FORMALIZATION IN BIOLOGY

J. H. WOODGER

The word 'formalization' seems to be used in two senses — a narrow and a wide sense. In my monograph *The Technique of Theory Construction* (1939) I distinguished four processes : (i) formalization; (ii) axiomatization; (iii) logical analysis; and (iv) symbolization. By 'formalization' I meant the process of *constructing the metatheory* of a scientific theory; this is formalization in the narrower or stricter sense. But the word also seems to be used in a much wider sense which includes at least axiomatization and symbolization. On the present occasion I shall assume that formalization in the wider sense is intended; because, at least as far as biology is concerned, there appears to be no need for formalization in the narrower sense.

Anyone who attempts to build a bridge between natural science on the one hand and logic and mathematics on the other is always confronted with difficult problems of exposition. In talking to his colleagues in natural science (especially in the biological sciences) he will be compelled to spend much time in explaining the mathematics, leaving little for the natural science. In talking to mathematicians it will be necessary for him to explain the natural science, leaving little time for the mathematics.

When I was invited to take part in a colloquium on scientific applications of mathematical logic held in Paris in August 1952 I decided that it would be better to try to deal with logical and mathematical problems which have arisen from such applications rather than with biological problems themselves. On the present occasion I shall try to interest you in a problem of biological definition, for the solution of which the theory of sets offers a suitable foundation. The biological topic concerned is not one which involves a great deal of specialized knowledge, but is at the same time one of general interest and importance because it is not confined to pure biology but occurs also in medicine, sociology and criminology. I refer to the antithesis between heredity and environment, and the problem of making clear the distinction between them, a problem to which I have devoted a good deal of time and attention. In biological books we read of hereditary (or inborn) characters and acquired characters. We are told that hereditary characters are those which are 'due to heredity' and acquired characters are those which are 'due to environment'. In order to deal with the problem within set-theory I shall refer to the classes of lives which are specified by reference to these characters and I shall call such classes *phenotypes*, — a word which is already in use in genetics. A phenotype is a sub-set of a species which is specified by reference to some observation or test other than a breeding experiment. A phenotype is a natural kind or sort in the sense that if P is a phenotype and x is anything which is *not* a member

of P, then $P \cup \{x\}$ is *not* a phenotype. Genetics is primarily an affair of parents and offspring. It is assumed that each parent makes a *physical contribution* — called a gamete — towards each offspring. The two gametes unite to form a fertilized egg or *zygote*. This zygote begins to develop in its environment and if this environment is favourable and the zygote healthy the process of development continues until the adult condition is reached and the offspring may itself become a parent. What develops out of a zygote in a given environment is here called a *life*.

When people say that a certain phenotype (or character characterizing members of that phenotype) is 'due to heredity' they appear to mean 'dependent upon the kind of egg *from which* development of members of that phenotype begins'; and when they say 'due to environment' they appear to mean 'dependent upon the kind of environment *in which* such development occurs'.

Now I assume that no phenotype can be hereditary in the sense of being dependent *only* on the kind of egg involved, because there are some environments in which no egg will develop, such as boiling oil or strong acid; clearly environments of these kinds must be excluded. Similarly I assume that no phenotype can be acquired in the sense of depending *only* on the kind of environment. Because a phenotype, members of which appear in an environment of a given kind *irrespective* of the kind of egg, from which development begins, would be as extraordinary as one, members of which appear irrespective of the kind of environment in which development takes place. Thus we expect a healthy child who is brought up in a home in which English is the principal language spoken, to speak English, irrespective of whether the child's parents came from China or Peru. But we do not expect the dogs and cats which are reared in such homes to speak English.

