Safety Science 69 (2014) 3–17

Contents lists available at SciVerse ScienceDirect

Safety Science

journal homepage: www.elsevier.com/locate/ssci

Safety management – A multi-level control problem ☆

Björn Wahlström [a,*], Carl Rollenhagen [b]

[a] Vattenfall SMI, AX-22920 Brändö, Finland
[b] Department of Philosophy, Royal Institute of Technology, SE-100 44 Stockholm, Sweden

ARTICLE INFO

Article history:
Available online 4 July 2013

Keywords:
Safety management
Control structures
Modelling safety
Methods and tools
Nuclear power

ABSTRACT

Activities in safety management build on a control metaphor by which control loops are built into the man, technology, organisational and information (MTOI) systems to ensure a continued safety of the operated systems. In this paper we take a closer look on concepts of control theory to investigate their relationships with safety management. We argue that successful control relies on four necessary conditions, i.e. a system model, observability, controllability and a preference function. The control metaphor suggests a division of the state space of the modelled system into regions of safe and unsafe states. Models created for selected subsystems of the MTOI-system provide a focus for control design and safety assessments. Limitations in predicting system response place impediments to risk assessments, which suggest that new complementary approaches would be needed. We propose that polycentric control may provide a concept to consider in a search for a path forward. We investigate approaches for modelling management systems and safety management. In spite of promises in the use of a control metaphor for safety management there are still dilemmas that have to be solved case by case. As a conclusion we argue that the control metaphor provides useful insights in suggesting requirements on and designs of safety management systems. The paper draws on experience from the Vattenfall Safety Management Institute (SMI), which started its operation in 2006.

1. Introduction

Members of societies are faced with risks emanating from both old and new technological systems. Accidents that have occurred in nuclear power plants, chemical production facilities, off-shore installations, and many other industrial branches, present a growing concern in society. Focus on technological factors, human factors, and the more recent conceptual innovation of safety culture, has provided safety managers with new tools and methods for safety management. Moreover, the safety paradigm has shifted from focus on parts towards a holistic conceptualisation of safety, but exactly what this means in terms of practical safety management is not always clear. One difficulty seems to be that the systems involved encompass very diverse areas of knowledge and modelling approaches, making it hard to find a common language for describing different aspects of system behaviour.

Many authors have argued that systemic approaches to safety should be applied (Bang Dyhrberg and Langaa Jensen, 2004; Kirwan, 2011). We generally support such approaches. In the present contribution we argue that a systemic approach entails an understanding of the specific characteristics that govern the behaviour of generic subsystems found in socio-technical systems. Already the need for considering people and organisations in system models calls for inputs from areas such as anthropology, psychology, social psychology, sociology, management science and economics. In a search for a common approach for modelling systems, subsystems and their interactions, we argue that a control metaphor may provide an overarching language that can be used both in design and analysis of safety management.

Applying a control metaphor for understanding system behaviour is not new. For example, it has been applied at several hierarchical levels for understanding safety of large complex systems by Rasmussen and Svedung (2000), Kjellén (2000) and Swuste et al. (2010), among many others. To illustrate, the society exercises control of companies that operate hazardous facilities through laws and regulations. Organisations implement policies and management systems to ensure that plants are designed, built, operated and maintained in a safe manner. Controls are implemented in many processes and they include both feedback and feed forward paths by which outcomes are monitored and correcting actions are initiated when deficiencies are found. Our contribution rests on the assumption that in order to efficiently use a control metaphor to support safety management, it is necessary to, at least in gross terms, identify a set of generic component classes in terms of man (people), technology, organisation and information (Rollenhagen, 2003).

☆ This is an extended version of a paper that was presented on the PSAM11 conference in Helsinki, 25–29 June, 2012.

[*] Corresponding author. Tel.: +358400448710; fax: +35898030781.
E-mail address: bjorn@bewas.fi (B. Wahlström).

0925-7535/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ssci.2013.06.002

The aim of this paper is to establish a basis for applying the control metaphor to create a frame of modelling, which can be used in analysis and design of systems and their controls to ensure that an acceptable safety can be reached. We are here mainly interested in process safety (Grote, 2012) for complex interconnected systems such as nuclear power plants, but the suggested approach can also be used for other hazardous systems. The advantage is that system models can be built using an approach that enables a consistent transfer of focus from the whole system toward increasing details. In this way it is possible to model various parts of a production system with their controls to provide evidence that the whole system can be operated safely.

If we can manage safety of a system with respect both to entirety and to details, it would be possible to define *necessary* conditions for safety, which in principle implies that we can build a reviewable safety case. The dilemma, however, is that we are still not able to build *sufficient* conditions for safety, because one can always argue that there is some unknown sequence of events, which would lead to an accident. With the proposed approach, we think that there are possibilities to argue that sufficient conditions can be claimed for at least some restricted parts of the production system and its controls.

The paper starts from a general discussion of threats, risk and safety, which touches on the question how safety can be built and demonstrated. We argue for the need of considering the four systems, man (people), technology, organisational and information (MTOI) all with their own modelling paradigms. We consider the design basis threat (DBT) concept, because this concept provides a set of initiating events that in many cases can give a starting point for system design. Of course it is also important to be aware of threats that are excluded from the analysis for some reason or another.

The third section considers in more detail requirements for a successful control of safety. Considering safety controls, one may separate between *state control* and *transition control*. State control represents the controls necessary to keep a system is in a safe region of a state space, and transition controls transfer the system back to a safe state if it has entered an unsafe region.

The fourth section discusses five control structures, which in the control of large complex systems are used and combined in various ways. These controls have their own characteristics, which are important to understand in modelling, designing, operating and maintaining their functionality. Controls are applied for different purposes and one may separate between main and supporting control tasks. Hierarchical controls are formed through an interconnected network of control loops that get their inputs from many diverse sources and which can influence both concrete and abstract entities.

From there we move to the general problem of modelling. To ensure a proper understanding of sociotechnical systems and their risks it is necessary to include at least four distinctly different systems, man (the M-system), technology (the T-system), organisation (the O-system) and information (the I-system). An important part of the modelling effort is to select a state space of the used models and to assess how these state spaces may be divided into three regions, one region of safe states, one region of unsafe states and one region where safety is undecided.

In the sixth section we discuss structural and mathematical prerequisites that place serious impediments on possibilities to predict system behaviour. In this section we also briefly discuss the concept of polycentric control. With this approach one may construct a set of "small worlds" for which controls can be designed and assessed. By the use of independent autonomous systems it is likely that less predictability would result, but we think that the balance nevertheless will be positive, since this approach has the benefit of building resilience into different parts of a system (Hollnagel et al., 2006).

In the seventh section we discuss management systems with the intent of bringing concepts from management science in line with the proposed control metaphor. To model the O-system in larger details, it is necessary to consider organisational structures that are defined through processes and functions. Towards the end of the section we argue, along with many others, that the feedback of experience may be the most important function within safety management. Organisational changes close the loop from feedback of experience to actual improvements.

In the eighth section, we take a closer look on implications for safety management. An important insight is that safety cannot be the only condition that influences preferences in the control loops. Effort should also be spent on how other performance criteria besides safety are prioritized and enter the controls. A specific question is to consider differences in preferences during major lifecycle phases such as design, construction, operation and decommissioning. Audits, assessments and reviews as well as regulatory oversight can be perceived as control loops that aim for obtaining indications on deviations from norms and standards to initiate correcting actions.

In spite of research efforts and development of safety management there are still a number of dilemmas that have to be addressed on a case to case basis. One is connected to limits about what we know and another is what can be considered to be safe enough. Risk profiles often have their centre of gravity at low probability, high consequence events, which lead to large uncertainties in calculated risk estimates. Selecting the focus for a modelling effort is a challenging task, but if the entire system can be covered together with the most important safety controls, at least some confidence in safety can be reached. Additional dilemmas are related to finding suitable balances in preferences as well as creating suitable models of decision making.

A conclusion of the paper is that we find the control metaphor helpful in many respects. Especially the consideration of safe and unsafe regions in a state space may provide a path forward. Polycentric control may also contribute to new insights for safety when the behaviour of an interconnected network of "safe" systems is investigated. A main conclusion is however that uncertainty in risk estimates, in spite of modelling efforts, will remain large enough to motivate an application of the precautionary principle in societal decision making.

2. Threats, risks and safety

The concepts of risk and safety are constructed through the consideration of threats to which many uncertainties are associated (Aven et al., 2011). If a threat is realised by an initiating event, it will normally come with consequences in terms of costs for the system operator (and the society). It is therefore in the interest for system operators and the society to implement and otherwise support measures, by which threats could be eliminated, isolated, controlled and/or mitigated. Risks management involve two interacting parts, an *analysis* part, where threats are identified and assessed and a *design/implementation* part, where risks are acted upon. Safety improvements may include changes in system design as well as implementing safety barriers, active safety systems and protective functions. The acceptability of building and operating potentially dangerous systems is usually controlled by society, where the operator is obliged to present a safety case with arguments for why the system can be considered safe.

2.1. Probabilities or possibilities

Quantification of risk will need assessments of the uncertainties involved. A starting point is to consider uncertainties associated to an initiating event $h \in H$ that is effecting the system at a time instant $t = t_0$. Depending on its initial state x_0 and the control inputs $u(t)$ for $t > t_{0,}$ the system can take different routes, which may or may not lead to disastrous consequences. The uncertainties come from at least four sources, (1) uncertainties associated with the initiating event, (2) uncertainties connected to the initial state, (3) uncertainties connected to controls that are applied after the initiating event and (4) uncertainties associated with our understanding of the system.

