this work, these factors are the joint presence and absence of a property among two individuals. The framework of exponential families is very general however. A factor can just as easily be the joint presence of a property among all the members of a category. Essentially, this is accomplished mathematically by using functions like the following in a way that is precisely parallel to the way the fij were used in the specification of the maximum entropy estimator:

f_C(a basic event) = 1.0 if either all of the members of category C have the property in this basic event or all the members of category C do not have the property in this basic event.

The role of functions like this one is discussed in detail in section IV.1.

IV Theoretical Perspectives on Similarity Based Likelihood Judgment

IV.1 The Ubiquity and Relevance of Exponential Families and Their Sufficient Statistics

In this section I will provide a brief introduction to a class of probability distributions called exponential families that underly most parametric statistical analysis. Understanding the nature of this abstract class of parametric probability distributions can provide a unifying framework in which to describe a number of important issues relevant to similarity based likelihood judgment. A skeletal sketch of these issues is conveyed by contemplation of the following facts:

1) All maximum entropy distributions based on a given set of expected value equality constraints are exponential distributions parameterized by the values of those constraints.

2) Every probability distribution defined on a discrete space is a multinomial distribution and every multinomial distribution is a distribution from some exponential family, so, extensionally speaking, saying that a distribution defined on a discrete space is from an exponential family is uninteresting by itself. However, the framework of the exponential family gives one a way of talking about particular sets of distributions on discrete spaces that are interesting. For example, the exponential family of distributions on the space of the 2^n yes/no patterns of n individuals with the property that the likelihood of any given pattern is determined by which individuals pairwise "agree" with one another in that pattern is the maximum entropy distribution that was estimated from the similarities in the data analysis of section III.4.5.4. The distribution on this space in which the likelihood of any given pattern is a function of which individuals agree in that pattern plus which individuals said "yes" in that pattern turns out to be the distribution that is generated by the, now familiar, Boltzmann machine neural network when it is running freely under the interpretation that the "firing" of a specific node in the network at a given time represents a vote for the patterns in which that node or individual says "yes" (see Hinton et al. '86). This relationship will be elaborated on below. To specify how often any two of the

n individuals agree, plus how often each individual says "yes" takes just (n(n-1)/2 + n) numbers. This is a small number as compared to 2^n-1 numbers required to specify an arbitrary distribution on this space.

3) If there is some particular subset or "category" of the n individuals which all say "yes" together some fraction of the time - more often than they would all say yes together just by dumb luck - then with an additional parameter their behavior can be factored into the distribution as well.

The expected values which parameterize exponential families and maximum entropy distributions are expected values of some function from values in the sample space on which the probability distribution is defined to real numbers. These functions are commonly called "statistics". R.A. Fischer called a statistic sufficient "when no other statistic which can be calculated from the same sample provides any additional information as to the value of the parameter to be estimated." Suppose that X1...Xn are independent identically distributed random variables, and that the true distribution of X1...Xn is some member of a known class of distributions P depending on a (possibly multi-dimensional) parameter THETA. In other words, each member of P is indexed by one and only one value of THETA. Let T be a function defined on X^n (the space given by n repetitions of X). Then T is said to be sufficient for the family P if the conditional distribution of X1...Xn given the value of T(X1,...,Xn) is equal to the conditional distribution of X1...Xn given THETA. Some familiar examples of families of distributions and their sufficient statistics are 1) the family of binomial distributions with n observations and the statistic whose value is the number of "successes", and 2) the family of normal distributions and the two dimensional statistic equal to the sample mean and sample variance. For a more formal definition one may turn to Lehmann '83.

In general, there is not a unique sufficient statistic for a family of distributions, and some sufficient statistics are uninteresting. For example, the complete set of observations themselves are always sufficient. A more interesting notion is that of a minimal sufficient statistic. It is difficult to define this notion precisely without a lot of mathematical machinery but the basic idea is that T1 is minimal sufficient for P if for any other statistic T2 that

is sufficient for P there is a (possibly non-invertible) function f such that T1 = f (T2). Example 1) and 2) in the last paragraph were in fact examples of minimal sufficient statistics.

Most familiar parametric families of probability distributions are examples of exponential families. This is true of the normal, poisson, and multinomial families to name a few. A family of probability measures on a domain X is called an exponential family if there are real valued functions T1...Tk on X such that the probability density of every member of the family is equivalent to an expression of the form

$p(x) = c0 \bullet \exp[c1 \bullet T1(x) + ... + ck \bullet Tk(x)]$ for some choice of constants c1,...,ck.

The constant c0 is then determined by the need to normalize the integral of this function to make it a probability density. The T1...Tj are called **affinely dependent** if there exist constants c0,c1,...cj, not all equal to zero, such that the equation $c0 + c1T1(x) + ...+ cjTj(x) = 0$ holds for all x in the domain. If this condition holds then some ci is non-zero and we can solve for Ti as a linear combination of the other Tk and a constant. If so, then Ti contains only redundant information and we can eliminate it in the sense that any linear equation that could be formed with the T's before can still be formed without Ti. A representation of an exponential family will be called minimal if the Ti mentioned in its definition are not affinely dependent. The basic properties of exponential families revolve around their relationship to their minimal sufficient statistics. One of these properties is that for every sample X1...Xn that is i.i.d. from an exponential family P, if T1...Tk are functions appearing in a minimal representation of P then the vector $(\Sigma T1(x)/n,...,\Sigma Tk(x)/n)$ is a sufficient statistic. Under some further assumptions about the richness of the family P, which are usually satisfied in practical cases of interest, this statistic will be minimal sufficient. When the parameter space of an exponential family is not artificially restricted, then statistical estimation is particularly simple since a maximum likelihood estimate is always defined, unique, and equal to the parameterization that equates the observed sample averages of the functions Ti with their expected values. Which is in turn equal to the distribution of maximum entropy that has those observed sample averages as its expected values of those statistics.

