4 A Definition of Complexity

> *Complexity is that property of a model which makes it difficult to formulate its overall behaviour in a given language, even when given reasonably complete information about its atomic components and their inter-relations.*[36]

This is a general definition, which is intended to have different interpretations in different contexts. It relates the difficulty in formalisation of the whole to that of the formalisation of its parts (typically a ‘top-down’ model compared to a ‘bottom-up’ one) in the language. It is only applicable in cases where there is at least a possibility of gaining significant information about the components, thus clearly separating ignorance from complexity. Different conceptions of complexity depending on the base language chosen, the type of difficulty focused on and the type of formulation desired within that language.

The important aspects of this approach are that:

- it applies to models rather than natural systems;
- complexity is distinguished from ignorance;
- it is relative to the modelling language it is expressed in;
- it relative to the identification of components and overall behaviour;
- complexity is a global characteristic of a model;
- you will get different kinds of complexities from different types of difficulty;
- complexity represents the gap between component knowledge and knowledge of global (or emergent) behaviour;
- since difficulty is sometimes comparative, complexity will be also.

Using the picture of modelling I presented in section 2.3 on page 33 complexity is the difficulty of finding a description of the overall behaviour of a model. This description is also a model description, so we are dealing with (at least) two models here: the model of the ‘components’ and their interactions and the model of the resultant overall behaviour. I will call the former the *object model* and the later the *overall model*. The complexity is

36.Some aspects of this have been pointed out independently by other authors, e.g. Grassberger [194] talks about complexity as “difficulty”, Kaufman [249] relativises it to the language of representation, Raccoon [365] talks about comparing a top-down vs. bottom up approach, etc. but I have not see these aspects combined in this way.

thus the difficulty in finding an overall model description for the resultant behaviour of the object model.

Perhaps the simplest case is where the object model is a data model and the goal is to find a description in some modelling language whose content subsumes the data model content and is as specific as possible. The difficulty of ‘finding’ this description depends upon how this is sought and the criteria for what would constitute an acceptable solution. This is the basic situation that occurs in many machine learning set-ups (as discussed in section 2.4 on page 38), where the problem is to design or discover process to find such descriptions. The situation is illustrated in figure 17 on page 73.

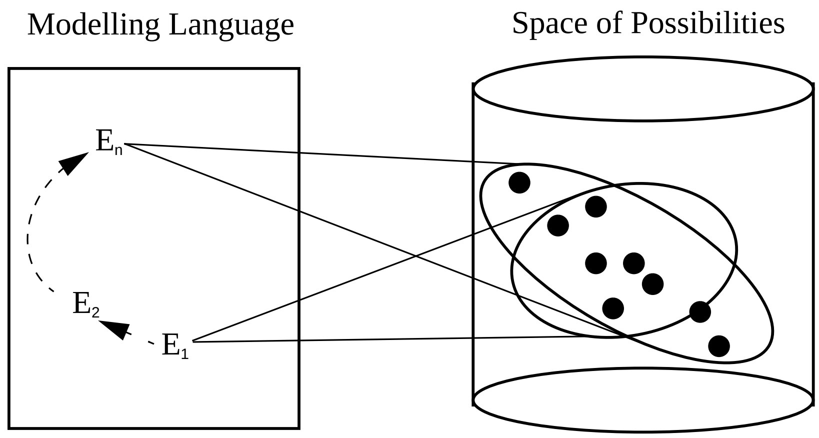

Figure 17. Finding a description to cover a data model

Then *measuring* this difficulty implies a mapping onto a numeric structure. The ‘search’ process occurs by acting on the model descriptions so that the measure of difficulty is a measure upon the modelling language. Thus we have a ‘chaining’ of models (as described in section 2.2 on page 28), Thus the complexity here is a numeric model of the difficulty of finding a model description that meets some criteria for matching (e.g. by what I called subsumption or approximation in section 2.2 on page 22) the data model of the phenomena under consideration.

