# Chapter 1: The Need

*Virtually all design is conducted in a state of relative ignorance of the full behaviour of the system being designed.*
*Henry Petroski*

## A State of (Relative) Ignorance

When a system is designed there is from the very beginning a need to know how it will function. Indeed, the very reason for system design is to construct an artefact that provides an intended functionality. In Henry Petroski's book about 'Design Paradigms,' from which the above epigraph is taken, the topic was engineering design, and the majority of examples were physical structures and static systems, such as bridges. The 'behaviour' of a bridge is seemingly simple: it just has to be there and to maintain its own structure in order to allow safe passage for whoever or whatever uses it. Yet even in this case there is a 'relative ignorance' of what may possibly happen, as numerous examples of collapsing structures have demonstrated.

While design may not require perfect knowledge it at least requires acceptable ignorance. Partial ignorance is unavoidable in principle as well as in practice, but it should be so little that the consequences are unnoticeable. In other words, we must be reasonably sure that the systems we build will do what they are supposed to do, that they will function reliably as intended, and that they additionally will not do anything they are not supposed to do. The latter, in the case of a bridge, means collapsing or falling down. Yet history is full of examples of just that happening from the Dee Bridge disaster (1847) to the spectacular collapse of the Tacoma Narrows in 1940 and more recently to the failure of the I-35W Mississippi river bridge (2007).

### *The Reasons for Ignorance*

There are several reasons for this relative ignorance. One is that the properties of the materials may not be known well enough, neither how they will behave under extreme conditions (pressure, temperature, wind, etc.) nor how they will behave over time (ageing, degradation). Another is that it is uncertain what external conditions the system will be subjected to. In the case of bridges, the conditions refer to weather, changes in the chemical composition of the air, quality of maintenance, changes in 'customer' characteristics (e.g., heavier cars and more traffic), etc. A third reason is that work usually is done under conditions of insufficient time, information, and resources. The only thing we can know for certain is that things will happen that we have never thought of, although that in itself is of limited comfort.

If the situation is difficult for structures and static systems, it is even worse for dynamic systems. (Note, by the way, that so far we only have referred to nominally technical systems, e.g., a bridge as a bridge. Even here, the influence of social systems is obvious, for instance in how well the technical system is maintained, how well the components meet the specifications or requirements, how well the system is designed and built, etc.) For dynamic systems it is necessary to consider also the state of the parts and components as well as the dependencies among them. Energy – and information – must be provided, mass and materials must be moved around, substances will be transformed, etc. This creates literally countless dependencies of which there is relative ignorance, but which nevertheless must work as planned in order for the system to fulfil its purpose. But as it is well nigh impossible to foresee all possible combinations, even for purely technological systems, surprises abound.

### *Ignorance, Risk, and Safety*

Petroski's lament can be extended from the design of technical systems to include accident investigation and risk assessment as well. Virtually all accident investigations and virtually all risk assessments are conducted in a state of relative ignorance of the full behaviour of the system being analysed – and in some cases even in a state of ignorance about the typical behaviour.

In relation to event investigations, we can rarely if ever get all the information we need about what happened. One reason is that the search for information is influenced by biases and practical constraints. One such bias is the What-You-Look-For-Is-What-You-Find (WYLFIWYF) principle, which means that we look for what we assume is important. This precludes us from finding anything that we do not look for – serendipity excepted. Another bias is 'illusory comprehension.' The fact that we can squeeze events into pre-existing explanatory frameworks, all of which imply causality, means that we see causality even though it may not really be there. Among the practical constraints is the all too frequent lack of time, which means that the search for information is stopped when an acceptable explanation has been found, even though this may be incomplete or incorrect.

In relation to risk assessment, one source of ignorance is the inescapable uncertainty about what the future will bring. An observation made by many philosophers, and often repeated by politicians, is that we cannot know with certainty what will happen in the future. The Danish philosopher Søren Kierkegaard (1813-1855) noted that while life can only be understood backwards, it must be lived forwards. Samuel Coleridge (1772-1834) somewhat more poetically noted that "the light which experience gives us is a lantern on the stern which shines only on the waves behind us." A second source of ignorance is that most of the models or representations we use are so oversimplified that their validity is questionable. In an event tree, for instance, it is assumed that the chosen representation principle (binary branching) is an acceptable representation of reality. But that is clearly not the case, both because a distinction between fail and succeed is relative rather than absolute, and because things rarely develop in the way that was expected. A third source is the lack of imagination that partly is innate, partly comes from familiarity and habituation. A textbook example of that is Alan Greenspan's characterisation of the 2008 financial crisis as a "once-in-a-century credit tsunami, … that … turned out to be much broader than anything I could have imagined."