Sufficient Statistics and Data Reduction

The computational significance of the existence of low dimensional sufficient statistics is twofold. On an algorithmic level, there is the capability for substantial data reduction leading to low storage requirements. On an informational level, the amount of data that is required to achieve statistical accuracy is greatly reduced. If the sample space on which a family of distributions is continuous, and all of the distributions are supported by the whole space then, under some mild regularity conditions, the very existence of a set of jointly sufficient statistics that do not grow in size with the size of the data sample already implies that the family of distributions is exponential. On a finite sample space, such as the one we are concerned with, all possible probability distributions are always finite multinomial. It is easy to show that every finite multinomial distribution can be written in exponential form. For example, suppose a sample space has three possible disjoint outcomes, A, B, and C, and a distribution on this space assigns probabilities .2, .3, and .5 respectively. Let IA, IB, and IC be functions defined on this domain that return 1.0 if a point of the domain is in A, B, or C respectively, and 0.0 otherwise. Then $p(x) = \exp[\ln(.2)IA(x) + \ln(.3)IB(x) + \ln(.5)IC(x)]$ is an equivalent exponential representation. In light of the discussion above however, it is clear that such a representation cannot be minimal since knowledge about the occurrence of IA(x) and IB(x) is sufficient to completely determine IC(x). The representation $p(x) = 0.5\exp[\ln(.4)IA(x) + \ln(.6)IB(x)]]$ is a minimal representation. This example is an instance of the general fact that a multinomial distribution with M possible outcomes is completely defined by at most M-1 independent parameters and has at most M-1 jointly sufficient statistics in its minimal representation. Very often it will be the case that there are known relationships between the various outcomes of a multinomial distribution, and it is in such circumstances that exponential families with fewer than M-1 terms are interesting (for discrete spaces such models are often termed "loglinear" and introduced with somewhat different terminology). The common reason for this is that the different outcomes represent combinations of a small number of "factors" that can assume two or more values. The factors are in the similarity based likelihood model are the

presence or absence of a biological property in a particular type of mammal. It is typical of most statistical modelling in such cases to investigate low order interactions between different factors, before positing higher order interactions. It is straightforward to understand the assumptions made by such models in terms of exponential families. To use the example at hand, let the set of distributions of interest represent the presence and absence of a property among a set of k mammals, m1 ... mk. A model including all first and second order factors could use k functions ti, 1...i...k, that record the presence or absence of the property for m1...mk and fij, defined as above, recording whether mi and mj match on a particular property. In general, there will be k(k-1)/2 second order functions fij. So the number of statistics (and free parameters) in this family of models is k + k(k-1)/2. As k grows large, this number is absolutely and proportionally drawfed by the number of parameters in the full interaction (general multinomial) model, $2^k - 1$.

Expected Values as Parameters

It is simple to evaluate an expression of the form p(x) = c0•exp[c1•T1(x) + ... + ck•Tk(x)] when the constants c0,c1,...,ck are known. As was stated above, the constants c1,...ck are determined by the expected values of the T1...Tk and the constant c0 is determined by normalization. It is theoretically true therefore that known values for $\Sigma T1(x)/n$, ... $\Sigma Tk(x)/n$ uniquely define and parameterize such a distribution, and this fact is made use of here for relating probability distributions and similarity values. When k is large it becomes computationally prohibitive to actually find the exact value of the maximum entropy distribution for these constraints. For small k, finding the ci can be done by an iterative process that cycles repeatedly through the individual equations, and converges reasonably quickly; for large k, the preferred approach in practice is not to solve for the ci at all, but rather to estimate the required probabilities using stochastic simulation. This technique is allied with that of the Hopfield nets and Boltzmann machines proposed as computational models of neuron like computation by Hinton et al '86. The computational technique actually dates back Mitropolis '53.

The function relating a set of avgTi constraints to a set of ci is partial. Every choice of real numbers for the ci from 1 to k results in a probability

distribution after the appropriate normalizing constant c0 is chosen. However, many choices of the numbers for E[Ti] are not mutually compatible, given the desired interpretation of the Ti. For this reason it is worth considering the possiblity that what people actually store as a representation of similarity is the "weights" and "thresholds" of a Boltzmann-like network. If that is the case then the "estimated similarities" would actually be produced by a type of monte-carlo estimation of their association values. All of this is naturally speculative, but it is fair to say that on the relative scale of these things, the similarity based likelihood strategy proposed here is readily implementable by biological models. In the next section I briefly review what a Boltzmann machine actually is and how it relates to these exponential models.

IV.2 Neural Network Models for the Maximum Entropy Estimator

The maximum entropy estimator described in section II.5 is closely related to the Boltzmann machine models of Hinton and Sejnowski '86, as well as a number of other associative memory models in the neural network literature (e.g. Hopfield '82) . I will first describe the basic elements of this network and then explain the connection with the maximum entropy estimator. My presentation will follow MacKay '91.

A Neural Network Implementation

Architecture: The network consists of N "neurons" which have binary activities $x_i = \pm 1$. Each neuron is connected to every other through symmetrical connections that are assigned real-numbered weights $w_{ij} = w_{ji}$. The neurons do not feedback to themselves so $w_{ii} = 0$.

Dynamics: The dynamics of an individual neuron x_i depend on the value of its "activation" a_i, which is described by the equation:

$a_i = \Sigma_j w_{ij}x_j - b_i$ where t_i, a real variable represents the threshold or bias of the neuron x_i.