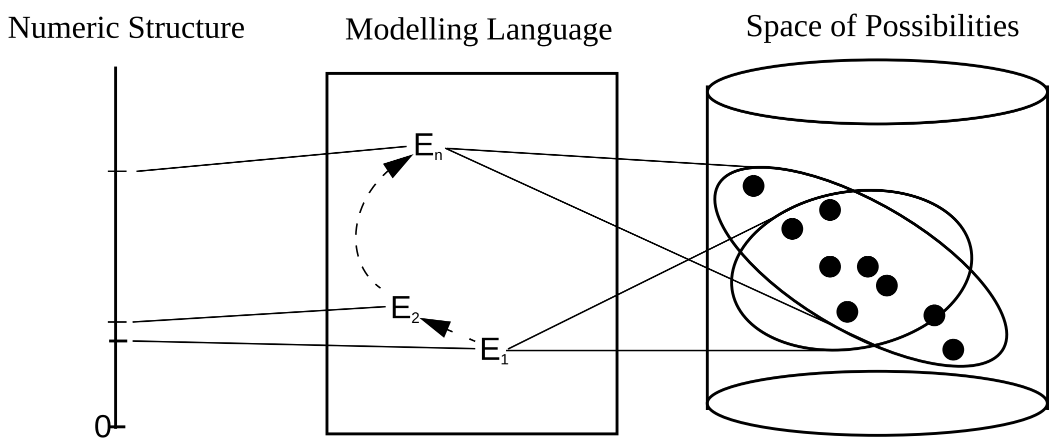

Figure 18. Measuring the difficulty of finding a model description

Of course, most modelling situations are more complicated that this. In general the object model will not be a simple data model but one that supports some inferences in the form of predictions from initial conditions. The method of finding a description for the resulting predictions might not be an inductive one but an analytic one based on known properties of the description. The criteria for the sought-for formulation of overall behaviour might include the acceptability of trade-offs between the error, specificity and cost of finding this description.

There are many different types of complexity involved in actual examples of scientific modelling. Conflation of these into a single “complexity” of scientifically modelling a certain system will generally result in confusion.

There might be:

- the complexity of the data: the difficulty of encoding of a data model compactly given a coding language;
- the complexity of the informal (mental) model: the difficulty in making an informal prediction from the model given hypothetical conditions;
- the complexity of using the formal model to predict aspects of the system under study given some conditions;
- the complexity of using the formal model to explain aspects of the system under study given some conditions.

Each of these will be relative to the framework it is being considered in (although this and the type of difficulty may be implicit). Further if attention is switched to the process of measurement involved in the production of the data or the process of improving the model then you get more corresponding complexities. Some sets of complexities will be easier to relate than others: there may be a close link between the complexity of the data and the complexity of the formal model to explain that data (relative to an encoding of that data into the appropriate formal language) but a more distant link between the complexity of the data relative to a formal language and the informal model used by the scientist in order to guide her search.

To take a more concrete biological example, the biological complexity of a model of an organism (the difficulty of explaining its functioning in terms of its genetic make-up) and a sort of environmental evolutionary complexity (the complexity of a model of its behaviour with respect to an environment in terms of the evolutionary pressures that created it) may be almost completely unrelated. Unfortunately, such different complexities are often conflated into one. This has considerably muddied the debate about whether there has been an increase in the complexity of organisms over the course of evolution (see [71, 221, 147] and section 6.6 on page 130).

In general, the complexity of natural systems or natural phenomena are closely associated with the complexity of their most appropriate models, but typically, in order to objectify the derived measure of complexity these models are relativised to some privileged framework (where this privilege is either argued for or established by the traditions and methodologies of the field). Which are the most suitable choice of models can be greatly influenced by the modelling language and the tolerance of error and vagueness. For more detail on this see Appendix 6 - Complexity and Scientific Modelling on page 199.

There are a number of aspects of this definition that need clarification. I will discuss these points below.