But if all this is admitted then it is clear that we cannot distinguish hereditary phenotypes from acquired phenotypes in the traditional way, since we are driven to admit that in *both* cases the kind of egg *and* the kind of environment is involved. In the first case to emphasize the kind of egg but must at the same time exclude certain kinds of environment. In the second case we want to emphasize the kind of environment but must exclude certain kinds of egg. How is this to be done? I propose to do it in the following way: First I define the notion of a *genetical system*. Next I explain what I mean by saying that a phenotype is *environmentally insensitive in* a genetical system. Finally I define an hereditary (or inborn) phenotype as one which is environmentally insensitive in *every* genetical system of which it is a member. Similarly, I define what is meant by saying that a phenotype is *zygotically insensitive in* a genetical system and then I define an acquired phenotype as one which is zygotically insensitive in *every* genetical system of which it is a member.

The chief difficulty is to define 'genetical system' and my first task must be to explain the notions by means of which this difficulty is to be overcome. I must mention first that the definitions to be given here all belong to an axiom

system an exposition of which cannot be given here (1). The primitives are functors which have variables whose values are individuals associated with them, but on the present occasion I must begin my exposition with the help of functors of higher type.

I use

$$\boldsymbol{L}_E(Z)$$

to denote the set of all lives which develop in an environment belonging to the set E from a zygote (fertilized egg) belonging to the set Z. For genetical purposes it is frequently necessary to specify sets of zygotes by reference to the kinds of *gametes* (or parental contributions) which have united to form them. I therefore use

$$\boldsymbol{U}(\alpha, \beta)$$

to denote the set of all zygotes which are formed by the union of a gamete belonging to the set α with one belonging to β (small Greek letters are here used as variables which have classes of gametes as their values). It frequently happens that we want to substitute expressions of this form for the variable 'Z' in expressions like '$\boldsymbol{L}_E$ (Z)'. In this way the notation can become rather cumbersome and it is then worth while to introduce an abbreviation by definition, thus I put:

$$\text{'}\boldsymbol{D}(\alpha, \beta, E)\text{' for '}\boldsymbol{L}_E(\boldsymbol{U}(\alpha, \beta))\text{'}$$

Next I use

$$\boldsymbol{G}_E(X)$$

to denote the class of all gametes which are formed by members of the class X of lives when they develop in members of the classe E of environments. Here again it frequently happens that we want to substitute expressions like '$\boldsymbol{D}(\alpha,\beta,E)$' for the 'X' in '$\boldsymbol{G}_E(X)$' and abbreviation again becomes desirable. I therefore put

$$\text{'}\boldsymbol{G}(\alpha, \beta, E)\text{' for } \boldsymbol{G}_E(\boldsymbol{D}(\alpha, \beta, E))\text{'}$$

We are now ready to undertake the rather difficult task of defining 'genetical system'. This is difficult not so much because the notion itself is difficult but because, like all definitions, this one must be so formulated as to exclude certain possibilities but not others, and in the present state of the analysis it is not easy to foresee all that may be required of the new notion of genetical system. What I have to offer therefore is not presented as something that is in any sense final. The suitability of a definition, like that of an axiom, can only be decided when its consequences have been well worked out, and I cannot claim to have carried this process very far with the definition of 'genetical system'. In what follows the definition is reached by a succession of steps.

(1) An account of this system will be published with the other contributions to an international symposium on the Axiomatic Method which was held at the University of California, Berkeley, December 26, 1957 to January 4, 1958.

1. A *genetical set* is any set consisting of a finite number of pair-wise mutually exclusive phenotypes, a finite number of classes of gametes and a finite number of classes of environments. It will be assumed that the classes which are members of a genetical set are all non-empty and finite.

2. Since a genetical set S does not include classes of zygotes among its members, we cannot speak of the zygote classes *belonging* to such a set, but we can speak of the classes of zygotes *associated* with it, meaning thereby the set of all classes Z of zygotes such that for some gamete classes α and β belonging to S we have $Z = \boldsymbol{U}(\alpha, \beta)$.

3. A genetical set S is *developmentally closed* if and only if for any gamete classes α and β belonging to S ,and any environment class K belonging to S, we have either $\boldsymbol{D}(\alpha, \beta, K) = \Lambda$, or there exists a phenotype P belonging to S such that $\boldsymbol{D}(\alpha, \beta, K) \neq \Lambda$ and $\boldsymbol{D}(\alpha, \beta, K) \subseteq P$.