The x_i are updated at random intervals according to the equation

$x_i(t+1)$ = 1, with probability (1.0 / [1.0 + exp[-ai]]),
-1, with probability 1.0 - (1.0 / [1.0 + exp[-ai]])

Distribution:

We define a global "energy" for the network relative to a given state or realization of a combination of x values by $E = \Sigma_i a_i$. It is well known that after this network has run for a while and reached "equilibrium" it will have a Gibbs distribution in which the probability of being in any state depends only on the ratio of the enegry of that state to the other states, and that for any state k, the probability of being in state k at equilibrium it will be

$pr(k) = \exp[-E_k] / \Sigma_j \exp[-E_j]$ where j ranges over all 2^N states.

The most convenient way to obtain the expected values of various events from such a network is by random sampling of the occurrence of the events during some time window of events.

Mapping from Similarity Space to Network Space:

The activation state (1,-1) of a node of the network represents the "yes property"/"no property" status of some individual in similarity space. If the unconditional probability of each of the individuals in our similarity space is identical - we don't have reason to believe than any individual is a priori more likely than any other individual to have a random unknown property - then the threshold values of the network, the b_i, should be set to identical constants. If every individual is as likely as not to have a random property, then the b_i would be set to 0.0. If it is less likely than an individual has a property than that the individual does not, then the b_i would be greater than 0. If the unconditional probabilities are all equal, whatever their value, the the "weights" of the Boltzmann network, the w_{ij}, should be set equal to the parameters c_{ij} of the associated maximum entropy distribution which has the correct expected values for the similarity statistics of individuals i and j. The Boltzmann machine can evaluate condtional probabilties defined relative to

this distribution by sampling (monte carlo). The sampling that takes place in the evaluation of conditional probabilities proceeds by letting the network run freely after the nodes which are being conditionalized on are "clamped" to the required values. This can be made clear with an example. Suppose we are interested in evaluating the conditional probability that individual i has a random property given that individual j does and individual k does not. The way that this conditional probability is evaluated is by momentarily "clamping" individual network node j so that it stays in the '1' state and "clamping" individual network node k so that it stays in the '-1' state. The network is then allowed to "run" with the same dynamics as before except that the updating of node j and node k is held in obeyance. The relative frequency with which node i will be "on" in the network, so running, has an expected value equal to the condtional probability of individual i having a random property given that individual j has this random property and individual k does not, where conditional probability is defined relative to the maximum entropy probability distribution compatible with the similarities.

Learning "Similarity" with a Boltzmann machine:

Hinton & Sejnowski '86 describes a learning algorithm for Boltzmann machines used as associative memories. The relevance of this learning algorithm to the current context is that the network could learn the statistical content of the similarity relationships, if such relationships were determined solely by property sampling, by learning to match the expectations of the network, as expressed by on/off patterns of nodes, to the sampled distribution of some environment or domain of properties. I will not describe this learning algorithm here. The reader is referred to Hinto & Sejnowski '86.

V Perspective and Conclusion

The theories proposed in this work were framed with the explicit goal in mind of establishing a conceptual link between similarity-based likelihood judgment and probabilistic inference. The most important empirical result in this work has been the finding that similarity based likelihood judgments, under the proper circumstances, may be viewed as coherent evaluations of conditional probabilities. These conditional probabilities are not conditional on all of a person's knowledge and beliefs. They are rather conditional on a small number of locally relevant sources of evidence. For similarity based likelihood judgment, these sources of evidence are provided by similarities between individuals. Although similarity based likelihood judgment may be understood as a form of judgment that parallels statistical inference, there are important differences. Some of these differences are the following.

a) In some situations, this reasoning pattern is applied in a way that is incompatible with a statistical interpretation (e.g. "inclusion fallacy") and its application can result in serious errors of judgment.

b) Although similarities may be interpreted as probabilistic beliefs, they do not arise solely through any conventional form of statistical sampling plan. Similarities may rather be estimated on the basis of heterogeneous sources of evidence. Some of these types of evidence were discussed in section II.3.

c) Similarity based likelihood judgments are made "on the fly" rather than at the conclusion of any pre-planned experiment.

The present work has provided a description of similarity based likelihood which involves a number of seemingly disparate principles.which have appeared at various points throughout the thesis. In actuality, most of these principles are closely related to one another, although the relationships have frequently been left implicit in the discussion. I will now list these principles and describe some of the relationships between them.

1) Similarity Postulate i - the probability that an individual i_0 has a property P will vary positively with the similarity of the pairs consisting of i_0 and each

of the individuals known to have P and negatively with the similarity of the pairs consisting of i_0 and each of the individuals known not to have P.

2) Similarity Postulate ii - in the absence of information other than the GIs, and if the likelihood of each of the individuals in the set $\{i_k, 0 \leq k \leq m\}$ having a property from a domain **P** is identical, the probability that i_0 has P will be a function of only the knowledge of which other individuals in the set $\{i_k, 0 \leq k \leq m\}$ have and do not have P and the set of values given by the similarity function applied to each pair of individuals in the set $\{i_k, 0 \leq k \leq m\}$, holding the domain **P** containing P constant.

3) Similarity Postulate iii - pairwise similarities between individuals relative to a domain are functions of beliefs about patterns of individuals having and not having properties of that domain. These patterns are described by the functions $p_{11}(A,B,D)$, $p_{10}(A,B,D)$, $p_{01}(A,B,D)$, and $p_{00}(A,B,D)$.

