International Journal of General Systems, 3(3), 145-174, 1977

System Identification, Approximation and Complexity

Brian R. Gaines
Man-Machine Systems Laboratory
Department of Electrical Engineering
University, of Essex, Colchester, Essex, U.K.

This paper is concerned with establishing broadly-based system-theoretic foundations and practical techniques for the problem of system identification that are rigorous, intuitively clear and conceptually powerful. A general formulation is first given in which two order relations are postulated on a class of models: a constant one of *complexity;* and a variable one of *approximation* induced by an observed behaviour. An *admissible model is* such that any less complex model is a worse approximation. The general problem of identification is that of finding the *admissible subspace* of models induced by a given behaviour. It is proved under very general assumptions that, if deterministic models are required then nearly all behaviours require models of nearly maximum complexity. A general theory of approximation between models and behaviour is then developed based on subjective probability concepts and semantic information theory The role of structural constraints such as causality, locality, finite memory, etc., are then discussed as *rules of the game.* These concepts and results are applied to the specific problem or stochastic automaton, or grammar, inference. Computational results are given to demonstrate that the theory is complete and fully operational. Finally the formulation of identification proposed in this paper is analysed in terms of Klir's epistemological hierarchy and both are discussed in terms of the rich philosophical literature on the acquisition of knowledge.

1 Introduction

The problem of inferring the structure of a system from observations of its behaviour is an ancient one with many ramifications. The literature on the subject is vast, having its roots in the philosophical problems of the nature of *reality* and our inference of it from sensations (Locke, 1690; Berkeley, 1710; Kant, 1781). Plato's (380BC) famous simile of the prisoners in the cave who, "like ourselves . . . see only their shadows, or the shadows of one another, which the fire throws on the opposite wall of the cave", epitomizes the inherent uncertainty and tenuous nature of inference processes that we all perform and take for granted in our everyday life.

Even if the metaphysical problems of *reality* are discarded for a more pragmatic approach that asks, not whether our inferred structures are *real*, but instead whether they at *work*, i.e. are useful, or valid in some weaker sense, deep philosophical problems remain. If our model structure accounts only for the observed behaviour on which it is based then it appears to suffer from the usual defect of deductive inference, that it is inferentially vacuous leaving us with only a re-description of the observations. Whereas if we demand that our structures be *predictive,* allowing us to account for further observations**,** then we come up against Hume's (1777; Popper, 1972) conclusive arguments that inductive inference can never be validated (at least, neither deductively nor inductively).

Much of the literature on the "philosophy of science" is concerned with the epistemological problems stemming from Hume's arguments. Reactions range from existentialist dread (Heidegger, 1949) emphasizing the personal and unique nature of our experience and evaluation of it, to the totally impersonal formal methodologies of incremental data acquisition and model structure adjustment of Carnap's logical probability and confirmation theory (Carnap, 1950; Carnap, 1952; Schilpp, 1963; Carnap and Jeffrey, 1971; Erwin, 1971; Swinburne, 1973).

Within this spectrum one has a range of existentialist positions (Blackham, 1961); the deep methodological studies of Brentano (1973) and Husserl (1965) leading to a variety of *phenomenological* analyses (Pivcevic, 1975); arguments that observational descriptions of phenomena are already *theory-laden* (Hanson, 1958; Polanyi, 1958; Gale and Walter, 1973); logical analyses of the role of *analogy* in model formation (Hesse, 1966; Dorrough, 1970; Uemov, 1970), of *simplicity* and *economy* in model selection (Post, 1960; Blackmore, 1972; Sober, 1975), and of *convention* in model utility (Lewis, 1969); the demonstrations that *deduction* is itself open to Hume's criticisms (Dummett, 1973; Haack, 1976); the detailed analysis of *specific flaws* in the texture of Hume's argument (Madden, 1971; Stove, 1973); the aphoristic *metaphysical reply* to Hume given by Wittgenstein (Dilman, 1973); the various *indications* of induction developed by Reichenbach (1949), Harrod (1956), Katz (1962), Black (1970), Rescher (1973), and others (Swinburne, 1974); Popper's (1959; 1963; Schilpp, 1974; Putnam, 1975) methodological point that hypothesized structures can never be verified, only *falsified,* and the Popper-Carnap controversy (Michalos, 1971) over falsification versus confirmation; the sociological models of scientific *revolutions* of Kuhn (1962) and the painstaking studies of how these social functions actually operate of Merton (1973); the more *structural* rationale of scientific method of Lakatos (1970) and Hesse (1974): the *anarchistic* counter examples of Feyerabend (1975), and the coolly cynical appraisal of the *arbitrariness* of the whole debate by Gellner (1974).

Many aspects of these philosophical debates find a more mathematical formulation in studies of the history (Hacking, 1975) and foundations of *probability theory* (Reichenbach, 1949; Savage, 1954; Foster and Martin, 1966; Kyburg, 1970; Ruzavin, 1970; Carnap and Jeffrey, 1971; De Finetti, 1972; Fine, 1973; Maxwell and Anderson, 1975), *semantic information theory* (Bar-Hillel and Carnap, 1953; Bar-Hillel, 1964; Hilpinen, 1970; Hintikka, 1970), *decision making and statistical inference* (Jeffrey, 1965; Levi, 1973; Menges, 1974), *computational complexity* (Martin-Lof, 1966; Kolmogorov, 1968; Willis, 1970; Schnorr, 1971; Chaitin, 1975; Lempel and Ziv, 1976) and *grammatical inference* (Feldman, 1972; Patel, 1972; Maryanski, 1974; Fu and Booth, 1975). The philosophical debate has also become rather more pointed in recent years because it has been possible to apply the arguments *operationally* to machines (Putnam, 1964) that exhibit reasoning by analogy (Kling, 1971), law discovery (Newell and Simon, 1972; Simon, 1973), inference from incremental observations (Solomonoff, 1964; Klir, 1975), expectation (Nelson, 1975), and many other manifestations of human learning behaviour (Andreae and Cleary, 1976).

With this background in mind one treads warily in the development of behaviour-structure inferencing techniques in system theory. Many of the philosophical problems may be evaded by assuming that the observed behaviour arose from one of a set of known possible systems. For example, Zadeh (1962) defines *system identification* as "the determination on the basis of input and output, of a system within a specified class of systems, to which the system under test is equivalent". Most practical studies of system identification (Eykhoff, 1974) operate within this framework and presuppose both a "specified class of systems" and a well-defined decision procedure for determining that a system is "equivalent" to one of these systems on the basis of its input-output behaviour.

I shall adopt this point of view in this paper and give a general systems theoretic formulation of the identification problem that encompasses previous specific formulations and yet is sufficiently

precise for some interesting features of identification to be determined. In particular the formulation deals very effectively with a major problem left open in the definition above—that of defining "equivalence" when the class of systems considered is acausal in some sense, e.g. non-deterministic or stochastic automata. The input-output behaviour of an acausal system is *not* uniquely related to its structure and some element of approximation (Wharton, 1974) is essential in the definition of equivalence. The formulation proposed allows for this in a very clear and general form that enables, for example, the problem of identifying stochastic automata, or grammars, to be rigorously defined and solved (Gaines, 1975a; Gaines, 1976a; Gaines, 1976d).

Out of the general formulation come two specific lines of development:-

a) How can the techniques be applied to specific classes of system and what theoretical and practical results may be obtained, i.e. given that the source of behaviour is a deterministic finite state automaton (DFSA) or a stochastic finite state automaton (SFSA) what are an appropriate definition of equivalence, and a practical implementation, and what results may be obtained about them, theoretically and experimentally?

b) How should the class of model systems be specified and what happens if the observed behaviour arises from a system *not* in this class? e.g. will the identification svstem behave reasonably and give some approximate and useful results, or does it fail completely. Here there are some surprising results that run counter to our (current) intuitions.

Both aspects of the identification problem are developed in this paper. In particular, the problem of identifying stochastic finite state automata and grammars is analysed theoretically and practical experimental results are given. These are closely related to the literature on computational complexity already noted. but have additional interest because they also establish close links with *subjective* foundations of probability (Savage, 1954; Smith, 1961; Smith, 1965; Aczel and Pfanzagl, 1966; Shuford, Albert and Massengill, 1966; Winkler and Murphy, 1968; Savage, 1971; Winkler, 1971; De Finetti, 1972; Hogarth, 1975; Shuford and Brown, 1975), in particular Pearl's (1975a; 1975d; 1975c; 1975b) recent results on economic bases for subjective probability and relations between approximation and complexity. Also the results of using classes of model systems different from the system observed are analysed and discussed for a variety of cases. The various aspects of the problem are also linked to the rich philosophical literature on the acquisition of knowledge referenced above This is made more explicit and precise than was previously possible by using the hierarchical classification of epistemological levels recently proposed by Klir (1976) as a framework for both the philosophical connotations and the identification techniques.

This paper is one of a series concerned with establishing broadly based system-theoretic foundations and practical techniques for the problem of system identification that are rigorous, intuitively clear and conceptually powerful. An informal introduction to the techniques and a range of experimental examples have been given in (Gaines, 1976a), and studies have been reported of applications to the inference of finite-state probabilistic grammars (Gaines, 1976d). It is expected that later papers will report particular applications to the analysis of sequential behaviour in humans and animals. The role of this paper is to provide the background and foundations for such studies.

2. A General Formulation of the Identification Problem

Our problem can be stated initially to be: "given a sample of the behaviour of some system, to determine the optimum model from some prescribed models that would account for its. Note that Zadeh's definition of the previous section has already been generalized. We do not need at this stage to define what is meant by *behaviour* the terms of reference by which we describe it, etc. In particular. the notions of *input* and *output* have been dropped. These are structural concepts that belong to the models *not* the behaviour, and in the examples given it will be shown that the input/output distinction can be *inferred* and need not be postulated. A similar remark has been made by Klir (1975) who calls such systems without the input/output distinction *neutral.*

In addition the notion of *equivalence* between behaviour and structure has been dropped. This, in the sense of a mathematical equivalence relation, is too powerful a notion for a theory that must encompass the modelling of acausal systems. We need instead some concept of *degree of approximation,* an order relation that allows us to say that one model accounts for the behaviour *better* than does another. When the more powerful equivalence *does* exist, for example, in modelling the behaviour of deterministic finite-state sources with DFSA then it allows for the elegant category-theoretic formulation of identification in terms of an adjunction between behaviour and structure developed by Goguen (1973; 1975), Arbib (Arbib and Manes, 1974; Bobrow and Arbib, 1974), and Ehrig (Ehrig and Kreowski, 1973; Ehrig, 1974).

However, determinism is a myth in the real world, although the success of mechanistic models of the universe in the eighteenth century has made it something of a holy grail for science until the present day and there is substantial evidence that the assumption of deterministic causality is deeply rooted in both human cognitive and perceptual processes (Gaines, 1976f). Many authors have argued for the replacement of deterministic causality with *probabilistic causality* (Rescher, 1970; Mackie, 1974; McClelland, 1975; Suppes, 1984). I have proposed elsewhere (Gaines, 1976b) that the process in science that Carnap (1950) calls *precisiation,* of replacing a phenomenal *explicandum* with a noumenal *explicatum* need not be interpreted in the narrow sense of precise deterministic explicata, and have demonstrated (Gaines, 1976e) that this is not only a metaphysical point but also a practical one—*the universe becomes incredibly complex and our models of it nonsensical if we assume determinism in the face of even a slight trace of acausal behaviour.*

2.1 Complexity and Admissibility

Having weakened *equivalence to* an order-relation *of approximation,* we face one residual problem with the definition above, that the concept of an *optimum* model is no longer appropriate. For example, imagine that you and I are each given the same sample of behaviour and asked to model if from the same class of models. "My model is a better approximation," I say, "Ah," you reply, "but mine is a far *simpler* model. Indeed, I am not sure that all yours does is not just to retain a *memory* of the behaviour, whereas mine, whilst a worse approximation, is clearly a far better representation. If the behaviour were actually being generated by the system corresponding to my model the degree of approximation I have achieved would be quite reasonable."

These are the key issues, that we do not rate all models as themselves equivalent. There is invariably an order relation on our prescribed class of models that gives rise to a trade-off between the degree of approximation and the preference for models. It is common to call this

ordering one of *complexity* with the preference being for the less complex models. For convenience I shall adopt this terminology, but with the warning that *the order relation is not intrinsic to the class of models.* We may both adopt the same class of models but what I regard as complex may for you be simple. The ordering of models in terms of complexity is arbitrary and depends upon our individual points of view.

Too much stress cannot be laid upon the fact that our model classes are incompletely specified for purposes of identification until we have defined an ordering upon them. It is a trap into which we may easily fall, particularly in general systems theory. Several classes of models open to us appear so general that we feel they must be adequate to account for *all* possible behaviours. And so they are, but that is not sufficient to allow us to suppose that we can base an all-embracing system science upon this class of systems. When we specify our order relation upon the models we may find that the behaviours of many important systems require complex models under our ordering, whereas, with a different ordering on the same class of models, they all become simple.

If this happens then it is probable that there will be a *scientific revolution* in which the order relation on our models is changed to make the observed world less complex. Since we normally wish to associate the ordering with some intrinsic feature of our models this will also lead to us viewing the models in a different way so as to emphasize some new aspect of their structure. For example, clearly every finite sample of behaviour (which is all we ever have) can be accounted for by a DFSA. However, so can it by an SFSA or a Turing machine—why choose one rather than another? The results of (Gaines, 1976e) show that the behaviour of a simple stochastic FSA requires, in general, a very complex deterministic FSA model, and, likewise, the behaviour of a simple push-down automaton requires very complex models with both DFSA and SFSA.

Given order relations of: complexity on our class of models, and of *approximation* on the extent to which a model accounts for a given observed behaviour, it is possible to give a precise formulation of identification in a general systems context: "given a sample of the behaviour of some system, to determine those models from some prescribed class of models that are *admissible* in that, for each one, any other model that gives a better approximation in accounting for the behaviour is more complex." The concept of admissibility is one borrowed from statistics (Weiss, 1961) and proves a powerful one in problems of control theory (Kwakernaak, 1965) and pattern recognition (Gaines and Witten, 1977) in situations where no definition of optimality is possible. Note that there is rarely a unique admissible model but instead a subset containing several admissible models. However, this subset has some interesting properties that do much to make up for the lack of a single unique model.

In the next subsection I shall put this definition of identification into mathematical form, go on to analyse some of its properties, particularly those related to computational complexity and then develop more specific features of it related to particular classes of identification problems.

2.2 Mathematical Formulation of Identification

The concepts developed in the previous sections may be formulated more formally in terms of: a set of possible observed behaviour, B; a set of models, M; the pointed monoid, $(\mathbf{Ord}_M, \leqq)$, of all order relations on M with one specified relation, $\leqq$, singled out; and a mapping, $f: B \rightarrow \mathbf{Ord}_M$, from the set of behaviours, B, to the set of order relations on M, $\mathbf{Ord}_M$. The quadruple $(B, M, \leqq, f)$, defines an *identification space.* The relation $\leqq$ is one of model complexity and if $m, n \in M$ are such that $m \leqq n$ we shall say that the model m is not more complex than n. Other considerations

being equal it will be assumed that the least complex possible model is preferred. Note. however, that $\leq$ may be only a partial order so that, in general, there will be a *set* of minimal models rather than a unique minimum. The mapping f is determined by the further order relation of approximation that each behaviour induces on the set of models. We shall write for $b \in B$, $\leq_b = f(b) \in \mathbf{Ord}_M$, and if $m, n \in M$ are such that $m \leq_b n$ we shall say that model m is not a worse approximation to behaviour b than is model n. The best models for b are thus those minimal in the order relation, $\leq_b$ which again need not be more than a partial order.

Now we are in a position to define a solution of the identification problem in terms of the product of the two order relations, $\leq_b^*$

$$\forall\, m, n \in M,\ m \leq_b^* n \Leftrightarrow m \leq n \ \text{ and } \ m \leq_b n \tag{1}$$

i.e. $m \leq_b^* n$ if and only if m is neither more complex nor a worse approximation than n. The minimal elements of the new order relation have the property that there are no other models that are both less complex and a better approximation than them. Even if both $\leq$ and $\leq_b$ are total orders it is likely that $\leq_b^*$ will be a partial order (we can trade more complexity for better approximation) and hence there will be in general no unique minimum model. The minimal elements are all *admissible* (Weiss, 1961) solutions to the identification problem because they cannot be decreased in complexity without worsening the approximation and cannot be improved in approximation without increasing complexity. Thus we may define the solution of the identification problem for a space *(B, M, ≤, f)* and an observed sequence $b \in B$ to be the *admissible subspace* determined by b, $M_b \subset M$, such that:

$$M_b \equiv \{m: \forall n \in M,\ n \leq_b^* m \Rightarrow m \leq_b^* n\} \tag{2}$$

i.e. if any model is better than one in M_b then it is equivalent to it under the order relation $\leq_{b^*}$, the usual requirement for minimality. Thus models in M_b are either equivalent or incomparable in terms of $\leq_b^*$ Note that, for rigour when dealing with infinite sets we should impose the constraint that the minimal elements under $\leq_b^*$ exist and belong to M, e.g. by taking a suitable closure.

M_b may be shown to have similar, but inverse, structures under $\leq$ and $\leq_b^*$:

Result 1. For comparable models, the relations $\leq$ and $\leq_b$ are antitone in M_b .i.e. $\forall m,n \in M_b$:*m* and *n* are comparable under $\leq$ and $\leq_b$, $m \leq n \Leftrightarrow n \leq_b m$

Proof Assume $m \leq n$—if $m \leq_b n$ then $m \leq_b^* n$ so that, since $n \in M_b$, $n \leq_b^* m$ which implies $n \leq_b m$. Conversely, assume $m \leq_b n$—if $m \leq n$ then $m \leq_b^* n$ so that, since $n \in M_b$, $n \leq_b^* m$ which implies $n \leq m$. Note the symmetric way in which the relations of complexity and approximation have entered the discussion and definitions, and the anti-symmetry between them in the admissible subspace of models shown by this result. Models in the admissible subspace are effectively unordered by $\leq_b^*$ but ordered in effectively identical, but inverse, ways by $\leq$ and $\leq_b$.

