4. Schemata

An adaptive system faces its principal challenge when the set of possible structures $\mathcal{A}$ is very large and the performance functions μ_E involve many local maxima. It is important then for the adaptive system to provide itself with whatever insurance it can against a prolonged search. It is clear that the search of $\mathcal{A}$ must go on so long as significant improvements are possible (unless the system is to settle for inferior performance throughout the remainder of its history). At the same time, unless it exploits possibilities for improved performance *while* the search goes on, the system pays the implicit cost of a performance less even than the best among *known* alternatives. Moreover, unexploited possibilities may contain the key to optimal performance, dooming the system to fruitless search until they are implemented. There is only one insurance against these contingencies. The adaptive system must, as an integral part of its search of $\mathcal{A}$, persistently test and incorporate structural properties associated with better performance. As with most insurance, this particular policy contains a limiting clause: useful properties must be identified to be exploited. The present chapter is concerned with this limitation.

Almost by definition useful properties are points of comparison between structures yielding better-than-average performance. The question then is: How are the structures in $\mathcal{A}$ to be compared? If the structures are built up from components, comparison in terms of common components is natural and the question becomes: How is credit for the above-average performance of a structure to be apportioned to its components? A more general approach uses feature detectors (see section 3.4) to make comparisons. Since one can find an appropriate detector for any effectively describable feature of structures in $\mathcal{A}$ (including the presence or absence of given components) this approach is well suited to present purposes.

To begin with let us see how comparisons can be developed when a finite set of detectors $\{\delta_i: \mathcal{A} \to V_i, i = 1, \ldots, l\}$ is given. In terms of the given detectors each structure $A \in \mathcal{A}$ will have a *representation* $(\delta_1(A), \delta_2(A), \ldots, \delta_l(A))$; that is, each structure A will be described by its particular ordered set of l *attributes* or detector values $\delta_i(A) \in V_i$, $i = 1, \ldots, l$. Thus, for a chromosome A, V_i can desig-

nate the set of alleles of locus i (see section 3.1) and the corresponding representation of A is the specification of the ordered set of alleles which make up the chromosome. For a von Neumann economy (section 3.2) the V_i can designate the possible levels of the ith activity so that the representation of a mixture of activities is simply the corresponding activity vector. Similar considerations apply to each of the remaining illustrations of chapter 3. Clearly, with a given set of l detectors, two structures will be distinguishable only insofar as they have distinct representations. Since, in the present chapter, we are only interested in comparisons let us assume that all structures in $\mathfrak{A}$ are distinguishable (have distinct representations) or, equivalently, that $\mathfrak{A}$ is used to designate distinguishable subsets of the original set of structures. For simplicity in what follows $\mathfrak{A}$ will simply be taken to *be* the set of representations provided by the detectors (rather than the abstract elements so represented).