4) Maximum entropy estimator part a) and b) - similarity based likelihood judgments involving a small set of related individuals, $\{i_k, 0 \leq k \leq m\}$, and properties that are drawn from a common domain are coherently describable as conditional probabilities relative to a single probability measure, pr, which is the maximum entropy distribution compatible with the pairwise similarities between these individuals relative to the fixed domain. These conditional probabilities are the conditional probability that the conclusion individual has some property given the known positive and negative cases of related individuals having the property.

5) Similarities as sufficient statistics - knowledge of the estimated pairwise similarities of a set of individuals with respect to a domain is sufficient, in a statistical sense, to estimate the probability that one of these individuals has a property chosen from this domain given some knowledge of other related individuals having and not having the property. This is to say that further information about the frequency with which n-place combinations of these individuals (for n>2) have simultaneously shared particular properties is not actually utilized.

Principles 1) through 5) are highly redundant. For example, principle 4) and principle 3) together imply principle 1). This was proved in section II.5 of the thesis. Given principle 3), principles 4) and 5) are essentially different ways of saying the same thing. Given principle 3) these principles also imply principle 2). Given principle 3), principle 2) implies principle 5), and therefore 4) and 1) as well.

It can readily be concluded from the preceding discussion that an important, perhaps THE important empirical claim at stake here is that principles 2) and 3) hold at once for what has been described as similarity based likelihood judgment. Because "similarity" is a concept that intrinsically has so many degrees of freedom about it, the answer is not necessarily a simple yes/no/maybe. The positive evaluation of the maximum entropy estimator in section III.4.5.5 indicates that we can provide an affirmative answer in at least one non-trivial sense. In order to provide this however, it was necessary to "bend" the concept of similarity into the form required by the model - both literally and metaphorically. Are principles 2) and 3) absolutely correct? It's hard to know. The discussion of the probabilistic representation of categorical information (within exponential families) in section III.7.2 , for example, contains an implicit suggestion about other factors, not described by principle 2), which would be likely to play a role in the type of experimental task examined here. This other information is not described by the GIs so it would not contradict the theory, but would rather be some form of "noise" in the evaluation of the theory. Future research will undoubtably instruct us as to whether or not the analysis proposed here fashions a truly useful theoretical tool from similarity, or whether similarity as an explanatory concept will eventually shatter into many small pieces. We'll see.

Bibliography

Akaike, H. (1974) "A new look at the statistical model identification," IEEE Trans. on Automatic Control 19: 716-723.

Aleliunas (1990) "A new normative theory of probabilistic logic," in H.E. Kyburg '90.

Bacchus, F. (1990) *Representing and Reasoning with Probabilistic Knowledge - A Logical Approach to Probabilities.* MIT Press. Cambridge, MA.

Bacchus, F., Kyburg, H.E., & Thalos, M. (1990) "Against conditionalization," *Synthese* 85: 475-506.

Bar-Hillel, Maya. (1982) "Studies of representativeness," in Kahneman et al. 1982., pp.69-98.

Barron, A., & Cover, T. (1991) "Minimum complexity density estimation," IEEE Trans. on Info. Theory 37(4): 1034-1054.

Bishop, Y.M., Fienberg, S.E., & Holland, P.W. (1975) *Discrete Multivariate Analysis: Theory and Practice.* MIT Press. Cambridge, MA.

Brooks, L. "Nonanalytic concept formation and memory for instances," in E. Rosch et al. '78.

Carey, S. (1985) *Conceptual Change in Childhood.* MIT Press. Cambridge, MA.

Cheeseman, P. (1988) "An Inquiry into computer understanding," *Computational Intelligence* 4(1): 58-66.

Cheeseman, P. (1983) "A method of computing generalized Bayesian probability values for expert systems," *Proceedings of the 6th International Joint Conference on AI (IJCAI-83).* Karlsruhe, Germany. pp. 198-202.

Christensen-Szalanski, J., & Beach, L. (1982) "Experience and the base-rate fallacy," *Organizational Behavior and Human Performance* 29: 270-278.

Dempster, A.,P. (1967) "Upper and lower probabilities induced by a multivalued mapping," *Annals of Mathematical Statistics* 38: 325-339.

Doyle, J. "A truth maintenance system," *Artificial Intelligence* 12: 231-272.

Efron, B. (1982) *The jackknife, the bootstrap, and other resampling plans.* Society for Industrial and Applied Mathematics. Philadelphia, PA.

Estes, W.K. (1976) "The cognitive side of probability learning," *Psychological Review*: 83, 37-64.

Fine, T. (1973) *Theories of Probability.* Academic Press, New York.

Gaerdenfors, P. (1988) *Knowledge in Flux.* MIT Press. Cambridge, MA.

Geffner, H., & Pearl, J. (1990) "A Framework for Reasoning with Defaults," in H.E. Kyburg (1990).

Gelman, S. (1988) "The development of induction within natural kind and artifact categories," *Cognitive Psychology* 20: 65-95.

Gelman, S. & Markman, E. (1986) "Categories and induction in young children," Cognition 23: 183-209.

Geman, S. & Geman, D. (1984) "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," IEEE Trans. on Pattern Analysis and Machine Intelligence 6: 721-741.

Gigerenzer, G., Hell, W., & Blank, H. (1988) "Presentation and content: the use of base rates as a continuous variable," *Journal of Experimental Psychology: Human Perception and Performance* 14: 513-525.

Gigerenzer, G., & Murray, D. (1987) *Cognition as Intuitive Statistics.* Lawrence Erlbaum. Hillsdale, NJ.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Kruger, L. (1989) *The Empire of Chance.* Cambridge University Press. Cambridge, UK.

Ginsberg, M.L. (ed.) (1987) *Readings In Nonmonotonic Reasoning.* Morgan Kaufmann. San Mateo CA.