2.2.1. *Example—identification of deterministic automata* It is these well structured subspaces that replace the unique solutions of, for example, problems amenable to the Nerode equivalence, (Nerode, 1958; Arbib, 1969) and it is of interest to see how such problems fit within the current framework. Consider the identification space *(D*,M, ≤ ,f)*. where D^* is the free monoid generated by a (finite) alphabet, D, of atomic *descriptors* (inputs or outputs); M is the set of irreducible, Mealy, finite state deterministic automata with a specified initial state having inputs

and outputs in D; the ordering of complexity, $\leq$, is determined by the number of states of the automata, so that $m \leq n \Leftrightarrow$ number of states of m less than or equal to the number of states of n; and the ordering $\leq_b$, induced by a sequence of observed input-output behaviour, $b \in D^*$, *is* in fact a binary classification in which m is maximal in the order unless it generates the sequence b exactly when it is minimal (where m starts in the specified initial state, receives the input sequence imbedded in b and emits an output sequence that is tested against that imbedded in b).

This is the standard case of deterministic automata inferencing from a sample of the input-output behaviour and the Nerode equivalence. or one of its generalizations, may be used to determine a minimal-state machines (Rabin and Scott, 1959) using high-speed algorithms (Hopcroft, 1971). This essentially splits the space of possible machines into three sets:

1) a unique machine (up to isomorphism) that is minimal in the ordering of approximation (an exact fit) and, subject to this, minimal in the ordering of complexity (minimal state).
2) a set of machines with fewer states than this that are maximal in the ordering of approximation, i.e. are simpler but do not fit the behaviour.
3) the remaining machines with more states, or with the same number of states that do not fit the behaviour.

The first two sets of machines together form the admissible subspace, M_b , for the problem. In this example the second set of machines is of little interest because there is no gradation of approximation. It would be possible to define a graded form of approximation in terms of some finer evaluation of the extent to which the outputs of a model, m, diverge from those of the behaviour, b. However, in conventional deterministic modeling it is the uniqueness of the "solution" obtained in the first set of admissible models that is of interest.

2.2.2. Example—identification of probabilistic automata The problem of identifying probabilistic automata will be treated in more detail in sections 3.3.1 and 4. However, it is useful to contrast the techniques I have previously described (Gaines, 1975a; Gaines, 1976a) with those for deterministic automata above. The main difference between the two problems is that a probabilistic automaton model gives not a specific output but instead a *distribution over possible outputs,* and a distribution over possible next states. We can evaluate the distribution over outputs with respect to the actual observed output by using one of the loss functions devised to elicit subjective probabilities (Aczel and Pfanzagl, 1966; Shuford et al., 1966; Winkler and Murphy, 1968; Savage, 1971; Winkler, 1971; Hogarth, 1975; Shuford and Brown, 1975), e.g. a loss of minus the log of the proposed probability for the output that actually occurs (this is zero if the actual output is predicted with probability 1 and positive otherwise). We can eliminate the effect of having only a distribution over next states by using *observable* automata only in which the actual output that occurs is sufficient to resolve the uncertainty as to the next state.

Consider the identification space, $(D^*, M, \leq, f)$, where: D^* is a free monoid as before; M is now the set of irreducible, observable, Mealy, finite-state probabilistic automata over D with a specified initial state: the ordering of complexity, $\leq$, is number of states as before; and the ordering, $\leq_b$, induced by a sequence of observed input-output behaviour, b, is determined by the natural numerical ordering on the sum of the losses when a model, m, is used to predict b (where m starts in the specified state. receives the input sequence imbedded in b and emits probability distribution over the outputs that are used in conjunction with the actual output to determine both the loss and the next state). The smallest loss gives the best approximation and zero loss

(minimum possible) corresponds to exact deterministic prediction of the outputs and hence to the deterministic modelling already discussed.

The admissible subspace for probabilistic identification does not split trivially as it did for deterministic modelling. For a given number of states there will generally be models that give a smaller loss (better approximation) than any models with fewer states. As the number of states in the model (the complexity) increases the loss will get less until it eventually becomes zero and a deterministic model has been found. However, I have shown elsewhere (Gaines, 1976e) that this (maximum-state or best approximation) admissible model, with a truly random source, will have on average about the same number of states as the number of observations (length of behaviour, *b)* and is a structurally meaningless *memory* of the observations.

It is also now of interest to look at the other extreme, not the maximum-state admissible model (perfect fit) but the minimal-state admissible model with, in fact only one state. A 1-state model can predict only a constant distribution over the descriptors, say μ (*d)* for $d \in D$, where:

$$\sum_{d \in D} \mu(d) = 1 \tag{3}$$

and, if there are *k(d) d's* in the behaviour *b*, the total loss will be:

$$P = -\sum_{d \in D} k(d) \log(\mu(d)) \tag{4}$$

which is well known to be minimized (Mathai and Rathie, 1975) when:

$$\mu(d) = k(d)/k \tag{5}$$

where:

$$k = \sum_{d \in D} k(d) \tag{6}$$

The mean expected loss under these conditions is:

$$(P / k)_{\max} = -\sum_{d \in D} \mu(d) \log(\mu(d)) \tag{7}$$

which is the (zero-order) Shannon entropy for the distribution.

Result 2. If we plot the approximation against the complexity for the admissible models we get a monotonically falling graph that intersects the abscissa (minimum loss) at about the length of the observed behaviour if it is a Bernoulli sequence, and intersects the ordinate (minimum states) at an estimate of the entropy of the observed behaviour if it is a Bernoulli sequence.

It is this first condition of maximal complexity. that gives an operational definition of the concept of *randomness of* even a single sequence (Kolmogorov, 1968) within the framework of computational complexity (Willis, 1970; Schnorr, 1971; Chalfan, 1986). It is the second measurement, of entropy, that is conventionally taken as a measure of the randomness of a sequence. Note, however. that the sequence need not be Bernoulli (zero-memory) for the first criterion (maximum complexity) to apply. The later discussion and results will show then that the shape of the entire plot of approximation against complexity for the admissible subspace, rather than just the intercepts on the axes, may be used to analyse the randomness *and the structure* of the sequence of behaviour.

In the next main section I shall first generalize the first part of the result above to show that the *expected complexity* of a randomly generated behaviour is nearly equal to the *size* of that behaviour under very weak assumptions, and, secondly, in relation to the second part of the result, I shall derive measures of approximation with certain convergence properties that give rise to an expected loss that is the *entropy* of a behaviour, again in the general case under very weak assumptions.

3. Complexity, Approximation and Partitioning of D-Sets

In considering the identification of sequential systems the observations are essentially ordered in time and it is natural to consider them to be some subset of a free monoid of atomic descriptions (as was done in sections 2.2.1 and 2.2.2). I shall make this assumption in sections 3.3.2 and 4 to focus upon some specific inference problems. However, important results may be derived before any specific structure upon observations is assumed, and, in the present section, no structure will be postulated. The concepts developed and results obtained are thus applicable to problems of identifying systems other than automata, e.g. to problems of optical pattern recognition where there is a spatial rather than a temporal coherence between observations, or to the problem of reconciling multiple observers of a system where these may be only a partial order on observations from different sources.

We will take the *behaviour* of a system to be a mapping, $b: E \rightarrow D$, from a (finite) set of *events, E*, to a set of *descriptors, D.* The event space will normally have some algebraic structure, such as an order relation, upon it. Note, however, that knowledge of this structure (if it exists) is not necessary to the results of the following section (it will not be introduced until section 3.3), and that the term *event* is not intended to have necessary temporal connotations, e.g. an event might be a configuration of surface elements making up a picture and the mapping from events to descriptors might specify the reflectance and hue of each element. It is convenient to adopt Goguen's (1974) neat terminology for such mappings and call the behaviour a D-set with E as its *support* and D as its *truth-set.* This establishes an important link to our other studies of the logic of automata (Gaines and Kohout, 1975), and the possible logical, algebraic, topological and arithmetic foundations of automata theory. In particular it establishes a link to the wide range of results on fuzzy and probabilistic systems and the relationships between them (Goguen, 1974; Arbib and Manes, 1975; Kaufmann, 1975; Zadeh, 1975; Gaines, 1976b; Gaines, 1976c; Zadeh, 1976; Gaines and Kohout, 1977). Such structural considerations will be touched on only briefly in this paper (section 5.1) but form an important direction in which to extend the results.

3.1 General Results in Complexity

The first part of Result 9 of the previous section may now be derived with less assumptions so that it applies to the more general formulation of system behaviour given in section 2.2 and above. It is possible to generalize the enumerative technique used in (Gaines, 1976e), as Pearl (1975d; 1975c; 1975b) has done independently in conjunction with Shannon's rate distortion theory, to show that nearly all behaviours are complex given certain very weak and intuitive assumptions about the number of models with a given complexity and the number of behaviours of a given size.

Assume that the complexity of a model is an integer in the range 1 to infinity and that the numberof distinct models with complexity, C, or less is $M(C)$. Then we may show that the mean

complexity of a complete set of models up to and including those of complexity C_{max} is itself of order C_{max}.

Result 3. If $M(C)$ grows at least exponentially with C then the ratio of the mean value of C, C_{mean}, to C_{max} in a complete set of models of complexity up to and including C_{max} is asymptotic to 1.

Proof The mean complexity is given by

$$C_{mean} = \frac{1}{M(C_{\max})} \sum_{C=1}^{C_{\max}} C \times (M(C) - M(C-1)) = C_{\max} - \sum_{C=1}^{C_{\max}-1} M(C)/M(C_{\max}) \tag{8}$$

Now if $M(C)$ grows at least exponentially with C then we have:

$$M(C)/M(C-1) \geq A \tag{9}$$

for some constant A, so that:

$$\sum_{C-1}^{C_{\max}-1} M(C)/M(C_{\max}) \leq 1/(A-1) \tag{10}$$

So that:

$$C_{mean} \geqq C_{max} \text{-}1/(\text{A} - 1) \tag{11}$$

Hence, the ratio of C_{mean} to C_{max} is asymptotic to 1.

Exponential growth is a common feature of most model sets, such as automata, since the addition of one more state multiplies the number of possible models by at least a constant. However. it is also possible to relate the need for exponential growth in the set of models with a similar rate of growth in the set of possible behaviours. Take the *size* of a behaviour to be the number of atomic descriptors necessary to describe it, i.e. the number of events in the behaviour. and let the number of distinct behaviours of size S be $B(S)$. Suppose now that for deterministic modelling it is impossible for a given model to be an exact fit (zero approximation) to more than one behaviour, i.e. there is a mapping from models to behaviours, not necessarily 1-1. Thus we must have that the maximum complexity of models necessary for behaviours of size S is bounded by:

$$M(C_{\max}) \geq B(S) \tag{12}$$

If we now suppose that $M(C)$ and $B(S)$ are both similar functions of C and S respectively such that Equation (12) implies:

$$C_{\max} \geq S - k \tag{13}$$

where k is a constant. Then, if all behaviours are equally likely, Result 3 implies:

$$C_{mean} \geq S - k' \tag{14}$$

where k' is a constant, provided $B(S)$ grows at least exponentially with S.

The *size.* defined in this way, of a **D-set** of N_e events over N_d descriptors is $(N_d)^{N_e}$ showing the required exponential growth. As noted previously, the action of a new element to a model will usually increase the number of models by at least N_d also. For many cases the rate of growth of models with complexity is a polynomial in N_d and C, times an exponential term of the form N_d^C so that Equations (13) and (14) do apply. Thus the result expressed in (14) is of wide applicability, loosely expressed:

Result 4. The mean complexity of model required over a uniform distribution of behaviours of a given size is asymptotically proportional to the size of behaviour provided the number of distinct models and the number of distinct behaviours grow in a similar fashion with respect to complexity and size, respectively, at least exponentially.

3.2 Approximation between *D-sets*

Having taken a behaviour, *b, to* be a **D-set** represented as a mapping from events to descriptors, *b: E →D,* we might now assume that a modeller of the observed behaviour also produces some behavioural **D-set**, $m: E \rightarrow D$, as an attempt to represent *b.* A deterministic modeller would produce a single, unique **D-set** and one could ask whether it was identical to *b.* However, if the class of models available was such that identity was not possible then it would be necessary to have some measure of the extent to which m *approximates b.* One obvious measure is the total number of events on which *m* and *b* disagree:

$$N(b,m) = \sum_{e \in E} (1 - \lambda(b(e), m(e))) \tag{15}$$

where λ is a two-argument function that takes the value 1 if its arguments are equal and 0 otherwise.

$N(b,m)$ *is* actually a *distance measure,* i.e. we can show:

$$N(x,x)=0 \tag{16}$$

$$N(x,y)=0 \Rightarrow x=y \tag{17}$$

$$N(x,y)=N(y,x) \tag{18}$$

$$N(x,y)+N(y,z) \geq N(x,z) \tag{19}$$

So that it is possible to speak of the measure of approximation as being the "distance" of the model from the behaviour. The result is dependent only on λ itself being a distance measure and hence generalizes to weighting schemes other than the simple one given above. If the **D-set** also had the structure of a monoid then the measure N could be seen as closely related to measures of *string approximation* (Sellers, 1974) used in studies of text editing (Wagner and Fischer, 1974) and the determination of genetic ancestors (Fitch and Margolias, 1967; Sankoff, 1972).

Measures of approximation such as *N* would be appropriate to a modeller that proposed just one behaviour to test against the observed behaviour. For example, in the context of modelling probabilistic automata, the modeller might put forward the behaviour having *maximum likelihood.* However, in general. an acausal modeller would propose not just one particular behaviour but rather a *set of possible behaviours,* and we need a measure of approximation that gives a distance from a set of behaviours to the observed behaviour. The minimum distance of one object from a set is one well known extension, but fails in this case because a modifier could generate all possible behaviours and hence ensure zero distance. If, a distribution over the set of proposed behaviours is also given, however. then the mean distance of the modeller's proposed behavioural **D-set** from that actually observed would seem to be a suitable measure.

There is an alternative, but equivalent, viewpoint that throws new light on the problem. A distribution over possible behaviours is equivalent to a set-of distributions over descriptors. one for each event. The modeller can then be seen to be proposing for each event not a predicted

descriptor but instead a *distribution* over possible descriptors. This move from (maximum likelihood) deterministic predictions to so-called subjective probabilities has been studied both theoretically and experimentally in recent years (Aczel and Pfanzagl, 1966; Shuford et al., 1966; Winkler and Murphy, 1968; Savage, 1971; Winkler, 1971; Hogarth, 1975; Shuford and Brown, 1975) in order to elicit more information from human beings and to provide formal foundations for *subjective probability theory* (Carnap, 1962; Good, 1962; Wright, 1962; Villegas, 1964; Vickers, 1965; Menges, 1970; Grofman and Hyman, 1973). It is possible to use the techniques and results developed in these studies directly inrelation to the current problem of acausal system identification. Indeed, the developments reported here might be seen as an extension of subjective probability theory to *sequentia1 processes.*

Thus suppose now that the modeller proposes, not a **D-set** of descriptors, but rather a set of distributions over descriptors. a mapping, $\mu{:}E \rightarrow [0,1]^D$ from events to a product space of numbers between 0 and 1 that sum to unity. I shall write $\mu(e,d)$ for the proposed value assigned to descriptor d at event e—we have:

$$\sum_{d \in D} \mu(e,d) = 1 \tag{20}$$

It is simple to extend the measure of approximation N (Equation 15) to apply to these distributions by averaging the value as previously defined over the distributions. Let:

$$NE(b, \mu) = \sum_{e \in E} \sum_{d \in D} (1 - \mu(e, d)) \lambda(b(e), d) \tag{21}$$

If the $\mu(e,d)$ were in fact used as generating probabilities to generate a single **D-set** at random to match against the behaviour, then NE would be the expected number of errors.

The table below illustrates the modelling process now envisaged and the calculation of the measure NE. For comparison a maximum likelihood proposed behaviour is also given and N is calculated.

Event	:	1	2	3	4	5	6	7	8	9
Behaviour	:	A	B	C	A	A	B	B	C	C
	A :	0.1	0.1	0.2	1	0.1	0.4	0.4	1	0.5
Model	B :	0.2	0.4	0.1	0	0.2	0.6	0.6	0	0.5
Distributions	C :	0.7	0.5	0.7	0	0.7	0	0	0	0
Max. Likelihood	:	C	C	C	A	C	B	B	A	A/B

so that NE = 0.9 + 0.6 + 0.3 + 0 + 0.9 + 0.4 + 0.4 + 1 + 1 = 5.5

and N = 1 + 1 + 0 + 0 + 1 + 0 + 0 + 1 + 1 = 5

Thus the modeller proposes for event 1, not the maximum likelihood prediction *C*, but instead the distribution (0.1, 0.2, 0.7) over (*A, B, C*), i.e. $\mu(1, A) = 0.1$, $\mu(1,B) = 0.2$, etc.

This formulation is interesting because it closely resembles the procedures used by de Finetti (1972) and Savage (Savage, 1971) to elicit subjective probabilities from human subjects. De Finetti notes that if a target sequence is generated by a Bernoulli source and the subject gives a vector of numbers representing a distribution over possible symbols at each occurrence, then

there is a loss function, that, when minimized by the subject, leads to him giving true probabilities. This is in our present terminology:

$$SE(b,\mu) = \sum_{e\in E} \sum_{d\in D} (\lambda(b(e),d) - \mu(e,d))^2 \qquad (22)$$

i.e. the sum of the squares of the differences between the proposed distributions and the actual event "distribution" (1 for the event which occurred and 0 for each of the others). Savage shows the same property for an alternative loss function:

$$LE(b,\mu) = \sum_{e\in E} \sum_{d\in D} -\lambda(b(e),d)\log_2(\mu(e,d)) \qquad (23)$$

i.e. the sum of minus the logarithms of the components in the distribution of the elements that actually occur in the behavioural **D-set.**

The convergence properties of these loss functions is readily demonstrated by assuming that the descriptor at event e is itself generated probabilistically by the same generating probabilities *p(e,d)* and proving that the minimum expected loss occurs where *μ(e,d)=μ(e,d)*. It has been shown (Aczel and Pfanzagl, 1966; Shuford et al., 1966) that there is an infinite family of such loss functions with the convergence property that a subject minimizing them is forced to give true probabilities in a probabilistic situation. De Finetti (1972) showed this happened experimentally and the procedure has been used to assess "good probability assessors" in meteorology (Winkler and Murphy, 1968; Winkler, 1971) and to get maximum information about students' knowledge in multi-choice examinations (Shuford and Brown, 1975). Pearl (1975a) has recently given more meaning to the various measures that may be used to elicit subjective probabilities by relating them to possible hypotheses that the subject might make about the distribution of future payoffs in what, to him, is a gambling situation. For example *SE* corresponds to an exponential fall in future expected payoffs and *LE* corresponds to the slower decay of a Cauchy density. The original measure proposed, *NE* of Equation (21) does not lead to the optimal modeller giving true probabilities, but is instead minimized by the modeller who gives *maximum-likelihood* estimates in a probabilistic situation, i.e. a distribution having the value 1 for the most likely event and 0 for all the others. Since, it again corresponds to well-defined and well-known pattern of decision making behaviour.

