Chapter 1

The Driving Forces

The focus of CSE is how humans can cope with and master the complexity of processes and technological environments, initially in work contexts but increasingly also in every other aspect of daily life. The complexity of the current technological environment is not only something that must be mastered but paradoxically also provides the basis for the ability to do so. This entangling of goals and means is mirrored in the very concepts and theories by which we try to understand the situation of humans at work. To set the context, this chapter gives an overview of the scientific developments of the 20th century that have shaped our thinking about humans and machines.

INTRODUCTION

This book could reasonably start by asking how Cognitive Systems Engineering – in the following abbreviated to CSE – came about. Yet more relevant than accounting for the *how* is accounting for the *why*, by describing the forces that led to the formulation of the basic ideas of CSE. Such a description is important both to understand what CSE is all about, and as a justification for the CSE as it is today. Although CSE was formulated more than twenty years ago (Hollnagel & Woods, 1983), the basic message is certainly not outdated and its full potential has not yet been realised.

A description of the driving forces has the distinct advantage of being based on hindsight, which makes it possible to emphasise some lines of development that are useful to understand the current situation, without blatantly rewriting history as such. The three main driving forces are listed below. As the discussion will show, the situation at the beginning of the 21st century is in many ways similar to the situation at the end of the 1970s.

- The first driving force was the growing complexity of socio-technical systems, which was due to the unprecedented and almost unrestrained growth in the power of technology, epitomised by what we now call computerisation or applied information technology. This development

started slowly in the 1930s and 1940s, but soon after gained speed and momentum so that by the end of the 1970s computers were poised to become the dominating medium for work, communication, and interaction – revolutionising work and creating new fields of activity.

- The second driving force was the problems and failures created by a clumsy use of the emerging technologies. The rapid changes worsened the conditions for already beleaguered practitioners who often had insufficient time to adjust to the imposed complexity. One consequence was a succession of real world failures of complex systems that made human factors, human actions, and in particular the apocryphal 'human error' more conspicuous.
- The third driving force was limitations of linear models and the information processing paradigm. Although the popularity of human-computer interaction was yet at an initial stage, the view of humans as information processing systems had been keenly adopted by the engineering and computer science communities, leading to a fragmentary view of human-machine interaction.

Computerisation itself was the outcome of a number of theoretical and technological developments that went further back in time. Some of these were helpful and provided the concepts, models and methods that made it possible to address the practical issues of the time – around the middle of the 20th century – while others were decidedly unhelpful, in the sense that they accidentally created many of the problems that practitioners had to struggle with. It is usually the case that the positive aspects of an innovation – be they technical or conceptual – attract attention and therefore often quickly are used in applications. In the initial enthusiasm the negative aspects are easily overlooked and therefore only become clear later – sometimes even much later. This may happen in a concrete or material sense such as with nuclear power or the general pollution caused by industrial production. It may also happen in an incorporeal sense such as when a certain way of thinking – a certain paradigm in the Kuhnian meaning of the term (Kuhn, 1970) – turns out to be a stumbling block for further development. As we shall argue throughout this book, the information processing paradigm represents such as case. The positive sides were immediately and eagerly seized upon, but the negative sides only became clear almost half a century later.

On Terminology

Before proceeding further, a few words on terminology are required. Due to the background and tradition of CSE, the focus is on human-machine systems rather than human-computer systems, where the term *machine* is interpreted broadly as representing any artefact designed for a specific use. For the same reasons, the human in the system is normally referred to as an *operator* or a

practitioner, rather than a user. Finally, a system is used broadly to mean the deliberate arrangement of parts (e.g., components, people, functions, subsystems) that are instrumental in achieving specified and required goals (e.g., Beer, 1964).

COMPUTERISATION AND GROWING COMPLEXITY

When we refer to a technological system, we invariably think of it in the context of its use. People are therefore always present in some way, not just in the sense of individuals – such as in the paradigmatic notion of the human-machine system – but also in the sense of groups and organisations, i.e., social systems. Technological systems are of interest because of how they are used, rather than as pieces of equipment that exists physically – made up of mechanical, electronic, hydraulic, and software components. Regardless of whether the application is autonomous – as in the case of a space probe or a deep-sea robot – or interactive (with all kinds of shadings in between), a technological system is always embedded in a socio-technical context. Every system has been designed, constructed, tested, and put into use by people. Every system requires maintenance and repair, although for some it may be practically impossible to do so. Every system produces something, or represents something, with an intended use, hence with an intended user. In system design, people apply all their powers of creativity and imagination to prepare for the eventual application and to guard against possible failures.

Although all systems thus in a fundamental sense are socio-technical systems, it is useful to distinguish between *technological system*, where technology plays a central role in determining what happens, and *organisations*, where humans mainly determine what happens. In CSE, organisations are themselves considered as artefacts, as something devised for a specific purpose, although they are of a social rather than physical nature.

Self-Reinforcing Complexity Cycle

The intertwining of technology and complexity can be illustrated as in Figure 1.1. An arbitrary starting point for the cycle is the technology potential, which can be used to modify the way things are done as well as to introduce new functions altogether. Some familiar examples are the use of numerically controlled machines, industrial robots, computer-assisted design, flexible manufacturing, office automation, electronic exchange of data and funds, decision support systems, and the Internet. The growing technology potential is invariably seized upon and exploited to meet performance goals or efficiency pressures. This is referred to as the Law of Stretched Systems, originally suggested by Lawrence Hirschhorn:

> Under resource pressure, the benefits of change are taken in increased productivity, pushing the system back to the edge of the performance envelope. (Woods & Cook, 2002, p. 141)

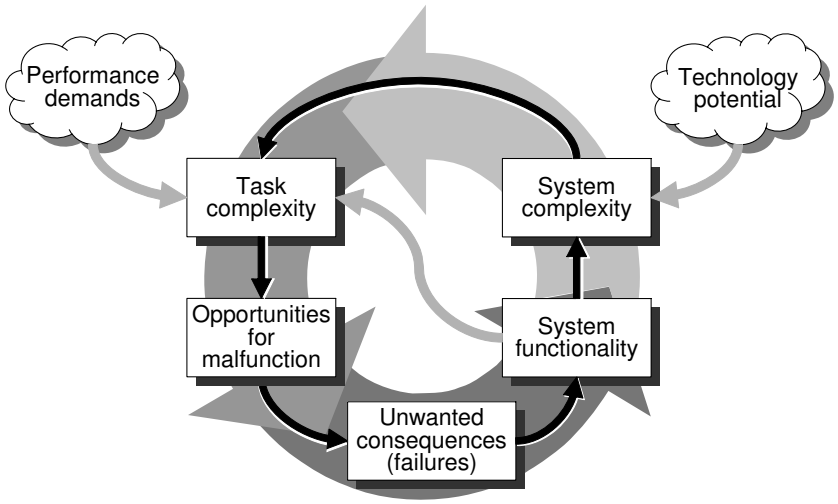

Figure 1.1: The self-reinforcing complexity cycle.