Girosi, F., Poggio, T., & Caprile, B. (1990) "Extensions of a theory of networks for approximation and learning: outliers and negative examples," A.I. Memo No. 1220, *Artificial Intelligence Laboratory, MIT.* Cambridge, MA.

Goodman, N. (1955) *Fact, Fiction, and Forecast.* Harvard University Press. Cambridge, MA.

Grandy, W., & Schick, H. (eds.) (1991) *Maximum Entropy and Bayesian Methods.* Kluwer. The Netherlands.

Halpern, J., & Rabin, M. (1987) "A logic to reason about likelihood," *Artificial Intelligence* 32(3): 379-406.

Harman, G. (1986) *Change In View.* MIT Press. Cambridge, MA.

Hawkins, R., & Bower, G. (eds.) (1989) *Computational Models of Learning in Simple Neural Systems.* Academic Press. San Diego.

Hinton, G. & Sejnowski, T. (1986) "Learning and relearning in Boltzmann machines," in D. Rumelhart & J. McClelland '86.

Hopfield, J. (1982) "Neural networks and physical systems with emergent collective computational abilities," *Proceedings of the National Academy of Science, USA* 79: 2554-8.

Jaynes, E.T. (1979) "Where do we stand on maximum entropy?" in Levine & Tribus '79. pp. 15-118.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases.* Cambridge University Press. Cambridge.

Kahneman, D., & Tversky, A. (1979) "Prospect theory: an analysis of decision under risk," *Econometrica* 47: 263-291.

Kahneman, D., & Tversky, A. (1982) "Subjective probability: a judgment of representativeness," in Kahneman et al. 1982, 32-47.

Koehler, J. (1990) "The status of base rates in probabilistic judgment," unpublished manuscript.

Kullback, S. (1959, 1978) *Information Theory and Statistics.* Peter Smith. Gloucester, MA.

Kyburg, H.E. (editor) (1990) *Knowledge Representation and Defeasible Reasoning.* Kluwer Academic. Netherlands.

Kyburg, H.E. (1983) *Epistemology and Inference.* University of Minnesota Press.

Kyburg, H.E. (1983) "The Reference Class," Philosophy of Science 50(3): 374-397.

Lehmann, E.L. (1983) *Theory of Point Estimation.* John Wiley. New York.

Levi, I. (1980) *The Enterprise of Knowledge.* MIT Press. Cambridge, MA.

Levine, R.D., & Tribus, M. (eds.) (1979) *The Maxium Entropy Formalism.* MIT Press. Cambridge, MA.

Levy, W. (1989) "A computational approach to hippocampal function," in R. Hawkins and G. Bower '89.

McCarthy, J., & Hayes, P. (1969) "Some philosophical problems from the standpoint of artificial intelligence," in Melzer and Michie (eds.), *Machine Intelligence 4*, Edinburgh University Press, Edinburgh. 463-502.

MacKay, D. (in Grandy & Schick 1991). "Maximum entropy connections: neural networks." pp.237-244.

Medin, D.L., & Schaffer, M.M. (1978) "A context theory of classification learning," Psychological Review 85: 207-238.

Medin, D.L., Wattenmaker, W.D., & Hampson, S. (1987) "Family resemblance, conceptual cohesiveness, and category construction," *Cognitive Psychology* 19: 242-279.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., & Teller, E. (1953) "Equation of state calculations for fast computing machines," *Journal of Chemical Physics* 6: 1087.

Nisbett, R.E., Krantz, D.H., Jepson, D., & Kunda, Z. (1983) "The use of statistical heuristics in everyday reasoning," *Psychological Review* (90): 339-363.

Nofosky, R. (1990) "Relations between exemplar-similarity and likelihood models of classification," *Journal Of Mathematical Psychology* 34(4): 393-418.

Nofosky, R. (1988) "Similarity, frequency, and category representations," *Journal of Experimental Psychology: Learning, Memory, & Cognition* 14: 54-65.

Nofosky, R. (1984) "Choice, similarity, and the context theory of classification," *Journal of Experimental Psychology: Learning, Memory, & Cognition* 10: 104-114.

Osherson, D., Smith, E., Wilkie, O., Lopez, A., & Shafir, E. (1990) "Category based induction," Psychological Review 97(2): 185-200.

Osherson, D., Stern, J., Wilkie, O., Stob, M., & Smith, E. (1991) "Default Probability," *Cognitive Science* 15: 251-269.

Pearl, J. (1988) *Probabilistic Reasoning In Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann. San Mateo.

Pollock, J. (1990) *Nomic Probability and the Foundations of Induction.* Oxford University Press. Oxford.

Press, W.H., Flannery, B.P., Teukolsky, S.A., & Vetterling, W.T. (1988) *Numerical Recipes in C.* Cambridge University Press, Cambridge.

Rips, L. (1975) "Inductive judgments about natural categories," *Journal of Verbal Learning and Verbal Behavior* 14: 665-681.

Rips, L., Shoben, E., & Smith E. (1973) "Semantic distance and the verification of semantic relations," *Journal of Verbal Learning and Verbal Behavior* 12: 1-20.

Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry.* World Scientific. Teaneck, N.J.

Rissanen, J. (1987) "Stochastic Complexity," *Journal of the Royal Statistical Society,Series B* 49(3): 252-265.

Rosch, E., & Lloyd, B. (editors) (1978) *Cognition and Categorization.* Lawrence Erlbaum. Hillsdale, NJ.

Ross, L. & Anderson, C. (1982) "Shortcomings in the attribution process: on the origins and maintenance of erroneous social assessments," in Kahneman et al. 1982., pp. 129-152.

Rumelhart, D., & McClelland, J. (eds.) (1986) *Parallel Distributed Processing: explorations in the microstructure of cognition. Vol. 1.* MIT Press. Cambridge, MA.