It has been argued that the concept of risk should be constructed from identified threats without a too early consideration of how uncertainties will be handled (Aven and Zio, 2011). This is a warranted approach, because decisions on when certain risks can be accepted or not, may favour either a probabilistic or possibilistic approach. A deterministic reasoning about uncertainties will favour possibilistic approaches, whereas the probabilistic approach is used in the probabilistic safety analysis (PSA) methodology. Using a probabilistic approach a common approach is to consider the pair (C and P), where C is a measure for the cost of consequences and P is the probability. Safety then implies that C and P should be small enough to be acceptable. A combined risk measure is usually given as $R = C * P$, but it may still be important to remember that safety also has a societal dimension (Rochlin, 1999). This societal dimension has to do with the framing, the role of emotions in perceptions and value judgements of risks Hermansson, 2012).

2.2. Models of how safety is constructed

The systems to be controlled consist of three very different subsystems that should be modelled based on their own assumptions. These systems have sometimes been termed *man*, *technology* and *organisation* (MTO) or *plant*, *people* and *processes* (PPP). Rollenhagen (2003) has previously suggested that in addition to these systems one should also include a fourth system *information*, which can be considered to include databases, instructions and documentation.

The field of safety engineering has increasingly built a knowledge base comprising of various safety principles and strategies (Möller and Hansson, 2008). Examples of such safety principles are a *graded approach to safety* and the concept of *defence in depth* (e.g. multiple and independent safety barriers). These principles are used both in design and as a basis for oversight in existing systems.

Barrier functions can be seen as controls that serve two purposes, (1) to protect the system from leaving a safe region and (2) to force a system back to safety when it has entered an unsafe region. Protection of the barriers can in turn be achieved by introducing the single failure criterion and applications of redundancy, separation, diversity and the grace rule.[1] We argue that these safety principles, in different ways, can be applied as controls in all four subsystems: man, technology, organisation and information.

2.3. Design basis threats

A commonly accepted principle for safety design at least in the nuclear field, is to suggest a strategy based on so called design basis accidents (DBA), with the interpretation that such events will serve as probing stones for robustness of system design. The DBA concept has been generalized to security as design basis threats (DBT), which the system should be able to cope with. The DBT concept has an immediate connection to initiating events $h \in H$ that may hit a system. An application of the DBT concept is to select a finite set of events $h_i \in H$ with $i \in \{1, \ldots, N\}$, which in the design process are used to verify that the system is able to cope with them.

The value of the DBT concept can be seen if we are able to construct a region $H_i \subset H$ with the property that any event $h \in H_i$ can be considered less serious than h_i. If this is possible, it is sufficient to analyse a finite set of events h_i to claim that a system will cope with all events $h \in H_1 \cup \ldots \cup H_N \subset H$. We are thus in our analysis able to cover not only specific events, but also connected regions in the event space. In this way we would be able to discuss whether or not a risk analysis can be considered reasonably complete.

2.4. A safety case

For systems where safety is a concern, it is also necessary to provide evidence to a third party that it is safe enough. The third party may be an independent regulator, a governmental office or even an internal safety office. A common way of provide this evidence is through a *safety case*. A safety case will at least in principle be structured around *claims* and *evidence* for the claims to be true.

More concretely a safety case will typically contain system descriptions and a set of design basis threats together with descriptions of how they are acted upon using favourable designs and various controls. Predictions of system responses are generated with simulations by computer codes that have been validated using experimental facilities. The safety case may include a probabilistic safety analysis (PSA), where transients have been modelled with fault trees and event probabilities.

2.5. Remaining threats

In any risk analysis it is necessary to set a cut level beyond which risks are accepted. This means that some reasonably objective estimate of risk has been possible to establish (Hermansson, 2012). To set such a level some agreed criteria have to be used. Uncertainties in risk estimates may imply that certain initiating events wrongly are placed in the category of rest risks.

For a system that have been protected by applying the defence in depth principle, the most serious remaining threat is that several of the safety barriers would be made non-functional by some single event, a common cause failure (CCF). The likelihood that some unknown CCF simultaneously would influence several of the safety precautions can never be completely removed, but it may be decreased using the principles of diversity and separation.

More generally one may suggest that the quality of a safety case could be assessed with respect to its *completeness*, *correctness* and *consistency* (C^3). The completeness of a safety case cannot be assured a priori, but the process of experience feedback should at least be able to ensure that it is updated when new knowledge is obtained. Correctness in turn can be assessed by assessing the validity of models and data used for generating predictions of system responses in specific situations. Consistency within a safety case aims at assuring that safety relevant scenarios have been modelled with a similar degree of accuracy.

Many proven strategies are regularly applied to ensure safety in complex systems, some of which have been briefly reviewed above. There are still many remaining problems facing analysis of complex systems. Particularly, we would need to know how contributions from human behaviour, organisational factors, culture, information issues etc. interact with technological factors. Since several knowledge domains are involved in this modelling and analysis it would be a great advantage if general principles and a common modelling language could be applied.

[1] The grace rule implies that a certain minimal time before manual actions are required is applied as a design principle for the technical system. In the nuclear field it is sometimes referred as the 30-min rule.

3. Control of safety

Assuming a system S, a generalised control task is characterised so that it can be manipulated with inputs u giving observed outputs y. The system S is assumed to be described with a state variable x, which has the property that by giving its state x_0 at time t_0 and the system input $u(t)$ from t_0 onwards, unambiguously define future outputs $y(t)$ (Zadeh and Desoer, 1963). The state x_0 can be perceived as integrating the history of past inputs $u(t)$ for $t < t_0$. Necessary conditions for successful control can then be given by applying the following general requirements:

– There should be a system model M, which can be used to make predictions (deterministic or probabilistic) about outcomes that selected actions would produce.
– The system should be observable, which means that it should be possible to determine the state of the system.
– The system should be controllable, which means that one should be able to manipulate (or control) the state of the system with the available input variables.
– There should be a preference relation, which makes it possible to separate between desired and non-desired outcomes.

The following discussion is based on the assumption that the system considered is deterministic, but it can as well be generalised to probabilistic systems. If this is the case one have to consider also uncertainties that are connected to initiating events, system initial state, controls applied and the validity of the model.

3.1. The system model

In the selection of a system model, the intent of the modelling has to be considered. System models are generally assumed to consist of subsystems, which have their own internal structure. Models can be seen as micro-explanations of macro-behaviour that is visible for the systems. The approach we suggest is well in line with for example Haimes (2012).

A further task in producing a system model is an assessment of system state and its components. These state components are reflected in the chosen model. Depending on the selected purpose of the model (explanation and/or control) various variables must be selected. When the aim is controlling safety, it is of course necessary to include those variables that are assumed (or empirically known) to be essential for safety. One of the crucial tasks for the control of safety is to find those variables that are most essential for safety.

Control of safety suggests a need for distinguishing between safe and unsafe regions of the state space, as well as control inputs that will maintain system state within or transfer it to a safe region. Conditions defined in the safety case for a system can for example provide one set of necessary conditions for safety, which may be used to define safe regions in the state space of the system. Similarly signals that indicate the occurrence of an initiating event may be used to define unsafe regions in the state space of the system.

3.2. Observability and controllability

Both observability and controllability are associated with specific system models. The state of the system, and inputs to and outputs from the systems determines if observability and controllability are satisfied. More precisely, if a system is observable, then its state can be perceived from its outputs. By a similar token, a system is controllable, if there are inputs, which take the system to desired states.

Measurability is a stronger condition in comparison with observability. To be able to measure a system state, suitable scales must be applied (cf. Mohaghegh and Mosleh, 2009b) and some measuring device must be applied. In the case that the system only satisfies the observability criterion, a state estimator has to be applied, which is collecting output information over time.

Controlling safety entails that control functions can keep the system states in desired regions of the state space and that controls can be used to prevent movements to unsafe states and/or by some means transfer the system to a safe state. A system model contains information about how to make these direct or indirect transitions from unsafe to safe states.

3.3. The preference relation

Preference relations are needed to value input to and outputs from the system. For example, input controls may put a burden on resources and this has to be weighed against various factors, such as the time a system stays outside its desired state regions. Some control inputs may also be applied for the purpose of providing additional resources.

The preference relation gives a way to calculate values of control inputs along the corresponding path of state transition, i.e. state trajectories, in a time interval from t_0 to t_1 in which the system is moved from an initial state x_0 to an end state x_1 with the control $u(t)$ for $t_0 < t \leqslant t_1$. In control of safety the state trajectory is important, because it should as far as possible be maintained in safe regions of the state space. Other component of the preference relation may be related to the efforts of applying selected inputs to the system.

In the context of controlling safety, one possible strategy is to use a simple preference relation that associates large costs to transitions that take the system outside its desired states. Moreover, due to uncertainties in model predictions and state estimates, it is often necessary to require margins in setting safety targets. Using the time integral of such margins, this could be used as measurement of safety in state transitions. Extending deterministic system models to probabilistic model can be used with application of probability distributions and expected values (Aven, 2011).

