Notes on the Foundations of System Safety and Risk

Peter B. Ladkin
Technische Fakultät, Universität Bielefeld
ladkin@rvs.uni-bielefeld.de

June 30, 2000

Draft Version: comments and corrections welcomed!

Disclaimer This document consists of parts of what is intended to be a longer work, in various different states of completion and maturity. Readers are kindly requested tolerate the unevenness of presentation.

Contents

Chapter 1

The Foundations of System Analysis

1.1 Preliminaries: The Importance of Reasoning

The Primacy of Reasoning in Prediction Assessing and ensuring the safety of artifacts and procedures is largely a matter of predicting what may happen in the future. If with perfect foresight one knew that an accident was not going to happen, then one knows with perfect foresight that perfect safety is ensured. But since the future has not happened yet, we cannot report on safety purely using observations. We must reason from the current and past situation in the world to a future situation. We must formulate what we know and attempt to use this information to predict the future as best we can. Safety assessment thus involves reasoning. Precise safety assessment involves precise reasoning. Formal logic is the study of precise reasoning; in some sense, rigorous system safety reasoning must be applied formal logic.

Expressive Limitations of Reasoning However, formal logic as practised by logicians and philosophical logicians and as applied by computer scientists and engineers is not a finished science. In fact, there are relatively few "*settled*" parts of formal logic: the propositional calculus, the predicate calculus, certain so-called "*constructive*" formulations of them, certain forms of tense logic, certain logics with modalities, certain forms of higher-order logic. Important parts of engineering reasoning that have no agreed formulation, or which have demonstrated unsuitabilities for the task which we wish they performed are: arithmetic itself, reasoning about parts and wholes, reasoning about obligations, reasoning about causality.

Computational Limitations of Reasoning Apart from these, there is the simple problem of how to handle reasoning. Reasoning may be simple or complex. It may be readily understandable or obscure to all but a chosen few. It can be very hard to construct reasoning that will determine whether a given assertion is canonically true from, or canonically refuted by, or canonically not decided by, certain other assertions. Even if the formulation of reasoning itself was demonstrably adequate, our ability to reason inside that formulation is limited by the complexity of its combinatorics.

Uncertainty We are very uncertain about many important things. Will this joint hold under this constant pressure for the next two years? Well, our reasoning says it should; most of them have in the past; specially constructed tests come out positive; but some of them, very few of them, have still failed under these requirements. We can't be certain the joint will work, but we think it is very, very likely and we're prepared to place a very high bet on the fact that it will.

Formal logic dealing with uncertain reasoning is often called probability logic. It's very hard, it's complex, and it doesn't tell you much about decisions. Formal methods of reasoning about decisions are also difficult. Still, we have to do it.

The Engineering Task Building safety cases for systems, and performing safety assessments of systems, are examples of this very hardest reasoning. Since the systems are being built and in use, we must use whatever imperfect techniques we have. By some measures, we have been very successful with these techniques. Complex commercial aircraft don't crash every day. By other measures, we have not been. Commercial aircraft still crash for easily avoidable and repeated causes. The complexity of systems is increasing enormously, defying our ability to apply the techniques we knew successfully to assess the safety properties of these new systems. Sometimes reliable and safe designs are replaced, degrading either reliability or safety properties or both in the process. And there are new systems, performing new functions. We have to do something.

Practical Application Doing what we did before, when it worked, can be a good guide. Thus, we follow standards and checklists. Not doing what we did before, when it didn't work, can also be a good guide. Thus, we analyse accidents. Making sure we don't make mistakes that we knew how to avoid is always a good idea. Thus, inspections, reviews, and in general checking our reasoning is a good way of avoiding mistakes in design. Inspections and appropriate maintenance is a good way of avoiding the consequences of unwanted change.

Requirements of Effective Reasoning Reasoning about systems is a large component of safety in design. In order to reason effectively, it is customary to have a language in which one can make assertions, with this language being bound in some way to the "*world*" and its states; to "*objective reality*", to notions of "*truth*" and "*falsity*" of assertions in the language evaluated on "*the actual situation*". I don't know any other way of doing it. Successful engineering involves techniques and procedures which many trained people - engineers - can use. So engineering reasoning about systems must nowadays take the form that reasoning about anything has taken in the last few hundred years. Logic, ontology, objects, properties and relations, assertions concerning them and relations of deduction and logical consequence, probability and probabilistic reasoning, the use of a formal language to make unambiguous assertions amongst trained practioners, assessment procedures for truth and falsity or likelihood of assertions, and so on.