4.1 Aspects of the Definition

4.1.1 Identity of a system

Identifying *whole* systems is problematic in general. Many natural systems are causally interconnected, making any division seem somewhat arbitrary from anything approaching a general viewpoint. This problem is not solved by the seeming ease with which we actually *do* assign useful identities to systems as this seems to result at least partially from a blend of psychological, pragmatic and accidental features. There may well be true philosophical aspects of this distinction and I will briefly discuss how ideas of complexity may be a part of this account (section 6.8 on page 132), but since we are restricting ourselves to particular representations of any system in a language this is not a problem for us; we *decide* the level of interaction of the system with the rest of the representation when we specify our representation and *assign* an identity to some portion thus making it our focus. This does not eliminate the problem, of course, as it then becomes critical *how* we model any real system if we want to discuss the complexity of an aspect of it, but such concerns are common to many aspects of using representations of real systems.

So for us a *system* may just be a collection of statements intended to describe it sufficiently. We necessarily decide questions of *system identity* as a by-product of representing the system to start with.

4.1.2 Atomic components

An atomic component is one that can not be reduced in terms of other components in the representation chosen. The above definition thus only refers to systems where a base layer of atomic components can be identified, because it is a definition that compares descriptions. The choice of modelling framework will include deciding what constitutes the components and the overall behaviour (see section 4.1.4 on page 78).

Not all representations will have atomic components. For example, one might consider a web of social cause and effect. Here every cause might itself have several sub-causes and every effect be composed of sub-effects. When tracing this web downwards one may not reach any ultimate cause. The representation of this might well be recursive, if not circular. However, if this is to be formulated in terms of a modelling language, there will have to be a limit on the number of symbols used. This limited model can be considered as an approximation of the representation.

Taking such a series of ever more detailed models might result in a series whose complexity grows unboundedly. In this case one might be justified in describing the complexity of the target system as infinite. On the other hand one might come to a definite limit, where increasing detail increased the complexity progressively less, indicating a limit, but this limit might not be unique. If one took a different series of system approximations one might well get a different perspective and limit. In this case one could not ascribe a single complexity to the system with respect to this framework, unless one reduced the scope of the study to a few aspects to ensure uniqueness (or at least close convergence). These different series of approximations could be viewed as different perspectives each giving a different complexity or else you could view it as simply undefined since it would not have identifiable atomic parts in this sense.

In other words, if the approximations do not converge to the intended representation then one would have a (greater than usual) problem with ascribing meaning to them in terms of the target system anyway. This problem with ascribing meaning would include the sense of the complexity of these approximations. To resolve this problem one would need to change or augment your representations in order to make their intention reliably determinable, one aspect of which is how they approximate to your intended model. Then, if this method of approximation also allows one to approximate a meaningful idea of complexity then this complexity can be taken as the limit of the complexities in the approximations. Otherwise it is doubtful whether such an identification could be made sufficiently stable to be at all useful[37].

Another possibility is that, although there are no “atomic” parts at one level, there is a meta-level description where there are such. For example a directed digraph may have no unique starting node, but there might be a description of it in terms of nodes and arcs where the symbols representing these might be basic.

4.1.3 Difficulty

There are many possible kinds of difficulties in formulating the overall behaviour of a system. Each will have a correspondingly appropriate measure of complexity (see section 4.3 on page 83 for examples of this).

If the size of the representation was the primary concern, i.e. the memory requirements for the storage of it, then this criteria of size represents your measure of complexity (see section 3.4.1 on page 57. For example, this might be particularly relevant for humans when they have are dealing with a fairly meaningless representation (like numbers) when they had no memory aids such as pencil and paper. The longer it is the

37.Several approaches in physics aim to explicitly capture this scaling behaviour, see Appendix 6 - Complexity and Scientific Modelling on page 199 and section 8.16 on page 145 on measures to capture hierarchical scaling behaviour.

more difficult it is to deal with. This difficulty is easily overcome with simple aids. It is also very sensitive to the language of representation. It is notable that humans tend to chose internal representations that circumnavigate this limitation in short-term memory or "chunking" as it is sometimes called[38].

If one has plenty of memory, *one knows the algorithm* and is in a situation where one has to deal with example problems of indefinite size, then the time the overall behaviour takes to compute compared to the problem size will present the major difficulty. This produces the time measure of computational complexity (see section 8.47 on page 162) and their corresponding classes (P, NP, NP complete etc.). This represents more of a difficulty than mere space problems, but it assumes that the problem of *finding an algorithm* has already been solved.