Some of the explicit motivations for putting technology to use are reduced production costs, improved product quality, greater flexibility of services, and faster production and maintenance. It need hardly be pointed out that these benefits are far from certain and that a benefit in one area often is matched by new and unexpected problems in another. Furthermore, once the technology potential is put to use this generally leads to increased system complexity. Although this rarely is the intended outcome, it is a seemingly inescapable side effect of improved efficiency or versatility. The increased system complexity invariably leads to increased task complexity, among other things (e.g., Perrow, 1984). This may seem to be something of a paradox, especially when it is considered that technological innovations often – purportedly – are introduced to make it easier for the users. The paradox is, however, not very deep and is just another version of the well-known irony of automation (Bainbridge, 1983). The growing task complexity generally comes about because adding functionality to a system means that there is an overall increase in complexity, even if there are isolated improvements. Another version of the paradox is the substitution myth, which will be discussed in Chapter 5.

As it can be seen from Figure 1.1, growing task complexity may also be a result of increased performance demands. Since our technological systems – including, one might add, our social institutions – often seem to be working

at the edge of their capacity, technology potential provides a way to increase capacity. This may optimistically be done in the hope of creating some slack or a capacity buffer in the system, thereby making it less vulnerable to internal or external disturbances. Unfortunately, the result invariably seems to be that system performance increases to take up the new capacity, hence bringing the system to the limit once more. Consider, for instance, the technological developments in cars and in traffic systems. If highway driving today took place at the speed of the 1940s, it would be very safe and very comfortable. Unfortunately, the highways are filled with more drivers that want to go faster, which means that driving has become more complex and more risky for the individual. (More dire examples can be found in power production industries, aviation, surgery, etc.)

Complexity and Unpredictability

Continuing through Figure 1.1, increasing task complexity together with increasing system complexity means more opportunities for malfunctions. By this we do not mean just more opportunities for humans to make mistakes, but rather more cases where actions have unexpected and adverse consequences or where the whole system malfunctions. Consider, for instance, the Space Shuttle. In addition to the Challenger and Columbia accidents, it seems to be the rule rather than the exception that a launch is delayed, in most cases due to smaller failures or glitches. Clearly, the Space Shuttle as a system has become so complex that it is close to being unmanageable (Seife, 1999, p. 5).

The larger number of opportunities for malfunction will inevitably lead to more actual malfunctions, more failures, and more accidents, which close the circle, as shown in Figure 1.1. The increase may either be in the frequency of already known malfunctions or in the appearance of new types of failure. The increasing number of failures in complex systems is a major concern for the industrialised societies and has itself been the motivation for numerous developments in the methods used to analyse, prevent, and predict such accidents (Hollnagel, 2004). At present, we will just note that a common response to incidents and accidents is to change the functionality of the system, typically by introducing additional barriers and defences. Although this in principle can be done without making the system more complex – in fact, it can sometimes be done by making the system simpler – the general trend is to add technology to make systems safer. Very often technological developments, such as increased automation, are promoted as the universal solution to accident prevention (the aviation industry being a good example of that). Another common solution is to introduce new barrier functions and defences to avoid future accidents, thereby making the systems more complex.

The net result of these developments is a positive feedback loop, which means that deviations will tend to grow larger and larger – resulting in more serious events and accidents (Maruyama, 1963). Although this interpretation may be overly pessimistic, the fact remains that we seem to be caught in a vicious circle that drives the development towards increasingly complex systems. One of the more eloquent warnings against this development was given in Charles Perrow's aptly named book *Normal Accidents* (Perrow, 1984), in which he argued that systems had become so complex, that accidents were the norm rather than the exception. It is in this sense that the growing technological complexity is a challenge as well as a motivation for CSE.

Figure 1.1 is obviously a simplification, which leaves out many nuances and opportunities for self-correction in the loop. In reality the situation is not always as bad as Figure 1.1 implies, since most complex systems function reliably for extended periods of time. Figure 1.1 is nevertheless useful to illustrate several important points.

- Systems and issues are coupled rather than independent. If we disregard these couplings in the design and analysis of these systems, we do it at our own peril.
- Events and relations must be understood in the context where they occur. It is always necessary to consider both dependencies to other parts of the system and to events that went before. This is particularly the case for human activities, which cannot be understood only as reactions to events.
- Control is fundamental in the definition of a cognitive system. Since all systems exist in environments that to some extent are unpredictable, there will sooner or later be a situation that was not considered when the system was designed. It can be difficult enough to keep control of the system when it is subject only to the 'normal' variability of the environment, but it becomes a serious challenge when unexpected situations occur. In order for the system to continue to function and maintain its integrity, control is necessary whether it is accomplished by the system itself or by an external agent or entity.

The positive feedback loop described above is a useful basis for understanding changes in human interaction with technology. Chapter 2 will provide a more thorough treatment of this issue and describe how technological developments have led to changes in the nature of work. For the moment we shall simply name three significant consequences of the growing complexity.

- The striving for higher efficiency inevitably brings the system closer to the limits for safe performance. Concerns for safety loom large and neither public opinion nor business common sense will accept efficiency

gains if they lead to significantly higher risks – although there sometimes may be different opinions about when a risk becomes significantly higher. Larger risks are countered by applying various kinds of automated safety and warning systems, although these in turn may increase the complexity of the system hence lead to even greater overall risks. It may well be that the number of accidents remains constant, but the consequences of an accident, when it occurs, will be more severe.