## Ignorance, Complexity, and Intractability

Ignorance of the future (and to some extent also ignorance of the past) is sometimes attributed to the degree of complexity of the systems we are dealing with, or simply to the purported fact that 'today's systems are – or have become – complex.' Complexity is, however, not a well-defined concept, as the following definitions exemplify:

- Mathematical complexity is a measure of the number of possible states a system can take on, when there are too many elements and relationships to be understood in simple analytic or logical ways.
- Pragmatic complexity means that a description, or a system, has many variables.
- Dynamic complexity refers to situations where cause and effect are subtle, and where the effects over time of interventions are not obvious
- Ontological complexity has no scientifically discoverable meaning, as it is impossible to refer to the complexity of a system independently of how it is described.
- Epistemological complexity can be defined as the number of parameters needed to describe a system fully in space and time. While epistemological aspects can be decomposed and interpreted recursively, ontological aspects cannot.

It might indeed be asked whether there is a state of relative ignorance because the systems we deal with are complex, or whether we call the systems complex because we do not have – and possibly cannot have – complete knowledge about them.

While we may entertain the hope that complete knowledge in principle is possible for technological systems (barring the vagaries of software), there is no reason for such optimism in the case of socio-technical systems. Here ignorance is a fact of life because it is impossible fully to define or describe the parameters in space or time even if we knew what they were. The main reason for this is however not that there are too many parameters, but rather that the systems are dynamic, i.e., that they continuously change.

In order for a system to be understandable it is necessary to know what goes on 'inside' it, to have a sufficiently clear description or specification of the system and its functions. The same requirements must be met in order for a system to be analysed and in order for its risks to be assessed. That this must be so is obvious if we consider the opposite. If we do not have a clear description or specification of a system, and/or if we do not know what goes on 'inside' it, then it is clearly impossible effectively to understand it, and therefore also to investigate accidents or assess risks.

The presence of the (relative) ignorance clashes with the assumptions of established safety analysis methods. These assumptions are a heritage from the large-scale technological systems for which the first safety assessment methods were developed in the late 1950s. Although the underlying assumptions rarely are stated explicitly, they are easy to recognize by looking at established methods, such as FMEA (Failure Mode and Effects Analysis), HAZOP (Hazard and Operability Study), Fault Trees, etc. The four main assumptions are:

- A system can be decomposed into meaningful elements (parts or typically components). Similarly, events can be decomposed into individual steps or acts. (The principle of decomposition is, of course, in conflict with the holistic principle that the whole is more than the sum of the parts.)
- Parts and components will either work or fail. In the latter case, the probability of failure can be analysed and described for each part or component individually. This is part of the rationale for focusing on the human error probability, and indeed for classifications of human errors.
- The order or sequence of events is predetermined and fixed as described by the chosen representation. If a different sequence of events needs to be considered, it is necessary to produce a new version of the representation, e.g., a new event tree or fault tree.
- Combinations of events are orderly and linear. They can be described by standard logical operators, and outputs are proportional to inputs.

Although these assumptions may be warranted for technological systems, it is highly questionable whether they apply to social systems and organisations, or to human activities. Models and methods that require that the system in focus can be fully described will for that reason not be suitable for socio-technical systems, neither for accident analysis and nor for risk assessment. It is therefore necessary to look for methods and approaches that can be used for systems that are incompletely described or underspecified. The two types of systems can be called tractable and intractable, respectively. The differences are summarised in Table 1.1 below.

*Table 1.1 Tractable and intractable systems*

| | Tractable system | Intractable system |
|---|---|---|
| Number of details | Description are simple with few details | Description are elaborate with many details |
| Rate of change | Low; in particular, the system does not change while being described | High: the system changes before a description can be completed. |
| Comprehensibility | Principles of functioning are completely known | Principles of functioning are partly unknown |
| Characteristic of processes | Homogeneous and regular | Heterogeneous and possibly irregular |

The differences between tractable and intractable systems can be illustrated by two examples. First consider a tractable system, such as a car assembly line. Here descriptions are (relatively) simple with only a small number of details. Work is meticulously planned and scheduled so that the assembly can be as efficient as possible and produce cars of a high quality. The rate of change is low, and usually the result of a planned intervention. Work is dominated by routine and is therefore homogeneous and highly regular. Finally, the comprehensibility is high, meaning that there is little, if anything, that is not understood in detail. The system is therefore tractable, which means that it can be specified in great detail and that decomposition is a natural approach to understand it better.