From a philosophical viewpoint the difficulty in *reducing* a big problem to ones involving more fundamental units is a more major problem. This is the basic analytic problem. There are different ways of tackling this, many of which involve changing the language of representation (see section 5 on page 86). It is this aspect that may be most important when studying the complexity of formal systems.

4.1.4 Formulating overall behaviour

What constitutes the *global* behaviour of a system must be, to some extent, decided by the observer. For example, the overall properties of propositions may be considered from the point of view of a proof theory of a logical language with respect to a language of its theories. In less restricted systems, like a system of differential equations, there may be a range of global behaviours to choose between.

This can not be an arbitrary choice, however. There are limitations to what could be considered a global description. I list some below.

- The properties of the parts must determine the overall behaviour to a large extent. This expresses something of the essential relation of parts and wholes. If they do not determine it, then the "complexity" of formulating the behaviour relative to these parts would be fairly meaningless.
- More general solutions should be preferred to mapping more specific properties of the system. A general analytic solution to a system of differential equations is

38.From Miller [325]

to be preferred over its value in a specific range.

- The language of representation can limit the choice of overall behaviour available for formulation.

These constraints still leave considerable choice as to what could be considered as "overall behaviour" and what the atomic components might be. In many cases it might be arbitrary which is which, for example looking at a language's syntax and semantics - you might consider the complexity the difficulty of determining the syntax from the semantics or vice versa,

Often the question of what it means to formulate the overall behaviour is closely linked to the concept of "difficulty" considered relevant, see section 4.1.3 on page 77 above. For example, if "difficulty" means the computational time taken to compute the result of a program, then the formulation of the overall behaviour is the results or trace of that computation.

4.1.5 Complexity vs. ignorance

The most basic distinction that this definition makes is between difficulty due to ignorance and difficulty due to complexity. You can only reliably attribute complexity to a system when there is a possibility of knowing a reasonable amount about its components, otherwise the apparent difficulty of formulation might be merely due to some simple but unknown mechanism. Specifying *complete* knowledge of the components and their interactions is frequently impractical and often unnecessary; you sometimes can have sufficient information to rule out any simple unknown factors but still be faced with complex global behaviour.

For example an element of randomness is frequently taken to be atomically simple in languages of representation because there is no possibility of obtaining information about its general process (e.g. error terms).

One corollary is in situations where a large amount of ignorance is inevitable (e.g. the universe at the instant of the big-bang). The above definition will just not apply to such situations as near complete information about their components will not be available. Whether this means that complexity is *inapplicable* to such cases or that we will just *never know* the extent of their complexity does not make any practical difference to the use of the concept. This is the core of the arguments in the previous section see section 3.3 on page 47 for more on the scope of the idea of complexity.

4.1.6 As a gap between the global and local

Often complexity represents a gap between the ease of representation of the component parts and the difficulty with respect to the overall behaviour of the system[39]. This is a key characteristic of *emergent* behaviour.

Thus a model of the brain may be seen as complex because of the difficulty of explaining its over-all behaviour from the interaction of its parts (usually considered to be its neurones). The extent of the irreducibility of this difficulty is due to the fact that we are largely ignorant of the parts' processes and interactions.

This is not to say that all complexity must result in a contrast of levels of description. It is just when the contrast between different descriptions is so marked that it is practically impossible to translate between descriptions (great complexity) that this produces emergence. An example is the gap between the mental and physical worlds [116][40].

4.1.7 The comparative nature of complexity

Since, given a similar framework, difficulty is comparative so is complexity. Much of the utility of a concept like complexity comes from estimating the degree of difficulty of a problem and in attempts at simplification. In order to be able to use these aspects the model of complexity chosen needs to be comparative; you need to be able to say that one representation is more complex than another. For example, you might wish to compare the complexity of different axiomatisations of a logic.