4. A genetical set S is *genetically closed* if and only if for any gamete classes α and β and any environment classes K belonging to S such that $\boldsymbol{D}(\alpha, \beta, K) \neq \Lambda$, we have $\boldsymbol{G}(\alpha, \beta, K) \neq \Lambda$ and $\boldsymbol{G}(\alpha, \beta, K) \subseteq \sigma$, where σ is the sum of all the gamete classes in S.

5. A genetical set S is *phenotypically connected* if and only if, for every phenotype P belonging to S, there exist gamete casses α and β of S, and an environment class K of S such that $\boldsymbol{D}(\alpha, \beta, K) \neq \Lambda$ and $\boldsymbol{D}(\alpha, \beta, K) \subseteq P$.

6. Finally, a *genetical system* is a genetical set which is phenotypically connected and both developmentally and genetically closed.

Suppose $\boldsymbol{G}$ is the class of all garden peas with green seedleaves, g the class of all gametes produced by such peas when they develop in members of the class $\boldsymbol{M}$ of all environments in which such peas develop normally to maturity (and excluding environments in which mutation involving colour of the seed leaves occurs), then the set $\{\boldsymbol{G}, \mathbf{g}, \boldsymbol{M}\}$ will be an example of the simplest kind of genetical system — the kind I call *genetical units*; namely those with one phenotype, one gamete class and one class of environments and no more. In this case we shall have :

$$\boldsymbol{D}(\mathbf{g}, \mathbf{g}, \boldsymbol{M}) \neq \Lambda \text{ and } \boldsymbol{D}(\mathbf{g}, \mathbf{g}, \boldsymbol{M}) \subseteq$$

and

$$(\mathbf{g}, \mathbf{g}, \boldsymbol{M}) \neq \Lambda \text{ and } \quad (\mathbf{g}, \mathbf{g}, \boldsymbol{M}) \subseteq \mathbf{g}.$$

Genetical systems, being sets, will be subject to the operations of addition and multiplication and to the relation of inclusion in the ordinary set-theoretical sense. Moreover, being sets of sets they can also be combined in another way which I call set-by-set addition multiplication and inclusion. Thus if $\{P, \alpha, E\}$ and $\{Q, \beta, K\}$ are genetical systems, then their ordinary sum will be $\{P, Q, \alpha, \beta, E, K\}$, but their set-by-set sum will be $\{P \cup Q, \alpha \cup \beta, E \cup K\}$. Their ordinary product will (if $P \neq Q, \alpha \neq \beta$ and $E \neq K$) be the null class and their set-by-set product will be $\{P \cap Q, \alpha \cap \beta, E \cap K\}$. Similarly it is possible to have one genetical system included in another in the ordinary sense and also in the sense in which $\{P, \alpha, E\}$ is included in $\{Q, \beta, K\}$ if $P \subseteq Q, \alpha \subseteq \beta$, and

$E \subseteq K$. The task of working out theorems concerning these operations on, and relations between, genetical systems has as yet hardly begun. Also the task of discovering under what conditions sums and products of genetical systems are themselves genetical systems has not yet been carried very far. It is of some interest to note, however, that the genetical systems studied in Mendelian genetics can all be represented as sums of genetical units or as set-by-set products of such sums. It is characteristic of Mendelian genetics that it deals almost exclusively with genetical systems having only one environment class. We can define the *system number* of a genetical system S as the ordered triple $<n, m, k>$ of natural numbers, where n is the number of phenotypes, m the number of gamete classes and k the number of environment classes of S. The cardinal number of S will then be the sum of these three numbers. Every genetic unit will have the system number $<1, 1, 1>$; the systems studied in Mendelian genetics under the heading ‚monohybridism' usually have the system number $<2, 2, 1>$; and those exhibiting dihybridism have the system number $<4,4,1>$. If in a genetical system S there exists a pair of gamete classes α and β and an environment class K such that $\boldsymbol{G}(\alpha, \beta, K) \neq \Lambda$ and $\boldsymbol{G}(\alpha, \beta, K)$ is included in the sum of exactly 2^l of the gamete classes of S, but there is no such combination of gamete and environment classes of S which yields the sum of a *higher* number of gamete classes of S, then we can say that S is of *order l*. It is possible to have two Mendelian systems both of system number $<4, 4, 1>$ but one being of order 1 (with monohybrids) and the other of order 2 (with dihybrids).