Then consider an intractable system, such as an emergency room (ER) in a hospital – or for that matter an emergency management room anywhere. Descriptions of such systems are elaborate and with many details since work is non-routine and the same situation rarely occurs twice. Intractable systems are heterogeneous rather than homogeneous. The rate of change is high, which means that the system – and its performance – is irregular and possibly unstable. Unlike a car assembly plant, work in an ER is difficult to plan because it is impossible to know when patients will arrive, how many there will be, and what kind of treatment they require. Finally, comprehensibility is low, because not everything is understood in detail. The system is therefore intractable, which means that it cannot be specified in detail, and that it does not make sense to decompose it.

## Systems Redefined

Systems are usually defined with reference to their structure, i.e., in terms of their parts and how they are connected or put together. Common definitions emphasise both that the system is a whole, and that it is composed of independent parts or objects that are interrelated in one way or another. Definitions of this type make it natural to rely on the principle of decomposition to understand how a system functions, and to explain the overall functioning in terms of the functioning of the components or parts – keeping in mind, of course, that the whole is larger than the sum of the parts.

It is, however, entirely possible to define a system in a different way, namely in terms of how it functions rather than in terms of what the components are and how they are put together. From this perspective, a system is a set of coupled or mutually dependent functions. This means that the characteristic performance of the system – of the set of functions – cannot be understood unless it includes a description of all the functions, i.e., the set as a whole. The delimitation of the system is thus not based on its structure or on relations among components (the system architecture). An organisation, for instance, should not be characterised by what it is but by what it does. Neither should it be characterised by the people who are in a given place (on the organisation chart or in reality) but by the functions they performs. One consequence of a functional perspective is that the distinction between a system and its environment, and thereby also the system boundary, becomes less important, cf., the discussion of foreground and background functions in Chapter 5.

## From Probability to Variability

One important development in the history of industrial safety was the transition that happened in response to the accident at the Three Mile Island nuclear power plant in 1979. This led to a change in focus from technology alone to human performance and the ways in which this could go wrong. Since established safety practices required that the probability of a failure or malfunction could be calculated, this spawned numerous proposals for how to calculate the probability of a 'human error.'

To cut a long story short, the so-called first generation of Human Reliability Assessment (HRA), represented by methods such as THERP (Technique for Human Error Rate Prediction) and HCR (Human Cognitive Reliability), all assumed that it was meaningful to refer to a human error probability, although they also acknowledged that the value or magnitude of this depended on external performance shaping factors. The human failure probability was nevertheless the coveted 'signal,' while the influence of the performance shaping factors was the 'noise.' This position was later effectively reversed in the so-called second generation HRA methods. In these methods, represented by ATHEANA (A Technique for Human Error Analysis), CREAM (Cognitive Reliability and Error Analysis Method), and MERMOS (Méthode d'Evaluation de la Realisations des Missions Opérateur pour la Sûreté), the influence of the performance conditions was seen as more important than the postulated human error probability. In other words, the influence of the performance shaping factors now became the signal, while the human error probability became the noise, to the extent that some methods even refrained from referring to the notion of human error at all.

A similar transition took place when the focus changed from the human factor to the organisation and/or safety culture. This happened in the mid-1980s, and is often linked to the disaster at the Chernobyl nuclear power plant and the loss of the space shuttle Challenger (both in 1986). In order to understand these accidents it became necessary to introduce new factors or conditions, although the basic thinking about safety remained the same. But even though the idea of safety culture was useful, it was nevertheless difficult to include the organisation in the calculation of failure probabilities. In practice, the search for a 'human error probability' was complemented by a search for an 'organisational error probability' or 'organisational failure rate' – although it was not usually expressed so bluntly. Describing it in this way, however, makes it clear that an 'organisational failure rate' is a meaningless concept. An organisation can neither fail nor function in the same way that a component can, i.e., it does not make sense to think about it in terms of the bimodal principle, as being either right or wrong. Indeed, an organisation – or a department in an organisation or a specific role – cannot really be thought of as a component in the first place.