4. Control structures in use

Control loops are implemented through control agents, which as inputs read information from the controlled system and apply its own actions as controls to the system (cf. Fig. 1). In the input part of a control loop an initial classification/judgement is made if there is a need for action. Further advice on actions to take will be communicated to the output part of the control loop. The control loops will get their means such as goals, norms and algorithms from the outside and they will rely on certain resources for their function. A control agent can be either in an active or in a passive mode. If it is in a passive mode, it can be activated through a trigger

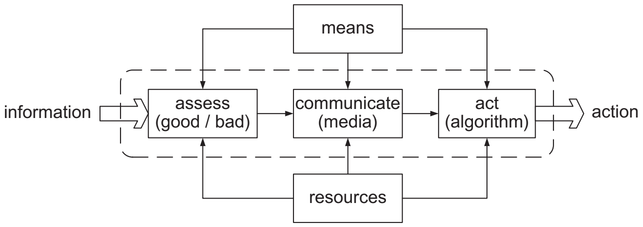

Fig. 1. Functions of a generalized control task that is implemented by a control agent. The agent collects information from the system as its input and calculates the necessary control actions by which new controls are given to the system.

signal, which may be an event of any kind. The control agent may be an automatic system, a single person, an organisational unit or any combination of these. A common flaw in control loops is that the feedback path for some reason has been broken (Marais et al., 2006).

4.1. Open loop control

Open loop control is the most simple of the control structures. Open loop control strategies do not use direct feedback from the system state. In practice, this control structure is used based on earlier experience, i.e. specific situations are handled with simple models comprising of input output combinations. The drawback of this control strategy is that it requires a large set of pre-calculated inputs for situations that may occur. The simplest form of this control is that a certain input produces a predetermined output.

This type of control structure was in the nuclear field applied for the control of emergency situations before the TMI accident. The rationale for this control structure is that emergency situations require accurate and rapid responses, which if calculated in advance, are likely to perform better than ad hoc responses. Instructions for the control of emergency situations are usually built in a top down manner by considering situations that can occur and symptoms that can be observed (Hale and Borys, 2012a). Some feedback from the system state can with this control structure be implemented, using a predetermined set of *if–then* rules.

4.2. Closed loop control

In closed loop control, a continuous feedback is used to determine control actions. Or in other terms; the control is adjusted continuously in small steps depending on system outputs. Of that reason, this type of control makes adjustment to specific situations possible (which is not the case in open loop control). The assumption behind closed loop control is that set points can be determined (i.e. there is a norm that describes target states of the system). By measuring the differences between actual and target values, suitable control actions can be calculated.

As should be apparent from above, closed loop control presents advantages as compared with open loop control. However, there are also disadvantages. For example, if there are large differences in time constants in the controlled system, short time adjustments may influence necessary long term adjustments. Also, the existence on non-linearity in system behaviour may make it necessary to change control parameters to depend on the state of the system.

Most practical control tasks involve systems with multiple inputs and multiple outputs. In such cases interactions between internal system variables often imply that changes in any of the inputs will influence several of the outputs. In this case a control influencing only one of the outputs has to be constructed through a coordinated action using several of the inputs.

4.3. Adaptive control

Adaptive control is a more refined structure. Here, the system model is evaluated continuously based on inputs and outputs. An additional control loop is used to evaluate and synthesize the best controller to be applied. This type of control has a particular advantage in noisy and/or changing environments. This model is related to what is usually referred to as double loop learning (Argyris and Schön, 1978).

Adaptive control is in one way or another used in most control systems. The adaptation may be manual, where control parameters are changed depending on system state. Organisational controls typically include adaptations that are based on a more or less systematic analysis on what to do if defined performance levels are not reached.

4.4. Learning control

Learning control can be used to change a whole control structure and not only local control parameters. But on what ground should one control structure be favoured in place on another? Several strategies are possible; to copy control structures used by others, to use continuous experience from previous own control structures, or to use trial and error in a search to find better control strategies.

Often, various problems associated with a used control strategy initiates search for improvements and innovation. In the safety field, one should however be aware of the fact that radical innovations may be associated with increased risk. This usually motivates a more incremental approach towards controls in safety critical organisations. Kontogiannis (2012) has proposed a classification of typical control flaws in control structures that can be used as an inspiration for discussion and increased awareness associated with control functions.

4.5. Hierarchical control

Control structures are commonly implemented in a hierarchical structure rather than in single loops. Higher level controls are then used to provide control parameters, set points and preference functions (Mesarović et al., 1970). By providing hierarchical control structures, the control design is portioned into independent task for individual subsystems, which are coordinated by control functions higher in the control hierarchy. Hierarchical controls can be perceived as an interconnection of several generalized control tasks (cf. Fig. 1).

A typical place where hierarchical controls are used is when a system has to go through distinctly different operational states. Then one may switch between local control structures based on controls from a higher level of coordinating controls. One other example is if human errors or technical failures bring the system into an emergency mode, which activates safety systems and other specific controls that initiates a transition to a safe region of the state space.

5. The construction of models

Building a model of some phenomenon implies that it gets a targeted focus. A model concentrates on something to be studied and leaves other things out. A given model has a restricted region of validity and care should be exercised not to use it outside that region. In short a model is a simplification with both strengths and weaknesses. System models used for safety should incorporate knowledge from many different fields (Mohaghegh and Mosleh 2009a). In this context we differentiate between four model classes, man, technology, organisation and information (MTOI). When a division between the model and its environment has been made, the focus of further modelling efforts are directed towards the internal behaviour of the model. Internal variables should at least in principle be associated to state variables for the selected model. In the following we go through the four MTOI-systems and make some initial proposals for state variables to be used and how they may interact. Such an exercise will also provide information that can be used in distinguishing between safe and unsafe states.

5.1. The M-system

As a first proposal for a state model of the M-system, we would propose values, motivation and competency as three major components. Competency may be further subdivided into sub-categories, for example specialists and generalists, procedural and declarative competence, explicit and tacit knowledge, etc. Values, competency and motivation will vary between individuals and change with time in response to control actions. There are means by which one can integrate these variables over the members of an organisation; for example, research in safety climate has suggested tools that can be used (Hahn and Murphy, 2008).

If we consider these major state components of the M-system, there are many subcomponents that may be suggested to provide increasing amounts of detail for describing interactions between variables. In an assessment of interactions between the MTOI systems, one may for example assume the following:

- high motivation and well-targeted competencies within organisations can make human errors less likely,
- organisational units that define norms of necessary competencies and evaluate gaps as compared to available competency are likely to reach high and targeted competencies,
- values that prioritise safety are likely to initiate concerns for latent safety deficiencies that may exist in the T-, O- and I-systems and are therefore likely to reduce problems.

Safe states of the M-system could be identified as high motivation and a combination of suitable competencies and values. It may for example be assumed that regular information exchange with people and organisations outside the system are important for maintaining and developing both motivation and competency. Necessary conditions for safety may be suggested and indications for how well they have been reached may then be developed. Unsafe states could in a similar way be described and instruments to generate warning signals could be created.

5.2. The T-system

A proposal as a state variable for the technical system can be obtained by considering physical state variables such as pressures, temperatures, concentrations, levels, and distances. For qualitative models one may for example use rough characterisations of conditions of systems, subsystems and components. Assuming that conditions are known, it should be possible to make some integration over all components to indicate some average condition for specified subsystems.

In the interaction between the other three systems, one may assume that:

- well-functioning interfaces between the M- and T-systems can make human errors less likely,
- if maintaining a good condition of a system is an explicit goal, it is less likely that maintenance debts will be created,
- if the T-system is described accurately in the I-system together with appropriate instructions, it can be assumed that errors in operation and maintenance can be made less likely.

A division between safe and unsafe states of the T-system can typically be obtained from the core chapters of a safety case. Safety technical specifications may for example provide a first approximation of a safe region in the state space of the T-system. Symptoms defined in disturbance and emergency instructions define unsafe regions and control algorithms by which safe states can be reached.

5.3. The O-system

Candidates for state variables of the O-system are, for example, values, goals and structure. Organisational values and goals are sometimes partly conflicting, which entails trade-offs and negotiations for setting priorities between them. Organisational structures are defined in responsibilities and authorities. A separation between the formal and informal organisation may also be introduced to account for the fact that formal prescriptions of a management system are not always followed to the point (Kennedy and Kirwan, 1998). Smartt and Ferreira (2012) propose a framework that seems useful in a more detailed modelling of the O-system.

In the interaction with the other three systems, one may assume the following:

- if the organisational structure is clearly described and well understood, then differences between the formal and informal organisations are less likely to develop,
- if the organisation has pursued completeness, consistency and correctness in the safety analysis report, unexpected situations are made less likely,
- if the organisation has made efforts to verify and validate used models, instructions and documentation, the likelihood of correct actions in disturbances and emergencies is increased.

A large difference between the formal and informal organisation may be considered as an unsafe state of the O-system, which should be possible to detect in audits and reviews. Trust and confidence between the M- and O-systems can be considered to be necessary for safety and deviations should be possible to detect in surveys and interviews.

5.4. The I-system

The state of the I-system is proposed to be constructed by considering what information is stored and how stored information is updated and accessed. Available safety standards can define a minimal set of information to be stored. Databases and documents should be structured to make it easy to access information needed in specific situations. This may for example be achieved through making copies of specific information available in places, where they are expected to be used and by building efficient search algorithms for accessing information that is stored in electronic media.