The extreme case of merely classifying systems as either complex or simple is a (sometimes less useful) special case of this (see section 6.7 on page 131). It is merely a much coarser ordering. There are those who reserve use the label of "complexity" for holistic systems[41]. This is fine but tends to deprive the word of many of its connotations. This is often a reaction to limited internal system-theoretic models of complexity which take one aspect of complexity with respect to only one framework for which some sort of privilege is claimed[42].

39.See Raccoon in [365] for a clear account of this.

40.Pribram in [360] even attributes some complexity to the 'gap' between the hemispheres of the brain!

41.Or non-mechanistic systems, see Rosen [389]

4.1.8 The existence of complexity

It may be argued (by an extreme optimist) that *all* difficulty in formulation of overall behaviour is due to ignorance (as Waxman did in [463]). In which case complex systems would not "really" occur, i.e. all complexity is only apparent. Counter-examples of a formal nature tend to discount this possibility (see the example in section 3.4.5 on page 60 above).

Another more common example is the game of chess. Here the component rules and their interaction are completely specified, but the questions of overall specification, like determining a perfect strategy are still very difficult.

4.1.9 Relativisation to a language

Many examples from the previous chapter (section 3 on page 44) illustrate how the complexity of a representation can change when the language it is expressed in changes. This is a vital property of complexity as one of the principal tactics (as far as humans are concerned) used to simplify representations is to change the language they are embedded in (section 5.7.4 on page 123).

I mean the scope of "language" to be fairly wide - it will include informal as well as formal languages, natural as well as symbolic. Of course, the more informal the language, the more informal the derived measure of complexity it will imply relative to some subset and difficulty. It is doubtful you could formally show a complexity comparison between two poems, unless the scope of the language they were compared in and related to was sufficiently specified.

4.2 Examples

In addition to the examples of common usage above in section 3.1 on page 44, I present a few examples to illustrate the workings of the above definition. As I will deal with formal languages below (section 5 on page 86) these will be concerned with other areas.

42.It is ironic that it is the holists who have reduced the meaning of complexity and simplicity to that of natural and formal systems.

4.2.1 The flight behaviour of a herd

Consider a hypothetical herd of animals, in which the flight behaviour of each individual is fairly well understood in terms of its reaction to danger, its wish to follow others of its kind, etc., where we are trying to understand the flight behaviour of the whole herd in terms of its overall direction and path. If it turned out that all the animals always followed one specified leader, or all went in a direction represented by the average of the separate directions they would have gone individually then we would be justified in calling the flight behaviour of the herd simple. If the behaviour turned on the precise configuration of the herd at the time of attack, so that the animals followed different individuals at different instances as the configuration developed, we would be justified in calling the behaviour more complex.

To illustrate how the language can make a difference consider a case where the language did not take into account the direction of travel (of individual or herd) but only the distance and average speed. In this case, the behaviour might be simple, despite the fact that the behaviour is complex if you had to take the direction into account.

The type of “difficulty” can also be crucial. Given a language which includes direction and speed, a predator might be concerned with the difficulty of predicting in which direction the herd will initially set off, while an ecologist might only be concerned with the eventual direction and distance a herd travelled until it settled down. The first might be difficult to predict, the second easy; in the first case the behaviour of the herd would be complex while in the second it would be simple.

4.2.2 Cellular automata

Cellular automata have become icons of complexity in physics (due to Wolfram’s promotion of them as such [471]). They are described in very simple local terms that are easy to compute (simple enough that one could imagine them modelling real physical properties) but have complex global properties[43].

Here the specification of the parts is easy (an initial binary pattern and a simple rule for determining each bits next state, e.g. for each bit: 1 if the sum of bits of it and its two neighbours is odd, 0 otherwise). The complexity of the result after a certain number of iterations given the initial rule and bit pattern is dependent on the language of such

43.For an account of this see Wolfram [473].

formulation. It may be easy to formulate some statistical results for a large class of such rules but very difficult to predict if certain sub-patters will occur.