Sakamoto, Y., Ishiguro, M., and Kitagawa, G. (1986) *Akaike Information Criterion Statistics.* D. Reidel. Dordrecht.

Salmon, W. (1984) *Scientific Explanation and the Causal Structure of the World.* Princeton University Press, Princeton, NJ.

Schwarz, G. (1978) "Estimating the Dimension of a Model," *Annals of Statistics* 6(2): 461-464.

Shafer, G. & Tversky, A, (1985) "Languages and designs for probability judgment," *Cognitive Science* 9: 309-339.

Shafer, G. (1976) *A Mathematical Theory of Evidence.* Princeton University Press. Princeton, NJ.

Shafir, E., Smith, E., & Osherson, D. (1990) "Typicality and reasoning fallacies," *Memory & Cognition.* 18(3): 229-239.

Shastri, L. (1988) "A connectionist approach to knowledge representation and limited inference," *Cognitive Science* 12: 331-392.

Shephard, R.N. (1958a) "Stimulus and response generalization: Deduction of the generalization gradient from a trace model," *Psychological Review* 65: 242-256.

Shephard, R.N. (1958b) "Stimulus and response generalization: Tests of a model relating generalization to distance in psychological space," *Journal of Experimental Psychology* 55: 509-523.

Shephard, R.N. (1957) "Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space," *Psychometrika* 22: 325-345.

Smolensky, P. (1986) "Information processing in dynamical systems: foundations of harmony theory," in D. Rumelhart and J. McClelland '86.

Tversky, A., & Kahneman, D. "Cumulative prospect theory: an analysis of decision under uncertainty," working paper.

Tversky, A., & Kahneman, D. (1982) "Causal schemas in judgments under uncertainty," in Kahneman et al. '82 pp.117-128.

Tversky, A., & Kahneman, D. (1973) "Availability: a heuristic for judging frequency and probability," *Cognitive Psychology* 4: 207-232.

Tversky, A. (1977) "Features of Similarity," *Psychological Review* 84: 327-352.

Tversky, A., & Gati, I. (1978) "Studies of similarity." (in Rosch et al. 1978).

Appendix A - Generation of the likelihood booklets.

There are two stages to the pseudo-random generation of a likelihood booklet: the first stage consists of choosing a set of seven mammals which will be used in the construction of all of the likelihood rating arguments. The second stage consists of the random selection of the arguments themselves. There is actually a third stage which consists of outputting a description of the physical form of the likelihood booklet in a language appropriate to a text formatting program, which will later operate on this description to produce the physical manifestation of the booklet. I will omit the description of this third stage however.

Both of the first two stages of booklet generation make reference to a qualitative internal representation of mammal similarity that is stored by the booklet generation program. The information in this representation is made use of in order to guard against "unnatural" arguments, as these were described in section II.2 of the text. I will first describe this representation.

Similarity databases

The program makes reference to four-databases organized around the abstract similarity domains of ancestral lineage, physical form, diet, and habitat. Each database represents a number of clusters of the 47 mammals. Each represented cluster expresses the intuitive idea that all of its member mammals are similar to one another with respect to the domain of that database. The set of clusters in each similarity database forms a non-overlapping exhaustive partition of the 47 mammals. If a given mammal was not considered similar to any of the other 47 mammals with respect to the domain of that database then that mammal forms its own cluster of one.

Choice of the 7 mammals

The "goal" of the random algorithm for choosing the set of 7 mammals to be used in generating the current likelihood booklet was to obtain a set of mammals having a balance of "similar" and "dissimilar" pairs relative to the 4 represented similarity aspects of ancestral lineage, physical form, diet, and habitat. The algorithm for selecting a set of mammals proceeds sequentially. If there have been k ($2 <= k <= 7$) mammals selected at the current stage then there are "k choose 2" = $k!/(2!)(k-2)!$ distinct pairs. Relative to each of the above similarity aspects, each of these "k choose 2" pairs is considered to be either similar or dissimilar. The selection algorithm seeks to select new mammals in such a way that the number of similar and dissimilar pairs are roughly balanced for each aspect. Specifically, define the "badness" of a set of k mammals as the sum over the 4 aspects of the square of the difference between the number of similar and dissimilar pairs for each aspect.

The constrained random algorithm to choose a set of 7 mammals proceeds as follows.

1. A random permutation of the 47 total mammals is generated.
2. The first two mammals on this list that met the following criteria were chosen to be included in the set of seven mammals: the pair of mammals are similar in ancestral lineage but not in physical form, diet, and habitat. The chosen mammals were removed from the current list of available choices and placed in the chosen set.
3. All of the mammals on the current list of available choices are examined to determine the "badness" of the set of mammals that would result if they were chosen for inclusion in the current set.
4. The mammal resulting in the minimum "badness" is chosen to be included in the set and removed from the available list. In case of ties, the ordering of the list is used as a tie-breaker.
5. Steps 3. and 4. are repeated until a set of 7 mammals is arrived at. This set is examined to insure that for every aspect, there are no less than 6 similar and 6 dissimilar pairs (out of a possible 21). If the current set passes this test it is accepted and the algorithm terminates, otherwise the algorithm returns to step 1. and the process repeats.

Formation of the arguments

A single argument is specified by the following set of choices:

i. an unfamiliar property aspect
ii. a conclusion mammal
iii. the number of mammals in a set of premise mammals that are considered "similar" to the conclusion
iv. the choice of the mammals to be included in the set described in choice iii.
v. the number of mammals in a set of premise mammals that are considered "dissimilar" to the conclusion.
vi. the choice of the mammals to be included in the set described in choice v.
vii. a choice is made of whether the set of similar premise mammals are to appear as the positive premises (those mammals having the property) or the negative premises (those mammals not having the property).