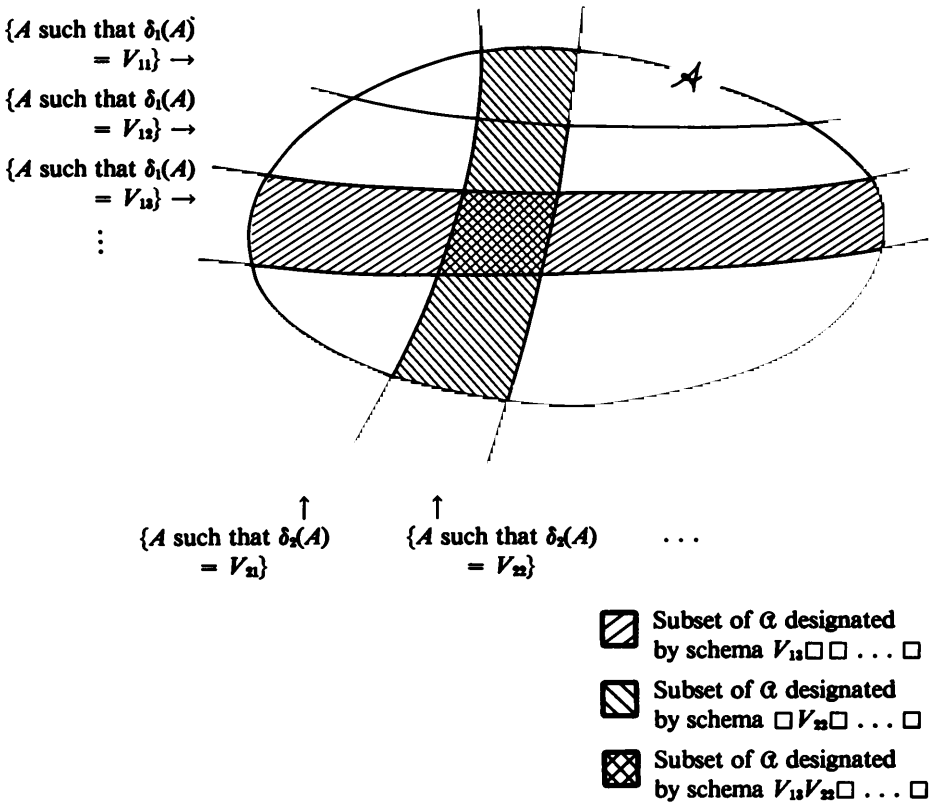

Fig. 9. Diagrammatic presentation of some schemata

Now our objective is to designate subsets of $\mathcal{A}$ which have attributes in common. To do this let the symbol "□" indicate that we "don't care" what attribute occurs at a given position (i.e., for a given detector). Thus $(v_{13}, \square, \square, \ldots, \square)$ designates the subset of all elements in $\mathcal{A}$ having the attribute $v_{13} \in V_1$. (Equivalently, $(v_{13}, \square, \ldots, \square)$ designates the set of all l-tuples in $\mathcal{A}$ beginning with the symbol v_{13}; hence, for $l = 3$, (v_{13}, v_{22}, v_{32}) and (v_{13}, v_{21}, v_{31}) belong to $(v_{13}, \square, \square)$, but (v_{12}, v_{22}, v_{32}) does not.) The set of all l-tuples involving combinations of "don't cares" and attributes is given by the augmented product set $\Xi = \Pi_{i=1}^{l} \{V_i \cup \{\square\}\}$. Then any l-tuple $\xi = (\Delta_{i_1}, \Delta_{i_2}, \ldots, \Delta_{i_l}) \in \Xi$ designates a subset of $\mathcal{A}$ as follows: $A \in \mathcal{A}$ belongs to the subset if and only if (i) whenever $\Delta_{i_j} = \square$, any attribute from V_j may occur at the jth position of A, and (ii) whenever $\Delta_{i_j} \in V_j$, the attribute Δ_{i_j} must occur at the jth position of A. (For example, $(v_{11}, v_{21}, v_{31}, v_{43})$ and $(v_{13}, v_{21}, v_{32}, v_{43})$ belong to $(\square, v_{21}, \square, v_{43})$ but $(v_{11}, v_{21}, v_{31}, v_{42})$ does not.) The set of l-tuples belonging to Ξ will be called the set of *schemata;* Ξ amounts to a decomposition of $\mathcal{A}$ into a large number of subsets based on the representation in terms of the l detectors $\{\delta_i : \mathcal{A} \to V_i, i = 1, \ldots, l\}$.