The most striking difference between *SE* and *LE* may be seen by contrasting them on the example given previously where *E=5.5*

$$\begin{aligned} SE &= (0.9^2+0.2^2+0.7^2)+(0.1^2+0.6^2+0.5^2)+(0.2^2+0.1^2+0.3^2)+(0^2+0^2+0^2)+(0.9^2+0.2^2+0.7^2)+ \\ &\quad (0.4^2+0.4^2+0^2)+(1^2+0^2+1^2)+(0.5^2+0.5^2+1^2) \\ &= 1.34+0.62+0.14+0+1.34+0.32+2+1.5= 7.26 \end{aligned}$$

$$\begin{aligned} LE &= -1\log_2(0.1)-\log_2(0.4)-\log_2(0.7)-\log_2(1)-1\log_2(0.1)-\log_2(0.6)-\log_2(0.6)-1\log_2(0)-1\log_2(0) \\ &= 3.32+1.32+0.51+0+3.32+0.74+0.74+\infty+\infty=\infty \end{aligned}$$

The logarithmic measure will not tolerate the situation where an event is given a valuation of zero but then occurs—the error then becomes infinite, whereas both *NE* and *SE* give large but finite errors in this situation. The logarithmic measure is also distinguished in that it depends only on the valuation given to the event which actually occurred regardless of the distribution over the other components. This has been taken by some writers as a desirable feature although

the argument seems dubious and there are more meaningful considerations that make the logarithmic measure attractive (see following section).

One important aspect of the move earlier in this section from the concept of a modeller proposing a distribution over possible behavioural **D-sets** to the concept of its proposing a set of distributions is that the loss measures may be regarded as having a component associated with each event. The overall loss is, in all three cases, the sum of the losses associated with each event. The component added for event may be described as the *surprise* caused by that event. All three measures agree that the surprise caused by an event given the valuation I which actually occurs is zero (e.g. event 4 in the example). They give varying weights to events which would occasion little surprise (e.g. event 3) or much surprise (e.g. event 1) and, as noted, the logarithmic rule expresses infinite surprise at an event that occurs when the valuation given to it is zero. This valuation of "surprise" is consistent with the model of decision-making based on *"potential surprise"* proposed by the economist Shackle (1955; 1969), and is particularly useful in on-line learning algorithms where a marked increase in the rate of surprise may be used to indicate the need for the recomputation of the model.

3.2.1 *Entropy—the expected loss for probabilistic behaviour* One can avoid the premature use of evocative terms such as *subjective probabilities* for the distributions proposed by the modeller in the previous section, preserving methodological neutrality until results, under certain circumstances, prove the terms justified. The theoretical and experimental studies of de Finetti, Savage *et al.,* indicate that *if* the actual event sequence is probabilistically generated then a modeller that is optimal (in the sense of minimizing the poorness-of fit measures, *SE* or *LE) will* be forced to propose the actual generating probabilities of events. This result is an important link between *subjective* and *physical* or *frequentist* accounts of probability theory. It is equally important as a link between our general approach to system identification and probabilistic modelling. However, the measures defined in the previous section and the identification techniques based on them do *not* in themselves entail a hypothesis of probabilistic acausality. The fact that they behave meaningfully and well when used with probabilistic systems is clearly desirable, even essential, but there is no converse argument that they are based on a hypothesis of probabilistic behaviour in the system modelled.

Clearly, we may now expect to obtain results for probabilistic modelling (optimality of identification techniques, decision criteria for selecting amongst admissible models, etc.) which do not necessarily apply in more general cases—indeed are not meaningful unless further hypotheses are made about the more general case. Clearly also, there are few hypotheses comparable in power and significance to that of a probabilistic generator. In the experimental studies, I have taken examples of asynchronous systems modelling where no probabilities are definable but it is possible to obtain weaker, structural rather than numeric, results for identification techniques based on the measures of approximation defined. An example of nonprobabilistic acausality will be given in section 4.3 where several samples of the behaviour of a deterministic system are identified—the acausality arising through the sampling process and having no numeric, probabilistic significance in the model. Such examples are important in demonstrating the generality of the approximation measures outlined and also illustrate the role of the *probability logic* (Rescher, 1969; Gaines, 1976b) underlying probabilistic models in representing more general acausal phenomena; Kalman (1957) has shown that whilst linear operational techniques may be applied to sampled-data or discrete systems separately, the most

appropriate representation of a sampled-data, discrete system is pseudoprobabilistic; I have shown elsewhere (Gaines, 1975b; Gaines and Kohout, 1975) the wider roles of probability logics in modelling *possibility* and *eventuality;* and close links have been established (Rescher, 1963; Danielsson, 1967; Miura, 1972) between probability topics and *modal logics* (Hughes and Cresswell, 1968; Snyder, 1971) of possibility, necessity and time (Prior, 1967).

However, it is also of interest to determine what happens in a truly probabilistic situation when a modeller does manage to propose precisely the optimal distributions that he is forced to converge towards when minimizing the loss functions *SE* and *LE*. Suppose the actual probability of occurrence of descriptor *d* at event *e is p*(*e,d*) and the distributions proposed by the modeller are such that μ(*e,d*)=p(*e,d*). The expected values of the loss functions of equations 21 through 23 *NE*, *SE* and *LE,* respectively, are:

$$\hat{NE}(bp) = \sum_{e \in E} \sum_{d \in D} p(e,d)(1 - p(e,d)) \tag{24}$$

$$\hat{SE}(bp) = \sum_{e \in E} \sum_{d \in D} p(e,d)(1 - p(e,d))^2 + (1 - p(e,d))(p(e,d))^2 = \sum_{e \in E} \sum_{d \in D} p(e,d)(1 - p(e,d)) \tag{25}$$

$$\hat{LE}(b,p) = \sum_{e \in E} \sum_{d \in D} p(e,d) \log_2(p(e,d)) \tag{26}$$

Equations 24 through 26 show that the expected value of the loss, or approximation measure, when the modeller matches the actual generating probabilities is actually an *entropy* (Aczel, 1971) function for all three measures. In particular, *LE* of Equation (26) is the familiar Shannon entropy whose significance in inductive inference has been emphasized by Watanabe (1969). The coincidence in values of *NE* and *SE* is interesting but spurious since as noted previously, whilst the condition μ(*e,d*)=p(*e,d*) gives a minimum for *SE* and *LE,* the corresponding condition minimizing *NE is:*

$$\mu(e,d) = \begin{cases} 1, if\ , d = d' \\ 0, otherwise \end{cases} \tag{27}$$

where *d'* is some descriptor such that:

$$p(e,d') = \max_d(p(e,d)) \tag{28}$$

a maximum likelihood estimate. Thus De Finetti's quadratic loss function *SE* of Equation (22) has the same expected loss as the more obvious definition of *NE* in Equation (21) (expected number of errors) but forces actual probability estimation on the modeller which *NE* does not. It is also interesting to note that *NE* and *SE* are the expected value of *cont*(*e*&*d*) and *LE* is the expected value of *inf*(*e*&*d*), linking these results to theories of semantic information and logical probability (Carnap, 1950; Carnap, 1952; Bar-Hillel and Carnap, 1953; Schilpp, 1963; Bar-Hillel, 1964; Hilpinen, 1970; Hintikka, 1970; Carnap and Jeffrey, 1971; Erwin, 1971; Swinburne, 1973).

To give physical significance to the values p(*e,d*) within the current framework (in which the events in *E* have so far been regarded as unique) we need to hypothesize some structure on the event space such that many different events may be regarded as equivalent. Then the descriptors at those events may be regarded as a sample space for some process generating descriptors with a

probability distribution $\mu(e,d)$ (where $\mu(e,d)$ $\mu(e',d)$ for equivalent e and e'). Such an equivalence relation between events makes it possible for events, under the equivalence relation, to be recurrent, and hence for a modeller to determine the generating probabilities. If there is more than one equivalence relation (i.e. all events are not equivalent to one another) then there is also some *structure* present to be determined by the modeller. The following section discusses this problem of the determination of this structure.

3.3 Structures on *D-sets*

We have come remarkably far without assuming any structure on the **D-sets** representing behaviour and proposed by modellers. This has been quite deliberate since I wished to effect a clear separation between: *(a)* concepts of complexity and randomness (section 3.1); *(b)* measures of approximation between behaviour and model (section 3.2); and (c) the modelling of sequential machines (this section). It will be noted that the developments in sections 3.1 and 3.2 are independent of one another and neither requires such postulates as: the behaviour is that of a sequential machine; events are ordered in time: etc. The concept that a *random* behaviour is one whose *complexity* is nearly equal to its size. and the result that the expected loss in approximating a random event is an *entropy* for it. have both been obtained under very weak assumptions and constraints. In particular neither involves notions of c*ausality* or of *computational complexity* in terms of sequential machines. In the context of the-general formulation of the problem of identification given in section 2.2 the first result relates to the intercept of the plot of approximation against complexity for admissible models with the axis of zero approximation. and the second result relates to its intercept with the axis of zero complexity and to the significance of the approximation at different complexities. Thus the two results are linked through the concept of an admissible subspace of models.

I emphasize the general status of these results because they will clearly have a role to play in any more specific formulations of the identification problem In a sense they will be *re-discovered* for each class of behavioural and model structures, and a major virtue of a general systems theoretic formulation is to guide this re-discovery and allow the fundamental status of these concepts and results to be seen independently of the class of problems for which they are obtained. Conversely, the essentially structural role of notions such as causality can be seen more clearly if they are separated from the numerical notions of complexity and entropy and only introduced at this late stage.

3.3.1 The "Rules ofthe Game" What then are the rules of any further constraints upon the behavioural D-set. that it is *ordered, monoidal,* etc.? In fact. they are best viewed as *rules of the game* whereby the modeller is allowed to use the descriptors attached to some subset of events as a basis for proposing distributions that apply to others. Consider the two extremes: the modeller has to propose distributions completely a *priori;* with complete knowledge of events the modeller can propose precisely the actual behaviour. Both these extremes are trivial and in any realistic problem formulation the modeller will have *partial knowledge* about some events that win influence his proposals about others.

I call the constraints upon the use of partial knowledge *rules of the game* because the modeller can usefully be viewed as playing a game against an opponent, or at least being constrained by a referee—he has to state what algorithm. within the rules of the game, he is using to propose a distribution and then his opponent or the referee evaluates it and reports back the loss, or degree

of approximation, achieved. This concept is useful because in practice, since we are concerned with the problem of formulating a model for an already known behaviour, Of identifying its underlying generative structure, the modeller actually knows all the behaviour and it is what he is allowed to use that is constrained. For example, if events are ordered (e.g. in time) and the modeller in proposing a distribution over descriptors for a particular event is allowed to use only the value of descriptors for events properly less than that one, then it is said to be *causal* (Windeknecht, 1967; Orava, 1971; Windeknecht, 1971). If there is a metric *topology* on the event space and the modeller in proposing a distribution for a particular event is allowed to use only the value of descriptors for events in a neighbourhood of specified size about that event then it is said to be *local.* If the algorithm used by the modeller in determining its proposals for each event is independent of the particular event then it is said to be *uniform.*

Non-uniform modelling can be justified if the event space has known inhomogeneities, e.g. boundaries in optical pattern recognition. *Uniform local* algorithms are appropriate to such operations as contour extraction in scene analysis (Moore, 1971; Duda and Hart, 1973) and has a natural interpretation in terms of Lie groups (Hoffman, 1966; Grenander, 1969). They have been studied, for example, by Minsky and Papert (1969) in determining the power of "diameter-limited perceptrons". *Uniform local causal algorithms* correspond to so-called *finite-memory* (Hellman and Cover, 1970; Witten, 1976) algorithms, and have been studied as *definite events* (Perles, Rabin and Shamir, 1963) *locally-testable languages* (McNaughton and Papert, 1971; Zalcstein, 1971), etc.

Local and finite-memory algorithms must not be confused with *finite-state* algorithms. The notion of state, as developed by Birkhoff (1927), Markov (1954; de Fiorino, 1966), Zadeh (1964; 1969), Salovaara (1974), and others, is a system concept of major importance in its own right and quite distinct from that of "finite memory". There is an ontological distinction between them in that the "finite memory" refers to a bounded locality accessible in the environment, in the observed behaviour, whilst the "finite-state" refers to a limited storage capacity in the algorithm, or model.

States are best introduced into tile current discussion by considering them as (conceptual) *tokens* that may be placed on events as an aid to proposing distributions. A finite-state algorithm has only a specified finite number of distinctions between tokens that it may make. The placing of tokens is governed by rules of (he game similar in their possibilities to those for proposing distributions. However, the placing of tokens and the proposing of distributions is now allowed to depend not just on the description at a specified set of events but also the tokens placed at them.

The terms *local* and *causal* may clearly be applied to the algorithm for placing tokens at events as before. However, an interesting new distinction arises since the descriptor actually at the event may, or may not, be abound to influence the token placed there (it would clearly be trivial to allow it to influence the proposed distribution for descriptors there). If the algorithm for allocating a token to an event is dependent only on descriptors and tokens at other events and gives a definite result it is said to be *deterministic.* If it gives a definite result only when the descriptor at the event is also specified it is said to be nondeterministic but *observable;* otherwise it is just *nondeterministic.* Thus, in the context of causal systems. a nondeterministic but observable algorithm cannot predict the next state precisely but knows what it is when the next

descriptor is known, i.e. it can always specify the current state precisely, not just as a distribution.

The concept of state is normally associated with causal processes, i.e. time sequences. If we specify that the events are either chain-ordered or incomparable so that the behaviour may be regarded as a submonoid of the free monoid of descriptors, and that a local neighbourhood of an event are the events immediately next to it in its chain, then a *local causal deterministic* algorithm is a deterministic sequential machine or *automaton,* and a *local causal observable* algorithm is an observable probabilistic automaton.

The role of states in general may be seen as that of *extending the locality* open to a local algorithm. Tokens may be placed to indicate features of the event space that are not apparent from inspection of the allowed local neighbourhood. For example. in pattern recognition the *global property* of being part of an oriented line, as opposed to the local property of being an oriented line segment, requires the placing of tokens to compute. It is the defect of the perceptron that it cannot place such state tokens that Minsky and Papert (1969) studied, demonstrating the resultant weakness with respect to the determination of global properties.

It will be noted that, whereas in a causal system there is an unambiguous sequence for the placing of tokens, the ordering in time, this does not exist in general and problems may arise in, for example the spatial assignment of tokens such that many, or no. final assignments are possible. It is interesting to speculate that this may correspond to some of the phenomena of spontaneous change in perception (Gregory, 1970) whereby, for example, a figure will be first perceived as one object and then as another—the assignment of local features to global percepts being ambiguous and a number of different, but self consistent, assignments being possible.

Informally also the concept of states as conceptual tokens placed in the environment is attractive because it is apparent that many problems of sequential inference are avoided in human and animal societies by placing actual *physical tokens* in the environment. The internal states necessary to an algorithm solving an essentially sequential inferencing problem may be seen as compensating for the lack of such physical sign posting in the actual environment. Since it will be clear that even simple sequential inferencing tasks can require massive computation it may well be that the mathematical models of induction thus generated are *not* a reasonable representation of human thought processes. *We mark the environment to* make inferencing simple and to avoid the need for internal states (short-term memory) as much as possible.

3.3.2 Probabilistic automaton structures If we now focus down upon local causal algorithms over descriptor monoids, i.e. probabilistic automaton. it would be useful to derive some overall convergence results similar to those for deterministic automaton modelling using the Nerode equivalence. These are primarily (Gaines, 1976e):

1) The number of states in the model is a monotonic non-decreasing function of the length of the observed sequence of behaviour. Thus the model cannot become simpler with further observations;

2) The number of states in the model of a sequence of behaviour generated by an M-state deterministic automaton cannot exceed M. Thus the model cannot grow more complex than the generating system.

The situation is far more difficult to analyse for probabilistic modelling because there is now no well-defined *best model* but instead an admissible subspace of possible models. However. some comparable results may be obtained. Any probabilistic automaton may be regarded as outputting a distribution over possible outputs at each step (this is not identical in concept to an ensemble of all possible transitions. but rather more artificial since we are following the state-distributions of a single automaton) Consider the problem or how a modeler might actually come to match a sequence of distributions. The two sources of difficulty are:

a) Since the sequence is not Bernoulli and the distributions change from event to event the modeller must be able to locate himself in the sequence. i.e. the events with different distributions must he observable. This is similar to the situation in deterministic modelling where no modeller can discriminate between two different structures if their reduced, observable forms are isomorphic. In practice, for a probabilistic model, this condition implies simply that, even though from a state we can predict only the probability of the next state, after the transition the output must be sufficient to indicate the actual state. Thus in analysing the match between source and model we need only consider the reduced form of the source—this appears in result (2) above as the number of states in a model cannot *exceed* those in the source rather than becomes eventually equal to their number.

b) The distributions themselves are not the actual outputs and an indefinitely large sample of actual outputs is necessary in order to estimate them. Thus the distributions in any *transient* behaviour of the source automaton cannot be precisely estimated, only those in its recurrent behaviour.

Combining these two factors we can see that it is realistic to consider matching precisely in a model only the *recurrent* behaviour of the *reduced form* of a probabilistic automaton. The recurrency is an additional constraint compared with deterministic modelling and clearly a reasonable one in the circumstances. Conceptually the modeller of an observable sequence of distributions is applying a Bernoulli sequence modelling strategy to each distinct distribution. However, he has *both to* discover the observation algorithm *and* estimate the distributions.

Consider the formulation of the identification problem given in section 2.2.2 with the identification space, $(D^*, M, \leq, f)$, where: M is the space of probabilistic Moore automata in reduced form; m $\leq$n if m has less states than n; the value f(b)= $\leq_b$ being defined for a sequence of behaviour b by $m \leq_b n$ if $LE(b, \mu_m) \; LE \leq (b, \mu_n)+\varepsilon s$ where μ_m and μ_n are the distributions arising from modelling m and n respectively, s is the length of b and ε is a small tolerance to allow for statistical fluctuations (an alternative f may be similarly defined based on SE instead of LE). Now consider the behaviour of the admissible subspace for a sequence of observations of increasing length of a finite state probabilistic source.