- A second important issue is the increased dependence on proper system performance. If one system fails, it may have consequences that go far beyond the narrow work environment. The increasing coupling and dependency among systems means that the concerns for human interaction with technology must be extended from operation to cover also design, implementation, management, and maintenance. This defines new demands to the models and methods for describing this interaction, hence to the science behind it.
- A third issue is that the amount of data has increased significantly. The sheer number of systems has increased and so has the amount of data that can be got from each system, due to improved measurement technology, improved transmission capacity, etc. Computers have not only helped us to produce more data but have also given us more flexibility in storing, transforming, transmitting and presenting the data. This has by itself created a need for better ways of describing humans, machines, and how they can work together. Yet although measurements and data are needed to control, understand, and predict system behaviour, data in itself is not sufficient. The belief that more data or information automatically leads to better decisions is probably one of the most unfortunate mistakes of the information society.

CONSPICUOUSNESS OF THE HUMAN FACTOR

Over the last 50 years or so the industrialised societies have experienced serious accidents with unfortunate regularity, leading to a growing realisation of the importance of the human factor (Reason, 1990). This is most easily seen in how accidents are explained, i.e., in what the dominant perceived causes appear to be.

It is commonly accepted that the contribution of human factors to accidents is between 70% – 90% across a variety of domains. As argued elsewhere (Hollnagel, 1998a), this represents the proportion of cases where the *attributed* cause in one way or another is human performance failure. The attributed cause may, however, be different from the actual cause. The estimates have furthermore changed significantly over the last 40 years or so, as illustrated by Figure 1.2. One trend has been a decrease in the number of accidents attributed to technological failures, partly due to a real increased

reliability of technological systems. A second trend has been in increase in the number of accidents attributed to human performance failures, specifically to the chimerical 'human error'. Although this increase to some extent may be an artefact of the accident models that are being used, it is still too large to be ignored and probably represents a real change in the nature of work. During the 1990s a third trend has been a growing number of cases attributed to organisational factors. This trend represents a recognition of the distinction between failures at the sharp end and at the blunt end (Reason, 1990; Woods et al., 1994). While failures at the sharp end tend to be attributed to individuals, failures at the blunt end tend to be attributed to the organisation as a separate entity. There has, for instance, been much concern over issues such as safety culture and organisational pathogens, and a number of significant conceptual and methodological developments have been made (e.g. Westrum, 1993; Reason, 1997; Rochlin, 1986; Weick, Sutcliffe & Obstfeld, 1999).

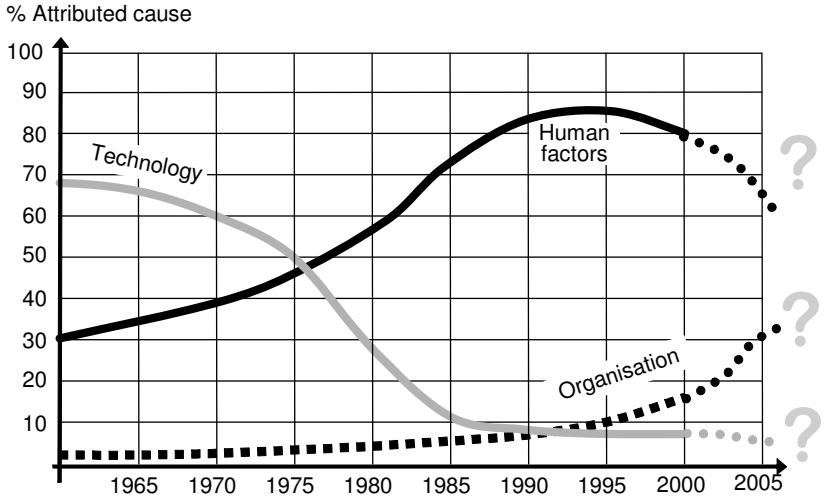

Figure 1.2: Changes to attributed causes of accidents.

The search for human failures, or human performance failure, is a pervasive characteristic of the common reaction to accidents (Hollnagel, 2004). As noted already by Perrow (1984), the search for human failure is the normal reaction to accidents:

> Formal accident investigations usually start with an assumption that the operator must have failed, and if this attribution can be made, that is the end of serious inquiry. (Perrow, 1984, p. 146)

Since no system has ever built itself, since few systems operate by themselves, and since no systems maintain themselves, the search for a human in the path of failure is bound to succeed. If not found directly at the sharp end – as a 'human error' or unsafe act – it can usually be found a few steps back. The assumption that humans have failed therefore always vindicates itself. The search for a human-related cause is reinforced both by past successes and by the fact that most accident analysis methods put human failure at the very top of the hierarchy, i.e., as among the first causes to be investigated.

THE CONSTRAINING PARADIGM

One prerequisite for being able to address the problems of humans and technological artefacts working together is the possession of a proper language. The development of a powerful descriptive language is a fundamental concern for any field of science. Basically, the language entails a set of categories as well as the rules for using them (the vocabulary, the syntax, and the semantics). Languages may be highly formalised, as in mathematics, or more pragmatic, as in sociology. In the case of humans and machines, i.e., joint cognitive systems, we must have categories that enable us to describe the functional characteristics of such systems and rules that tell us how to use those categories correctly. The strength of a scientific language comes from the concepts that are used and in precision of the interpretation (or in other words, in the lack of ambiguity). The third driving force of CSE was the need of a language to describe human-technology coagency that met three important criteria:

- It must describe important or salient functional characteristics of joint human-machine systems, over and above what can be provided by the technical descriptions.
- It must be applicable for specific purposes such as analysis, design, and evaluation – but not necessarily explanation and theory building.
- It must allow a practically unambiguous use within a group of people, i.e., the scientists and practitioners who work broadly with joint human-machine systems,

In trying to describe the functioning and structure of something we cannot see, the mind, we obviously use the functioning or structure of something we can see, i.e., the physical world and specifically machines. The language of mechanical artefacts had slowly evolved in fields such as physics, engineering, and mechanics and had provided the basis for practically every description of human faculties throughout the ages. The widespread use of models borrowed from other sciences is, of course, not peculiar to

psychology but rather a trait common to all developing sciences. Whenever a person seeks to understand something new, help is always taken in that which is already known in the form of metaphors or analogies (cf., Mihram, 1972).

Input-Output Models

The most important, and most pervasive, paradigm used to study and explain human behaviour is the S-O-R framework, which aims to describe how an organism responds to a stimulus. (The three letters stand for Stimulus, Organism, and Response.) The human condition is one of almost constant exposure to a bewildering pattern of stimuli, to which we respond in various ways. This may happen on the level of reflexes, such as the Patella reflex or the response of the parasympathetic nervous system to a sudden threat. It may happen in more sophisticated ways as when we respond to a telephone call or hear our name in a conversation (Cherry, 1953; Moray, 1959; Norman, 1976). And it happens as we try to keep a continued awareness and stay ahead of events, in order to remain in control of them.