Schemata provide a basis for associating combinations of attributes with potential for improving current performance. To see this, let "improvement" be defined as any increment in the average performance over past history. That is, if $\mu(\mathcal{A}(t))$ is the performance of the structure $\mathcal{A}(t)$ tried at time t, the object is to discover ways of incrementing

$$\bar{\mu}(T) = \frac{1}{T} \sum_{t=1}^{T} \mu(\mathcal{A}(t)).$$

(A more sophisticated measure would give more weight to recent history, using

$$\mu'(T) = \left(\sum_{t=1}^{T} c_t \mu(\mathcal{A}(t))\right) / \left(\sum_{t=1}^{T} c_t\right), \quad c_t > c_{t'} \quad \text{for } t > t',$$

but the simple average suffices for the present discussion.) Though $\bar{\mu}(T)$ can be incremented by simply repeating the structure yielding the best performance up to time T this does not yield new information. Hence the object is to find *new* structures which have a high probability of incrementing $\bar{\mu}(T)$ significantly. An adaptive plan can use schemata to this end as follows: Let $A \in \mathcal{A}$ have a probability $P(A)$ of being tried by the plan τ at time $T + 1$. That is, τ induces a probability distribution P over $\mathcal{A}$ and, under this distribution, $\mathcal{A}$ becomes a sample space. The performance measure μ then becomes a random variable over $\mathcal{A}$, $A \in \mathcal{A}$ being tried with probability $P(A)$ and yielding payoff $\mu(A)$. More importantly, any schema $\xi \in \Xi$ designates an event on the sample space $\mathcal{A}$. Thus, the restriction $\mu \mid \xi$ of μ to

the subset designated by ξ, is also a random variable, $A \in \xi$ being tried with probability $(P(A))/(\sum_{A' \in \xi} P(A'))$ and yielding payoff $\mu(A)$. In what follows ξ will be used to designate both an element of Ξ and the corresponding random variable with sample space ξ, the particular usage being specified when it is not clear from context. As a random variable, ξ has a well-defined average μ_ξ (and variance σ_ξ^2) where, intuitively, μ_ξ is the payoff expected when an element of ξ is randomly selected under the marginal distribution $(P(A))/(\sum_{A' \in \xi} P(A'))$.

Clearly, when $\mu_\xi > \bar{\mu}(T)$, instances of ξ (i.e., $A \in \xi$) are likely to exhibit performance better than the current average $\bar{\mu}(T)$. This suggests a simple procedure (a bit too simple as it turns out) for exploiting combinations of attributes associated with better-than-average performance while further searching $\mathcal{A}$: (i) try instances of various schemata until at least one schema ξ is located which exhibits a sample average $\hat{\mu}_\xi > \bar{\mu}(T)$; (ii) generate *new* instances of the (observed) above-average schema ξ, returning to step (i) from time to time (particularly when the increasing overall average $\bar{\mu}(T)$ comes close to $\hat{\mu}_\xi$) to locate new schemata $\{\xi'\}$ for which $\hat{\mu}_{\xi'} > \bar{\mu}(T)$. In effect, then, credit is apportioned to a combination of attributes in accord with the observed average performance of its instances. This procedure has some immediate advantages over a fixed random (or enumerative) search of $\mathcal{A}$: it generates improvements with high probability while gathering new information by trying new $A \in \mathcal{A}$; furthermore, the new trials of the above-average schema ξ increase confidence that the observed average $\hat{\mu}_\xi$ closely approximates μ_ξ. It is oversimple because each instance $A \in \mathcal{A}$ tried yields information about a great many schemata other than ξ—information which is not used.

Given l detectors, a single structure $A \in \mathcal{A}$ is an instance of 2^l distinct schemata, as can be easily affirmed by noting that A is an instance of any schema ξ defined by substituting "□"s for one or more of the l attribute values in A's representation. Thus a single trial A constitutes a trial of 2^l distinct random variables, yielding information about the expected payoff μ_ξ of each. (If l is only 20 this is still information about a million schemata!) Any procedure which uses even a fraction of this information to locate ξ for which $\mu_\xi > \bar{\mu}(T)$ has a substantial advantage over the one-at-a-time procedure just proposed.