A result equivalent to (1) above could be that the maximum number of states in the admissible models is monotonic non-decreasing. However, this is not so because in the short term the particular sequence generated may be such as to justify complex models. However, there is a result equivalent to (2):

2') For any ε, for a given probabilistic source, the maximum number of states of an admissible model cannot eventually exceed that of the source as the sequence of observation increases. This follows because the properties of LE are such that averaged

over a long sequence of recurrent states any modeller cannot do better than put forward precisely the distributions of the reduced form of the source and hence the source itself will have at least as low a value of *LE* as any model with a greater number or states.

We still cannot show that the maximum number of states is precisely that of the source, even in reduced form. because the modelling procedure cannot necessarily identify the transient behaviour of the source. Our definition of $\leq_b$ based on *LE* and s means that in the long term the transient behaviour will have a decreasingly small effect so that eventually the admissible models neglect it. The maximum number of states in an admissible model will then correspond to the number of states in the recurrent part of the automaton generating the observed behaviour (there is clearly no well-defined *recurrent part* in general since we may enter different recurrent parts after the same initial transient). What of the admissible models with less than the maximal number of states? These correspond to the "lumping" of states in a Markov process and will inevitably give higher values for the entropy of the process and hence *LE*. They are best approximations to the source in the sense that they minimize the deviation in behaviour from that of the actual source.

In the next section some experimental results on an actual computer implementation of a probabilistic automaton modelling algorithm are given that illustrate the points made in this section.

4 Some Computational Studies

The approach to the problem of identification developed in the previous sections is intended to be completely operational and hence open to computer implementation and experimental study and application. Even the most general formulation of section 2.2 is mathematically precise in terms of order relations and open to computational study for any identification technique. Probably the most interesting level at which to undertake such computational study at the present time, however, is that of probabilistic automaton inference. This is currently an unsolved problem with many interesting recent case studies, generally in terms of probabilistic grammatical inference (Feldman, 1972; Patel, 1972; Maryanski, 1974; Fu and Booth, 1975) that is also of great practical importance for the analysis of real world data such as human and animal behaviour patterns (Vleck, 1970; Dawkins and Dawkins, 1973; Dawkins and Dawkins, 1974; Suppes and Rottmayer, 1974).

The main reason that the problem of inferring the structure of a discrete probabilistic system from its behaviour has not been solved (in the sense that the equivalent problem for deterministic systems has) is that discussed in section 2, that the nature of a "solution" is not well defined—we have had no decision procedure to determine that a particular probabilistic automaton is a "correct" representation of a given sequence of behaviour. Thus, the most successful studies of the problem have tended to be those that have concentrated on the methodology of *approximation,* of deciding when a particular automaton or grammar is *reasonably correct.* In particular the recent studies of Maryanski (1974), part of a programme of research directed by Booth (Patel, 1972; Fu and Booth, 1975), provides a series of challenging problems and results for other workers. The concepts of an admissible subspace (section 2.2), and the specific measures of approximation (section 3.2), developed in this paper provide the basis for a formulation and solution of the problem of probabilistic structure inference that is precisely

defined, and hence *complete* in a sense that previous formulations have not been, and it is clearly of interest to evaluate the operation of the techniques on actual data.

The following sub-section describes ATOM, a suite of programs embodying the identification techniques described in this paper. A reanalysis of some of Maryanski's data using ATOM, is given in (Gaines, 1976d), and a variety of informal examples in (Gaines, 1976a). The examples in the paper have been selected both to demonstrate the basic techniques and also to illustrate the hierarchical analysis of epistemological levels in system identification proposed by Klir (1976; Klir and Uttenhove, 1979) and discussed in section 5.1 of this paper.

4.1 The ATOM Identification System

The identification system to be described is one in which the class of models M is that of finite-state probabilistic automata with the relationships, and $\leq_b$ defined as in section 3.3 by the number of states in the model and the logarithmic approximation measure, *LE,* respectively. The computational algorithms to determine the admissible models are one of a suite of programs called "ATOM" written in the interpretive string-handling language BASYS (Facey and Gaines, 1973; Gaines and Facey, 1975) on a time-shared PDP11/45 (the algorithms have also been used in FORTRAN on a PDP10 with KA10 processor and run some 50 times faster). ATOM provides facilities for interactively entering observed data and forming on-line predictions from models, and so on. However, for the automata modelling studies it is generally used in restartable background batch mode since computational runs of hundreds of hours may be required.

A behaviour to be modelled is input to ATOM as a string of arbitrary character strings separated by spaces or end-of-line terminators. Thus:

```
MARY HAD A LITTLE LAMB ITS FLEECE WAS ?
```

is a sequence of behaviour consisting of 9 symbols, and:

```
I=2
P=A(I)
J=P/I+7
```

is a sequence of behaviour consisting of 3 symbols. This acceptance of free format strings is particularly helpful in some examples such as natural language processing and automatic programming.

ATOM assumes that all the symbols are automaton *outputs* unless it is separately informed that a certain set of symbols are *inputs* and/or another set are *delimiters.* All the modelling schemes treat these two classes in a similar fashion: *inputs* are not brought into the string approximation measurement, i.e. one does not evaluate the extent to which input symbols are predicted correctly: *delimiters* are taken to indicate that the string before the delimiter may not be used to predict that after it. In the automaton models a delimiter causes a reset to a single state of the model. Otherwise both inputs and delimiters are treated as any other symbols in the string of behaviour. Note that the availability of delimiters enables separate samples of behaviour (separate sentences say in a grammatical inferences problem) to be freely concatenated together, separated by delimiters, to form a single sequence of behaviour. Note also that the modelling process does not necessitate inputs and delimiters being specified in this way. If they are then the computation is faster, but if they are not then their nature may be inferred from the results - i.e. inputs arc "outputs" that cannot be predicted and delimiters are those which appear as a general reset - examples of such inferences will be given later.

The automation identification subprogram in ATOM generates. for a given behaviour, the admissible subspace of either Mealy or Moore probabilistic automata. as requested, commencing with l-state models (Mealy) or *k*-state models (Moore) where *k* is the number of different symbols in the behaviour). The actual output of the programme is thus the set of best-approximation 1-state models, the set of best-approximation 2-state models. etc. The search ceases when no more admissible models are found, but in practice this condition rarely arises since the search space for larger models grow rapidly with the number of states and the programme is terminated by lack of time rather than by completion of the search. However, since the simpler admissible models are output first, the modelling is always complete up to models with the number of states at which it was terminated.

The search procedure is essentially simple because *only the space of non-deterministic automata has to be searched,* not that of probabilistic automata, i.e. the transitions are initially regarded as being only present or absent. When a non-deterministic model of the behaviour has been generated the actual transition probabilities are filled in from the relative frequencies of the transitions in the particular model with the given behaviour. This is legitimate because these values are known to minimize the *LE.* The value of *LE* for the model/behaviour pair is then calculated and used to ascertain the approximation relative to previous models generated. If the approximation is the same, or better, than that of the best models previously formed then the new model is added to the set of potentially admissible models. Any models with the same number of states but a poorer fit on both criteria are discarded. The search then continues. Whenever the models with a given number of states has been searched then the remaining best models with that number of states are filed as being admissible. The normalized values *of NE* (assuming maximum-likelihood estimates) and *LE* are filed with the model.

The generation of models is basically an exhaustive enumeration of all possible observable nondeterministic automata. However, some care is necessary to avoid duplication and to take advantage of any structural features of the sample behaviour (e.g. some symbols never following other symbols). Models are generated using the actual behaviour to fill in state transitions. The initial model is a l-state automaton and, if *N*-state models are being searched, *N* is taken as a bound on the number of states. The initial state has to be the first and only state. Each symbol in turn in the behaviour is then examined. If it corresponds to an existent transition no action need be taken. If there is no transition corresponding to it then one is entered to the first state and a marker placed on a stack. The state is then advanced to its next value and the next symbol checked.

Eventually a model has been formed and may be evaluated for *LE.* A backtrack is then made by taking off the stack the last transition entered ands if it is to state *k*, changing it to be to state *k* + 1 and continuing as before. However, if state l was a new state then it is removed and backtracking performed again, or if *k* was the last state and not new, and *k* is less than the allowed maximum number of states, then a new state is added and the transition entered to this. Eventually backtracking is no longer possible and all models with the allowed number of states have been generated without duplication and without considering transitions not necessitated by the behaviour being modelled.

"Delimiter" symbols are taken to cause a reset to a single state (usually the initial state since strings with delimiters normally commence with one). "Input" symbols are not taken into

account in the calculations or *LE.* The following sections contain examples of ATOM automaton modelling in action.

4.2 Identification of a Five-State Stochastic Automaton

Figure 1a shows a sequence of 200 descriptors generated by a 5-state binary stochastic source (read in the natural order of English text, from left to right, line to line). This was input to ATOM and the admissible subspace of models computed up to 6 states based on a normalized *LE* measure of approximation. A deterministic model was also computed.

Figures 1a-f: Identification of a 5-state stochastic automaton.

B B A B B A B A B A B A B A B B A B B A B B A B B A B B A B
B A B A B A B B A B A B A B A B A B B A B B A B A B A B A B
A B A B A B A B A B A B B A B B A B B A B A B A B A B A B A
B A B A B A B B A B B A B B A B B A B B A B B A B B A B B A
B A B A B A B B A B A B A B A B A B A B A B B A B B A B B A
B A B B A B B A B B A B A B A B A B A B A B B A B A B A B A
B A B A B A B B A B B A B A B A B A B A B A B B A B A B A B A

Figure 1a: The behaviour sequence of 200 descriptors.

Figure 1b shows the ATOM output of 1 through 6 state admissible models. The first line gives NE (as fraction and percentage): followed by PLOGP (normalized *LE);* the type, number of states and number in order of search of the model; and the name of data file. The following lines describe the state transitions of the model: for each state, the transition under each descriptor and the number of times it occurs (in brackets), e.g.:

$$1{:}B \rightarrow 1(34)\ A \rightarrow 2(83)$$

means that in state 1 a B leads to state 1 34 times and an A leads to state 2 83 times. The final line after each model shows the total number of models searched—note that no 6-state models exist that are better then the 5-state best.

Figure lc is a plot of approximation against complexity (PLOGP or *LE* against number of states) for the admissible models. and Figure 1d is a set of diagrams of the models themselves with descriptor associated with the transition subscripted with the number of times the transition occurs. Note how *LE* commences near its maximum value of unity for a l-state (Bernoulli sequence) model, corresponding to the total number of A's in the sequence being close to that of B's. It drops sharply by one half for a 2-state model that shows essentially that A is always followed by B. It drops substantially again for a 3-state model that adds the constraint that BB is followed by A. Going to a 4-state model produces no significant drop but there is one in going to a 5-state model. In fact this gives an accurate model of the source which was a 2-state system generating (AB)* that had a probability of 0.2 in one of its states of transiting to a 3-state system generating (ABB)* that in its turn had a probability of 0.3 of returning to the original system from one of its states. Going on to a 6-state system produces no significant drop in *LE* and in fact *LE* reaches zero only with a (deterministic) model of 172 states.

```
83/200 (41.50%)    PLOGP=0.978      MEALY:1:1     FIVE.DAT
* 1: B -> 1(117) A -> 1(83)

1 MODELS WITH 1 STATES - BEST PLOGP 0.978

33/200 (16.50%)    PLOGP=0.499      MEALY:2:8     FIVE.DAT
  1: B -> 2(84)
* 2: B -> 2(33) A -> 1(83)

9 MODELS WITH 2 STATES - BEST PLOGP 0.499

33/200 (16.50%)    PLOGP=0.402      MEALY:3:51    FIVE.DAT
  1: B -> 2(84)
* 2: B -> 3(33) A -> 1(50)
  3: A -> 1(33)

85 MODELS WITH 3 STATES - BEST PLOGP 0.402

31/200 (15.50%)    PLOGP=0.385      MEALY:4:529   FIVE.DAT
  1: B -> 2(57)
* 2: B -> 3(29) A -> 4(27)
  3: B -> 3(4) A -> 1(56)
  4: B -> 3(27)

969 MODELS WITH 4 STATES - BEST PLOGP 0.385

19/200 (9.50%)     PLOGP=0.317      MEALY:5:6215 FIVE.DAT
  1: B -> 2(34)
  2: B -> 3(24) A -> 4(10)
  3: A -> 1(33)
  4: B -> 5(50)
* 5: B -> 3(9) A -> 4(40)

13378 MODELS WITH 5 STATES - BEST PLOGP 0.317


216347 MODELS WITH 6 STATES - BEST PLOGP 0.317
```

Figure 1b: ATOM analysis of behaviour—1 through 6-state admissible models.

In terms of regular events the actual sequence is of the form (BBA | BA)* and the useful models produced are:

1-state: (A*B*)* LE=0.978
2-state: (BB*A)* LE=0.499
3-state: (BBA | BA)* LE=0.402
5-state: (BBA | BA)* LE =0.317

The drop in LE between 3 and 5 states is caused not by an improvement in the language structure in these terms but by the decoupling inferred, that a BBA is more likely to be followed by a BBA, and a BA by BA, than the alternative possibilities, i.e. a better indication of the language structure would be ((BBA)* (BA)*)*.

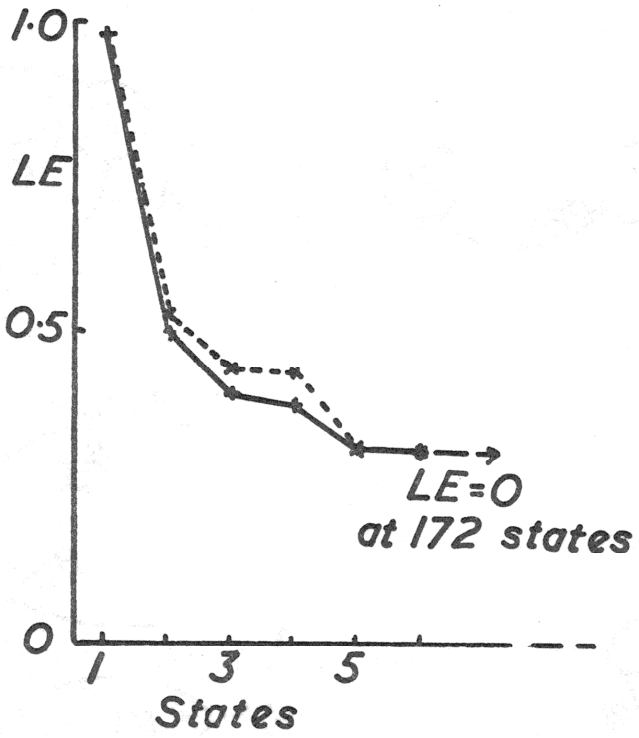

Figure 1c: Plot of approximation against complexity (logarithmic measure, *LE*, against number of states) dashed plot from validation sample.

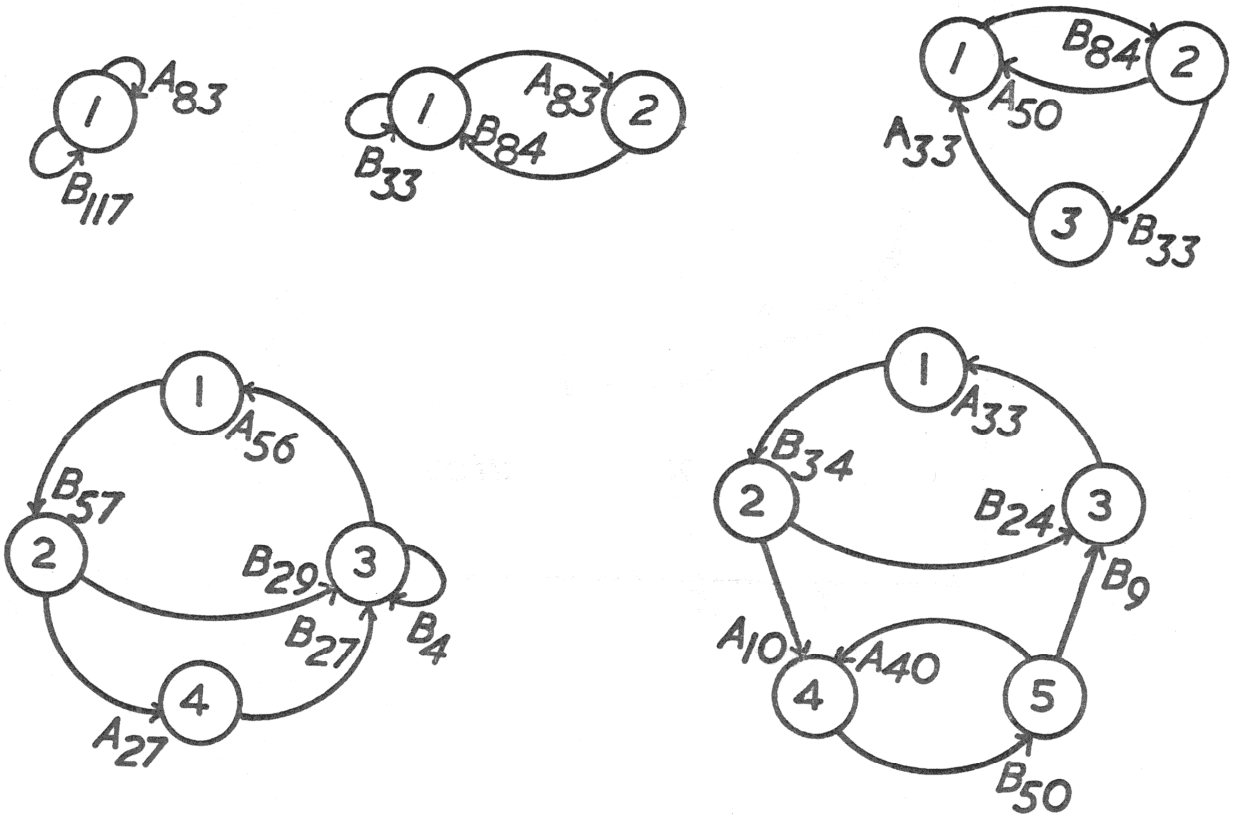

Figure 1d: Admissible models produced by ATOM up to 5 states.