In the interaction between the other three systems, one may assume the following:

- if the I-system has been well structured, then users within the M-system are supported in their tasks and one may assume that errors would be less likely,
- if the information stored in the I-system is reasonably complete, understandable and used, it can be assumed that change activities will not diminish safety,
- assessments and reviews of the I-system have the possibility to detect problems and initiate improvements for a better performance.

A division between safe and unsafe states of the I-system should at least in principle be possible to make by considering processes, activities and task as defined in the management system in order to check if actors within the M-system can obtain necessary background information and appropriate guidance in situations that may occur. The I-system should contain descriptions of all four subsystems together with information on their present states. Possible safety threats may emerge if important information is not updated when changes in the M-, T- or O-systems are made.

5.5. How the models can be used

The models above can be used in two tasks, (1) to design and implement specific control loops and (2) to assess the functionality of existing control loops. The design and implementation of a new control loop starts with a selection of suitable components to include from any of the subsystems in the MTOI-system. This selection of components gives a first suggestion for a state space, which may go through iterations with the model to obtain a satisfactory state representation. The next step is to assess characteristics of safe and unsafe states. Algorithms for maintaining the system in a safe state and protection algorithms that take the system back to a safe regions after unwanted excursions may then be suggested. When controller inputs and outputs have been selected, observability and controllability can be checked. A final step in the design and implementation effort is to validate the model and the functioning of the control algorithms.

In the assessment of the functionality of an existing control loop, it is necessary to ensure that the model is valid, that the information collected to the input of the loop is correct, that the algorithm is suited to its purpose and that the action part of the control loop is functional. If the control loop relies on external resources, one can check that they are sufficient. Important is also to consider how and when erroneously functioning control loops can be detected. For important control loops one may install specific error detecting and correcting controls.

In this connection it is important to stress that modelling can be done on several hierarchical levels. On the highest level, system overall system performance is the focus and it will depend on many different controls of which the safety management system is one. Safety management could in turn be seen as an activity that is subdivided in the activities of risk assessments, experience feedback and change management. We may therefore be interested in how these activities are controlled separately and together to ensure that the safety management is efficient. Below we consider a simple example to illustrate major steps of and important points in discussing models and their controls.

5.6. An illustrative example

To illustrate the reasoning above, we will use the experience feedback from events that takes place in an organisation. If we assume that a small group of people (ERF) has been formed for this purpose, then a simple description of their task is that they should act on an input stream of internal and external events, analyse them and generate recommendations for system modifications. To perform their task they have access to events and system descriptions, the competency to select and analyse important events and the skills to recommend modifications to be made. We could now be interested in either (1) system performance, where the ERF-group acts as a control loop to initiate modifications or (2) in the performance of the ERF-group, with its own controls, which ensure proper functionality of the group (cf. Table 1).

If we assume that we are interested in the performance of the ERF-group, we should model it as our system to be controlled and ask ourselves what its internal state may be. Because it is a group of people, a starting point could be values, motivation and competency (cf. Section 5.1). Regular audits and managerial reviews are two control loops by which the performance of the ERF-group can be assessed. These controls would likely operate with different input information and have different ways to influence the state of the ERF-group. A typical audit will use the instructions for the ERF-group and check if they are working according to them. Senior managers in a review may have information on the resources the ERF-group has used and the quality of investigation reports they have produced and they may in addition have information from benchmarks from other similar systems to compare with.

A check of the four criteria for successful control would in this case suggest the following:

- System model. Audits take a very simplistic approach to how performance of the group is generated and the observations they make are rather crude. Managerial reviews have better possibilities to assess important state components of the ERF-group, but they may have to rely on additional inputs from self-assessments, surveys, interviews, etc.
- Observability. With the initial proposal of a state for the ERF-group it is very doubtful if it is possible to ensure observability, but suitable simplifications of the model together with observations that senior managers can obtain should help.
- Controllability. If we assume that reports are the main output of the ERF-group controllability should be assured, but if the reports are just filed with no actions taken, the whole intent with the ERF-group is brought to nothing.
- A preference relation. The preference relation for the controls of the ERF-group is difficult to establish, because the performance of the group has to be judged using at least both quality and numbers. Audits and managerial reviews should therefore have some way to classify observations they make and assess the need for modifications in instructions, resources and other conditions that influence the performance of the ERF-group.

6. Structural and mathematical prerequisites

Classical safety science assumes that a certain degree of predictability of system behaviour can be achieved. This predictability has been challenged already in the transfer from modelling only technical systems to include also models of people and organisations. In this section we argue that irresolution about future is inherent in all models we use. We have already argued that models are simplified representations of reality and that they therefore can only give weak directions on how responses of a real system will develop. This modelling accuracy is the major impediment to risk analysis in its present forms, which means that new and complementary approaches should be sought. In the final part of this section we discuss the concept of polycentric control that may make it possible to focus in turn on restricted parts of the MTOI-system when controls are designed and safety is assessed.

6.1. Gödel and Turing

The contributions of Kurt Gödel and Alan Turing to mathematical theory have been immense. We think that their discoveries have an application on safety science too. To explain the theorem of Gödel, one may say that the set of theorems T as generated by a set of axioms A either contains improvable theorems or the set of theorems is contradictory. A simple application for safety is to consider this result in a set of instructions, either there are situations that cannot be handled with the instructions or there are conflicting instructions in the set. This means that we have to accept that any system of instructions will be incomplete, i.e. there are situations, which may require new bottom-up constructions (Hale and Borys, 2012b).

Turing's theorem has to do with a universal machine, which can be used to emulate a programmable computer. A basic problem for such a machine is whether or not the machine will halt for a given program. Turing showed that no program could solve the halting problem. In our interpretation this implies that we cannot a priori predict if a given algorithm will hit a given state or not. Certainly we may run the algorithm and observe what will happen up to a

Table 1
Control loops at different hierarchical levels that are associated with the activity of event investigation.

System	Control loop	Goal	Inputs	Output	Algorithm	Unsafe states
Production system	Safety management	Ensure that agreed safety performance is reached	Own operational experience International operational experience	Recommendations for improvements	Compare key performance indicators with defined targets	Bad safety climate Bad motivation
	Management reviews	Review efficiency of different functions	Operational records	Reallocation of resources	Comparison with targets Interviews With Responsible Person	Under resourced functions
Safety management	Risk analysis	Provide risk estimates, suggest solutions	Event sequences Reliability data Possible remedies	Risk estimates Alternative solutions Recommendations	Assessment based on fault trees and reliability data	Competency deficiencies Bad motivation
	Experience feedback system	Collect and analyse experience, generate recommendations	Event descriptions System descriptions Interviews	Recommendations	Instructions for event analysis	Too little resources Inappropriate tools Bad organisational climate
	Change management	Decide and act on recommendations	Recommendations Risk estimates Alternative solutions	Decision Project initiation and execution Collection of obtained experience	Instructions for change management and project execution	Deficient understanding of system Lack of resources
Experience feedback system	Validation of collected data	Regular review of functions	Important events that have been missed	New competency needed?	Gap analysis	Unreliable data Deficient screening process
	Event calibrations	Separate between important and less important events	Classifications used by others Discussions with internal experts	OK Observations Deviations	Comparisons of reports Comparisons of used resources	Lack of competency Deficient understanding of system
	Self-assessment	Continuously ensure a good performance	Needed competency in analysis Available competency	Good practices Observations Recommendations	Comparison with guidelines	Self-assessments not active Inappropriate tools
Activities within experience feedback system	Assess needs for resources	Ensure the availability of best practices	Rest lists Difficult events analysed	OK Request for additional resources	Gap analysis	Lack of integrity
	Assess efficiency of used tools	Ensure availability of best technology	International searches Connections to colleagues	OK Update Tools	Comparisons between available tools	Lack of competency
	Assess needs for new competency	Close competency gaps	Problems in recent event analyses Contacts with external colleagues	Requests with motivation	Assessment of fields of competency Attrition and retirements	Unsuitable competency Lack of resources

time t_1, but we still cannot say what will happen if the run is continued from there onwards.

These results from mathematical theory place impediments on the commonly seen need to predict how a system will behave in different situations. The only way to predict how a system will behave is to let it loose and observe. Whatever set of control algorithms we develop, there will be situations that the algorithms do not cover. One may suggest that a simulation model could be developed and run with various scenarios to investigate whether or not unsafe states will be reached. We will however show below that also this approach is hampered with its own impediments.

6.2. Fractals and chaos

A common belief is that deterministic systems are predictable, but chaos theory has shown that this is not true. Many nonlinear dynamic systems in physics, engineering, biology and economics show behaviour that is highly sensitive to changes in initial conditions (Mandelbrot and Hudson, 2005). Rapidly advancing research in the field has brought pictures of fractal sets, showing increasingly fine structures when zooming into larger details. The theory points to two impediments that challenge the view that systems could be simulated accurately enough to make predictions of safe or unsafe behaviour. Firstly, if the system under investigation shows chaotic behaviour in some parts of its state space, then small errors in an assessment of its initial state may produce vastly different behaviour. Secondly, any prediction will due to numerical inaccuracies be relevant only for a short time into the future.

Fractals and chaos has connections with complexity theory in which emergence and self-organising behaviour are important areas of study (Waldrop, 1992; Zexian, 2007). Complex systems may include thousands of components; include feedback and feed forward loops; nonlinear interactions between their parts and so on, making any assessment of causes and consequences unreliable. We therefore argue that a far reaching modelling aiming at a detailed simulation at a systems level have too many uncertainties to be viable. Detailed modelling has for most applications to be restricted to a small part of the system, i.e. a *small world*.