Exploiting this tremendous flow of information poses a much more clearly defined challenge than the one which started the chapter. Schemata have advanced our understanding, in this sense, but the new problem is difficult. The amount of storage required quickly exceeds all feasible bounds if one attempts to record the average payoff of the observed instances of each schema sampled. Moreover, the information will be employed effectively only if it is used to generate new $A \in \mathcal{A}$ which, individually, test as many above-average schema as possible. The adaptive

system is thus faced with a specific problem of compact storage, access, and effective use of information about extremely large numbers of schemata. Chapter 6 ("Reproductive Plans and Genetic Operators") sets forth a resolution of these difficulties, but a closer look at schemata (the remainder of this chapter) and the optimal allocation of trials to sets of schemata (the next chapter) provides the proper setting.

Let us begin with a concrete, but fairly general, interpretation of schemata stemming from the earlier discussion of control and function optimization (section 3.5, p. 57). Consider an arbitrary bounded function $f(x)$, $0 \leqq x < 1$, and assume that x is specified to an accuracy of one part in a million or, equivalently, that values of x are discretely represented by 20 bits. Define $\mathcal{A}$ to be the set of 2^{20} discrete values of x represented with 20 detectors $\{\delta_j : \mathcal{A} \rightarrow \{0, 1\}, j = 1, \ldots, 20\}$ where $\delta_j(x)$, $x \in \mathcal{A}$, assigns to x the value of the jth bit in the binary expansion of x. The schema $1\square\square\ldots\square$ then is just the right half-plane $\frac{1}{2} \leqq x < 1$, while the schema $\square\square0\square\ldots\square$ is a set of four strips $\{0 \leqq x < \frac{1}{8}, \frac{1}{4} \leqq x < \frac{3}{8}, \frac{1}{2} \leqq x < \frac{5}{8}, \frac{3}{4} \leqq x < \frac{7}{8}\}$ and the schema $1\square0\square\ldots\square$ is the intersection of the two previous schemata $\{\frac{1}{2} \leqq x < \frac{5}{8}, \frac{3}{4} \leqq x < \frac{7}{8}\}$ (see Figure 10).

With this representation there are 3^{20} distinct schemata since any 20-tuple over the set $\{0, 1, \square\}$ defines a schema. (More technically, the schemata are simply hyperplanes, of dimension 20 or less, in the 20-dimensional space of detector-value combinations.) Note that there are many points, such as $x = \frac{13}{16} = .11010\ldots0$, which are instances of all three of the schemata just singled out. Note also that f has a well-defined average value f_ξ on each schema ξ (for any weighting of the values $f(x)$, as by a probability distribution). Clearly, for any x, knowledge of $f(x)$ is relevant to estimating f_ξ for any schema for which $x \in \xi$. Moreover, observations

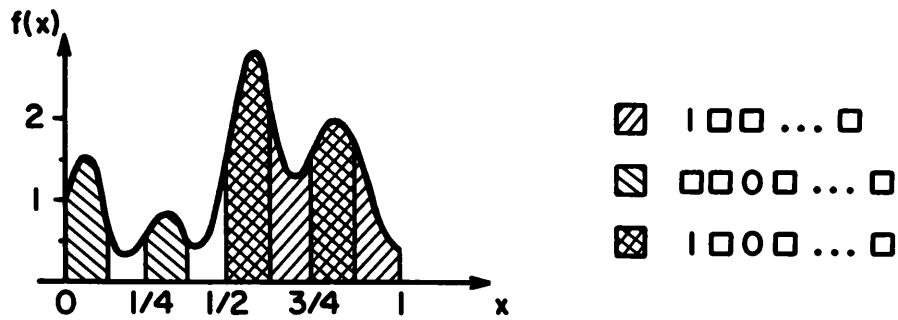

Fig. 10. Some schemata for a one-dimensional function

of f for relatively few x will enable f_ξ to be estimated for a great many $\xi \in \Xi$. Even a sequence of four observations, say $x(1) = .0100010\ldots0$, $x(2) = .110100\ldots0$, $x(3) = .100010\ldots0$, $x(4) = .1111010\ldots0$, enables one to calculate three-point estimates for many schemata, e.g. (assuming all points are equally likely or equally weighted), $\hat{f}_{1\square\square\ldots\square} = (f(x(2)) + f(x(3)) + f(x(4)))/3$ and $\hat{f}_{\square\square\square\square01\square\ldots\square} = (f(x(1)) + f(x(3)) + f(x(4)))/3$, and two-point estimates for even more schemata, e.g., $\hat{f}_{\square1\square\square01\square\ldots\square} = (f(x(3)) + f(x(4)))/2$ and $\hat{f}_{11\square\ldots\square} = (f(x(2)) + f(x(4)))/2$.