Figure 1e shows the pattern of "surprise" (minus log of predicted probability for descriptor that actually occurs) for the first 70 descriptors in the sequence for each of the admissible models. It is apparent how the surprise caused by the deterministic transitions (within the two subsystems) drops to zero except for those where transitions may occur from the BBA subsequence to the BA subsequence. or vice versa. At these points the surprise at transitions within a subsystem is low whereas that at transitions between subsystems is high. Note that the "surprise" patterns shown are retrospective—they result from "replaying" the entire behaviour through the final model based on it. A different pattern would emerge if the surprise were calculated incrementally with the modeller calculating a model on the basis of the sequence so far observed. However, I shall not consider the incremental dynamics of such learning situations in this paper.

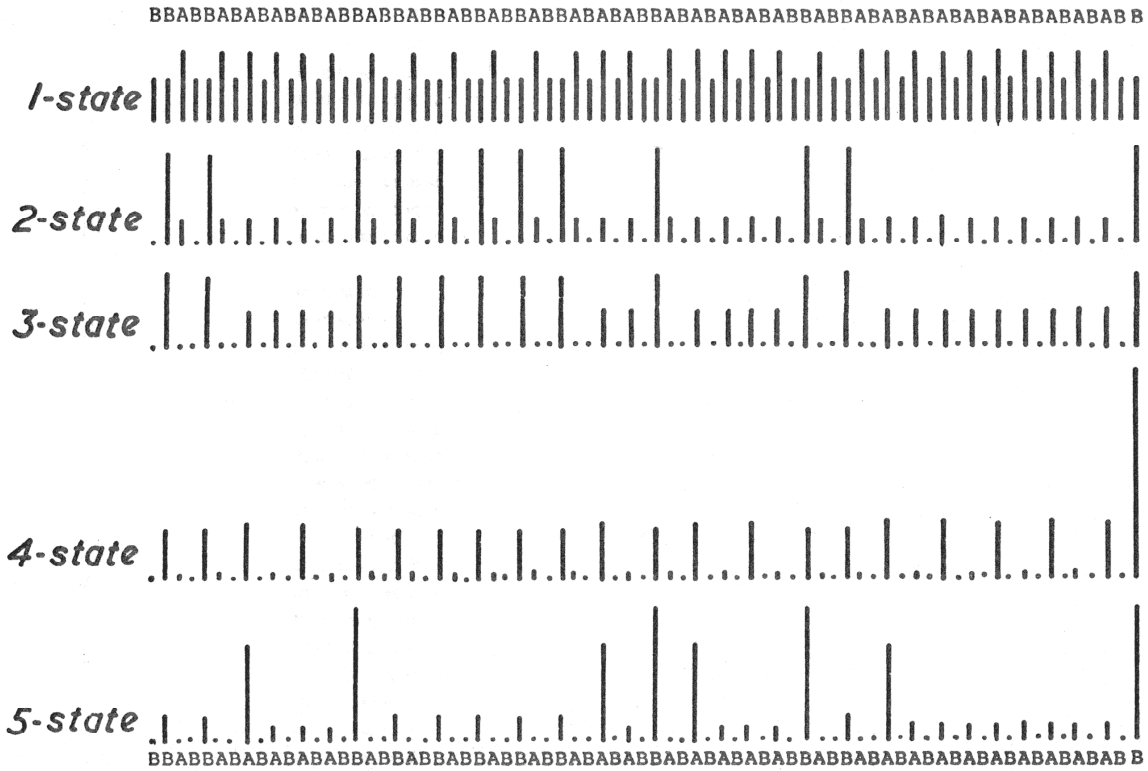

Figure 1e: Surprise at each of first 70 descriptors for 1 to 5-state admissible models.

I have used the term *significant* for the changes in *LE* in the previous paragraph. However, apart from the visual appearance of Figure lc where changes may clearly be seen, this term has not been justified. Some studies of mathematical significance in entropy change and uncertainty analysis has been made (Miller and Madonna, 1954; Garner and McGill, 1956; Kullback, 1968) but much more foundational work is needed to provide formal significance of the ATOM results, e.g. taking into account the number of models searched and the length of sequence analysed. One heuristic approach that is useful with the current system is to give ATOM a second, independent sample from the same source and get out values of *LE* for the admissible models generated from the original source, i.e. validate the models against an independent sample. This is the source of the dashed plot in Figure 1c that was obtained by evaluating the second sample shown in Figure 1f against the models of Figures 1b or 1d. Note how the 5-state, "correct", model stands out as not losing any significance whereas the dashed curve of approximation for the other models lies above that based on the sample that generated the models.

B A B A B A B B A B B A B B A B B A B B A B A B B A B B A B
B A B A B A B A B B A B A B A B A B A B A B A B A B A B A B
A B A B A B A B A B A B A B A B A B A B A B A B A B A B B A
B B A B B A B B A B A B A B A B A B A B B A B B A B B A
B B A B B A B B A B B A B B A B B A B A B A B A B A B A B
A B A B A B A B A B A B B A B A B A B A B B A B B A B B
A B B A B A B A B B A B B A B B A B B A

Figure 1f : Independent second sample of behaviour for validation.

4.2 A Transient Behaviour from a Five-State Automaton

The test against an independent sample suggested and demonstrated in the previous section makes sense if the source is in a recurrent phase. However, ATOM also provides meaningful models of transient sequences. Consider, for example, similar behaviour to that already analysed but where the BA to BBA transition only occurs once. The sequence $(BA)^{15}(BBA)^{10}$ was input to ATOM and analysed up to 6-state models to produce the plot of approximation against complexity shown in Figure 2a.

Figure 2a-c: Identification of a transient behaviour

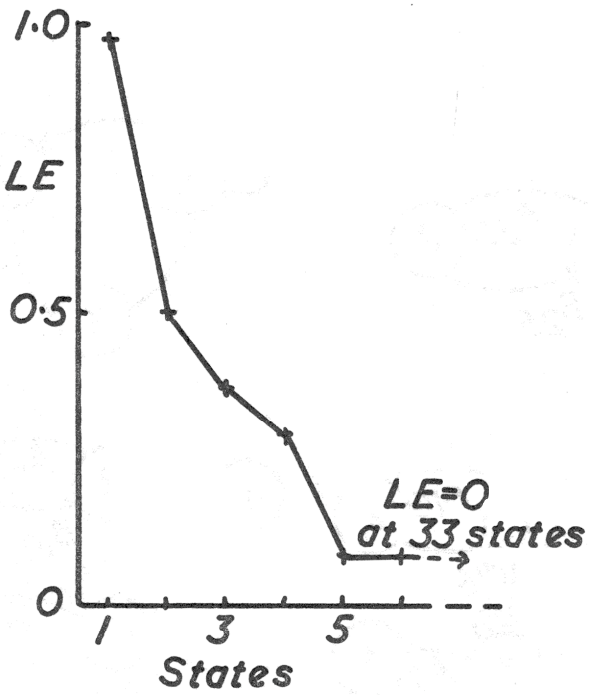

Figure 2a: Plot of approximation against complexity (logarithmic measure, LE, against number of states)

A turnover at 5 states is again apparent and the models of Figure 2b again show improving hypotheses as to the actual sequence structure. Note, however, in this case the final improvement in approximation at 5 states where the model actually breaks into two parts for the first time- there is only a single transition from state 1 to state 3 and there is no return from states 3, 4 or 5 to states 1 or 2. The patterns of surprise shown in Figure 2c also show up this clear change when a 5-state model is reached, with a very clear marker at the point of transition from the transient sequence to the recurrent sequence.

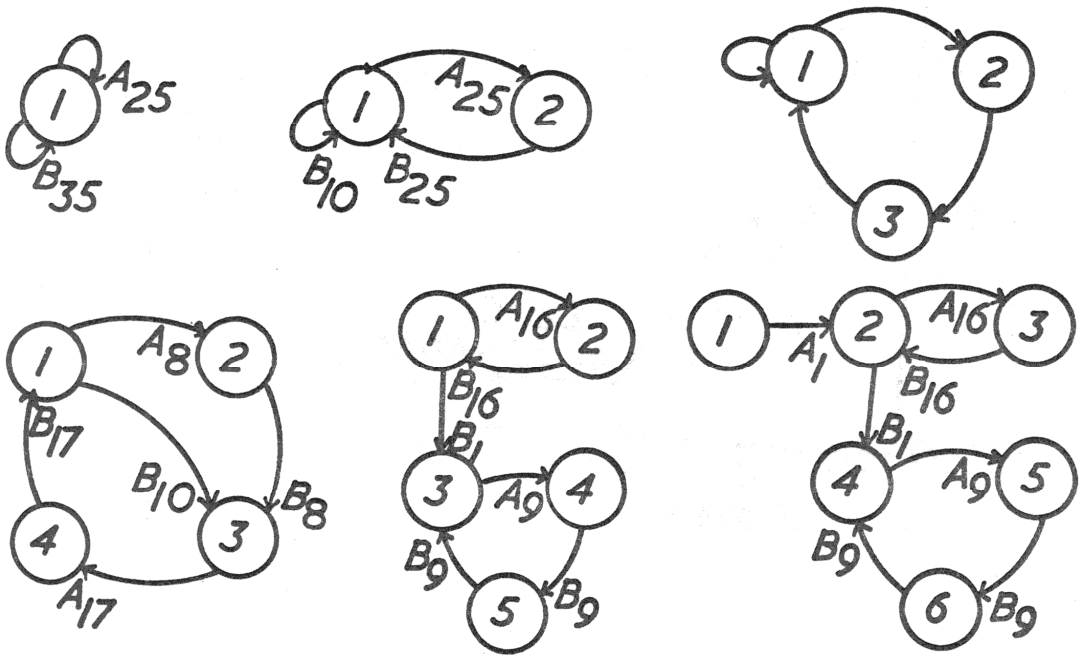

Figure 2b: Admissible models produced by ATOM up to 6- states

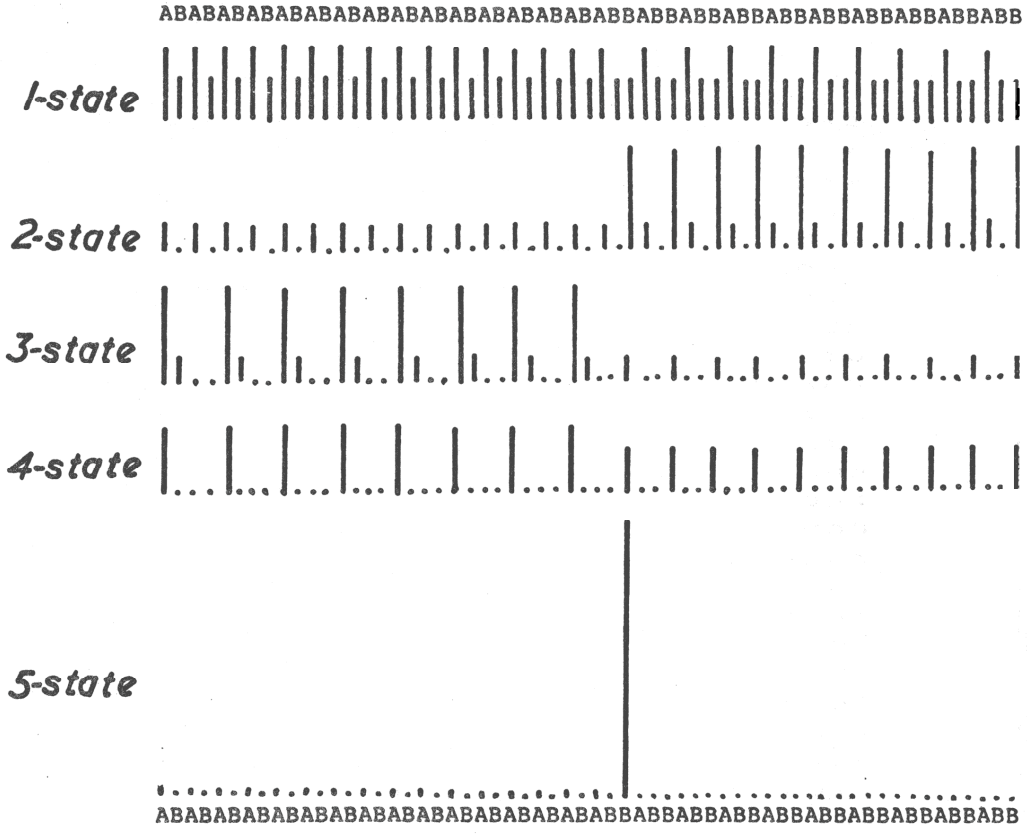

Figure 2c: Surprise at each descriptor for 1- to 5-state admissible models.

Thus ATOM can provide a very clear analysis of a sequence of behaviour that shows an isolated, non-recurrent change, where concepts of stationarity do not apply.

4.3 Miscellaneous Studies-Delimiters, Inputs and Outputs

Previous publications (Gaines, 1976a; Gaines, 1976d) contain a range of examples of ATOM inferring models from behaviour in a variety of applications including: stochastic grammatical inference; the analysis of human behaviour: and the derivation of programs from their traces—this last being an application of grammatical inference studied previously (Biermann, 1972; Biermann, Baum, Krishnaswomy and Petry, 1973; Crespi-Reghizzi, Melankoff and Lichten, 1973) only for cases where there are no errors in the trace and use of the Nerode equivalence gives an equivalent deterministic automaton flowchart for the program trace. There are two main features of interest in these other studies that are relevant to the general formulation of this paper-the hypothesizing or inferring, of *delimiters* and of *inputs.*

To illustrate the role of delimiters, Figure 3a shows the observed behaviour of a system as a sequence of 101 descriptors from the set {A, B, C, D}. In fact the behaviour is that of a 3-state deterministic machine generating the repetitive sequence (CBA)*, sampled in short segments of arbitrary length with the symbol D inserted as a delimiter at segment boundaries. It consists overall of 19 samples of behaviour concatenated together between delimiters.

Figures 3a-d: Identification of a sampled deterministic automaton.

A D B A D A C B D C B A C B D B A C B A C D C D A C B A D
A C B D B A C B A C B A C D B A D A C B A C B A C B A D A
C B A C B A C B A D C D B A C B A D B A C D B D C B A C D
A C B A C B A C B A C D D A

Figure 3a The behaviour-sequence of 101 descriptors.

Figure 3b is a plot of approximation against complexity for admissible models of the sequence and a turnover at 4 states is apparent. The transition diagram of Figure 3c shows the 4-state admissible model and the dominant CBA cycle is apparent. Superimposed on this are some less frequent "noise" transitions of which the most notable are those involving D since it can be seen that D may occur in any state, always leads to state 1, and from this state any descriptor may occur. The fact that D always leads to a single state shows that it acts as a *reset* input to the automaton causing a return to a single state. Since this implies in its turn that there is no memory in the automaton of the behaviour prior to D such a reset input is also called a *delimiter.* The occurrence of a D transition from any state and the exit to any state from the reset state suggests that state-independent. i.e. asynchronous, sampling is taking place. Note that the overall model is a *Moore* automaton in which transitions correspond to states: state 1 to D; state 2 to B; state 3 to A; and state 4 to C.

Figure 3d shows the surprise pattern for the initial sequence of observations from which it can be seen that D and the descriptor after it are both surprising but the rest of the descriptors, within the sampled behaviour pattern, are not. Thus ATOM has effectively derived the original structure of the system generating the behaviour. Note that this example involves no truly probabilistic phenomena. The sampled system is completely deterministic. The sampling itself is not random but only asynchronous and hence non-deterministic—longer samples were allowed deliberately in the later part of the sequence. As demonstrated in the previous example, the ATOM algorithm copes with non-probabilistic indeterminacy and derives the correct structure without strong probabilistic assumptions.

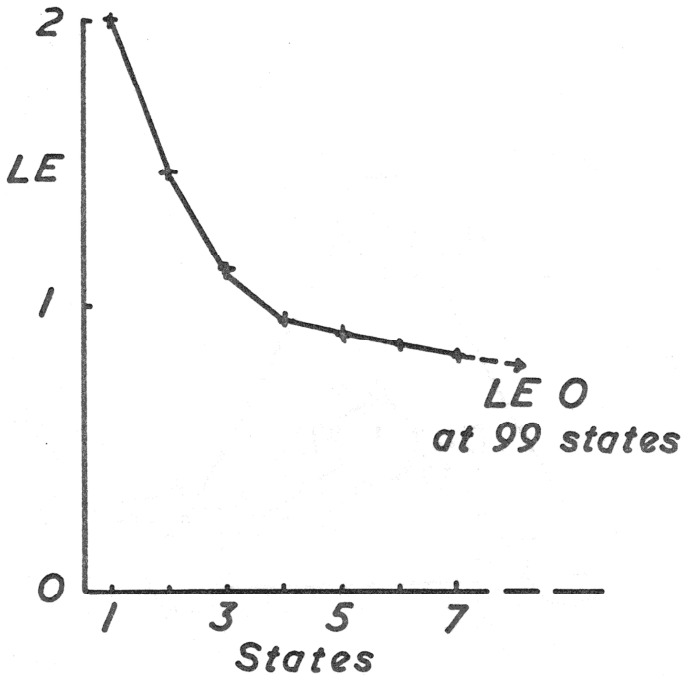

Figure 3b: Plot of approximation against complexity (logarithmic measure, *LE*, against number of states).

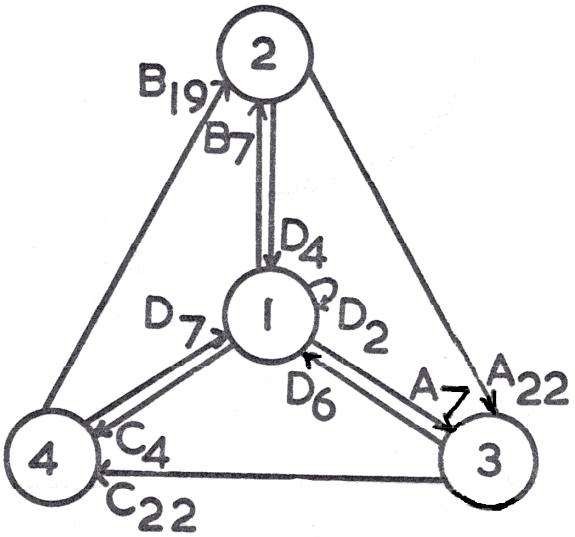

Figure 3c: 4-State admissible model.