In Any Case, Rigor and More Rigor It is often said that to assess and ensure safety, one must think of everything. It may be added that one should think of everything carefully. Situations should be thoroughly checked. Desire for rigor has led to *formalisms*, ritualised ways of speaking and reasoning that are easily reproducible in standard ways, and that show weaknesses and possible problems. Formal languages enable us to enumerate what we can say, and use of a formal language enables us to specify what we see. We can exhaustively enumerate possibilities and check them, using indirect techniques if exhaustive enumeration is too exhausting. We can identify mistakes we may make in expression and reasoning, in principle. We can identify and correct omissions. And we can do all this in a principled way that allows us to avoid repeats.

The Application of Rigor Rigor can be applied in two main ways when dealing with artifacts. It can be applied in reasoning about the artifact itself, and it can be applied to the management and other human behavior in the environment in which the artifact is placed. Both are important; regarded indeed as essential in modern system safety. We shall restrict ourselves here to the first: reasoning about the artifact, its design, its properties and its environment.

1.2 Formal Causal System Analysis

Artifactual systems are built by human beings according to causal principles. Parts are designed in order to have certain influence on other parts: this influence is causal. Analysis of the operation of these systems must therefore be a form of causal analysis. The methods hitherto used in analysis of the safety properties of systems are mostly forms of causal analysis, but without any specific or rigorous notion of what constitutes a cause or causal factor. This means that they are ultimately fouded on intuitive judgement of cause, and this intuition must be built up by consensus and experience in the engineering community. The fact remains, however, that a method based on an undefined and unclarified notion is not truly rigorous. The acid test of objectivity is that the criteria for judgement are explicit and can in principle be used by third parties that are not privy to the socially-learned intuition to check the results of reasoning. Intuition may speed things up, but it should not provide the fundament if alternatives are available.

We illustrate system analysis methods. These methods

- enable causal analyses of artifacts,
- are based on an explicit formal notion of causal factor, which experience has shown can be assessed “in the field” with relative accuracy with a modicum of training,
- are intended for safety analysis of designs: Causal System Analysis (CSA),
- are intended also for causal analysis of accidents: Why-Because Analysis (WBA).

1.3 What is a System?

Examples Things called systems are varied. Sociologists speak of social systems [Luh91]; political scientists speak of system-theoretic influences [Jer97]; other sociologists speak of complex technical systems [Per84, Sag93]; aircraft builders speak of physical systems and subsystems; ecologists of predator-prey systems or environmental cycles of substances; and of course there are computer systems.

What Do They Have In Common? I propose that systems contain *objects* which engage in *behavior*. This behavior, through the objects which constitute a system, may be influenced considerably by the behavior of objects which are not part of the system, through their relations with objects that are in the system. These objects are said to belong to the *environment* of the system. Besides this, there are other objects which have no perceptible influence on and no important relations with system objects, and vice versa. We say that these objects belong to the *world*.

For example, if I consider my bicycle to be a system with all its components (complete with rider), then the environment would include the streets and paths I am riding on and the immediate influences on their state, such as the weather or an overflowing river. Since I am in Germany, the Great Wall of China belongs to the world, not to the system or the environment.

One way of picturing objects divided in this way is as a Venn diagram, such as Figure 1.1, in which points represent objects.

Such a diagram may be misleading, in that it shows world objects and system objects sharing a boundary. "Sharing a boundary with" may be (misleadingly or not) identified with a relation, and system objects by definition only have relations with environment objects. Hence a diagram such as Figure 1.2 more closely visually represents what we are aiming at.

The System Boundary Although the entire universe can be considered as a single system (the collection of objects = everything; relations and properties = all relations and properties), this is not the system mostly considered by engineers. One mostly considers smaller parts of the universe; hence one may make a distinction between those objects that belong to the system and those that do not. This distinction leads to the notion of the *system boundary*, namely the distinction between what objects and behavior are to

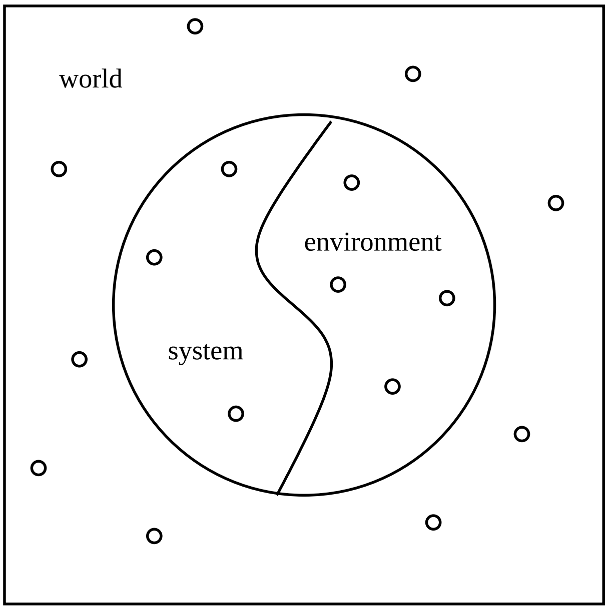


Figure 1.1: The World of Objects

be considered part of the system and which not. The system boundary may correspond to something real, or it may simply be some kind of metaphor. Which is the case will depend very much on what kind of thing the system is.