6.3. Probability distributions

Probability distributions are important when probability is used to model uncertainty. The validity of risk estimates relies on the assumptions that selected probability distributions are valid models of actual randomness. This is not always the case, which is illustrated by recent examples from economics (Taleb, 2004).

Another complication is that random processes are not necessarily stationary, which means that available observation time may not be enough for obtaining reliable estimates of the probability distributions involved. Even if the source of randomness in a system can be assumed to be white noise, the signal may go through filters and there may be interactions through feedback and feed forward loops that influence probability distributions.

6.4. Autonomous controllers at various levels

The arguments above stress a need to restrict the system model to a "small world", where we can be reasonably confident that simplifications give a restricted but still valid representation of reality. If this could be assured we may discuss how controls for this restricted system should be designed and how the resulting entirety can be assessed. Polycentric control that has been proposed as an alternative to traditional hierarchical control structures (Woods and Branlat, 2012), may provide a solution to problems of modelling and prediction.

Allowing restricted parts of a system to act autonomously with their own preferences and control structures, we can treat them as independent actors in their own right. As compared with the control structure described in Section 4.5, the only difference is that actors are allowed to have their own control objectives, which in turn may imply the need for considering also conflicting goals (Isaacs, 1965; Axelrod, 1984; Sigmund, 2010). If we consider an interconnected net of assuredly well-functioning units we may argue that they, possibly amended with additional coordinating controls, should form a well-functioning entirety.

The difficulty with this kind of loosening of the control structure is that there is no assurance that the loci of control will act for the benefit of the interconnected system. For example, there have been accidents, which have been attributed to a race between automatic and manual controls. Polycentric control as a concept is however attractive, because it prima facie models how real systems are structured. One may argue that present design methods in the control of complex systems actually bring as a result something that may be called polycentric control.

We argue, in spite of the added uncertainty, that this approach is worth pursuing because it opens up a possibility to verify structures and data used for models and controls. This means that it would be possible to validate their behaviour in selected transient situations. According to this approach selected subsystems would be considered in their own right and would be provided with their own protective controls for maintaining their states within allowed boundaries and with safety systems for bringing them back from excursions into unsafe states. Before polycentric controls can be introduced as a design principle for safety critical systems, it is necessary that the concept is thoroughly assessed and amended with new design rules.

7. Management systems

Two earlier constructs, quality systems and organisational handbooks, are today combined into one integrated management system (IAEA, 2006). In addition to performance and process safety, the management systems often also address environmental protection and occupational safety. The management system has two main functions, (1) it describes the O-system in some detail and (2) it defines organisational control loops that are implemented in the control of the system. Organisational structure defines responsibility and authority of organisational units and positions, which can be oriented as work processes, functional units or some combination of them. Feedback of operational experience is an important process, which can signal needs for improved performance and organisational change. Organisational adaptation and change are important mechanisms, which close the loop from feedback of operational experience to improved performance.

7.1. Organisational structure

Organisational structure can be seen as the line of responsibility and authority that defines delegation and reporting from the CEO through organisational units and down to single individuals. It can also be seen as the hierarchical subdivision of processes and functions into organisational units from a general level into an increasing degree of detail. Responsibility has to do with specific controls that are given to individuals and/or organisational units and authority has to do with the controllability criterion for these controls.

Organisational preference relations on the highest hierarchical level are commonly defined in mission, value and vision statements. Goals and objectives on lower levels are further broken down according to the organisational structure and written into procedures and instructions for the organisational units. It has been argued that preferences can be structured in a means-ends hierarchy, where means on a higher level define ends on a lower (Elrod and Hubbard, 1979).

Several control agents may participate in one control loop with tasks of information collection, decision-making, communication and control actions (cf. Fig. 1). Organisational control loops may be targeted at different tasks, where information is collected both from physical sensors and through questionnaires and interviews (Guldenmund, 2007). Control can be exercised both through physical devices and through more abstract entities such as goals, instructions and resource allocations. Models can be focused on organisational units and their control tasks, to assess if they have instructions and resources available to cope with situations that may occur.

7.2. Processes, activities and tasks

Processes are subdivided into interconnected subprocesses, activities, tasks and actions, which are given to organisational units and individuals through instructions at various levels in the organisation. In control terms an action can be considered as a simple open loop control, where a control agent is executing an action when triggered by an event or condition. Processes are often modelled using the structured analysis and design technique (SADT), which also have been applied in the safety field (Hale et al., 1997).

Planning and follow up is one important class of processes that is used at all organisational levels. Depending on the used time frame in the planning, it may be called strategic or operative. A plan can be seen as an open loop control that addresses *strengths*, *weaknesses*, *opportunities* and *threats* (Rumelt, 2011). The follow up part of a plan relies on performance feedback collected to close the quality circle with its components of *plan-do-check-act*. Planning relies on models that are used to link actions to expected outcomes. To support planning and follow up, organisations often define a set of key performance indicators (Kongsvik et al., 2010).

A specific class of tasks are connected to disturbances and emergencies. These could be seen as organisational controls that with the algorithms of specified instruction transfer the system from an unsafe state to a safe region of state space. In this connection it has been argued that there is a benefit to allow that the decision locus changes from what is applied in normal operation (La Porte and Consolini, 1991).

7.3. Functional units

Organisations that operate large systems are often divided into functional units by separating between operation, maintenance and various support functions. Operation is responsible for the 24/7/365 working of the system and maintenance is responsible for corrective and preventive actions to ensure that the system is kept working over extended periods. Support functions may include technical support, human resources, procurement, finances and stockpiles, which are necessary for operation and maintenance in their tasks. Operation and maintenance are in direct contact with the system, which has been said to execute controls at the *sharp end*. The support functions are more distant from the system and is therefore said to execute controls at the *blunt end* (Reason, 1998).

Failures in the controls at the sharp end are often seen immediately, whereas failures at the blunt end often are hidden and may remain so for long time intervals. Control actions in the sharp end have to develop in real time, where similar control actions in the blunt end seldom are time restricted. Instructions for the sharp end should typically be followed to the point, where instructions at the blunt end have a more guiding nature.

Functional units depend on diverse fields of knowledge. An organisation may either have required fields of knowledge available internally or missing knowledge may be bought from external suppliers. If knowledge is maintained internally, it may be necessary to build contacts to external competence centres for the organisational units involved. For competence bought from the outside it is instead necessary to build the competency to act as an intelligent customer. Human resources have an important control task to maintain necessary competency by compensating for attrition and retirements.

7.4. Feedback of operational experience

The process of collecting and acting on feedback of operational experience may be one of the most important safety related processes within a management system. One reason is that newly appointed managers often are change oriented, sometimes with too little concerns for possible downsides of proposed changes (cf. Columbia Accident Investigation Board, 2003). A second reason is that applications of new technologies or new management structures may contain initial design flaws that are detected after the start-up of the system. A third argument is associated with the inherent unpredictability in future behaviour, which was discussed in Sections 6.1–6.3.

The feedback of operational experience relies on several subprocesses, such as information collection, analysis, issuing recommendations, decision making and implementing selected measures. Each of the involved subprocesses may have their own deficiencies, which are preventing its functionality. A general observation from many accidents is that organisations often have been aware of the deficiencies that later in the analysis were identified as root causes to the accident (Cooke and Rohleder, 2006). Such an experience would suggest an in depth assessment of the controls involved in the feedback of operational experience.

A common lesson from incidents and accidents is that prescriptions found in the management systems are not always followed to the point. In some cases this can be blamed on ambiguities in instructions, but it can often be seen as a consequence of organisational culture (Schein, 1992), i.e. norms, attitudes, beliefs, preferences, practices and habits within the organisation. To control organisational culture may however be futile due to difficulties in fulfilling the four requirements for successful control.

7.5. Organisational adaptations and change

Organisational adaptations take place continuously in small steps as a result of controls in the planning and follow-up processes. Organisational changes are larger changes that close the loop from feedback of experience to actual improvements in performance. Organisational change should always be implemented in carefully planned steps, where risks connected to the changes are thoroughly assessed. Irrespective of the size of the change, it is important that they are reflected also as changes in the management system.

The need for changes in control structures may develop as the result of changes in the environment. Such changes may be due to innovations in technology and/or in organisational designs. Technical innovations may for example due to better methods for calculations allow for decreased margins or for scaling up of selected systems. They may also make certain tasks easier to perform, due to better tools and better accessibility of information. Organisational innovations may increase personnel commitment and efficiency through changes in the division of labour and/or in the reward systems. In the case radical changes are necessary, guidance from research in organisational learning may be helpful (Easterby-Smith et al., 2000).

8. Applications to safety management

Safety management may loosely be seen as the organisational controls that are important for safety. According to the principle of a graded approach to safety they should be looked at in more detail as compared to other controls defined in the management system. This would suggest a classification of work processes and functions that are defined in the management system with respect to their importance for safety (IAEA, 2012). Such a classification would also make it easier to assess needs for resources and managerial oversight.