Although the S-O-R paradigm is strongly associated with behaviourism, it still provides the basis for most description of human behaviour. In the case of minimal assumptions about what happens in the organism, the S-O-R paradigm is practically indistinguishable from the engineering concept of a black box (e.g. Arbib, 1964), whose functioning is known only from observing the relations between inputs and outputs. The human mind in one sense really is a black box, since we cannot observe what goes on in the minds of other people, but only how they respond or react to what happens. Yet in another sense the human mind is open to inspection, namely if we consider our own minds where each human being has a unique and privileged access (Morick, 1971).

That the S-O-R paradigm lives on in the view of the human as an information processing system (IPS) is seen from the tenets of computational psychology. According to this view, mental processes are considered as rigorously specifiable procedures and mental states as defined by their causal relations with sensory input, motor behaviour, and other mental states (e.g. Haugeland, 1985) – in other words as a Finite State Automaton. This corresponds to the *strong* view that the human *is* an IPS or a physical symbol system, which in turn 'has the necessary and sufficient means for general intelligent action' (Newell, 1980; Newell & Simon, 1972). The phrase 'necessary and sufficient' means that the strong view is considered adequate to explain general intelligent action and also implies that it is the only approach that has the necessary means to do so. In hindsight it is probably fair to say that the strong view was too strong.

The strong view has on several occasions been met by arguments that a human is *more than* an IPS and that there is a need of, for instance, intentionality (Searle, 1980) or 'thoughts and behaviour' (Weizenbaum,

1976). Interestingly enough, it is rarely doubted whether a human is *at least* an IPS or whether, in the words of cognitive science, cognition is computational. Yet for the purpose of describing and understanding humans working with technology, there is no need to make assumptions of what the inner mechanisms of cognition might be. It is far more important to describe what a cognitive system *does*, specifically how performance is controlled.

Although the S-O-R framework is no longer considered appropriate as a paradigm in psychology, it can still be found in the many models of human information processing and decision making that abound. It seems as if the study of human behaviour has had great difficulty in tearing itself away from the key notion that behaviour can be studied as the relation between stimulus (input) and response (output), even though the fundamental flaw of this view was pointed out more than one hundred years ago:

> The reflex arc idea, as commonly employed, is defective in that it assumes sensory stimulus and motor response as distinct psychical existences, while in reality they are always inside a co-ordination and have their significance purely from the part played in maintaining or reconstituting the co-ordination. (Dewey, 1896, p. 99)

Translated into the current terminology, Dewey made the point that we cannot understand human behaviour without taking into consideration the context or situation in which it takes place. Specifically, he made the point that it is wrong to treat the stimulus (input) and the response (output) as separate entities with an independent existence. They are both abstractions, which achieve their reality from the underlying paradigm and therefore artefacts of the paradigm. Consequently, if the principle of the S-O-R paradigm is abandoned, the need to focus on input and output becomes less important.

The Shannon-Weaver Communication Model

We are by now so used to the input-output model that we may no longer be aware of its peculiarities, its strengths, and its weaknesses. Seen from the perspective of the behavioural sciences, the ubiquitous graphical form of the input-output model can be traced to the seminal book on information theory by Shannon & Weaver (1969; org. 1949). The beginning of this book introduced a diagrammatic representation that, perhaps fortuitously, provided the sought for ‘image’ of the S-O-R model. Because of that we may refer to the Shannon-Weaver model as the ‘mother of all models’.

As shown by Figure 1.3, the model illustrates how a sender, called an information source, generates a message. The message is changed into a signal by the transmitter and sent through an information channel to the receiver. In the receiver, the signal is again changed into a message, which finally reaches the destination. The communication model was originally

developed to describe how something like a telephone system worked. If, however, we throw the telephone away the result is the prototypical situation of person A speaking to person B. The message is what person A, the sender, wants to say. The signal is the sound waves generated by the vocal system of the sender (the transmitter) and carried through the air (the channel) to the ear of person B (the receiver). The ears and the brain of the receiver transform the sounds into a meaning, which then (hopefully) has the desired effect on the receiver's behaviour. The description can obviously be applied in the reverse order, when the receiver becomes the sender and vice versa. This leads to the paradigmatic case of two-way communication.

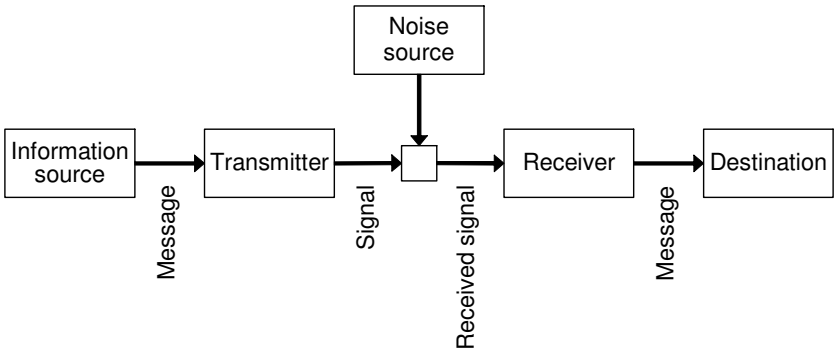

Figure 1.3: The Shannon-Weaver model of communication.

Prototypical Information Processing

The basic Shannon-Weaver model is doubled in Figure 1.4 to represent two-way communication. Basically the communication from sender to receiver is repeated, but now with person B, the original receiver, as the sender and person A, the original sender, as a receiver. This, of course, corresponds to the case of a dialogue, where – speaking more generally – system A and system B continuously changes their role. In order to make the first change 'work', so to speak, a second change has been introduced. This depicts how the message that has been received by system B is interpreted or processed internally, thereby giving rise to a new message that in the case of a dialogue is sent to system A, the original sender. The same change is also made for system A, who was the sender but who now has become the receiver.

It is not difficult to see the correspondence between this extended model and the common information processing model. Whereas information theory was interested in what happened to the signal as it was transmitted from sender to receiver, the emerging cognitive psychology was interested in what happened between receiving a message and generating a response. In this

case the Shannon-Weaver model could be used to describe the internal processes of the mind as a series of transformations of information. An example of that is George Sperling's classical studies of auditory memory (Sperling, 1963 & 1967), which described how the incoming sounds were sent through a number of systems and transformed on the way until it reached the level of consciousness.