Teleological Systems and Others Define a *teleological system* to be a system with a purpose or goal. This purpose may be the elicitation of certain behavior, or the attainment of a certain state, of system or environment or both (but not of "world" since the constituents of world are outside the mutual influence of system and surroundings, by definition). Artifacts are typical examples of teleological systems. A car is designed with the purpose in mind to transport people in a particular manner. A computer system is designed with the purpose of performing certain sorts of calculations in a certain manner. Examples of non-teleological systems are environmental systems such as an industrial effluent cycle or a predator-prey system. The international political system is also an example of a non-teleological system, although individual component systems, the governments of countries, are teleological. We shall be dealing mostly with teleological systems: those

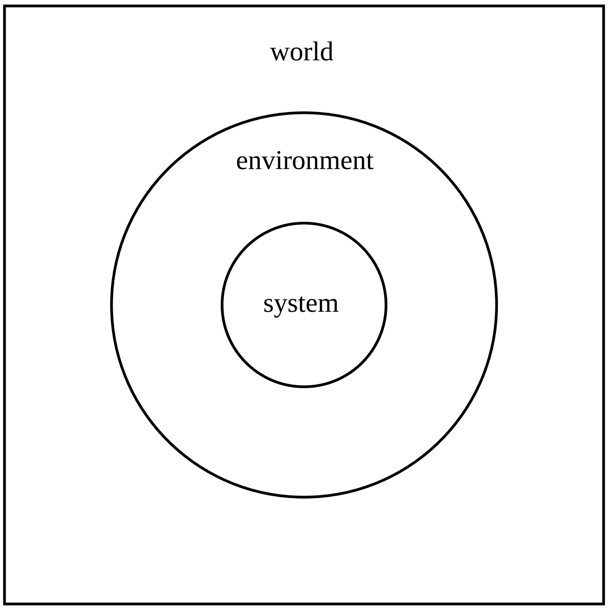


Figure 1.2: The World of Objects As It Should Be

for which goal states or goal behaviors of system and environment can be identified. Constructing teleological systems is the primary goal of engineering.

The Boundary Assumption For Teleological Systems In a teleological system, goals are somehow stipulated. One characteristic method of stipulating the boundary between teleological system and environment is to consider which features of the universum one can more or less control, and which not, and to make the decision on this basis. Call the assumption, that the decision to place the boundary is made more or less consistently with this control criterion, the *boundary assumption*.

Examples of the Boundary For example, the state of a runway surface may be controlled to some degree: one can clear it of debris, or excess water or snow, and direct traffic elsewhere until such time as it has been accomplished. In contrast, one has little or no control over the weather at the airport and hence the dynamic conditions of the air through which landing and departing aircraft fly. Under this criterion, it would be appropriate to consider the runway condition part of the system, and the weather not, and the (expected) behavior of the system varies accordingly. Although one is obliged simply to wait until bad weather changes for safer flying conditions, it would be an inept (or impoverished) airport manager who simply waited for the snow to melt from hisher runway.

1.4 Objects and Fluents

What Are Objects? Objects are, roughly, anything which may be denoted by a noun or noun phrase. More broadly, anything which may be the value of a quantifier [Qui64]. So, you and I are objects, real numbers are objects, the water enclosed inside notional boundaries specified as straight lines between three fixed geographical points is an object. But vanity is doesn't seem to be an object, and neither does humor, nor willpower, and neither is the value of a specific memory location in a computer over time, since this value is constantly something different.

Fluents These quantities can be considered to be objects if one performs certain operations known to logicians and ontologists (people who worry about what objects are, and what objects there are). We thus introduce *fluents*, which are *things which take values over time*. Using the notion of fluents, the number of things we can consider objects can increase considerably. If we think binary numbers are objects, then the value of a memory location over time is a fluent taking binary numbers as values. If we think that the exercise of vanity corresponds to excitation of certain parts of the human brain, then we can consider Fred's vanity to be that function over time which measures this excitation in an appropriate way. And of course now that we have these fluents, we can define further fluents that take these fluents as specific values; and so on iteratively. In short, with common objects and fluents, someone who wishes to talk about systems can indeed talk about as many objects as heshe wishes.