We now come to the problem: What does it mean to say that a phenotype P is *environmentally insensitive in* a genetical system S? This can be most easily explained by reference to a *matrix* of the relation $\boldsymbol{L}_E(Z) \subseteq P$ when the values of the variables 'E' and 'P' are confined to the environment classes and phenotypes, and the values of 'Z' to the associated zygote classes, of S. Under these circumstances this is a *functional relation* in the sense that each pair of values of the independent variables 'E' and 'Z' there will be only one corresponding value of 'P'. For suppose that S is a genetical system, Z an associated zygote class and E and P members of S such that

$$\boldsymbol{L}_E(Z) \neq \Lambda \text{ and } \boldsymbol{L}_E(Z) \subseteq P$$

then if Q is a phenotype of S distinct from P such that

$$\boldsymbol{L}_E(Z) \subseteq Q$$

we must have $\boldsymbol{L}_E(Z) \subseteq P \cap Q$. But a genetical system is a genetical set and by definition the phenotypes of a genetical set are mutually exclusive; consequently $P \cap Q = \Lambda$ and so $\boldsymbol{L}_E(Z) = \Lambda$, which is contrary to our hypothesis. There can, therefore, be no such Q distinct from P in S. Returning now to the problem of explaining environmental insensitivity, suppose

$$S_1 = \{P, Q, R, \alpha, \beta, E_1, E_2, E_3\}$$

is a genetical system and

$$Z_1 = \boldsymbol{U}(\alpha, \alpha); Z_2 = \boldsymbol{U}(\alpha, \beta) \text{ and } Z_3 = \boldsymbol{U}(\beta, \beta)$$

are its associated zygote classes. Suppose, further, that S_1 has the following $\boldsymbol{L}$-matrix:

$\boldsymbol{L}$	Z_1	Z_2	Z_3
E_1	P	Q	R
E_2	P	Q	Λ
E_3	P	Λ	Λ

Here we put the designations of the associated zygote classes at the tops of the columns and the designations of the environment classes at the left hand ends of the rows. If $\boldsymbol{L}_{E_i}(Z_j) \subseteq X$ we write 'X' in the intersection of the jth column with the ith row. It will be noticed that 'P' occurs in *every* row of the column headed 'Z_1' and in *no other column*. 'Q', however, occurs in column Z_2 but only in two of its rows and 'R' occurs in only one row of the column headed 'Z_3'. Now when a phenotype designation occurs in *every* row of every column in which it occurs at all I say that the phenotype so designated is *environmentally insensitive in* the genetical system concerned.

Now suppose that the matrix of S_1 is as follows:

$\boldsymbol{L}$	Z_1	Z_2	Z_3
E_1	P	P	P
E_2	Q	Q	Λ
E_3	R	Λ	Λ

In this case 'P' occurs in *every* column of the row labelled 'E_1' but in *no other* row; 'Q' occurs in row E_2 but not in each of its columns; and 'R' occurs in only one column of row E_3. We define a phenotype as being *zygotically insensitive in* a genetical system S if and only if its designation occurs in *every* column of every row of the $\boldsymbol{L}$-matrix of S in which it occurs at all. This in our last example only P is zygotically insensitive in S_1.