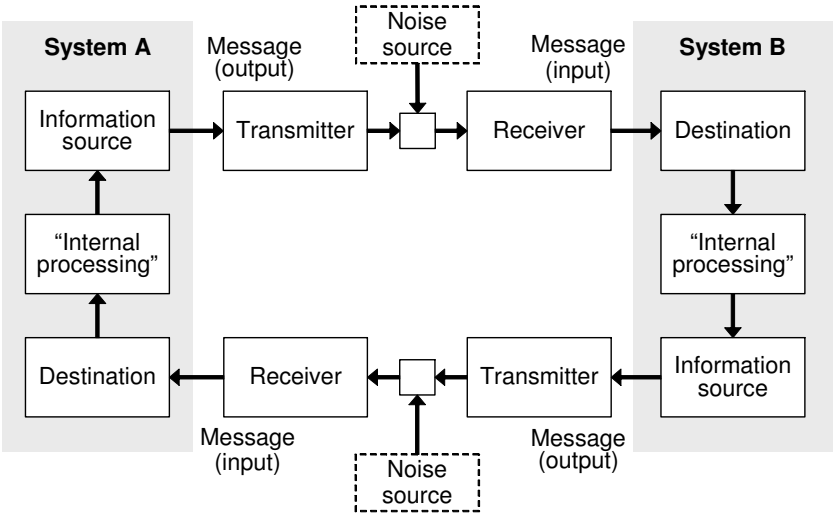

Figure 1.4: The extended Shannon-Weaver model.

In the present description there is no reason to go further into the details of how the model was developed. Several accounts have been provided of that (Attneave, 1959), although the basic model itself soon was taken for granted and therefore dropped out of sight. As we shall argue in the following, there were important consequences of adopting this paradigm, some of which are only now becoming clear. Whereas the Shannon-Weaver model is appropriate to describe the transmission of information between two systems, it is not necessarily equally appropriate to describe how two people communicate or two systems work together. The paradigm emphasises interaction, where CSE emphasises coagency. Although coagency requires interaction, it does not follow that it can be reduced to that.

FROM HUMAN-MACHINE INTERACTION TO JOINT SYSTEMS

One reason for the massive influence of the information processing paradigm was that it made it possible to describe what happened in the human mind between stimulus and response without being accused of resorting to

mentalism. It may be difficult today to appreciate the attraction of this, but fifty years ago it was certainly a determining factor. This has been mentioned by George Mandler as one of the two characteristics that set cognitive psychology apart from previous schools; the other was the recognition that cognition and consciousness were different concepts with only a partial overlap.

> By considering the various subsystems within the mental organisation as essentially independent entities connected by theoretically specifiable relationships, it has opened up theoretical psychology to a pluralism that is in sharp contrast to the monolithic theories of the 1930s and 1940s. (Mandler, 1975, p. 12)

Examples of the IPS metaphor can be found in the many models that have been proposed as explanations of human performance from the 1970s and onwards. A large number of these were mostly *ad hoc* explanations, since the IPS metaphor made it tantalisingly simple to suggest new mechanisms and structures to fit current problems. Gradually, however, several models emerged that achieved the status of consensus models. These models captured something essential about human performance, and expressed it in a manner that was both comprehensible and tremendously useful. Figure 1.5 shows an example of a model based on the limited capacity IPS. This model has an impeccable pedigree, going back to the fundamental research carried out by people such as Broadbent (1958), Cherry (1957), and Moray (1970). The important feature of this model, as formulated by Wickens (1987), was to point out that the limited attention resources could be considered for different modalities and that there were specific high-compatibility links between stimulus formats and central processing operations.

Information processing models can have different levels of detail and sophistication in how they account for the O in the S-O-R paradigm. Common to them all is that they start by some external event or stimulus (process information or the evaluation of a routine action) and end by some kind of response. In an IPS model the internal mechanisms are typically described in far greater detail than in an S-O-R model – at some point focusing on the *O* almost to the exclusion of the *S* and the *R*. It is nevertheless easy to appreciate that they have two fundamental similarities: the sequential progression through the internal mechanism, and the dependency on a stimulus or event to start the processing. Neisser (1976) caricatured the classical information processing view by describing the stages as ‘processing’, ‘more processing’, and ‘still more processing’. Despite the fact that this was done before most of the information processing models were formulated and gained general acceptance, Neisser’s criticism had little practical effect.

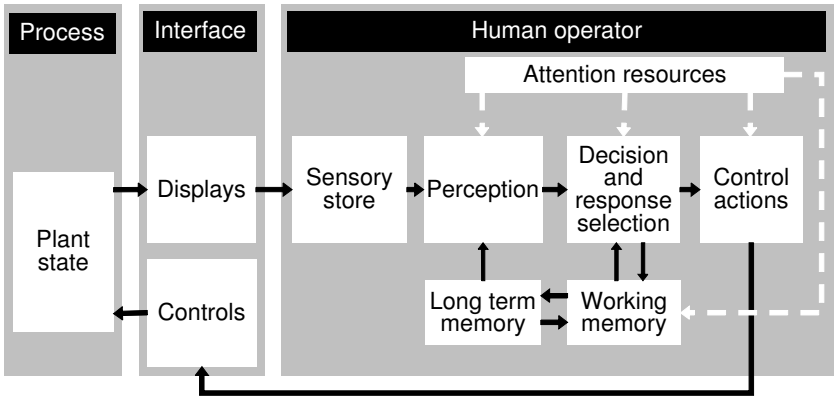

Figure 1.5: A limited capacity IPS model.

The Cognitive Viewpoint

In the 1980s, a growing number of people started to question the wisdom of considering the human as an IPS, in either the strong or the weak version. This led to the proposal of a model, which became known (in Europe, at least) as the cognitive viewpoint. This viewed cognition as active rather than re-active, as a set of self-sustained processes or functions that took place simultaneously, and changed the focus from the internal mechanisms of performance to overall performance as it could be observed. Behaviour was no longer seen as simply a function of input and system (mental) states, and the complexity of the inner 'mechanisms' was acknowledged to be too high to be captured adequately by a single theory.