Each argument was formed by a process of making each of these choices randomly in the order that they appear. The distribution of each of these choices was uniform subject to the obvious constraints of availability. In the representation scheme of the argument generating algorithm, each of the unfamiliar aspects {bone structure, dentition, digestion, thermal regulation, fluid regulation} is paired with one of the familiar similarity aspects {physical form, diet, habitat}. The pairings were (bone structure, physical form), (dentition, diet), (digestion,diet), (thermal regulation,habitat), (fluid regulation, habitat). In step i. of the algorithm above, an unfamiliar property aspect is chosen at random. Given the choice of the unfamiliar property aspect, the specific determination of the predicates "similar" and "dissimilar" which were mentioned in steps iii.-vii., is determined according to the qualitative similarity cluster representation of the familiar aspect that is paired with the unfamiliar aspect chosen for that argument. The argument set algorithm

constructs a total of 60 arguments by sequentially making choices i.-vii. and then checking that the argument form so chosen is not too much like an argument form already chosen. Two argument forms are considered too much alike if choices i-vi were identical. Finally, the order of premises within each argument and the order of arguments in the total set were randomly permuted

The Logic of This program

In section II.2 I showed an example of an "unnatural" argument. It was stated that this type of argument should not appear in the likelihood booklets, though no definition for what exactly constituted an unnatural argument was offered. The likelihood booklet generation program just described embodies, in effect, what I take to be a form of sufficient criteria for determining that an argument will not be unnatural. This criteria is that, for me, either all of the positive premise mammals or all of the negative premise mammals are similar to one another and the conclusion mammal in some way that is relevant to the unfamiliar property aspect. Precautions were taken to insure that this sufficient criteria did not go too far, and make every argument a "ringer" with, for example, all of the positive premise mammals and the conclusion mammal as felines, and all of the negative premises as rodents. One of these precautions was that the set of seven mammals was guaranteed to contain a pair of mammals that were biologically related and which would frequently appear as premises of the opposite polarity (this precaution was carried out by step ii. of the mammal selection algorithm). In general, the program described above tried to produce booklets as randomly as possible while steering between the extremes of unnaturalness and bluntness.

Appendix B- Details of Error Model fitting and evaluation

The error models tested for experiment-I describe distributions on the 10 bin histogram 0-9, 10-19,...,90-100. The data points in each bin represent the number of discrepancies between related pairs that had an absolute difference in that range. The data from each of the twenty subjects was considered separately. The actual quality of the fit was determined by considering the match between model predictions and the observed counts of related pair discrepancies in the 0-9,10-19, and in a composite 20-100 bin. The quality of the fit was determined by a chi-square test with two degrees of freedom, and the fit was deemed unacceptable if the match fell on the upper 5% of that distribution.

The parameters of the model distribution were fit according to the following procedure. Every model was required to precisely predict the observed expected square difference between a given subject's related pairs of judgments from the first and second session. A noise model itself is considered to be the noise distribution for a single judgment for some given argument, distributed around the mode or the expected value of that argument. If the noise is zero-mean additive then the prediction made by the model about the quantity of the expected square discrepancy between related pairs of argument ratings is equal to twice the canonical variance. For the two models which did not have a mixing component this meant that their variance was set to precisely one half the observed value of this quantity.

For the models with a mixing component of ∂, the value of the prediction that they make about the expected square discrepancy between related pairs will be equal to $(1-\partial)^2 \cdot 2VAR(N) + (1 - (1-\partial)^2) \cdot Var(J)$ where VAR(N) is the variance of the additive component and VAR(J) is the variance of the judgments themselves. Since this formula is to be equated to the observed square discrepancy between related pairs, we can solve for VAR(N) as a function of a given ∂. The procedure that was adopted for fitting the two parameters of this mixture model was to perform a one-dimensional search for the value of ∂ which provided the best fit, the value of VAR(N) being determinate for any given ∂. The quality of any given fit was determined by the chi-square distribution mentioned above.

Appendix C - Derivation of the Correlation Formula For Mixtures

The correlation between a set of judgments and predictions is given by the formula 1):

1) $E[(J_i - E[J_i])(P_i - E[P_i])] / (VAR(J) \cdot VAR(P))^{1/2}$

Assuming that $E[J_i] = E[P_i]$ this formula reduces to 2):

2) $(E[J_iP_i] - E[J]^2) / (VAR(J) \cdot VAR(P))^{1/2}$

Let the notation M_i (for the "Mode" of judgment i) be a value such that the distribution of J_i as a random variable is equal to a mixture of a) a $(1 - \partial)$ part that is equal to M_i plus a zero mean independent random noise variable with variance V(N) and b) an ∂ part that is an independent random variable with the same distribution as the collection of judgments. Then the expectation of J_i is given by line 3).

3) $E[J_i] = P_i = (1 - \partial)M_i + \partial E[J]$

The variance of P can now be computed as 4).