Safety is the main attribute for safety oriented organisations, but one cannot disregard the fact that it has to be placed in relation also to other performance attributes. Cost benefit analysis for example can be of large help in this connection. Safety management may also take very different forms during the life cycle of a system, which may imply that different safety management systems should be built for the phases of design, construction, operation and decommissioning. Activities within safety management should be reviewed at regular intervals to assess their efficiency. Audits, assessments and reviews as well as regulatory oversight can be seen as specific control loops that may need their own arrangements to be efficient. In this section we take a closer look on some implications for activities within safety management and for the construction of safety management systems.

8.1. Costs and benefits

Safety always comes with some costs. If for example system deficiencies are found they could usually be corrected in two or several possible ways, which have different costs for their implementation. It is therefore necessary to weight safety improvements in relation to the costs of their implementation. This comparison of costs and benefits is also seen in societal decisions, where hazardous technologies are considered to be acceptable only if they provide societal benefits that are higher than the costs involved in exposing society with their risks.

Cost benefit analysis more generally can be applied to activities within safety management. The concept of a graded approach to safety can be seen as an implicit cost benefit analysis and we argue that more explicit considerations sometimes can be useful. One may for example compare alternative modification projects that are expected to improve safety. If some quantification of increased safety can be given on a ratio scale, one may argue that a project twice as expensive as compared to another should give at least a twice as large improvement in safety.

This approach in using costs and benefits enter the considerations at several levels within an organisation. A typical process for the allocation of resources is to use a budgeting procedure, where organisational units indicate needs and get allotments. Within their own allotments they have a relative freedom to act on what they perceive as being the best use of available resources. Cost and benefits of alternative allocations of resources can be of large help as a guide for decisions on how resources should be allocated at different hierarchical levels in the organisation.

8.2. Life-cycle considerations

Major life-cycles such as design, construction, operation and decommissioning of a system have very different goals, in spite of the fact that safety is the overarching value. The design phase is concerned with creating a system that can be operated and maintained in a safe way. Construction is involved in ensuring that major requirements for systems, structures and components can be assured for the finalised system. The construction phase has in addition to cope with modifications in the design that are made necessary when earlier design deficiencies are detected. Safe operation and maintainability are typical requirements on design and construction, but operating and maintaining in practice will always generate new ideas for how thing could be improved. Decommission will typically be carried out according to some master plan, but practical difficulties will always emerge when such plans are implemented.

When a new system is planned it would be important to take a life-cycle outlook to identify major risks that have to be managed in different phases. If a problem is identified and resolved on the drawing board it is always cheaper and safer as compared to making modification in a system that already has been built. Careful reviews of outputs of design and construction activities are therefore important activities to assure that safety can be achieved in later lifecycle phases of the system (Falk et al., 2012).

8.3. Activities within safety management

A common question regarding activities within safety management is whether or not they are reasonably complete and efficient. One may argue that a broad and deep search always is likely to find something to improve. The type of deficiencies found would therefore give some bottom up indications of completeness. In a top down consideration completeness is related to the completeness of the original risk analysis that was made for the system. Have all possible situations been assessed? Have the consequences of errors and malfunctions in systems, structures and components been evaluated? A second level of inquiry could then address questions such as: What evidence do we have that important control loops within the MTOI-system are functioning as planned?

To be efficient the safety management activities should have the ability to detect even small deficiencies and correct them. Detected problem can here give some clues to the sensitivity of the activities that are searching for hidden deficiencies. An assessment of system changes made, can in turn give clues to whether or not they have been functional.

8.4. Audits, assessments and reviews

Audits, safety reviews and similar activities exemplify control loops that are aimed for finding and improving unsafe activities and practices. In a typical audit, various norms are used as a benchmark for observations. Deviations and observations provide inputs for possible improvements to be made. In a control theoretical perspective, one of the most crucial components for improving a system based on audits is that actions are taken when deviations are found. However, successful auditing practices, at least in principle, should be based on a system model that explains why some norms exist and why they are perceived as important for safety. In lack of such a model, an audit may be perceived as less useful for the participants.

Organisational reviews should be a part of controls performed by senior management to evaluate the overall performance of the organisation, its processes and functional units. If problems are detected they may result in adaptations in the line organisation, processes and instructions. In addition to their own reviews the senior management may call in peer reviews that are carried out by outsiders, who due to their own experience have a working knowledge of managing similar organisations. Such practices have the benefit of looking at organisational performance with fresh pairs of eyes.

8.5. Regulatory oversight

Regulatory oversight is another example of a control loop where the society's representatives seek to provide evidence that a system is operated in accordance with laws and regulations (Wahlström, 2007). A common regulatory prescription is that organisations should have a learning culture. However, adapting a control strategy that aims for efficient learning, one may argue that regulatory norms are necessary, but not sufficient for developing a learning safety culture; organisations should, aim for transcending mandatory regulations in the meaning of "following the regulations, but do even more for safety".

Usually, an important principle associated with regulatory control is that system operators have the undivided responsibility for safety. Although regulatory strategies vary, it is usual that a regulator gives freedom regarding exactly how specific regulations should be satisfied. In control theoretical terms, the norms may be underspecified with respect to what specific control actions that should be used, provided that the control actions (and the control structure) keep the system into the desired states (usually provided by the norms).

Regulatory oversight is an activity outside direct control of the management of the hazardous system. That fact does not mean that this activity would be unimportant. Instead the regulator may be considered as an important stakeholder with whom safety can be discussed and evaluated. Already the requirement to present arguments and evidence that an acceptable safety has been reached, initiates a second check of their validity. Trust is an important attribute in regulatory communication both with license holders and with the society. Guidance for assessing regulatory effectiveness has been given by IAEA (2002) and OECD/NEA (2011).

9. Dilemmas in control of safety

We have discussed safety management with the intent of suggesting methods and tools for design and assessment of implemented controls. In spite of the general applicability of suggested approaches, there are still a set of remaining dilemmas that have to be solved on a case by case basis. One is connected to limits to what we actually know. Another is connected to the fact that risks associated with hazardous systems often have their centre of gravity in low probability-high cost events. There are also dilemmas connected to the selection of the system to model. The search for safety indicators and the concept of safety culture have some dilemmas involved and agreeing on proper preference relations for the controls requires the resolution of various balances. Decision making in organisations and the society have their own practices, which may not always be in line with agreed safety goals.

9.1. Limits to what we know

Safety science has advanced from relatively simple models of reality to more elaborate constructions for how applications of safety engineering may improve the systems. Early models were transparent and could often be validated through simple experiments. This situation has changed. Models that were developed for technical systems were based on advances in mathematics, physics, chemistry and engineering. With the need also to include human and organisational factors in the models, it proved far more difficult to select suitable state variables and structures of influence within the systems. The result was that used models often were without scientific rigor and they were often proposed more based on beliefs than on empiric evidence.

We can never know if there are additional hidden deficiencies in our systems due to unknown or unexpected interactions between parts in the system. We cannot know if our postulated initiating events give a sufficient coverage of the space of possible events that may occur. We may however argue that many of the unknowns we know about, often are less serious than the challenges that have been considered in the design basis accident or design basis threat (DBA/DBT) scenarios. That will leave us with the unknowns we do not know about. To some extent even scenarios we do not know about may, according to the thinking of resilience engineering (Hollnagel et al., 2006), be acted upon by general measures installed in the control loops.

9.2. What is safe enough

The question what is safe enough was formulated in the late 1960ies (Starr, 1969), but it is still as relevant today as it was that time. Many different approaches have been proposed to address the question, such as for example the quality adjusted life years (QALY). This approach would propose a societal cost benefit process to select between risk reducing actions (Vanem, 2012).

Advances in probabilistic safety analyses (PSA) have brought additional opportunities in comparing risks, but experience has shown that differences in modelling approaches very seldom provide that possibility. On the other hand it is possible to compare risks that are assessed within one set of models and thus to set priorities for actions to reduce risks within that modelling frame (Holmberg and Pulkkinen, 2001).

The basic question of what is safe enough cannot be resolved in a rational decision making process, because different societies have their own values and preferences, which may vary with time. Therefore each society has to approach this question in its own political process. However, the increasing globalisation of the world has brought the need for some harmonization of approaches to risks that have a global dimension.

9.3. Low probability high cost events

Systems may be associated with low probability events, which may give rise to very high costs (Haimes, 2009). Such low probability-high cost events are often associated with uncertainties regarding how much investments should be made to further decrease the risks. A control strategy based on control actions that are able to cope with a large scope of events (even those that are judged as very unlikely) is therefore often recommended – this would, for example, entail that proper margins are built into the controls. Ensuring that a risk analysis is reasonable complete, correct and consistent is one of several strategies used to respond to this dilemma. If risks within a specific domain are assessed and different event chains lead to the same ultimate consequence, quantitative risk estimates and the costs of alternative changes may be used to select between them.

9.4. Selecting the system to model

An important task is to select the system to model. It can be a single component in a system that is maintained by some group of people, it can be an industrial facility with its people and organisation, it can be an international company with controls from a corporate level or it can be a regulatory agency in a country.

The control metaphor, the system model and its state space together with the four necessary conditions for successful control can support both design and assessments of important systems and their controls. The state space of the system model gives an important opportunity to consider characteristics of safe and unsafe regions in the state space. By a proper consideration of the restrictions in this "small world" one may be reasonably confident that the model is a valid representation of reality.

We propose that MTOI-model together with the control metaphor is used as a general frame for modelling safety. We see for example the so-called Swiss cheese model (Reason, 1998) as an instance of our more general MTOI-model. If we use the full spectrum of metaphors, analogies, qualitative and quantitative models with different levels of aggregation, we think it is possible to advance our understanding of risk and safety. The selection of focus for the modelling effort may perhaps then be more an art than a science, but we think it still is useful provided that assumptions are challenged and the process is documented.