The cognitive viewpoint introduced yet another change, namely the notion of the internal representation of the world in which the actions took place. This was formulated as follows:

> Any processing of information, whether perceptual or symbolic, is mediated by a system of categories or concepts which, for the information processing device, are a model of his (its) world. (De Mey, 1982, p. 84)

The internal representation, the system of categories, is characteristic for each system rather than generic and common to all systems. This means that two systems, or two persons, may have a different 'model of the world', i.e., different ideas of what is important as well as different knowledge and expectations. Specifically, as many have learned to their dismay, system designers and system users may have completely different ideas about how an artefact functions and how it shall be used.

The cognitive viewpoint also describes human performance as iterative or cyclic rather than as sequential. Cognition does not necessarily begin with an external event or stimulus; neither does it end with an action or response. This is in good correspondence with the perceptual cycle proposed by Neisser (1976). In CSE the perceptual cycle has been combined with the principles of the cognitive viewpoint to provide the basic cyclical model (Figure 1.7). The cyclical view explicitly recognises that meaningful human action is determined as much by the context (the task and the situation) as by the inherent characteristics of human cognition. Cognitive systems do not passively react to events; they rather actively look for information and their actions are determined by purposes and intentions as well as externally available information and events. The mistake of the sequential view is easy to understand because we are so readily hoodwinked into *observing* events and reactions and *interpreting* them using our deeply rooted model of causality. Yet an observable action does not need to have an observable event as a cause; conversely, an observable event does not necessarily lead to an observable action.

THE CLASSICAL HUMAN-MACHINE VIEW

Human information processing tended to focus on the 'inner' processes of the human mind and to describe these isolated from the work context in the tradition established by Wilhelm Wundt (cf., Hammond, 1993). This trend became stronger as time went by, partly because the proliferation of computers suddenly created a new population of users that were non-professional in the sense that they had not been specifically trained to deal with complex technology. This led to human-computer interaction as an independent field of study, which was practiced by people who had little or no experience from the processes and industries where human-machine interaction traditionally had been pursued. New generations of researchers and developers readily adopted the established mode of thinking and focused primarily on the interaction between the user and the computer with little concern for what might exist beyond that, except as an application layer. In practice this meant that there was no process over and above the human-computer interaction. The applications were in most cases driven by inputs from the user rather than having their own dynamics. Thus office and administrative applications came to dominate over process industries, power plants and aviation.

The essence of the classical human-machine view is shown in Figure 1.6, which, by using the simplest possible representation of each system as input, processing, and output functions, clearly shows the two main characteristics of the classical view. First, that the interaction is described exclusively as the exchange or transmission of input and output. Second, that humans and

machines are described in the same fashion, using the finite state machine as a basis. Technically speaking, the classical human-machine view represents a closed-loop control system in the tradition of the Shannon-Weaver paradigm.

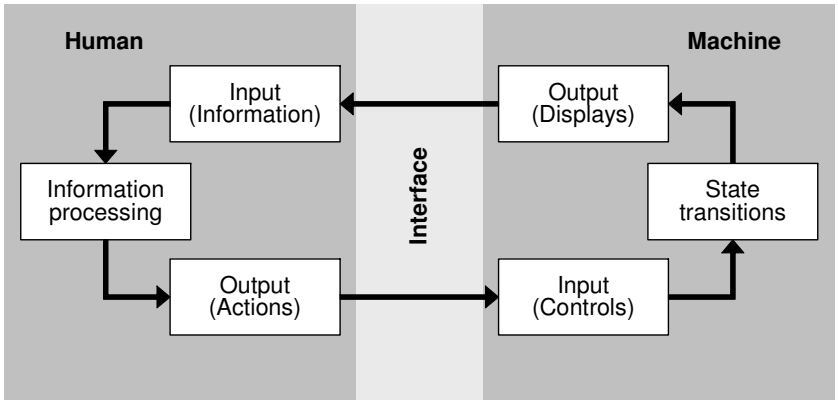

Figure 1.6: The classical human-machine view.

The Disintegrated View

The decomposed human-machine view became so accepted that the distinctive issue became interaction *with* the interface (computer) rather than interaction *through* the interface. The separation between humans and machines achieved the status of a real problem as shown by Norman's notions of the gulfs of evaluation and execution (Norman & Draper, 1986) and the very idea of usability engineering (Nielsen, 1993), and it became almost impossible to see it for what it really was – an artefact of the psychological application of the Shannon-Weaver model. Interface design became an important issue in and of itself where, for instance, the graphical user interface was seen as a problem in its own right rather than as something that played a role in how a user could interact with and control a process or application. The most recent example of this view is the notion of perceptual user interfaces (Turk & Robertson, 2000), which steadfastly continues the existing tradition, for instance, by focusing on the perceptual bandwidth of the interaction (Reeves & Nass, 2000). Yet the Shannon-Weaver communication paradigm was designed in order to deal with functioning of a communication *system*. In the context of human and machine working together we should be more interested in how the joint system performs than in how the parts communicate. While the communication between the parts and the interaction between humans and machines remain important topics, making these the dominating perspective or focus misses the more important

goal of understanding how the joint system performs and how it can achieve its goals and functions.

With the gradual change in terminology from human information processing to cognition and cognitive functions, the processes that mediated the responses to events became known as 'cognition in the mind'. Although cognitive science embraced the belief that cognition was computational, it was eventually realised that 'cognition in the mind' did not occur in a vacuum, but that it took place in a context or that it was 'situated'. Actions were no longer seen as exclusive outcome of mental activities, but rather as closely intertwined with artefacts and formations of the environment – including other people, of course. This became known in its own right as 'cognition in the world', and the use of tools and artefacts became a central activity in the study of applied cognition (Hutchins, 1995). Complementing 'cognition in the mind' with 'cognition in the world' overcame one limitation of human information processing but retained the focus on the cognition of the individual as a substratum for action rather than looking to the quality of action and the ability of the joint system to stay in control.

The disintegrated view reflects the assumptions of the sequential information processing paradigm in two different ways. The first is that the predominant models for cognition in the mind are sequential or procedural prototype models (Hollnagel, 1993b). The second is that actions are seen as responses to events, mediated by internal processes and structures. This view has several specific consequences, which are acknowledged to be important for the understanding and study of humans working with technology.