4.3 Relationship to Some Other Formulations

Broadly, existing measures of complexity can be seen as either:

1. *A special case of the approach given above in Section 4 above;*
2. *A relativisation of the approach either to some physical attribute (e.g. scale) or some 'privileged' framework to 'objectivise' it.*
3. *A weak characterisation of complexity better suited to some other descriptive label, such as 'information'.*

I consider a few existing formulations of complexity below, for some others see Appendix 1 - A Brief Overview of Some Existing Formulations of Complexity. The strength (or otherwise) of their claims as important or general measures of complexity thus rests with the critically of the type of difficulty they measure, the range of systems they apply to and the appropriateness of the underlying descriptive language.

4.3.1 Number of inequivalent descriptions

Extending the approaches to complexity taken by Rosen and Pattee (see section 6.7 on page 131), Casti [88] defines complexity as the number of nonequivalent descriptions that an observer can generate for a system it interacts with (see also section 8.29 on page 152). The observer must choose a family of descriptions of the system and an equivalence relation on them - the complexity is then the number of equivalence classes the family breaks down into given the equivalence relation.

This can be seen as a special case of my definition where the language of modelling is implicitly determined by the family of descriptions, their equivalence relation, and the relevant *difficulty* (being that of the coherence of the descriptions). If one had only one possible description that was completely coherent, one could claim that (with respect to the language) one had the representation of the system. If there were more then there were evidently incoherent aspects in the description that admit to alternative but not simultaneous description. See section 8.29 on page 152 for examples.

The trouble with this is that it is very difficult to ascribe any coherent meaning to the resulting measure, especially if one goes beyond a categorisation of one/many/infinite. In many languages (especially infinite ones) once one has two inequivalent descriptions of something then one can immediately construct a third inequivalent description by mixing parts of the first two, then construct a fourth and so on indefinitely. In other, suitably expressive languages one can always construct inequivalent descriptions given *any* particular subject and non-trivial equivalence relation, in which case there would be no simple systems. This makes using this particular measure of complexity difficult to use to make fine complexity judgements.

4.3.2 Effective measure complexity

EMC [194] can be seen as the difficulty of predicting the future values of a stationary series, as measured by the size of regular expression of the required model. See section 8.16 on page 145 for other related examples of this kind of approach.

4.3.3 Computational complexity

Computational complexity (section 8.47 on page 162) can be seen as the asymptotic difficulty (in terms of computational resources - typically time or storage space) in computing the output relative to the size of the input, given the specification of the problem. Here the difficulty is that of time. The overall description is the results of the computation, the bottom-up description is that of the problem, where the system is the computer program. Again this is a special case of my definition when one already has the ideal program for calculating the results. It is thus a residual complexity concerned with the practicalities of computation once one has solved the analytic aspects.

4.3.4 Algorithmic information complexity

Again, this can be seen as a special case. The algorithmic complexity is the minimal difficulty of storing a program to reproduce any particular pattern/number/index[44]. The overall description is the program and the bottom-up description is the original pattern and the language of description is (typically) a Turing machine. This is thus an indication of

44.Alternatively this could be taken as an estimation of the time taken to find an expression in a bottom-up enumerative search.

the amount of essential information (i.e. incompressible with respect to a Turing machine) inherent in the pattern. For more on this seesection 8.2 on page 136.

As argued above (section 3.4.1 on page 57) information is at best a weak measure of complexity unless applied at an abstract level.

4.3.5 Shannon entropy

Shannon Entropy, can be seen as the difficulty of guessing a message passing down a channel given the range of possible messages. The ideas is that the more difficult it is to guess, the more information a message gives you. This was not intended as a measure of complexity, but has been used as such by subsequent authors. See section 8.36 on page 156 for related approaches.

4.3.6 Crutchfield's "topological complexity"

The number of states in the smallest model in the minimal language where this is finite, is called the "topological complexity" in [122]. This is the difficulty of finding (in a bottom-up search procedure) a correct model, once the appropriate language has been chosen. Here the notion of complexity is relativised to the hierarchy of formal languages. This is not necessarily uniquely defined as the formal languages are only in a partial order. Also when it comes to stochastic languages, a minimal model roughly corresponding to "anything can happen" could generate any data. See section 8.32 on page 153 for related approaches.