9.5. Safety indicators and safety culture

Measuring the state of a system is, in control theoretical terms, one of the important functions to achieve successful control. Safety indicators provide one such measure of system states and have been a continuous discussion at least since the TMI accident (IAEA, 2000; Cooper, 2000). Unfortunately, many indicators are still based on ad hoc models rather than a system model of safety (Øien et al., 2011a). An approach in finding leading safety indicators would be to use validated models and their associated state spaces as a basis. Such a discussion can be supported by considerations of how different variables influence each other (Øien et al., 2011b).

Safety culture, a concept that has escaped a clear definition, was coined in the aftermath of the Chernobyl accident (IAEA, 1991). The concept of safety culture can be seen a reaction against the previous somewhat single-minded idea that safety basically is a technical issue. The concept of safety culture recognizes that development of safety always occurs in an organisational and cultural context. Due to the all overarching nature of the safety culture concept, and in considerations of some recent discussions of the concept (e.g. Haukelid, 2008; Silbey, 2009), it seems difficult to ensure that the four necessary conditions for control can be fulfilled. One reason is that there are few models of safety culture that place the concept in relation to risk – rather it is assumed that safety culture is somehow connected to risk (usually through behaviour). By the same token, indicators of safety culture should ideally be based on a model that describes how various "cultural" factors relate to risk. In our mind all-encompassing models (Davoudian et al., 1994; Mohaghegh and Mosleh, 2009a) have their own merits as structure for thought, but they seem to be too complex to be used for practical applications.

Recent regulatory interventions suggest that a deterioration of safety culture has become a common focus of regulatory concern. To what extent this focus is based on thorough assessments is hard to say. One may argue that normal variability in the man, technical, organisational and information systems are expected to generate deviations (Hollnagel et al., 2006). If these deviations should be seen as objective indicators of a deterioration of safety culture is a different matter (Mohaghegh and Mosleh, 2009b).

9.6. Balances in preferences

The preference relation (cf. Section 3.3) has to do with how values are used to set trade-offs between used resources and the quality of control. The preference relation is usually a function in which a suitable balance of state, resources and time is sought. There are many different balances to be considered in the definition of a preference function. One approach in discussing balances between values has been the competing values framework (Quinn and Rohrbaugh, 1983). This model differentiates between two dimensions *internal–external* and *control–flexibility*, to characterize how organisations focus themselves. Such models can support modelling of how values influence other variables within the O-system (Colley et al., 2013).

Trade-offs also has to be made between the selections of controls that differ in terms of their flexibility. A flexible control system is generally suitable for situations in which it is difficult to specify all situational features of importance. For example, rules and regulations are always underspecified since it is impossible to foresee all situations. Human activities, regarded as a control system, must then be flexible enough to cope with unforeseen situational circumstances. On the other hand, many situations may also be controlled by applications of more strict control so the balance between strict control and flexible control is a difficult issue. Associated to this later issue of balance, we also find questions regarding tradition and change – a learning organisation must be prepared to change, but at the same time maintain what has proven to be functional (Wahlström, 2011).

Hollnagel (2009) has suggested the ETTO principle to be used to find a balance between performing a job with high quality and at the same time uphold efficiency (for example in terms of time). It is also argued that thoroughness and efficiency can form a four-field, in which organisations move with time (Marais and Saleh, 2008). This balance is related to the urge for conservative decision making, which has to do with the precautionary principle (Sandin et al., 2002), particularly in situations where large cost differences exists between erroneously accepting one or the other of two hypotheses.

An important balance in preferences has also to do with how short-term and long-term goals are weighted. Short-term goals may be important to create resources for investments that are necessary for pursuing long-term goals. Economics suggest discounting to be used for comparing costs and benefits that that occur at different times, but the difficulty then is to select the discount rate to be used.

9.7. Decision making in organisations

One dilemma in the control of safety of large systems is that nobody can have a full understanding of all details. According to commonly used organisational principles this is taken care of by having a group of senior managers, where each has a specified area of responsibility. This implies that each of them can bring forward their own concerns, when resources are allocated between competing needs. One may however argue that this is only possible if the members of the senior management group have the understanding necessary to evaluate pros and cons of proposed actions that are put on the table.

The senior management group has in its work to rely on a cadre of experts, who assemble supporting information as a basis for decisions to be made. The first requirement is that the information set collected should be reasonably complete with respect to possible threats and their consequences. A second requirement is that it should truthfully report on conditions and available resources within the system. Each set of information has its own assumptions and preconditions that may be more or less explicit. In practice two competing proposals can very seldom be compared side by side. Decisions will therefore always be influenced by subjective judgement.

A confounding factor for decisions made in senior management groups is the remuneration, which the members get. That means that personal gain may influence the decisions made. Individuals reaching high management positions may also be more disposed towards risk taking than people in general. A possible approach in this connection may be to consider the ethical basis of decisions on risk and safety (Ersdal and Aven, 2008).

9.8. A risk adverse society

An assessment of societal approaches to risk indicates in our mind an increasing concern towards various risks. This may have to do with large-scale accidents that have occurred and the media response they have got. It may also have to do with a sense of outrage that common people have in response to accident investigations in which misconduct and greed has been revealed to have influenced important decisions on safety.

If the society asks for improved approaches in the control of safety, it would imply that better system models and controls are developed. How such efforts should be targeted is a tricky question to answer. Risk analysis and safety engineering, which at one level assess a broad spectrum of threats and on another level go down into details of systems and subsystems, should however be able to respond to this challenge.

Even in that approach it would still be important not to oversell safety of the systems, because there will always be uncertainties that at some occasion may show up as incidents and accidents. This message would therefore be important to communicate to managers, regulators, politicians, media and the general public.

10. Conclusions

In conclusion, the control metaphor is helpful for design and assessment of safety management system. The control metaphor provides a conceptual framework that puts focus on models, input and output variables, state spaces, etc. Especially the discussion of safe and unsafe regions of the state spaces can provide a fruitful approach. Our discussion above puts an emphasis on systems modelling on all hierarchical levels from controls implemented with international agreements down to controls of single safety critical components.

Models are simplifications of reality. Choosing the granularity level of a model regarding safety is a difficult task. Very complex models can be developed for safety, but models must at the same time be usable in practice for building and assessing safety management systems. The modelling structure we have discussed should be perceived as a general framework that can provide a base for more specific models. The concepts of Man, Technology, Organisation and Information offer a framework, which together with the control metaphor can focus on how systems and their controls interact.

We have suggested that polycentric control may be a useful concept in setting up systems for safety management. Polycentric control suggests the inclusion of specialised control agents at different hierarchical levels, which have a task of detecting selected safety threats and responding to them with protective actions. Taken together this would suggest an integrated control structure to be applied, where low-level automatic functions signal concerns, which are responded to with higher-level control loops. The burden of additional uncertainty due to many autonomous control loops, should in principle be offset by the benefit of targeted modelling and validated models.

It has been argued by Perrow (1984), that complexity and tight coupling among components in system present challenges in the control of safety. In the future, we will certainly continue to face accidents that were difficult to predict. However, even so, safety management must of course continue to develop its principles and good practices. Resilience engineering exemplifies one of several new paradigms that hopefully may provide deeper insight in efficient safety management.

Application of a precautionary principle (Sandin et al., 2002) is one of several principles that can guard a system in situations of large uncertainties. For new systems, or in situations when changes are made in old systems, the precautionary principle is central. However, development of new technology must still continue and therefore it is strongly recommended that a combination of proactive safety management should complement the reactive experience feedback strategy to a greater extent than found in many systems of today. This is not a new challenge; many have spoken about the need for a proactive stance. To be proactive, however, models have to be developed that complement existing strategies – models that are rich enough to contain the contributions from factors with a bearing on people, technology, organisation, culture, information, etc., but at the same time practical enough to be a real asset for safety managers. Creating a common language derived from control theory is one of several possible roads for future safety management.

Acknowledgements

The support of Vattenfall AB in enabling this research through interactions with their nuclear and hydro power plants is greatly appreciated. The opinions expressed in the paper are the authors only and do not necessarily reflect any Vattenfall policies or practices.