- Actions are treated as a series of discrete events rather than as a continued flow of events. Yet it takes but a moment's reflection to realise that what we do always is part of one or more ongoing lines of action and that one therefore should not consider actions one by one.
- Users are seen as single individuals. In practice, however, humans rarely work alone. Humans are always involved with and depending on other humans, even though they may be removed in time and space.
- The proactive nature of actions is neglected and the focus is on response rather than on anticipation. Yet human action is more often than not based on anticipation rather than simple (or even complex) responses.
- The influence of context is indirect and mediated by input. Yet in reality we know that context has a decisive influence on what we do and how we behave, even if that influence sometimes may be hard to spell out.
- Models are structural rather than functional. For instance, information processing models focus on how information is stored and retrieved rather than on the ability to remember and recall.

CHANGING THE PARADIGM

The many problems of the disintegrated view can only partly be overcome by compensating for them via more complicated model structures and functions. Sooner or later the fundamental flaws will have to be confronted. Instead of trying to solve the specific problems one by one, the solution lies in understanding the common root of the problems, and to overcome this by proposing an alternative, integrated view. This is more than a play with words, but signifies a fundamental change in the view of how humans and technology work together. The integrated view changes the emphasis from the interaction between humans and machines to human-technology coagency, i.e., joint agency or agency in common. Agency is here used as a verb describing the state of being in action or how an end is achieved, i.e., what a system (an *agent*) does.

We have argued above that the gulfs of evaluation and execution exist only because humans and machines are considered separately, as two distinct classes or entities. While it is undeniable that we, as humans, are separate from machines, the *physical separateness* should not lead to a *functional separateness*. The physical separateness was reinforced by the Shannon-Weaver paradigm, which was developed to describe the communication. Yet for CSE it is more important to describe the functioning of the joint cognitive system, hence to join human and machine into one.

Figure 1.7 illustrates the focus on joint system performance by means of the cycle that represents how the joint cognitive system maintains control of what it does. The cyclical model is based on the ideas expressed by Neisser's description of the perceptual circle (1976), and the basic cycle of planning, action, and fact finding in the 'spiral of steps' description of purposeful action (Lewin, 1958). The model aims to describe the necessary steps in controlled performance, regardless of whether this is carried out by an artefact, a human being, a joint cognitive system, or an organisation, and it is therefore also called a contextual control model (Hollnagel, 1993b).

The cyclical model has several specific consequences for the study of how humans and machines can work together. These are considerably different from the consequences of the sequential view, and deliberately so. The net outcome is that the cyclical view offers a better basis on which to study human-technology coagency.

- Actions are seen together. The cycle emphasises that actions build on previous actions and anticipate future actions. Human behaviour is described as a coherent series of actions – a plan – rather than as a set of single responses, cf., Miller, Galanter & Pribram, 1960.
- Focus on anticipation as well as response. Since the cyclical model has no beginning and no end, any account of performance must include what

went before and what is expected to happen. The cyclical model effectively combines a feedback and a feedforward loop.

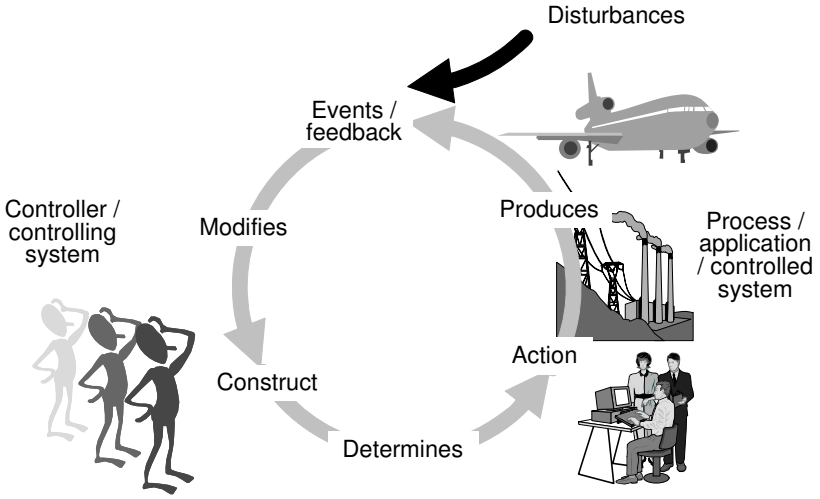

Figure 1.7: Basic cyclical model (COCOM).

- Users are seen as parts of a whole. The cyclical model focuses on coagency, on how users and environments are dynamically coupled, and on how actions and events are mutually dependent.
- Influence of situation or context is direct. In the cyclical model, context can affect the user's way of working – specifically how events are evaluated and how actions are selected, representing the fact that users may have different degrees of control over what they are doing.
- Models are functional rather than structural. The cyclical model makes minimal assumptions about components, hence about information processing. The emphasis is on performance rather than internal processes.

After the hegemony of the human information processing paradigm it may come as a surprise that the perspective on human action as being tied together, as continuous rather than discrete, is far from new. The criticism that CSE directs against the human information processing view is very similar to the criticism that functionalism directed against structuralism in the 1920s. This can be illustrated by the following quotation from a psychology textbook by H. A. Carr:

> Every mental act is thus more or less directly concerned with the manipulation of experience as a means of attaining a more effective

> adjustment to the world. Every mental act can thus be studied from three aspects – its adaptive significance, its dependence upon previous experience, and its potential influence upon the future activity of the organism. (Carr, 1925; quoted in Schultz, 1975, p. 161)

In 1925 mainstream psychology was focused on what happened in the mind and the controversy was therefore about mental acts. At that time there were no ergonomics, no human factors, and certainly no human-machine interaction. At the present time, psychology has developed in many different directions, one of them being the practically oriented study of humans and machines. The focus is nominally on actions and human performance, but as the preceding discussion has shown it is actually still about the mental acts, although they now are called human information processes or cognitive functions. If we replace the term 'mental act' with the term 'cognitive function', Carr's criticism of structuralism can easily be applied to the information processing paradigm and cognitive science. The lesson to be learned from this is not to forget the relativism of the paradigms we rely on.

DEFINITION OF A COGNITIVE SYSTEM

CSE was formulated in the early 1980s as a proposal to overcome the limitations of the information processing paradigm that already then had become noticeable, although not yet quite obvious. A cognitive system was at that time defined as:

- being goal oriented and based on symbol manipulation;
- being adaptive and able to view a problem in more than one way; and
- operating by using knowledge about itself and the environment, hence being able to plan and modify actions based on that knowledge.