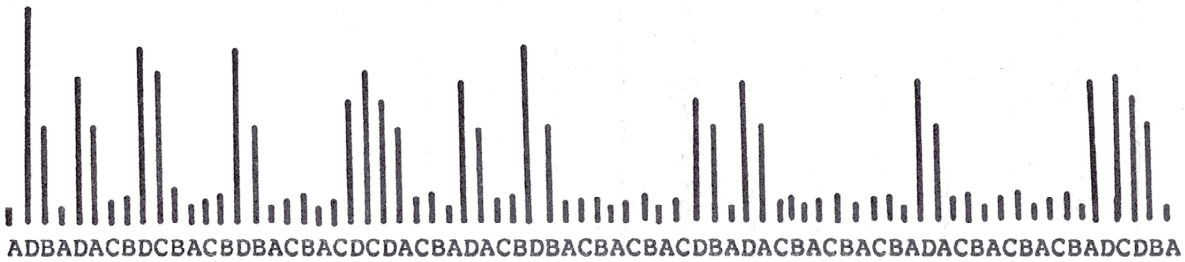

Figure 3d: Surprise at each descriptor in initial sequence of behaviour.

This example may also be used to illustrate a further feature of ATOM. Rather than inferring that a descriptor is a delimiter as a result of the ATOM analysis, we can specify in advance that it is to be treated as one. ATOM then only generates models in which that descriptor acts as a reset input, i.e. all transitions from it lead to only one state. This saves computation in searching the space of models, e.g. in deriving the 4-state model above over 10^8 4-state Mealy models were searched but if D is specified to be a delimiter this is to be reduced to 10^5 (if, additionally, only Moore models are searched this drops to 1). Hence, *a priori* structural knowledge may be traded directly for computing effort. However, if the structural hypothesis is incorrect then, as demonstrated in (Gaines, 1976e), the results obtained may be, not just an approximation, but instead totally meaningless.

The ability to specify a descriptor to be a delimiter is useful in allowing ATOM to be used to analyse behaviour that is a sub-monoid (i.e. a set of strings) rather than just a single string. If AB, CAA and ABC, are three separate observed behaviours for which a single common model. or grammar, is to be inferred, then they may be fed to ATOM as the single string, /AB/CAA/ABC/, with / specified as a delimiter. This technique has been used extensively in grammatical inference with ATOM and is discussed in (Gaines, 1976d).

Certain descriptors may also be specified to be inputs to ATOM and then the surprise at their occurrence is not added in to the measure of approximation, so that neither *NE* nor *LE* take the predictions of these descriptors into account when evaluating the models. Conversely an input/output distinction may be inferred from the surprise patterns of ATOM's models in which descriptors giving rise to zero surprise are taken to be *outputs,* the others being inputs. This form of inference is illustrated in (Gaines, 1976a) where sample of the input/output behaviour of a deterministic, non-autonomous automaton is fed to ATOM without the input output distinction being specified and this is readily inferred from the admissible models produced.

These computational studies illustrate clearly that the concepts developed in this paper are operational and can be applied to give clear analyses of actual data, to perform an automatic transformation from behaviour to model and thus identify the system giving rise to the behaviour.

5 Discussion

In section 1 I mentioned a wide range of philosophical studies relevant to the problem of system identification. In this final section I shall link some of the concepts and results of these studies to specific aspects of the approach to identification developed in previous sections and exhibited in ATOM. These links are intended to be suggestive and analogical rather than a formal correspondence. There is an important sense in which the formal, mathematical approach to identification stands alone independent of its interpretation—a single phenomenon within it may assume many different roles under different interpretations and may be consistent with a variety of conflicting philosophical positions. Conversely there is an equally important sense in which each philosophical position stands *per se* and is wider, and more enduring, than an approach to identification couched in current mathematical and system-theoretic concepts.

System theorists are naturally wary of tangling with philosophy and the philosophical literature. System studies have generally arisen from two main sources: applications of technology, particularly automation and computers, on the one hand: and biological and environmental modelling on the other. These practical foundations are somewhat remote from the fundamental

questions of philosophy, questions that to a large extent are dynamic and foundational just because no reasonable person would ask them. As Descartes remarks, "There is nothing so absurd or incredible that it has not been asserted by one philosopher or another". It is tempting to join Mach in his wistful remark, "I am a scientist and not a philosopher", but one also remembers that it is in his role of questioning established beliefs, as a philosopher of science, that his work had most impact on later developments such as relativity theory (Blackmore, 1972).

Conversely, philosophers are wary of becoming too enmeshed in the science and mathematics of their time. Particularly in this epoch we have come to accept change as the natural order of things, and to assume that the scientific thought of today will not be that of tomorrow. Nonstandard as a technical term in logic and mathematics has also become something of a laudatory adjective. Pluralism, relativism and a willingness to take a legalistic. rather than theological. approach to scientific debate, arguing from axioms rather than beliefs, are the bases of modern scientific thoughts. They do not give the appearance of a base on which to lay foundations, unless *laissez faire* is itself seen as a metaphysical position.

These disclaimers aside, however, there is much to be gained by bringing to bear every possible approach, every insight of philosophy, system theory and practice, upon the fundamentally intractable problem of knowledge acquisition, of system identification. In the next section I shall discuss the approach developed in this paper in the context of Klir's hierarchical model of system epistemology, and then discuss both the approach and Klir's model in the context of other philosophical studies.

5.1 Klir's Epistemological Hierarchy

Figure 4 shows Klir's (1976) hierarchy of epistemological levels of systems.

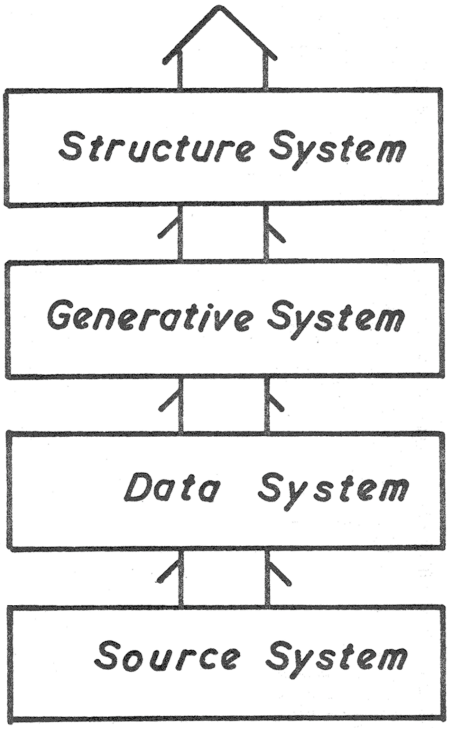

Figure 4: Klir's epistemological hierarchy

The lowest level is one of *source systems,* effectively one of data definition whereby the way in which behaviour will be described is defined and agreed. In terms of section 2.2 of this paper, this is the level of definition of *B*, the set of possible behaviours. More specifically. in terms of section 3, it is the level of definition of *D* and of possible $b \in B$ mapping from some event space *E* to *D*, i.e. of possible **D-sets**. Thus level zero in the hierarchy might be said to define a domain of *discourse* for the description of behaviour. In terms of the example of section 4.2, a statement at this zero level domain of discourse is any string of A's and B's.

The next level in the hierarchy is one of *data systems,* effectively one of system observation whereby the actual behaviour of some system is described in terms of the agreed domain of discourse at level zero. In terms of this paper, this is the level of definition of actual members of B of a **D-set** representing behaviour. Thus level one of the hierarchy might be said to define the *observed* (or *required*) behaviour of the system. In terms of the example of section 4.2, the data system is that string of A's and B's shown in Figure 1a.

The next level of the hierarchy is one *of generative* systems, effectively one of a stationary model for a system. In the context of system identification the model will arise as a hypothesis about the generation of a behaviour described at level one. Klir notes that the model may be deterministic or stochastic and that its invariance (or stationarity) is with respect to a set of variables that may include space and time. Thus the uniform applicability of these concepts to both sequential behaviour and spatial patterns, as noted in section 3, is intrinsic to Klir's presentation also. In terms of section 2.2 of this paper, level two of the hierarchy is one of the models from a set M appropriate to a behaviour at level one. This paper may now be seen as concerned with the relationships between levels one and two in Klir's hierarchy. In terms of the example of section 4.2, any of the automata shown in Figure 1d is a generative system at level two.

The next level in the hierarchy is one of *structure* systems. effectively one at which the models themselves are seen to have internal structure and hence to be analysable in other terms. This level may be regarded as one at which a number of atomic models are taken as co-existent, with the system being described in terms of a relationship between them. Alternatively, it may be regarded as one of non-stationary models that are changing along some dimension such as space or time. There is no formal discussion of structure at level three in this paper. However, in section 4.2 I have noted informally that the 5-state model produced may be viewed as a 2-state system loosely coupled to a 3-state system, and this is a statement at the level three domain of discourse of Klir's hierarchy.

The hierarchy proceeds to *meta-levels,* etc., but the four lowest levels are most relevant to this paper. In particular the clear separation of the source system from the data system, and of the generative system from the structure system, are very helpful in defining the boundaries of the current study. I have defined the source level in very general terms as a class of D-sets and have not been concerned with properties, or appropriateness, at this level. I have defined the set of models also in very general terms and have been concerned only with the behaviour they generate, not their internal structures. The focus on the interface between levels one and two has been quite deliberate because of the previous lack of a sufficiently precise formulation of how this relationship should be established and evaluated. This has previously made the rigorous formulation of the problem of *system identification* (behaviour-model transformation) impossible certainly for uncertain, e.g. stochastic or non-determinate, systems. The formulation in terms of

approximation and complexity in this paper, and the specific definitions of approximation and complexity for monoids. etc., resolve this problem.

Technically, and practically, the study must be extended to levels zero and three to be of real value. The general **D-set** with its total lack of relations between data terms is unrealistic for many practical problems. We have topological or metrical constraints upon the semantics of our data terms which should be taken into account at higher levels. For example, a measurement of 2V at a control system transducer is *between* one of 1V and one of 5V; it may even be said to be *nearer* the 1V reading than the 5V one; it might even make sense to talk of its being *twice* the 1V value; and so on. Thus real measurements are rarely the arbitrary, unrelated symbols that are all that are hypothesized in the definition of a **D-set** used in this paper.

However, it should be noted that any restriction placed on *B*, or **D-sets**, reduces the ontological neutrality of our definition. Such a restriction reflects a preconception about possible behaviours and hence about possible structures underlying them, e.g. we may condition ourselves to observe only causal, or even only linear, systems. The trade off, as illustrated in section 4.3, is between restrictions and speed-of-modelling data-requirements. A preconception may prevent us from discovering a property of the data, but lack of it, when it is justified, may both increase our computational burden and lead us to less precise conclusions than are justified. We cannot *afford* to re-discover that a system is linear every time we wish to check its transfer function, but if we measure only the linear describing function we shall never realise that a nonlinearity has appeared.

The determination of the structure at level three of the models at level two is an important problem in its own right. In deterministic automata theory the Hartmanis-Stearns (Hartmanis and Stearns, 1966) and Krohn-Rhodes (Arbib, Krohn and Rhodes, 1968) decomposition techniques lead essentially to structural representations at level three of models at level two. Similar techniques mar be applied to the probabilistic automaton models of section 4.2 to give a formal basis for the statement that the best 5-state model corresponds to coupled 2-state and 3-state models. The hierarchical models of systems developed and analysed over a long period by Mesarovic (Mesarovic, 1960; Mesarovic, Macko and Takahara, 1970; Mesarovic and Takahara, 1975) may be similarly automatically derived from model structures at level two, and much of "system theory" is concerned with such structural transformations (Klir, 1972). There is the same inherent arbitrariness in transformations between levels two and three that I have already noted between levels one and two. There is no one structural decomposition of a model that is correct but many of varying preferability. In recent years this relationship has begun to be analysed also in terms of (structural) *complexity* (Zeigler, 1974; Zeigler, 1975).

There is a direct relationship between measures of structural complexity at level three and the *model complexity* at level two assumed in this paper. The ordering of complexity over models required in section 2.2 corresponds to a level three construction of the least-complex structure for the model at level two, e.g. a *minimum-state* structure. Thus the counting of states that is taken as a basis for evaluating complexity in the specific automata-theoretic studies of this paper corresponds to a particular concept of the nature of systems at level three. However. it is only one possible measure even in the context of automata—we may well feel than an 8-state automaton that can be regarded as a certain form of product of three 2-state automata is rather less complex than one which cannot. This is one reason for defining complexity to be only a

partial ordering over models since this type of consideration does not necessarily lead to a simple numeric measure of complexity giving a chain order.

Such considerations are clearly of great importance but they do not affect the discussion of this paper where no constraint is placed on the ordering of model complexity, except to reinforce the statement in section 2.1 that the order relation is not intrinsic to the class of models. We do not normal envisage a model without also assuming a structure but, in fact. the same model may be regarded as having many different structures and complexities.

5.2 The Wider Context

The wider philosophical context for Klir's epistemological hierarchy and the operation of ATOM is far too wide to cover in this paper. It ranges through the whole gamut of philosophical positions and concepts mentioned in section 1, and does so necessarily in the sense that focusing on only one aspect itself distorts the overall picture. Much of the relevant discussion has taken place in the context of the behavioural sciences (Koestler and Smythies, 1969; Borger and Cioffi, 1970), including biology (Monod, 1972; Ayala and Dobzhansky, 1974; Lewis and Bohm, 1974), psychology (Wann, 1964; Wolman and Nagel, 1965; Care and Landesman, 1968; Mischel, 1969) sociology (Schutz, 1967; Blalock, 1971; Schutz and Luckmann, 1973; Comte and Andreski, 1974; Giddens, 1974), and economics (Shackle, 1955; Shackle, 1969; McClelland, 1975), because the problem of identification in the technological sciences is so much more readily defined—we are already aware of the structure, or structural possibilities, of systems that we ourselves have designed. The frontiers of the physical sciences stand between these extremes and have aspects of both (Feigl, Scriven and Maxwell, 1958; Feyerabend, Feigl and Maxwell, 1966; Radner and Winokur, 1970; Shanin, 1972), from the deep epistemological problems of particle physics (Körner, 1957; Bastin, 1971) and time (Gale, 1967; Gold and Bondi, 1967; Zeman, 1971) to the logical, postulational systems of dynamics (Bhatia and Szego, 1967; Auslander and Gottschalk, 1968). All of these differing subject areas have their own, varied but related, approaches to the problem of knowledge acquisition, and the problem of system identification, of behaviour - structure inference, is common to them all.

As noted in section 1, the formal approach to the inference of system properties developed in this paper, and by others, avoids rather than resolves many difficult problems. Hume's inductive scepticism leaves us with no logical justification of our processes of knowledge acquisition and hence what we actually do becomes somewhat arbitrary. This arbitrariness shows up in the present study at a number of levels: the choice of D-sets and their structures (if any); the choice of models; the ordering relation of complexity; the induced order relations of approximation; and so on. The effect of additional hypotheses to resolve some of the arbitrary decisions necessary has been discussed, and can be taken into account formally once the *decision* has been taken to include them.

Dilman (1973, p.19) argues that. although such decisions are *groundless,* they are, "not arbitrary in the sense that so much in our life hangs together with them." Wittgenstein (1972) emphasizes this view many times stating that, "Scepticism is not irrefutable, but obviously nonsensical when it tries to raise doubts where no question can be asked", and. "what the law of causality is meant to exclude cannot even be described". It is the basis of his remark, "If anyone said that information about the past could not convince him that something would happen in the future, I should not understand him." (Wittgenstein, 1953). This line of argument establishes a direct link

between Simon's (1973) "normative theory of law discovery" and our immediate predictive use of the "laws," even though they are only "patterns in data." The historian may be interested in models as providing a rational explanation of past events. but usually we intend them as guides to the future.

None of this, or any variant of the counter arguments to Hume (Swinburne, 1974) actually refutes the need for arbitrary (or, at least groundless) decisions, and, as a rejoinder to one line of argument that occurs at a number of levels. it is worth noting the metatheoretic result that: *the introduction of any hypothesis into the problem of identification allows the generation of counter-examples that do not satisfy the hypothesis and are modelled in a meaningless way.* This is rather too vague for formal proof at present but there are now sufficient examples to make it a reasonable inductive assertion. We cannot expect to introduce useful constraints on the inference process that are also ontologically neutral, and if we make an ontological commitment then we are in danger of not comprehending what are actually simple phenomena.

This is one of the problems, with its connotations of arbitrary decisions and commitment, that has lead to so-called existentialist philosophical studies, although the diversity of positions taken is such that one cannot speak of an existentialist "school." These studies may appear too introspective, often verging on solipsism and nihilism, to those whose standard is the classical philosophy of science. However, there is much to be gained in the understanding of the foundations of system theory, particularly in its application to the behavioural sciences. through the works of. for example, Merleau-Ponty (1964), Sartre (1943), Heidegger (1949) and Husserl (1965). Perhaps the motivation for studying the questions thus raised best comes from Becker's (1932) rather more humorous Yale lectures on the "heavenly city of eighteenth-century philosophers." To laugh at the preconceptions of the past in successive historic progression generates the momentum necessary to carry history into the future and allow us, as if retrospectively, to view the preconceptions of the present.

The authors noted above are associated with the *phenomenological* approach to the study of knowledge. whose proponents again range from the mystical to the scientifically respectable, notably Mach (Bradley, 1971; Blackmore, 1972). Phenomenology is concerned with the study of experience with a view to bringing out their "essences," their underlying reason." For Husserl (1965; Pivcevic, 1970), it was to be an analysis "free from presuppositions," as contrasted to the a priori bases for most metaphysics. Thus phenomenological analysis is concerned initially with the lowest level of Klir's hierarchy, one that is neglected, to some extent unrecognized, in classical systems theory. We are used to the terms in which behaviour is to be described being specified for us in advance and tend to regard the problem of identification as commencing beyond this point-a general method should be able to cope with any domain of discourse at a lower level. However, the terms of reference accepted at level zero propagate upwards through the hierarchy and dominate all higher-level concepts. The nature of such effects is best seen through the phenomenological case histories discussed by Spiegelberg (1976, e.g. the phenomenology of "force"), and Roche (1973). This last is of particular interest also for its comparison of phenomenological analysis with the "conceptual analysis" of logical positivism (Kolakowski, 1968).