Although HRA nominally looked for the probability of a 'human error,' the focus was actually on how human performance of a specific function might fail to reach its objectives rather than whether the human as such failed. In practice, the terminology nevertheless (mis)led people to focus on error probabilities. While it may be justified in the technological domain to see the performance of a function as synonymous with the state of a component, it is clearly not so in the case of human functions or the functions of a socio-technical system.

The differences in perspective become clear when a system is defined in terms of how it functions rather than in terms of its architecture and components. In this case the question is whether the *functioning* achieves its purposes. But this cannot be simplified to a question of whether the system is in a 'normal' state or a 'failed' state. It is instead a question of the variability of functioning and whether the outcome is acceptable under the existing conditions. But as soon as we say variability, we also acknowledge that any 'failure' will be temporary, hence reversible. We should consequentially try to understand how likely the variability of a system's performance is, and how the variability of multiple functions may interact to produce an unintended – and in most cases also unwanted – outcome.

What we are interested in is, however, not whether a function will be variable, since by definition they all are. Instead we are interested in whether the variability will be so large that the function will be unable to provide the desired outcomes. This can be due either to the variability of a single function or – more likely and also more importantly – the combination of the variability of multiple functions over time and over space. There will of course always be cases (even in complex socio-technical systems) where the variability of a single function (or a single activity) is so large that an adverse outcome is inevitable. But even in that case it is of limited use to say that the function has failed or that the component or entity has malfunctioned and to calculate the probability that this happened or will happen. In most cases the (unwanted) outcomes are due to interactions among individual functions, hence combinations of the effects of their variability. That being the case, it is clearly necessary to find ways to identify the potential for variability and to analyse how this may combine to produce the strong signals or the unwanted effects.

## Conclusions

Virtually all accident investigations and risk assessments are conducted in a state of relative ignorance of the full behaviour of the system. This condition contrasts with the fact that all established approaches to risk assessment require that it is possible to describe the system and the scenarios in detail, i.e., that the system is tractable. Unfortunately, all socio-technical systems are more or less intractable, which means that the established methods are not suitable. Since it is not reasonable to overcome this problem by making system descriptions so simple that they become tractable, it is necessary to look for approaches that can be used for intractable systems, i.e., for systems that are incompletely described or underspecified.

Resilience Engineering provides the basis for such approaches. Resilience Engineering starts from a description of characteristic functions, and looks for ways to enhance a system's ability to respond, monitor, learn, and anticipate. By emphasising that safety is something a system does rather than something it has, the unavoidable state of relative ignorance can be reduced by focusing on what actually happens. To do so requires a set of concepts, a terminology, and a set of methods that make it possible to describe work-as-done rather than work-as-imagined. That is what the FRAM is about.

## Comments to Chapter 1

The quote at the beginning is from page 93 in Petroski, H. (1994), Design Paradigms. Cambridge University Press.

The first presentation of the What-You-Look-For-Is-What-You-Find (WYLFIWYF) principle was at a Resilience Engineering workshop in Rio de Janeiro in December 2005. An explanation and examples can be found in Lundberg, J., Rollenhagen, C. & Hollnagel, E. (2009). What-You-Look-For-Is-What-You-Find – The consequences of underlying accident models in eight accident investigation manuals. *Safety Science*, *47*, 1297-1311.

The full Coleridge quotation from 1835 is "*If men could learn from history, what lessons might it teach us! But passion and party blind our eyes, and the light which experience gives us is a lantern on the stern which shines only on the waves behind us*." Søren Kierkegaard wrote that "*Livet forstås baglænds, men må leves forlænds*" in 1843.

The lack of imagination in designing, analysing, and managing both technological and socio-technical systems can easily have disastrous consequences and the importance of sufficient – or requisite – imagination can hardly be overstated. An early argument for that can be found in Westrum, R. (1991). *Technologies and society: The shaping of people and things*. Belmont, CA: Wadsworth.

Complexity, particularly as in complex systems, is often invoked as a *deus ex machina* to 'explain' why the use of modern technology is marked by so many unexpected events. Many also seem captivated by the so-called Complexity Sciences, although this may turn out to be the triumph of hope over experience. A good discussion of complexity is provided by Pringle, J. W. S. (1951). On the parallel between learning and evolution. *Behaviour*, *3*, 175-215.

There are several descriptions or surveys of HRA models, that also include part of their history. For example Kirwan, B. (1994). *A guide to practical human reliability assessment*. London: Taylor & Francis, or Hollnagel, E. (1998). *Cognitive reliability and error analysis method (CREAM)*. Oxford: Elsevier Science Ltd.