The picture is not much changed if f is a function of many variables $x_1, \ldots, x_d$. Using binary representations again, we now have $20d$ detectors (assuming the same accuracy as before), 3^{20d} schemata, and each point is an instance of 2^{20d} schemata. In the one-dimensional case the representation transformed the problem to one of sampling in a 20-dimensional space—already a space of high dimensionality—so the increase to a $20d$-dimensional space really involves no significant conceptual changes. Interestingly, each point $(x_1, \ldots, x_d)$ is now an instance of 2^{20d} schemata rather than 2^{20} schemata, an exponential (dth power) increase. Thus, for a given number of points tried, we can expect an exponential (dth power) increase in the number of schemata for which f_ξ can be estimated with a given confidence. As a consequence, if the information about the schemata can be stored and used to generate relevant new trials, high dimensionality of the argument space $\{0 \leqq x_j < 1, j = 1, \ldots, d\}$ imposes no particular barrier.

It is also interesting in this context to compare two different representations for the same underlying space. Six detectors with a range of 10 values can yield approximately the same number of distinct representations as 20 detectors with a range of 2 values, since $10^6 \cong 2^{20} = 1.05 \times 10^6$ (cf. decimal encoding vs. binary encoding). However the numbers of schemata in the two cases are vastly different: $11^6 = 1.77 \times 10^6$ vs. $3^{20} = 3.48 \times 10^9$. Moreover in the first case each $A \in \mathfrak{A}$ is an instance of only $2^6 = 64$ schemata, whereas in the second case each $A \in \mathfrak{A}$ is an instance of $2^{20} = 1.05 \times 10^6$ schemata. This suggests that, for adaptive plans which can use the increased information flow (such as the reproductive plans), many detectors deciding among few attributes are preferable to few detectors with a range of many attributes. In genetics this would correspond to chromosomes with many loci and few alleles per locus (the usual case) rather than few loci and many alleles per locus.

Returning to the view of schemata as random variables, it is instructive to determine how many schemata receive at least some given number $n < N$ of trials when N elements of $\mathfrak{A}$ are selected at random. This will give us a better idea of the *intrinsic parallelism* wherein a sequence of trials drawn from $\mathfrak{A}$ is at the same time a (usually shorter) sequence of trials for each of a large number of schemata

$\xi \in \Xi$. It will be helpful in approaching this calculation (and in later discussions) to more carefully classify the schemata. A schema will be said to be *defined on* the set of positions $\{i_1, \ldots, i_h\}$ at which $\Delta_{i_j} \neq \square$. If $V = \bigcup_i V_i$ has k elements and we consider all schemata over V, i.e., $\Xi = \{V \cup \{\square\}\}^l$, then there are k^h distinct schemata defined on any given set of $h \leqq l$ positions. Moreover, for any given set of h positions, *every* $A \in V^l$ is an instance of one of these k^h schemata. That is, the set of schemata defined on a given set of positions partitions $\mathfrak{A}$, and each distinct set of positions gives rise to a different partition of $\mathfrak{A}$. (For example, if $V = \{0, 1\}$ and $l = 4$, the set of schemata defined on position 1 is $\{0\square\square\square, 1\square\square\square\}$, where $0\square\square\square$ abbreviates $(0, \square, \square, \square)$ etc. It is clear that every element in $\mathfrak{A} = V^l$ begins either with the symbol 0 or else the symbol 1, hence the given set partitions $\mathfrak{A}$. Similarly the set defined on position 2, $\{\square0\square\square, \square1\square\square\}$, partitions $\mathfrak{A}$, and the set defined on positions 2 and 4, $\{\square0\square0, \square0\square1, \square1\square0, \square1\square1\}$ is still a different partition of $\mathfrak{A}$, a refinement of the one just previous.) There are $\binom{l}{h}$ distinct ways of choosing h positions $\{1 \leqq i_1 < i_2 < \cdots < i_h \leqq l\}$ along the l-tuple, and h can be any number between 1 and l. Thus there are $\sum_{h=1}^{l} \binom{l}{h} = 2^l - 1$ distinct partitions induced on $\mathfrak{A}$ by these sets of schemata. It follows that any sequence of N trials of $\mathfrak{A}$ will be simultaneously distributed over each of these partitions. That is, *each* of the 2^l *sets* of schemata defined on the 2^l distinct choices of positions receives N trials.