Seen in retrospect this definition clearly bears the mark of its time, which was characterised by a common enthusiasm for expert systems, knowledge representation, and quasi-formal principles for information and knowledge processing. Although the definition emphasised the importance of what a cognitive system did over detailed explanations of how it was done, it still reflected the widely accepted need to account in some way for the underlying means.

CSE was, perhaps immodestly, proposed as 'new wine in new bottles'. The new wine was the idea that the construct of a *cognitive system* could function as a new unit of analysis. This was based on the realisation that (1) a common vocabulary was needed and (2) cognition should be studied as cognition at work rather than as functions of the mind. The new bottles were the changed of focus in the domain – looking at the human-machine

ensemble and the coupling between people and technology, rather than the human *plus* machine *plus* interface agglomeration. The cognitive systems synthesis arose at the intersection of what were normally boundaries or dividing lines between more traditional areas of inquiry, such as technological versus behavioural sciences, individual versus social perspectives, laboratory experiments versus field studies, design activity versus empirical investigation, and theory and models versus application and methods.

Over the years the original definition of a cognitive system has changed in two important ways. First, the emphasis on overt performance rather than covert functions has been strengthened. In the CSE terminology, it is more important to understand *what* a joint cognitive system (JCS) does and *why* it does it, than to explain *how* it does it. Second, the focus has changed from humans and machines as distinct components to the joint cognitive system. There is consequently less concern about the human-machine interaction and more about the coagency, or working together, of humans and machines. There is also less need of defining a cognitive system by itself and to worry about definitions of cognition, cognitive functions, cognitive processes, etc. Since the joint cognitive system always includes a human, there is no real need to develop a waterproof definition of cognition, or indeed to argue about whether cognition – or even intelligence – can exist in artefacts.

The revised definition of a cognitive system is *a system that can modify its behaviour on the basis of experience so as to achieve specific anti-entropic ends*. The term entropy comes from the Second Law of Thermodynamics where it is defined as the amount of energy in a system that is no longer available for doing work; in daily language it is used to mean the amount of disorder in a system. Systems that are able locally to resist the increase in entropy are called anti-entropic. In basic terms it means that they are able to maintain order in the face of disruptive influences, specifically that a cognitive system – and therefore also a joint cognitive system – is able to *control* what it does. (In cybernetics, control is defined as steering in the face of changing disturbances (Wiener, 1965; org. 1948), cf. also the discussion of the Law of Requisite Variety in Chapter 2.)

Most living organisms, and certain kinds of machines or artefacts, are cognitive systems. In particular it must be noted that organisations are cognitive systems, not just as an agglomeration of humans but by themselves. Cognitive systems appear to have a purpose, and pragmatically it makes sense to describe them in this way. In practice, the purpose of the JCS is often identical to the purpose of the human part of the system, although larger entities – such as organisations – may be seen as having purposes of their own.

As shown by the above definition, a JCS is not defined by what it *is*, but by what it *does*. Another way to characterise the focus of CSE is to note that in the case of a JCS, such as a human-machine ensemble, it is characteristic

that the machine is not totally reactive and therefore not totally predictable (the same, of course, goes for humans, but this is less surprising). This lack of predictability is central for CSE. (The argument can in principle be extended to cover other combinations, such as human-human co-operation and some cases of sophisticated machine-machine co-operation.)

Consider, for instance, driving an automobile. Although seemingly a very simple thing to do, the driver-vehicle system can reasonably be seen as a JCS. First, the vehicle usually cannot be controlled by single actions but requires a combination of several actions. (This may apply even to functions that have nothing to do with driving, such as using the radio.) Second, the performance or function may not be entirely predictable, due to ambiguities in the design of the interface or because it is not clear how the buttons and controls function. Third, driving is the control of a dynamic process indeed it is steering in the cybernetic meaning of the word, which furthermore takes place in a dynamic and unpredictable environment.

The unpredictability of an artefact is generally due either to the inherent dynamics of the artefact or incomplete or insufficient knowledge of the user *vis-à-vis* the artefact – either permanently or temporarily – due to lack of training and experience, confusing interface design, unclear procedures, garbled communication, etc. In practical terms, CSE is interested in studying JCS that have one or more of the following characteristics:

- The functioning is non-trivial, which generally means that it requires more than a simple action to achieve a result or to get a response from the artefact. For more complex artefacts, proper use requires planning or scheduling.
- The functioning of the artefact is to some degree unpredictable or ambiguous for any of the reasons mentioned above.
- The artefact entails a dynamic process, which means that the pace or development of events is not user-driven. The general consequence is that time is a limited resource.

A situation of particular interest is the one where the machine controls the human – which in a manner of speaking happens whenever the user loses control. In these cases the controlled system by definition is not reactive and certainly not predictable, since users rarely do what the designers of the artefact expect them to.

The Scope of CSE

One of the motivations for the development of CSE was to provide a common set of terms by means of which the interaction between people and machines could be described. People, obviously, are natural cognitive systems, while machines in many cases can be considered as artificial

cognitive systems. In 1981, as well as today, the three most important issues for CSE were: (1) how cognitive systems cope with complexity, for instance by developing an appropriate description of the situation and finding ways to reach the current goals; (2) how we can engineer joint cognitive systems, where human-machine are treated as interacting cognitive systems, and (3) how the use of artefacts can affect specific work functions. In a single term, the agenda of CSE is how we can design joint cognitive systems so that they can effectively control the situations where they have to function.

An important premise for CSE is that *all work is cognitive*. There is therefore no need to distinguish between cognitive work and non-cognitive work or to restrict cognitive work to mean the use of knowledge intentionally to realise the possibilities in a particular domain to achieve goals. Everything we do require the use of what we have 'between the ears' – with the possible exception of functions regulated by the autonomic nervous system. The cognitive content of skills becomes obvious as soon as we try to unpack them or apply them under unusual circumstances, such as walking down a staircase in total darkness. The fact that we habitually are able to do a great many things without thinking about them or paying (much) attention to them does not make them non-cognitive. It rather demonstrates that there can be different levels of control in performance. Similarly, CSE considers only the use of artefacts, without making a distinction between 'cognitive tools' and 'non-cognitive tools'. An artefact, such as a bicycle, may have been developed to support a predominantly manual function but anyone who has ever tried to teach a child to ride a bicycle will be keenly aware that this involves a very high level of cognition. There is, consequently, no requirement to have a specific discipline of cognitive design dedicated to the design 'cognitive work' and 'cognitive tools'. The engineering of cognitive systems will do nicely on its own.