References

Argyris, C., Schön, D., 1978. Organizational Learning: A Theory of Action Perspective. Addison Wesley, Reading, Mass.
Aven, T., 2011. Some recent definitions and analysis frameworks for risk, vulnerability, and resilience. Risk Analysis 31 (4), 515–522.
Aven, T., Zio, E., 2011. Some considerations on the treatment of uncertainties in risk assessment for practical decision making. Reliability Engineering and System Safety 96, 64–74.
Aven, T., Renn, O., Rosa, E.A., 2011. On the ontological status of the concept of risk. Safety Science 49, 1074–1079.
Axelrod, R., 1984. The Evolution of Cooperation. Basic Books.
Bang Dyhrberg, M., Langaa Jensen, P., 2004. Organizations in context: proposal for a new theoretical approach in prescriptive accident research. Safety Science 42 (10), 961–977.
Colley, S.K., Lincolne, J., Neal, A., 2013. An examination of the relationship amongst profiles of perceived organizational values, safety climate and safety outcomes. Safety Science 51 (1), 69–76.
Columbia Accident Investigation Board, 2003. Report, vol. I, August. http://caib.nasa.gov/news/report/pdf/vol1/full/caib_report_volume1.pdf (29.11.12).
Cooke, D.L., Rohleder, T.R., 2006. Learning from incidents: from normal accidents to high reliability. System Dynamic Review 22 (3), 213–239, Fall.
Cooper, M.D., 2000. Towards a model of safety culture. Safety Science 36 (2), 111–136.
Davoudian, K., Wu, J.-S., Apostolakis, G., 1994. Incorporating organizational factors into risk assessments through the analysis of work processes. Reliability Engineering and System Safety 45, 85–105.
Easterby-Smith, E., Crossan, M., Nicolini, D., 2000. Organizational learning: debates past, present and future. Journal of Management Studies 37 (6), 783–796.
Elrod, E., Hubbard, C.L., 1979. Applying means – ends decision trees. Business, 17–25.
Ersdal, E., Aven, T., 2008. Risk informed decision-making and its ethical basis. Reliability Engineering and System Safety 93, 197–205.
Falk, T., Rollenhagen, C., Wahlström, B., 2012. Challenges in performing technical safety reviews of modifications – a case study. Safety Science 50, 1558–1568.
Grote, G., 2012. Safety management in different high-risk domains – all the same? Safety Science 50 (10), 1983–1992.
Guldenmund, F.W., 2007. The use of questionnaires in safety culture research – an evaluation. Safety Science 45 (6), 723–743.
Hahn, S.E., Murphy, L.R., 2008. A short scale for measuring safety climate. Safety Science 46 (7), 1047–1066.
Haimes, Y.Y., 2009. On the complex definition of risk: a systems-based approach. Risk Analysis 29 (12), 1647–1654.
Haimes, Y.Y., 2012. Modeling complex systems of systems with phantom system models. Systems Engineering 15 (3), 333–346.
Hale, A., Borys, D., 2012a. Working to rule, or working safely? Part 1: a state of the art review. Safety Science, j.ssci.2012.05.011.
Hale, A., Borys, D., 2012b. Working to rule or working safely? Part 2: the management of safety rules and procedures. Safety Science, j.ssci.2012.05.013.
Hale, A.R., Heming, B.H.J., Carthey, J., Kirwan, B., 1997. Modelling of safety management systems. Safety Science 26 (1–2), 121–140.
Haukelid, K., 2008. Theories of (safety) culture revisited – an anthropological approach. Safety Science 46 (3), 413–426.
Hermansson, H., 2012. Defending the conception of "Objective Risk". Risk Analysis 32 (1), 16–24.
Hollnagel, E., 2009. The ETTO Principle: Efficiency–Thoroughness Trade-Off. Ashgate.
Hollnagel, E., Woods, D.D., Leveson, N., 2006. Resilience Engineering: Concepts and Precepts. Ashgate.
Holmberg, J., Pulkkinen, U., 2001. Experience from comparing two PSA-studies, Nordic Safety Research, NKS-36.
IAEA, 1991. Safety Culture, INSAG-4. International Atomic Energy Agency, Vienna.
IAEA, 2000. Operational Safety Performance Indicators for Nuclear Power Plants, TECDOC-1141. International Atomic Energy Agency, Vienna.
IAEA, 2002. Guidelines for IAEA International Regulatory Review Teams (IRRTs), Services Series No. 8, International Atomic Energy Agency, Vienna.
IAEA, 2006. The Management System for Facilities and Activities, GS-R-3. International Atomic Energy Agency, Vienna.
IAEA, 2012. Use of a Graded Approach in the Application of the Safety Requirements for Research Reactors, SSG-22. International Atomic Energy Agency, Vienna.
Isaacs, R., 1965. Differential Games; A Mathematical Theory with Applications to Warfare and Pursuit, Control and Optimization. Dover Publication, Mineola, NY.
Kennedy, R., Kirwan, B., 1998. Development of a hazard and operability-based method for identifying safety management vulnerabilities in high risk systems. Safety Science 30 (3), 249–274.

Kirwan, B., 2011. Incident reduction and risk migration. Safety Science 49 (1), 11–20.

Kjellén, U., 2000. Prevention of Accidents through Experience Feedback. Taylor Francis, London.

Kongsvik, T., Almklov, P., Fernstad, J., 2010. Organisational safety indicators: some conceptual considerations and a supplementary qualitative approach. Safety Science 48, 1402–1411.

Kontogiannis, T., 2012. Modelling patterns of breakdown (or archetypes) of human and organizational processes in accident dynamics. Safety Science 50 (4), 931–944.

La Porte, T.R., Consolini, P.M., 1991. Working in practice but not in theory. Journal of Public Administration Research and Theory 1, 19–47.

Mandelbrot, B.B., Hudson, R.L., 2005. The (Mis) Behaviour of Markets. A Fractal View on Risk, Ruin and Reward. Profile Books, London.

Marais, K.B., Saleh, J.H., 2008. Conceptualizing and communicating organizational risk dynamics in the thoroughness–efficiency space. Reliability Engineering and System Safety 93 (11), 1710–1719.

Marais, K., Salen, J.H., Leveson, N.G., 2006. Archetypes for organizational safety. Safety Science 44, 565–582.

Mesarović, M.D., Macko, D., Takahara, Y., 1970. In: Theory of Hierarchical Multilevel Systems. Academic Press.

Mohaghegh, Z., Mosleh, A., 2009a. Incorporating organizational factors into probabilistic assessment of complex socio-technical systems: principles and theoretical foundations. Safety Science 47, 1139–1158.

Mohaghegh, Z., Mosleh, A., 2009b. Measurement techniques for organizational safety causal models: characterization and suggestions for enhancements. Safety Science 47, 1398–1409.

Möller, N., Hansson, S.O., 2008. Principles of engineering safety: risk and uncertainty reduction. Reliability Engineering and System Safety 93 (6).

OECD/NEA, 2011. Improving nuclear regulation. NEA Regulatory Guidance Booklets, vols. 1–14, NEA/CNRA/R(2011)10. Nuclear Energy Agency. Organisation for Economic Co-operation and Development.

Øien, K., Utne, I.B., Herrera, I.A., 2011a. Building safety indicators: Part 1 – theoretical foundation. Safety Science 49 (2), 148–161.

Øien, K., Utne, I.B., Tinmannsvik, R.K., Massaiu, S., 2011b. Building safety indicators: Part 2 – application, practices and results. Safety Science 49 (2), 162–171.

Perrow, C., 1984. Normal Accidents: Living with High-Risk Technologies. Basic Books, New York.

Quinn, R.E., Rohrbaugh, J., 1983. A spatial model of effectiveness criteria: towards a competing values approach to organizational analysis. Management Science 29, 363–377.

Rasmussen, J., Svedung, I., 2000. Proactive Risk Management in a Dynamic Society. Swedish Rescue Services Agency, Karlstad, Sweden.

Reason, J., 1998. Managing the Risk of Organizational Accidents. Ashgate Publishing Company, Brookfield, VT.

Rochlin, G.I., 1999. Safe operation as a social construct. Ergonomics 42 (11), 1549–1560.

Rollenhagen, C., 2003. Att utreda olycksfall: teori och praktik. Studentlitteratur, Lund.

Rumelt, R., 2011. Good Strategy, Bad Strategy; the Difference and Why it Matters. Profile Books, London.

Sandin, P., Peterson, M., Hansson, S.O., Rudén, C., Juthe, A., 2002. Five charges against the precautionary principle. Journal of Risk Research, 287–299.

Schein, E., 1992. Organizational Culture and Leadership. Jossey Bass, San Francisco.

Sigmund, K., 2010. The Calculus of Selfishness. Princeton University Press, Oxford.

Silbey, S.S., 2009. Taming prometheus: talk about safety and culture. Annual Review of Sociology 35, 341–369.

Smartt, C., Ferreira, S., 2012. Constructing a general framework for systems engineering strategy. Systems Engineering 15 (2), 140–151.

Starr, C., 1969. Social benefits versus technological risks. Science 165, 1232–1238.

Swuste, P., van Gulijk, C., Zwaard, W., 2010. Safety metaphors and theories, a review of the occupational safety literature of the US, UK and The Netherlands, till the first part of the 20th century. Safety Science 48 (8), 1000–1018.

Taleb, N.N., 2004. Fooled by Randomness; The Hidden Role of Chance in Life and the Markets. Penguin Books.

Vanem, E., 2012. Ethics and fundamental principles of risk acceptance criteria. Safety Science 50 (4), 958–967.

Wahlström, B., 2007. Reflections on regulatory oversight of nuclear power plants. International Journal of Nuclear Law 1 (4).

Wahlström, B., 2011. Organisational learning – reflections from the nuclear industry. Safety Science 49, 65–74.

Waldrop, M.M., 1992. Complexity; The Emerging Science at the Edge of Order and Chaos. Touchstone, New York.

Woods, D.D., Branlat, M., 2012. How Human Adaptive Systems Balance Fundamental Trade-Offs: Implications for Polycentric Governance Architectures. <http://www.resilience-engineering-asso.org/ACTES/2011/Papers/5.pdf> (29.11.12).

Zadeh, L., Desoer, C., 1963. Linear Systems Theory. McGraw Hill, New York.

Zexian, Y., 2007. A new approach to studying complex systems. Systems Research and Behavioral Science 24 (4), 403–416.