On the assumption that elements of $\mathfrak{A}$ are tried at random, uniformly (elements equally likely) and independently, we can use the Poisson distribution to determine the number of schemata receiving at least $n < N$ trials. The basic parameter required is the average number of trials per schema for any set of schemata defined on h positions. The value of this parameter is just N/k^h since there are k^h schemata defined on a fixed set of h positions. The Poisson distribution then gives

$$r(n, N) = \sum_{n'=n}^{\infty} ((N/k^h)^{n'}/n'!) \exp(-N/k^h)$$

as the proportion of schemata defined on the positions $i_1, \ldots, i_h$ and receiving at least n out of the N trials.

This can be directly generalized to give a lower bound in the case where the distribution over $\mathfrak{A}$ is no longer uniform. Let $\chi_\epsilon(i_1, \ldots, i_h)$ designate the fraction of schemata defined on $i_1, \ldots, i_h$ for which the probability of a trial is at least ϵ/k^h, let γ_h be the proportion of the $\binom{l}{h}$ sets of schemata defined on h positions for which $\chi_\epsilon(i_1, \ldots, i_h) > \beta_0$ and, finally, let $\gamma_0 = \min_h \gamma_h$. Then the expression above, by a simple manipulation, yields

$$\begin{aligned} r(n, \epsilon, N) &= \sum_{h=1}^{l} \gamma_h \binom{l}{h} \beta_0 k^h \sum_{n'=n}^{\infty} ((\epsilon N/k^h)^{n'}/n'!) \exp(-\epsilon N/k^h) \\ &\geqq \gamma_0 \beta_0 \sum_{h=1}^{l} \binom{l}{h} k^h \sum_{n'=n}^{\infty} ((\epsilon N/k^h)^{n'}/n'!) \exp(-\epsilon N/k^h). \end{aligned}$$

Values for this bound can be obtained from standard tables for the Poisson distribution, but the following representative cases will give some feeling for the numbers involved. Setting $k = 2$ and $l = 32$ (so that $\mathcal{A}$ contains $2^{32} \cong 4.3 \times 10^9$ distinct elements) we get:

(n, ϵ, N)	$r(n, \epsilon, N)$	
$(8, 1, 16)$	$> 700\gamma_0\beta_0$	
$(8, \frac{1}{2}, 16)$	$> 50\gamma_0\beta_0$	(If the distribution of
$(4, \frac{1}{16}, 32)$	$> 20\gamma_0\beta_0$	trials is uniform, $\gamma_0 = \beta_0 = 1$.)
$(8, 1, 32)$	$> 9000\gamma_0\beta_0$	
$(8, \frac{1}{4}, 32)$	$> 50\gamma_0\beta_0$	

Even for the small values of N considered here it is clear that a great many schemata will receive a significant number of trials. Moreover the figures given are quite conservative since at least one schema defined on each distinct set of positions must receive at least 1 trial in k^h, whereas the bound assumes none receive more than 1 in ϵk^h.