4) $VAR(P) = E[(P_i - E[P])^2]$
$= E[((1 - \partial)M_i + \partial E[J] - E[P])^2]$
$= E[((1 - \partial)M_i + \partial E[J] - E[J])^2]$
$= (1 - \partial)^2 E[(M_i - E[J])^2]$
$= (1 - \partial)^2(VAR(J) - VAR(N))$

The numerator of line 2) is now re-written as

5) $(1 - \partial)E[(M_i + N)((1 - \partial)M_i + \partial E[J])] + \partial E[J]^2 - E[J]^2$
$= (1 - \partial)(E[(M_i + N)((1 - \partial)M_i + \partial E[J])] - E[J]^2)$
$= (1 - \partial)((1 - \partial)E[M_i^2] + \partial E[M_i]E[J]) - E[J]^2)$
$= (1 - \partial)^2(E[M_i^2] - E[J]^2)$
$= (1 - \partial)^2(VAR(J) - VAR(N))$

and the denominator of line 2) can be written as

6) $(1 - \partial)(VAR(J)(VAR(J) - VAR(N)))^{1/2}$

Recombining 5) and 6) after cancelling related terms gives

7) $(1 - \partial) [(VAR(J) - VAR(N)) / VAR(J))]^{1/2}$

Appendix D -
The free parameter calibration procedure for the maximum entropy estimator

In this appendix I describe how the 25 "free" parameters are actually adjusted to fit the data provided by 21 judgments of overall similarity and 59 rated arguments (= 60 - 1 "left out"). The basic procedure starts out with a beginning value of these free parameters and adjusts them to minimize an error function. The adjustment stops when no more improvement in the error function is being produced. The basic error function is as follows.

1) Error of total fit = sw•"Error of similarity it" + "Error of likelihood fit".
where sw is the meta-parameter mentioned above.

The meta parameter is a fixed constant throught the computation being currently described. This error is defined in terms of the seperate errors for similarity matching and likelihood rating matching.

2) Error from similarity fit = Σ_{ij} D((a•Sim(i,j) + b) - SIM(i,j,OVER))
where D is a differentiable error norm to be discussed momentarily, a and b are two of the free parameters, Sim(i,j) is the "true" similarities which are a function of the current state of the model, and SIM(i,j,OVER) is the current subject's rating of the overall biological similarity of mammal pair (i,j).

3) Error from likelihood fit = Σ_k D((c• Condit(arg_k)+d) - Rating(arg_k))
where D is as before, c and d are two of the free parameters, and Condit(arg_k) is the condtional probability for argument k derived from the current model's version of pr.

Observe that if D is a differentiable function, then partial derivatives of the total error function with respect to each of the 25 free parameters will be differentiable if and only if Condit(arg_k) and Sim(i,j) are differentiable with respect to the other 21 paramets. The nature of the other free parameters is as follows. There exist 21 constants c_{ij} and a special constant c_0 such that

1) $pr(.be._k) = c_0 \cdot EXP[\Sigma_{ij}\ c_{ij} \cdot f_{ij}(.be._k)]$ and

2) $1/c_0 = \Sigma_k \, EXP[\ \Sigma_{ij}\ c_{ij} \bullet f_{ij}(.be._k)]$

The special constant c_0 is not a free parameter but the c_{ij} are. Given the 21 c_{ij}, the values of $pr(.be._k)$ for the 2^7 are determined. Given the values of the 2^7 $pr(.be._k)$, $Condit(arg_k)$ is determined for each k by the conditionalization formula. Therefore, the partial derivative of $Condit(arg_k)$ with respect to each of the $pr(.be._k)$ exists and the partial derivative of each of the $pr(.be._k)$ with respect to the c_{ij} from formula 1) and 2). So by several applications of the "chain rule", the partial derivative of the error of the likelihood fit with respect to each of the c_{ij} exists. A given set of values for the 2^7 $pr(.be._k)$ also determine the Sim(i,j). Specifically, the Sim(i,j) involve summing up a certain set of $pr(.be._k)$. Therefore, the partial derivative of the error of the similarity fit with respect to the c_{ij} exists and therefore so do the partials of the overall error of fit with respecto the c_{ij}.

The algorithm which was used to adjust the free parameters from a starting value to a final value was an algorithm commonly known as "conjugate gradient descent". A program implementing this algorithm and a description of why it works can be found in Press et al. '88, pp. 309-323.

Two passes were made with the minimization algorithm. In the first pass, the error norm used (the value of the variable 'D' above) was simply the square of the discrepancy. After the minimization routine had run its course using this error norm, it was restarted from its final value using the following error norm obtained from the paper by Girosi et al. '90.

$D(x) = x^2 - (1/\beta) \ln(1 + EXP[\ \beta x^2 - K\ ]\ )$ where β and K are fixed constants.

This error norm is particularly appropriate when dealing with outliers such as the uncorrelated "mixture judgments" described in the analysis of experiment I. It is shaped like a gutter. At the bottom it curves like a cup but after rising for a while the curve levels off and becomes flat. Error norms with this shape protect against the contingency that outlier data points will have a large effect on judgments. If ∂ is the percentage of these "mixture" judgments or outliers, then the parameter K should be set to $\ln((1 - \partial)/\partial)$. The value of K that was used in experiment II was given by this expression for a ∂ of 0.1, something like a median value from the

analysis of experiment I. Given a choice of K, the value of β can be chosen so as to determine where the curve levels off. The derivative of the D(x) with respect to x is given by the following equation.

$$D'(x) = 2x(1 - (EXP[\beta x^2 - K] / (1 + EXP[\beta x^2 - K])))$$

It can be observed that as x grows large, this term goes to zero. A value of β was chosen so that, roughly, D(x) reaches 90% of its asymptotic maximum at $x = \pm 20$ (relative to the 100 point scale).

The only thing left to specify is the starting values for the free parameters. The parameters a and c are started at 1.0. The parameters b and d are started at 0.0. The starting values of the c_{ij} were obtained from the values of the SIM(i,j,OVER) by linear transformation. There initial values were as follows

$$c_{ij} = \Sigma_{ij} 3.0 (SIM(i,j,OVER) - 1/21\Sigma_{ij} SIM(i,j,OVER)).$$

A theoretical analysis suggesting that the c_{ij} might be expected to have something like these values may be found in the paper MacKay '91.

3409-46