Against these definitions it might well be objected that they are not formulated within the object language because they involve reference to the

designations of the phenotypes concerned. For that reason the following formulations are offered. Let us use '*P* **Phen** S' as an abbreviation for '*P* is a phenotype and a member of the genetical system S'; and '*P* $\boldsymbol{EI}$ S' as an abbreviation for '*P* **Phen** S and *P* is environmentally insensitive in S'. We now define this relation as follows:

P $\boldsymbol{EI}$ S if and only if *P* **Phen** S and for every *Z*, if *Z* is an associated zygote class of S and there is an environment class *K* of S such that $\boldsymbol{L}_K(Z) \neq \Lambda$ and $\boldsymbol{L}_K(Z) \subseteq P$, then for *every* environment class *M* of S we have $\boldsymbol{L}_M(Z) \neq \Lambda$ and $\boldsymbol{L}_M(Z) \subseteq P$.

Similarly, using '*P* $\boldsymbol{ZI}$ S' as an abbreviation for '*P* **Phen** S and *P* is zygotically insensitive in S' we define this relation as follows:

P $\boldsymbol{ZI}$ *S* if and only if *P* **Phen** S and for every environment class *K* of S, if there is a zygote class *Z* associated with S such that $\boldsymbol{L}_K(Z) \neq \Lambda$ and $\boldsymbol{L}_K(Z) \subseteq P$, then for *every* zygote class *W* associated with S we have $\boldsymbol{L}_K(W) \neq \Lambda$ and $\boldsymbol{L}_K(W) \subseteq P$.

Using the *Principia Mathematica* arrow notation we now say:

P is *hereditary* if and only if $\overleftarrow{\textbf{Phen}}$ '$P \subseteq \overleftarrow{\boldsymbol{EI}}$' *P*

and

P is *acquired* if and only if $\overleftarrow{\textbf{Phen}}$ '$P \subseteq \overleftarrow{\boldsymbol{ZI}}$' *P*.

It will be evident that, if these definitions are adopted, then *all* the phenotypes of a genetical system which has only one environment class will be environmentally insensitive in that system. Now it has already been mentioned that the vast majority of Mendelian systems have only *one* environment class; consequently it is not surprising, in the light of what has been said, that the phenotypes in Mendelian systems are commonly said to be hereditary. It is possible that, if more systems were studied which have more than one environment class, we might find that some phenotypes which were formerly labelled as hereditary are in fact neither hereditary nor acquired and that yet others were both hereditary and acquired. For there appears to be no reason, when these notions are analysed more carefully than is usually the case, for supposing that they are either mutually exclusive or together exhaustive,

I should like to end on a note of interrogation by briefly mentioning a problem for which I have not yet succeeded in finding a solution. As we are dealing with finite classes there will, presumably, be genetical systems which are *maximal* in the sense that if any new members are added to their gemete or environment classes the result is *not* a genetical system because it is not developmentally closed. But there is a serious difficulty about defining maximal genetical systems in this way. Suppose, for example, that we have two genetical systems: $S_1 = \{P, \alpha, K\}$ and $S_2 = \{P, \beta, M\}$ such that $\alpha \neq \beta$ and $K \neq M$. Now it may be the case that the set-by-set sum of S_1 and S_2 namely: $S_3 = \{P, \alpha \cup \beta, K \cup M\}$ is also a genetical system and a step nearer a maximal system than either S_1 or S_2. But there is another possibility. It may

be the case that the ordinary sum of S_1 and S_2, namely: $S_4 = \{P, \alpha, \beta, K, M\}$ is not developmentally closed; not because either S_1 or S_2 is maximal in the above sense, but for another reason. It may be because, although we have $\boldsymbol{D}(\alpha, \alpha, K) \subseteq P$ and $\boldsymbol{D}(\beta, \beta, M) \subseteq P$, we do *not* have *either* $\boldsymbol{D}(\alpha, \alpha, M) \subseteq P$ *or* $\boldsymbol{D}(\beta, \beta, K) \subseteq P$. When this happens $\boldsymbol{D}(\alpha, \alpha, K)$ and $\boldsymbol{D}(\beta, \beta, M)$ are said to be *phenocopies* of one another. Thus the existence of phenocopies seems greatly to complicate the task of defining maximal genetical systems. At present I am unable to see clearly how to deal with this problem; neither am I at all clear about its importance.

UNIVERSITY OF LONDON

(*Reçu le 15 mars 1958*)