Once above the level of existence and the phenomenon, the links between the problem of identification, Klir's hierarchy, and philosophical discussions are more obvious. The *structuralism* (Ehrmann, 1970) and *presuppositional* (Wilson, 1975) analyses of modern

linguistic philosophy and the linguistic orientation of logical positivism clearly have their places in what may be regarded as grammatically orientated inductive inference process. It is easy to integrate this approach to the problem of identification with what Giedymin (1975) terms the six doctrines of "strict positivism." However, equally there is no need to accept these doctrines—the order relations of section 2.2 define a form of rationality and hence the basis for a "rationality principle" (Popper, 1972; Koertge, 1975), but this need be none that we know or accept at present.

Even with all the low level choices made, with source level, data level, and the class of models at generative level together with their orderings, well determined, there are still decisions to be made. For example, the regular events corresponding to each of the 1 through 5-state admissible models of the behaviour analysed by ATOM in section 4.2 are each viable hypotheses as to the structure of the data. It was assumed in that section that we recognize the 5-state model as that which is "correct," but it may well be that the improvement in approximation from 1 to 2 states is enough to satisfy us—"facts are always presented at a sacrifice of completeness and never with greater precision than fits the needs of the moment" (Blackmore, 1972, p. 174). Having hypothesized that the event structure is of the form (BB*A)*, however, how do we proceed? Are we, with Popper (1959), to search for the sequence AA that refutes our hypothesis or, with Putman (1975), to search for the sequence BBB that is predicted but not yet seen. In ATOM terms, finding the first would force us back to the 1-state model, whereas not finding the second would push us on to the 3-state model. There is an interplay not only between the consistency of data and theory but also between the complexity of a theory and its value in approximating the data—Newtonian mechanics and special relativity are both *admissible* theories on all current data in the formal sense defined in this paper.

Finally, the discussion of the preceding section serves to illustrate the role of programs such as ATOM as themselves sources of phenomena relating to inductive inference and knowledge acquisition The *gedanken experiment* of having a *confirmation machine* (Erwin, 1971) is itself a source of stimulating discussion, but it is better by far to actually *have* one and to see it in action, both on data for which one "knows the answer" and on data for which one does not. There is nowadays no reason why any theory of knowledge acquisition should not be put into operational form and allowed, or required, to compete in real and artificial worlds. If we cannot understand *automata* then how can we hope to comprehend *ourselves,* and if we cannot make sense of what we *can* do then how may we say what we *should* do?

Summary and Conclusions

The problem of system identification has effectively been defined as: "given a sample of the behaviour of some system, to determine the admissible subspace of some prescribed class of models that would account for it." Admissibility is itself defined in terms of two order relations on models: a static order of complexity; and a dynamic order of approximation induced by the behaviour. An admissible model is such that any other with better approximation has greater complexity. It has been shown that this framework encompasses the identification of both deterministic and stochastic finite state automata, or grammars, but that it also encompasses non-temporal "behaviour" in, for example, pattern recognition.

The behaviour of the plot of approximation against complexity for admissible models has been investigated. It has been shown (section 3.1) that for zero approximation (perfect fit) nearly all

behaviours require models of nearly maximum complexity. A general basis for the measurement of approximation has been introduced related to work on logical, or semantic probability and the elicitation of subjective probabilities (section 3.2), and it **i**s shown that, with probabilistic sources, the approximation of an admissible model may be regarded as an entropy measure of the behaviour (section 3.3). The role of structural constraints upon models has been examined and concepts of causality, locality and uniformity have been introduced as "rules of the game" (section 3.3.1). The way in which the placement of *state* "tokens" allows local algorithms to recognize global properties has been discussed together with the concept of observability. Finally, the sense in which an identification algorithm can determine the structure of a probabilistic automaton has been analysed (section 3.3.2).

A computer implementation of an algorithm for determining admissible models of sequential behaviour, ATOM, has been described (section 4.1) and its operation illustrated through the analysis of some sample behaviours, including coupled automata (section 4.2) and non-recurrent sequences (section 4.3). Plots of the approximation against complexity, admissible models, and the "surprise" of the models at each description in the behaviour have been given. In particular, the capability of ATOM to infer, or specify in advance, special properties of the descriptors, such as being delimiters or inputs has been discussed and illustrated (section 4.4).

In the discussion section the behaviour of ATOM and the general approach to identification proposed have been analysed in terms of Klir's epistemological hierarchy of systems. It has been shown that the problem solved is one of the interface between levels one and two, of data systems and generative systems (section 5.1). The relationship of these to both lower and higher levels, and in particular the relationship of the model complexity used to structural complexity at level two have also been discussed. Finally, a brief summary has been presented or the relationship of these studies to various philosophical developments (section 5.2).

This paper gives a precise and complete basis for the analysis and solution of one of the major problems of system identification, that of behaviour-model transformation, effectively the movement from level one to level two of Klir's hierarchy. In that sense it "solves" the technical problem of structural, or grammatical inference, for nondeterministic and stochastic automata. and for a wide class of similar problems. The practicality of the solution is limited by the computational effort required, but it may well be that which we find ourselves searching large numbers of models we should ask whether we have the right formulation—would another model space be more appropriate?; and whether we are solving the right problem—would a sub-optimal solution be equally acceptable? Both questions are analysable within the framework of this paper.

On the other hand. I have attempted to indicate the way in which a formal solution of the interrelation of levels one and two relates to consideration at level zero, the *phenomenological* level, and level three, the *structural* level, and how all these levels relate to the various approaches to knowledge acquisition studied in various philosophical contexts. The availability nowadays of powerful interactive computers makes it possible to implement the system-theoretic methodologies imbedded in those philosophical positions and investigate them not as theories but as practice. It is through observation of, and interaction with, the autonomous operation of our theories, as computer algorithms competing with one another in the acquisition of knowledge of real and artificial data, that we have most to gain at present.

Acknowledgements

Many people have influenced the course of development of this paper and the work reported in it John Andreae. now at the University of Canterbury. New Zealand, and Alan Watson, at the Psychological Laboratory, Cambridge, did much to motivate the problem through wide-ranging discussions and their own approaches. ATOM itself developed in a competitive atmosphere in my own laboratory with the stimulation of colleagues, notable Peter Facey and Ian Witten. I am grateful to Michael Arbib, Richard Dawkins, Joe Goguen, George Klir, Ladislav Kohout, Steve Matheson and Roger Moore, for helpful discussions and comments, and to Judea Pearl of UCLA for detailed criticism of the work.

References

Aczel, J. (1971). On different characterizations of entropies. Behara, M., Krickeberg, K. and Wolfowitz, I., Ed. **Probability and Information Theory: Lecture Notes in Mathematics, 89**. Berlin, Springer.

Aczel, J. and Pfanzagl, J. (1966). Remarks on the measurement of subjective probability and information. **Metrika 11** 91-105.

Andreae, J.H. and Cleary, J.G. (1976). A new mechanism for a brain. **International Journal Man-Machine Studies 8** 89 -119.

Arbib, M.A. (1969). **Theories of Abstract Automata**. Englewood Cliffs, N.J., Prentice-Hall.

Arbib, M.A., Krohn, K. and Rhodes, J.L. (1968). **Algebraic Theory of Machines, Languages, and Semigroups**. New York, Academic Press.

Arbib, M.A. and Manes, E.G. (1974). Foundations of system theory: decomposable systems. **Automatica 10** 285-302.

Arbib, M.A. and Manes, E.G. (1975). A category-theoretic approach to systems in a fuzzy world. **Synthese 30** 381-406.

Auslander, J. and Gottschalk, W.H. (1968). **Topological Dynamics**. New York, Benjamin.

Ayala, F.J. and Dobzhansky, T.G. (1974). **Studies in the Philosophy of Biology: Reduction and Related Problems**. London, Macmillan.

Bar-Hillel, Y. (1964). **Language and Information: Selected Essays on Their Theory and Application**. Reading, Massachusetts, Addison-Wesley.

Bar-Hillel, Y. and Carnap, R. (1953). Semantic information. **British Journal for the Philosophy of Science 4** 147- 157.

Bastin, T., Ed. (1971). **Quantum Theory and Beyond**. Cambridge, Cambridge University Press.

Becker, C.L. (1932). **The Heavenly City of the Eighteenth Century Philosophers**. New Haven, Yale University Press.

Berkeley, G. (1710). **The Principles of Human Knowledge**. London, Collins.

Bhatia, N.P. and Szego, G.P. (1967). **Dynamical Systems: Stability Theory and Applications**. Berlin, Springer.

Biermann, A.W. (1972). On the inference of Turing machines from sample computations. **Artificial Intelligence 3** 181-198.

Biermann, A.W., Baum, R., Krishnaswomy, R. and Petry, F.E. (1973). Automatic program synthesis reports. Computer and Information Sciences Research Center, Ohio State University. OSU-CISRC-TR-73-6.

Birkhoff, G.D. (1927). **Dynamical Systems**. New York, American Mathematical Society.

Black, M. (1970). **Margins of Precision: Essays in Logic and Language**. Ithaca, NY, Cornell University Press.

Blackham, H.J. (1961). **Six Existentialist Thinkers**. London, Routledge & Kegan Paul.

Blackmore, J.T. (1972). **Ernst Mach: His Work, Life, and Influence**. Berkeley, University of California Press.

Blalock, H.M. (1971). **Causal Models in the Social Sciences**. Aldine-Atherton.

Bobrow, L.S. and Arbib, M.A. (1974). **Discrete Mathematics**. Philadelphia, Saunders.

Borger, R. and Cioffi, F., Ed. (1970). **Explanation in the Behavioural Sciences**. Cambridge, Cambridge University Press.

Bradley, J. (1971). **Mach's Philosophy of Science**. London, Athlone Press.

Brentano, F.C. (1973). **Psychology From an Empirical Standpoint**. London, Routledge and Kegan Paul.

Care, N.S. and Landesman, C., Ed. (1968). **Readings in the Theory of Action**. Bloomington, Indiana University Press.

Carnap, R. (1950). **Logical Foundations of Probability**. London, Routledge & Kegan-Paul.

Carnap, R. (1952). **The Continuum of Inductive Methods**. Chicago, University of Chicago Press.

Carnap, R. (1962). The aim of inductive logic. Nagel, E., Suppes, P. and Tarski, A., Ed. **Logic Methodology and Philosophy of Science**. pp.303-318. Stanford, CA, Stanford University Press.

Carnap, R. and Jeffrey, R.C. (1971). **Studies in Inductive Logic and Probability**. Berkeley, University of California Press.

Chaitin, G.J. (1975). A theory of program size formally identical to information theory. **Journal ACM 22** 329-340.

Chalfan, K. (1986). An expert system for design analysis. Kowalik, J.S., Ed. **Coupling Symbolic and Numerical Computing in Expert Systems**. Amsterdam, North-Holland.

Comte, A. and Andreski, S. (1974). **The Essential Comte: Selected from Cours de Philosophie Positive**. London, Croom Helm.

Crespi-Reghizzi, S., Melankoff, M.A. and Lichten, L. (1973). The use of grammatical inference for designing programming languages. **Communications ACM 16** 83-90.

Danielsson, S. (1967). Modal logic based on probability theory. **Theoria 33** 189-197.

Dawkins, M. and Dawkins, R. (1974). Some descriptive and explanatory stochastic models of decision-making. McFarland, D., Ed. **Motivational Control Systems Analysis**. pp.119-168. London, Academic Press.

Dawkins, R. and Dawkins, M. (1973). Decisions and the uncertainty of behaviour. **Behaviour 45** 83-103.

De Finetti, B. (1972). **Probability, Induction and Statistics: The Art of Guessing**. London, Wiley.

de Fiorino, A.C. (1966). Generalized Markov algorithms and automata. Caianiello, E.R., Ed. **Automata Theory**. New York, Academic Press.

Dilman, I. (1973). **Induction and Deduction: A Study in Wittgenstein**. Oxford, Blackwell.

Dorrough, D.C. (1970). A logical calculus of analogy involving functions of order 2. **Notre Dame Journal of Formal Logic 11** 321-336.

Duda, R.O. and Hart, P.E. (1973). **Pattern Classification and Scene Analysis**. Cheltenham, Wiley.

Dummett, M.A.E. (1973). The justification of deduction. **Proceedings British Academy 59** 1-34.

Ehrig, H. (1974). **Universal Theory of Automata**. Stuttgart, Teubner.

Ehrig, H. and Kreowski, H.J. (1973). Systematic approach of reduction and minimization in automata and system theory. Technische Universitat Berlin. 73-16.

Ehrmann, J., Ed. (1970). **Structuralism**. Garden City, Anchor Books.

Erwin, E. (1971). The confirmation machine. Buck, R.C. and Cohen, R.S., Ed. **PSA 1970: In Memory of Rudolph Carnap**. pp.306-321. Dordrecht, Holland, Reidel.

Eykhoff, P. (1974). **System Identification**. London, Wiley.

Facey, P.V. and Gaines, B.R. (1973). Real-time system design under an emulator embedded in a high-level language. **Proceedings DATAFAIR 73**. pp.285-291. London, British Computer Society.

Feigl, H., Scriven, M. and Maxwell, G., Ed. (1958). **Concepts, Theories and the Mind-Body Problem**. Minneapolis, University of Minnesota Press.

Feldman, J.A. (1972). Some decidability results on grammatical inference and complexity. **Information and Control 20** 244-262.

Feyerabend, P., Ed. (1975). **Against Method**. London, NLB.

Feyerabend, P.K., Feigl, H. and Maxwell, G., Ed. (1966). **Mind, Matter and Method: Essays in Philosophy and Science in Honor of Herbert Feigl**. Minneapolis, University of Minnesota Press.

Fine, T.L. (1973). **Theories of Probability: An Examination of Foundations**. New York, Academic Press.

Fitch, W.M. and Margolias, E. (1967). Construction of phylogenetic trees. **Science 155** 279-284.

Foster, M.H. and Martin, M.L. (1966). **Probability, Confirmation, and Simplicity: Readings in the Philosophy of Inductive Logic**. New York, Odyssey Press.

Fu, K.S. and Booth, T.L. (1975). Grammatical inference: Introduction and survey - Part II. **IEEE Transactions on Systems, Man & Cybernetics SMC-5** 409-423.

Gaines, B.R. (1975a). Approximate identification of automata. **Electronics Letters 11** 444-445.

Gaines, B.R. (1975b). A calculus of possibility, eventuality and probability. Department of Electrical Engineering Science, University of Essex. EES-MMS-FUZI-75.

Gaines, B.R. (1976a). Behaviour/structure transformations under uncertainty. **International Journal Man-Machine Studies 8**(3) 337-365.

Gaines, B.R. (1976b). Foundations of fuzzy reasoning. **International Journal of Man-Machine Studies 8**(6) 623-668.

Gaines, B.R. (1976c). Fuzzy reasoning and the logics of uncertainty. **Proceedings of the Sixth International Symposium on Multiple-Valued Logic, 76CH1111-4C**. pp.179-188. New York, IEEE.

Gaines, B.R. (1976d). Inference of Stochastic Grammars: A Formulation and Solution. Department of Electrical Engineering Science, University of Essex. EES-MMS-AUT-76.

Gaines, B.R. (1976e). On the complexity of causal models. **IEEE Transactions on Systems, Man & Cybernetics SMC-6**(1) 56-59.

Gaines, B.R. (1976f). The role of randomness in system theory. Department of Electrical Engineering, University of Essex. EES-MMS-RAN-76.

Gaines, B.R. and Facey, P.V. (1975). Some experience in interactive system development and application. **Proceedings Institute of Electrical and Electronics Engineers 63**(6) 894-911.

Gaines, B.R. and Kohout, L.J. (1975). The logic of automata. **International Journal of General Systems 2** 191-208.

Gaines, B.R. and Kohout, L.J. (1977). The fuzzy decade: a bibliography of fuzzy systems and closely related topics. **International Journal of Man-Machine Studies 9**(1) 1-68.

Gaines, B.R. and Witten, I.H. (1977). Stability and admissibility of adaptive threshold logic convergence. **IEEE Transactions on Computers C-26**(1) 88-91.

Gale, G. and Walter, E. (1973). Kordig and the theory-ladeness of observation. **Philosophy of science 40** 415-432.

Gale, R.M., Ed. (1967). **The Philosophy of Time**. Garden City, N.Y., Anchor Books.

Garner, W.R. and McGill, W.J. (1956). The relation between information and variance analyses. **Psychometrika 21** 219-228.

Gellner, E., Ed. (1974). **Legitimation of Belief**. Cambridge, UK, Cambridge University Press.

Giddens, A. (1974). **Positivism and Sociology**. London, Heinemann.

Giedymin, J. (1975). Antipositivism in contemporary philosophy of social science and humanities. **British Journal for the Philosophy of Science 26** 275-301.

Goguen, J.A. (1973). Realization is universal. **Mathematical Systems Theory 6** 359-374.

Goguen, J.A. (1974). Concept representation in natural and artificial languages: axioms, extensions and applications for fuzzy sets. **International Journal Man-Machine Studies 6** 513-561.

Goguen, J.A. (1975). Discrete-time machines in closed monoidal categories. **Journal of Computer and System Sciences 10** 1-43.

Gold, T. and Bondi, H., Ed. (1967). **The Nature of Time**. Ithaca, N.Y., Cornell University Press.

Good, I.J. (1962). Subjective probability as the measure of a nonmeasurable set. Nagel, E., Suppes, P. and Tarski, A., Ed. **Logic Methodology and Philosophy of Science**. pp.319-329. Stanford, CA, Stanford University Press.

Gregory, R.L. (1970). **The Intelligent Eye**. London, Weidenfeld & Nicolson.

Grenander, U. (1969). Foundations of pattern analysis. **Quarterly Applied Mathematics 27** 1-55.

Grofman, B. and Hyman, G. (1973). Probability and logic in belief systems. **Theory and Decision 4** 179- 195.

Haack, S. (1976). The justification of deduction. **Mind 85** 112-119.

Hacking, I. (1975). **The Emergence of Probability: A Philosophical Study of Early Ideas about Probability, Induction and Statistical Inference**. Cambridge, Cambridge University Press.

Hanson, N.R. (1958). **Patterns of Discovery: An Inquiry into the Conceptual Foundations of Science**. Cambridge, Cambridge University Press.

Harrod, R.F. (1956). **Foundations of Inductive Logic**. London, Macmillan.