The figures just given also hint at a compact way for storing a great deal of information about schemata. Suppose that the object is to store the relative rank of a large number of schemata where, say, ξ ranks higher than ξ' when $\mu_\xi > \mu_{\xi'}$. Let us limit the number of ranks to M (e.g., by dividing the range of μ into M intervals and assigning ξ the same rank as ξ' if their average payoffs, μ_ξ and $\mu_{\xi'}$, fall in the same interval). Now with a set of M elements, $\mathcal{B} = \{A_j \in \mathcal{A}, j = 1, \ldots, M\}$, it is possible to represent the rank of a given schema ξ by the *number* of instances of ξ, $A_j \in \xi$, in the set. That is, if ξ has rank $m < M$, there will be m instances of ξ in $\mathcal{B}$, $\{A_{j_h} \in \xi, h = 1, \ldots, m\} \subset \mathcal{B}$. Note that there is no requirement that $A_j = A_{j'}$ for $j \neq j'$, so that given some other schema ξ' we may have $A_j \in \xi'$ but $A_{j'} \notin \xi'$. Thus the same instances used to represent the rank of ξ can be used *in differing numbers* to represent the ranks of other schemata.

For example, given the 8 individuals

$$A_1 = 000010$$
$$A_2 = 001111$$
$$A_3 = 011000$$
$$A_4 = 011011$$
$$A_5 = 101100$$
$$A_6 = 110011$$
$$A_7 = 111011$$
$$A_8 = 111101$$

we see that the schema $\xi = \square 110\square\square$ is assigned rank 3 by the three instances A_3, A_4, A_7, while the schema $\xi' = \square\square\square 01\square$ is assigned rank 4 using $A_4 \in \xi$ and $A_7 \in \xi$ with two other instances, and $\xi'' = 11\square\square\square\square$ is assigned rank 3 using $A_7 \in \xi$ and two other instances.

If we set $M = 32$ then the above calculation for $r(8, 1, 32)$ indicates that some sets of size 32 drawn from $\mathcal{A}$ (randomly generated ones in this case) can assign a rank $m \geqq 8$ to 9000 distinct schemata (for $k = 2$, $l = 32$). The problem then is one of using this potential to represent the relative ranking of the sample averages μ_ξ for a large set of observed schemata. Once again we must wait upon the discussion of reproductive plans in chapter 6 to see that this can be done.

Summarizing: Given a set of detectors $\{\delta_i : \mathcal{A} \to V_i, i = 1, \ldots, l\}$ the elements $A \in \mathcal{A}$ each have a *representation* $(\delta_1(A), \ldots, \delta_l(A))$ in terms of the ordered set of l *attributes* $\delta_i(A) \in V_i$, $i = 1, \ldots, l$. Each $\xi \in \Xi = \Pi_{i=1}^{l} \{V_i \cup \{\square\}\}$ designates a particular subset of $\mathcal{A}$, namely all elements of $\mathcal{A}$ for which the corresponding representations match all positions in ξ which are not "$\square$"s. Given a set of observations $\mathcal{A}(1), \mathcal{A}(2), \ldots, \mathcal{A}(t)$ from $\mathcal{A}$, the average payoff μ_ξ of the observed instances $\mathcal{A}(t') \in \xi$ is apportioned to ξ as its credit for the performances of the $A \in \mathcal{A}$ possessing the corresponding set of attributes. Since each $A \in \mathcal{A}$ is an instance of 2^l schemata it constitutes a valid sample point of 2^l distinct subsets of (or events on) $\mathcal{A}$. This suggests the existence of algorithms which, by testing many possibilities with a single trial, are *intrinsically parallel* and which store the relative rankings of μ_ξ for a great many schemata by selecting a small set $\mathcal{B} \subset \mathcal{A}$. The algorithms introduced in chapter 6 will realize these possibilities. Later (chapter 8) dependence on the detectors $\{\delta_i\}$ will be eliminated by subjecting the detectors themselves to adaptation.