Hartmanis, J. and Stearns, R.E. (1966). **Algebraic Structure Theory of Sequential Machines**. Englewood Cliffs, N.J., Prentice-Hall.

Heidegger, M. (1949). **Existence and Being**. Chicago,, H. Regnery.

Hellman, M. and Cover, T. (1970). Learning with finite memory. **Annals of Mathematical Statistics 41**(3) 765-782.

Hesse, M.B. (1966). **Models and Analogies in Science**. Notre Dame, Notre Dame University Press.

Hesse, M.B. (1974). **The Structure of Scientific Inference**. London, Macmillan.

Hilpinen, R. (1970). On the information provided by observations. Hintikka, J. and Suppes, P., Ed. **Information and Inference**. pp.97-122. Dordrecht, Reidel.

Hintikka, J. (1970). On semantic information. Hintikka, J. and Suppes, P., Ed. **Information and Inference**. pp.3-27. Dordrecht, Reidel.

Hoffman, W.C. (1966). The Lie algebra of visual perception. **Journal Mathematical Psychology 3** 65-98.

Hogarth, R.M. (1975). Cognitive processes and the assessment of subjective probability distributions. **Journal American Statistical Association 70** 271-294.

Hopcroft, J. (1971). An n log n algorithm for minimizing states in a finite automaton. Kohavi, Z. and Paz, A., Ed. **Theory of Machines and Computations**. pp.189-196. New York, Academic Press.

Hughes, G.E. and Cresswell, M.J. (1968). **An Introduction to Modal Logic**. London, Methuen.

Hume, D. (1777). **Enquiries Concerning Human Understanding**. Oxford, Clarendon Press.

Husserl, E. (1965). **Phenomenology and the Crisis of Philosophy**. New York, Harper & Row.

Jeffrey, R.C. (1965). **The Logic of Decision**. New York, McGraw-Hill.

Kalman, R.E. (1957). Nonlinear aspects of sampled-data control systems. **Proceedings Symposium on Nonlinear Circuit Analysis**. pp.273-313. New York, Polytechnic Institute of Brooklyn.

Kant, I. (1781). **Critique of Pure Reason**. London, George Bell.

Katz, J.J. (1962). **The Problem of Induction and its Solution**. Chicago, University of Chicago Press.

Kaufmann, A. (1975). **Introduction to the Theory of Fuzzy Subsets**. New York, Academic Press.

Kling, R.E. (1971). A paradigm for reasoning by analogy. **Proceedings 2nd International Joint Conference in Artificial Intelligence**. pp.568-585. London, British Computer Society.

Klir, G.J., Ed. (1972). **Trends in General Systems Theory**. New York, Wiley.

Klir, G.J. (1975). On the representation of activity arrays. **International Journal General Systems 2** 149-168.

Klir, G.J. (1976). Identification of generative structures in empirical data. **International Journal of General Systems, 3** 89-104.

Klir, G.J. and Uttenhove, J.J. (1979). Computerized methodology for structure modelling. **Annals of Systems Research 5** 29-66.

Koertge, N. (1975). Popper's metaphysical research program for the human sciences. **Inquiry 18** 437-462.

Koestler, A. and Smythies, J.R., Ed. (1969). **Beyond Reductionism: New Perspectives in the Life Sciences**. London, Hutchinson.

Kolakowski, L. (1968). **The Alienation of Reason: A History of Positivist Thought**. Garden City, N.Y., Doubleday.

Kolmogorov, A.N. (1968). Logical basis for information theory and probability. **IEEE Transactions on Information Theory IT-14** 662-664.

Körner, S., Ed. (1957). **Observation and Interpretation in the Philosophy of Physics, with Special Reference to Quantum Mechanics**. New York, Dover Publications.

Kuhn, T.S. (1962). **The Structure of Scientific Revolutions**. Chicago, University of Chicago Press.

Kullback, S. (1968). **Information Theory and Statistics**. New York, Dover Publications.

Kwakernaak, H. (1965). Admissible adaptive control. **Proceedings IFAC Symposium on The Theory of Self-Adaptive Control Systems**. London, Society of Instrument Technology.

Kyburg, H.E. (1970). **Probability and Inductive Logic**. New York, Macmillan.

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. Lakatos, I. and Musgrave, A., Ed. **Criticism and the Growth of Knowledge**. Cambridge, Cambridge University Press.

Lempel, A. and Ziv, J. (1976). On the complexity of finite sequences. **IEEE Transactions on Information Theory IT-22** 75-81.

Levi, I. (1973). **Gambling with Truth: An Essay on Induction and the Aims of Science**. Cambridge, MA, MIT Press.

Lewis, D.K. (1969). **Convention: A Philosophical Study**. Cambridge, MA, Harvard University Press.

Lewis, J. and Bohm, D., Ed. (1974). **Beyond Chance and Necessity: A Critical Inquiry into Professor Jacques Monod's Chance and Necessity**. Paris, The Teilhard Centre for the Future of Man.

Locke, J. (1690). **An Essay Concerning Human Understanding**. London, Hills.

Mackie, J.L. (1974). **The Cement of the Universe: A Study of Causation**. Oxford, Clarendon Press.

Madden, E.H. (1971). Hume and the fiery furnace. **Philosophy of Science 38** 64-78.

Markov, A.A. (1954). **Theory of Algorithms**. Moscow, Academy of Sciences of the USSR.

Martin-Lof, P. (1966). The definition of random sequences. **Information and Control 9** 602-619.

Maryanski, F.J. (1974). Inference of Probabilistic Grammars. University of Connecticut.

Mathai, A.M. and Rathie, P.N. (1975). **Basic Concepts in Information Theory and Statistics: Axiomatic Foundations and Applications**. New York, Wiley.

Maxwell, G. and Anderson, R.M. (1975). **Induction, Probability, and Confirmation**. University of Minnesota Press.

McClelland, P.D. (1975). **Causal Explanation and Model Building in History, Economics, and the New Economic History**. Ithaca, Cornell University Press.

McNaughton, R. and Papert, S. (1971). **Counter-Free Automata**. Cambridge, MA, M.I.T. Press.

Menges, G. (1970). On subjective probability and related problems. **Theory and Decision 1** 40-60.

Menges, G. (1974). **Information, Inference and Decision**. Dordrecht, Reidel.

Merleau-Ponty, M. (1964). **The Primacy of Perception**. Evanston, IL, Northwestern University Press.

Merton, R.K. (1973). **The Sociology of Science: Theoretical and Empirical Investigations**. Chicago, University of Chicago Press.

Mesarovic, M.D. (1960). **The Control of Multivariable Systems**. New York, Wiley.

Mesarovic, M.D., Macko, D. and Takahara, Y. (1970). **Theory of Hierarchical, Multilevel Systems**. New York,, Academic Press.

Mesarovic, M.D. and Takahara, Y. (1975). **General Systems theory: Mathematical Foundations**. New York, Academic Press.

Michalos, A.C. (1971). **The Popper-Carnap Controversy**. The Hague, Nijhoff.

Miller, G.A. and Madonna, W.G. (1954). On the maximum likelihood estimate of the Shannon-Wiener measure of information. Operational Applications Laboratory, ARDC. AFCRC-TR-54-75.

Minsky, M.L. and Papert, S. (1969). **Perceptrons: An Introduction to Computational Geometry**. Cambridge, MA, MIT Press.

Mischel, T. (1969). **Human Action**. New York, Academic Press.

Miura, S. (1972). Probabilistic models of modal logics. **Bulletin Nagoya Institute Technology 24** 67-72.

Monod, J. (1972). **Chance and Necessity**. London, Collins.

Moore, D.J.H. (1971). A theory of form. **International Journal Man-Machine Studies 3** 31-59.

Nelson, R.J. (1975). On machine expectations. **Synthese 31** 129-139.

Nerode, A. (1958). Linear automata transformations. **Proceedings American Mathematical Society 9** 541-544.

Newell, A. and Simon, H.A. (1972). **Human Problem Solving**. Englewood Cliffs, NJ, Prentice-Hall.

Orava, P.I. (1971). Notion of dynamical input-output systems: causality and state concepts. **International Journal System Science 5** 793-806.

Patel, A.R. (1972). Grammatical Inference for Probabilistic Finite State Languages. Ph.D. University of Connecticut.

Pearl, J. (1975a). An economic basis for certain methods of evaluating probabilistic forecasts. School of Engineering and Applied Science, UCLA. UCLA-ENG-7561.

Pearl, J. (1975b). On the complexity of computing probabilistic assertions. School of Engineering and Applied Science, UCLA. UCLA-ENG-7562.

Pearl, J. (1975c). On the complexity of imprecise causal models. School of Engineering and Applied Science, UCLA. UCLA-ENG-7560.

Pearl, J. (1975d). On the complexity of inexact computations. School of Engineering and Applied Science, UCLA. UCLA-ENG-0775.

Perles, M., Rabin, M.O. and Shamir, E. (1963). The theory of definite automata. **IEEE Transactions on Electronic Computers EC-12** 233-243.

Pivcevic, E. (1970). **Husserl and Phenomenology**. London, Hutchinson.

Pivcevic, E. (1975). **Phenomenology and Philosophical Understanding**. Cambridge, Cambridge University Press.

Plato (380BC). **The Republic**. Baltimore, Penguin Books.

Polanyi, M. (1958). **Personal Knowledge: Towards a Post-critical Philosophy**. Chicago, University of Chicago Press.

Popper, K.R. (1959). **The Logic of Scientific Discovery**. London, Hutchinson.

Popper, K.R. (1963). **Conjectures and Refutations: The Growth of Scientific Knowledge**. London, Routledge & Kegan Paul.

Popper, K.R. (1972). **Objective Knowledge: an Evolutionary Approach**. Oxford, Clarendon Press.

Post, H.R. (1960). Simplicity and scientific theories. **British Journal for the Philosophy of Science 11** 32-41.

Prior, A.N. (1967). **Past, Present and Future**. Oxford, Clarendon Press.

Putnam, H. (1964). Robots: Machines or artificially created life. **Journal of Philosophy** (61) 668-691.

Putnam, H. (1975). The 'corroboration' of theories. Putnam, H., Ed. **Mathematics, Matter, and Method**. pp.250-269. Cambridge, Cambridge University Press.

Rabin, M.O. and Scott, D. (1959). Finite automata and their decision problem. **IBM Journal of Research and Development 3** 114-125.

Radner, M. and Winokur, S., Ed. (1970). **Analyses of Theories and Methods of Physics and Psychology**. Minneapolis, University of Minnesota Presss.

Reichenbach, H. (1949). **The Theory of Probability: An Inquiry into the Logical and Mathematical Foundations of the Calculus of Probability**. Berkeley, University of California Press.

Rescher, N. (1963). A probabilistic approach to modal logic. **Acta Philosophica Fennica 16** 215-226.

Rescher, N. (1969). **Many-Valued Logic**. New York, McGraw-Hill.

Rescher, N. (1970). **Scientific Explanation**. New York, Free Press.

Rescher, N. (1973). **The Primacy of Practice: Essays Towards a Pragmatically Kantian Theory of Empirical Knowledge**. Oxford, Blackwell.

Roche, M. (1973). **Phenomenology, Language and the Social Sciences**. London, Routledge.

Ruzavin, G.I. (1970). Probability logic and its role in scientific research. Tanavec, P.V., Ed. **Problems of the Logic of Scientific Knowledge**. pp.212-265. Dordrecht, Holland, D. Reidel.

Salovaara, S. (1974). On set theoretical foundations of system theory-a study of the state concept. **Acta Polytechnica Scandinavica 15** 1-74.

Sankoff, D. (1972). Matching sequences under deletion insertion constraints. **Proceedings National Academy Science USA 69** 4-6.

Sartre, J.P. (1943). **L'Etre et le Néant**. Paris, Gallimard.

Savage, I.J. (1971). Elicitation of personal probabilities and expectations. **Journal American Statistical Association 66** 783 -801.

Savage, L.J. (1954). **The Foundations of Statistics**. New York, Wiley.

Schilpp, P.A. (1963). **The Philosophy of Rudolf Carnap**. La Salle, Illinois, Open Court.

Schilpp, P.A. (1974). **The Philosophy of Karl Popper**. La Salle, Illinois, Open Court.

Schnorr, C.P. (1971). A unified approach to the definition of random sequences. **Mathematical System Theory 5** 246-258.

Schutz, A. (1967). **The Phenomenology of the Social World**. Toronto, Northwestern University Press.

Schutz, A. and Luckmann, T. (1973). **The Structures of the Life-World**. London, Heinemann.

Sellers, P.H. (1974). An algorithm for the distance between two finite sequences. **Journal Combinatorial Theory (A) 16** 253-258.

Shackle, G.L.S. (1955). **Uncertainty in Economics**. Cambridge, Cambridge University Press.

Shackle, G.L.S. (1969). **Decision, Order, and Time in Human Affairs**. Cambridge, Cambridge University Press.

Shanin, T. (1972). **The Rules of the Game: Cross-disciplinary Essays on Models in Scholarly Thought**. London, Tavistock Publications.

Shuford, E. and Brown, T.A. (1975). Elicitation of personal probabilities and their assessment. **Instructional Science 4** 137-188.

Shuford, E.H., Albert, A. and Massengill, H.E. (1966). Admissible probability measurement procedures. **Psychometrika 31** 125-145.

Simon, H.A. (1973). Does scientific discovery have a logic? **Philosophy of Science 40** 471-480.

Smith, C.A.B. (1961). Consistency in statistical inference and decision. **Journal Royal Statistical Society B 23** 1-25.

Smith, C.A.B. (1965). Personal probability and statistical analysis. **Journal Royal Statistical Society A 128** 469-499.

Snyder, D.P. (1971). **Modal Logic and its Applications**. New York, Van Nostrand Reinhold.

Sober, E. (1975). **Simplicity**. Oxford, Clarendon Press.

Solomonoff, R.J. (1964). A formal theory of inductive inference. **Information and Control 7** 1-22, 224-254.

Spiegelberg, H. (1976). **The Phenomenological Movement: A Historical Introduction**. Hague, Nijhoff.

Stove, D.C. (1973). **Probability and Hume's Inductive Scepticism**. Oxford,, Clarendon Press.

Suppes, P. (1984). **Probabilistic Metaphysics**. New York, Blackwell.

Suppes, P. and Rottmayer, W. (1974). Automata. Carterette, E.C. and Friedman, M.P., Ed. **Handbooks of Perception, Vol. 1, Historical and Philosophical Roots of Perception**. New York, Academic Press.

Swinburne, R. (1973). **An Introduction to Confirmation Theory**. London, Methuen.

Swinburne, R., Ed. (1974). **The Justification of Induction**. Oxford, Oxford University Press.

Uemov, A.I. (1970). The basic forms and rules of inference by analogy. Tanasec, P.V., Ed. **Problems of the Logic of Scientific Knowledge**. pp.266-311. Dordrecht, D. Reidel.

Vickers, J.M. (1965). Some remarks on coherence and subjective probability. **Philosophy of Science 32** 32-38.

Villegas, C. (1964). On qualitative probability sigma-algebras. **Annals of Mathematical Statistics 1964** 1787-1796.

Vleck, C.A.J. (1970). Multiple probability learning. **Acta Psychologica 33** 207-232.

Wagner, R.A. and Fischer, M.J. (1974). The string-to-string correction procedure. **Journal ACM 21** 168-173.

Wann, T.W., Ed. (1964). **Behaviorism and Phenomenology: Contrasting Bases for Modern Psychology**. Chicago, University of Chicago Press.

Watanabe, S. (1969). **Knowing and Guessing: A Quantitative Study of Inference and Information**. New York, Wiley.

Weiss, L. (1961). **Statistical Decision Theory**. New York, McGraw-Hill.

Wharton, R.M. (1974). Approximate language identification. **Information and Control 26**(3) 236-255.

Willis, D. (1970). Computational complexity and probability constructions. **Journal ACM 17** 241-259.

Wilson, D. (1975). **Presuppositions and Non-Truth-Conditional Semantics**. London, Academic Press.

Windeknecht, T.G. (1967). Mathematical Systems theory: causality. **Mathematical System Theory 1** 279-288.

Windeknecht, T.G. (1971). **General Dynamical Processes: A Mathematical Introduction**. New York, Academic Press.

Winkler, R.C. and Murphy, A.H. (1968). Good probability assessors. **Journal of Applied Meteorology 7** 751-758.

Winkler, R.L. (1971). Probabilistic prediction: some experimental results. **Journal American Statistical Association 66** 625-688.

Witten, I.H. (1976). The apparent conflict between estimation and control-a survey of the two-armed bandit problem. **Journal Franklin Institute 301** 161-189.

Wittgenstein, L. (1953). **Philosophical Investigations**. Oxford, Blackwell.

Wittgenstein, L. (1972). **Tractatus Logico-Philosophicus**. London, Routledge & Kegan Paul.

Wolman, B.B. and Nagel, E. (1965). **Scientific Psychology: Principles and Approaches**. New York, Basic Books.

Wright, G.H.v. (1962). Remarks on the epistemology of subjective probability. Nagel, E., Suppes, P. and Tarski, A., Ed. **Logic Methodology and Philosophy of Science**. pp.330-339. Stanford, CA, Stanford University Press.

Zadeh, L.A. (1962). From circuit theory to system theory. **Procceedings IRE 50** 856-865.

Zadeh, L.A. (1964). The concept of state in system theory. Mesarovic, M.D., Ed. **Views on General Systems Theory: Proceedings of the Second Systems Symposium at Case Institute of Technology**. New York, John Wiley.

Zadeh, L.A. (1969). The concepts of system, aggregate, and state in system theory. Zadeh, L.A. and Polak, E., Ed. **System Theory**. pp.3-42. New York, McGraw-Hill.

Zadeh, L.A. (1975). **Fuzzy Sets and Their Applications to Cognitive and Decision Processes**. New York, Academic Press.

Zadeh, L.A. (1976). A fuzzy algorithmic approach to the definition of complex or imprecise concepts. **International Journal Man-Machine Studies 8** 249-291.

Zalcstein, Y. (1971). Locally testable languages. Department of Computer Science, Carnegie Mellon University. Technical Report.

Zeigler, H.P. (1974). A conceptual basis for modelling and simulation. **International Journal General Systems 1** 213-228.

Zeigler, H.P. (1975). Simulation based structural complexity of models. **International Journal General Systems 2** 217-223.

Zeman, J., Ed. (1971). **Time in Science and Philosophy**. Amsterdam, Elsevier.