1.5 State, Events and Behavior

Behavior Objects have behavior. We shall consider behavior to be how the properties of an object and its relations with other objects change over time. More generally, behavior is how a collection of objects (including the constitution of the collection itself) changes over time.

Change Consider a complete description of what properties objects have and what relations they have to each other. Such a description is for many reasons impossible to obtain, but let us not worry about that yet. This description could well be true at a certain moment of time. Some time later, this description could well be false. This is what we refer to as change over time.

State We shall say that such a complete description of objects with their properties and relations to each other at a moment of time is a *state description*, and the actual configurations of objects, properties and relations it describes is called a *state*. Figure 1.3 shows an example of a system state. The fluent x takes the value 2; the object *Valve1* has the property *Open* (we take this to mean that *Valve1* is *open*); the quantity of reactant (we take *Quantity*, denoted in subsequent figures also as Q, to be the function which gives as value the quantity of its argument; its argument is *reactant*) is 100 units.

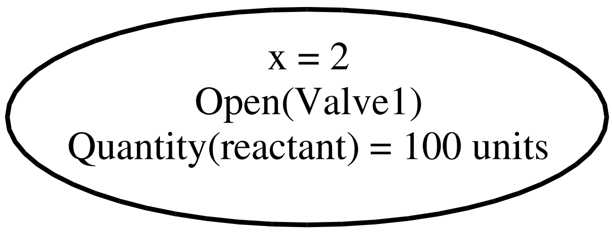


Figure 1.3: A System State

One limitation: it is important for technical reasons that the objects, properties and relations noted in a state contain no explicit reference to time and change themselves.

Justification of This Notion of State One may well ask, why this apparently somewhat restrictive notion of state? Why cannot state be anything at all? The answer is, that it is not as restrictive as it may at first seem, for the following reasons.

- One cannot know the future, hence if one wants the state to include determinate predications, one cannot include predictions of the future in the current state. This rules out including future temporal properties in a description of state;

- Temporal properties in the past can be included by the simple device of "history variables": one introduces a fluent which retains all the information about the past that needs to be retained by the system, or needs to be known by someone reasoning about the system. An "audit trail", if you like. The use of history variables in formal system description techniques is ubiquitous.

- Set theory formulated in first-order logic (also called "first-order set theory", or "Zermelo-Frankel set theory" after two of its founders) suffices for describing the component structure of systems, as well as all the mathematics one needs to describe the discrete or analog behavior of the system. Some mathematicians and some computer scientists do not like first-order set theory for various reasons. To accomodate the wish to avoid set theory, one can instead provide equivalent descriptions in *higher-order logical type theory*. There are very few properties known, if any, that cannot be accomodated in one or the other of these formal languages.

Comparing States Given two states (or two state descriptions), we may compare them to see what is the same and what is different. We shall consider *change* simply to be what is different, and a description of change to be a specification of the differences (sometimes it will also be important to specify what is the same in the comparison, sometimes not). Figure 1.4 shows an example of change as a comparison between two states. *Valve1* is not open in the first state, and in the second it is open. The other predicates have not changed: The fluent x still has the value 2; the quantity of reactant (denoted by the predicate Q with the argument *reactant*) is still 100 units.

Events and Event Types The change illustrated in Figure 1.4 does not describe much about either state. If I were part of the system of which the objects mentioned in the states are also part of, it doesn't say what clothes I'm wearing or where I am or what time it is. It potentially describes many individual system changes, namely all those in which the specific objects

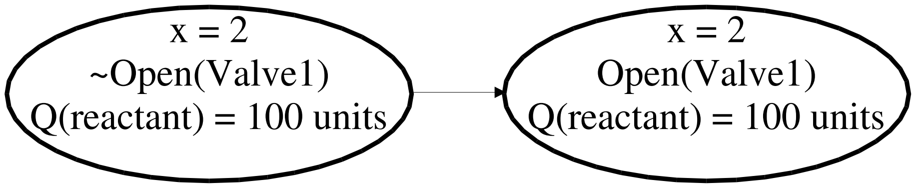


Figure 1.4: A State Change

described change in the way described. Let us call an individual change, which occurs at a specific time, an *event*. Then the state change in Figure 1.4 describes many events, namely all those in which the objects change as specified. We say it describes an *event type*. An event belongs to this type just in case it is an event in which x remains 2, *Q(reactant)* remains at 100 units, and *Valve1* changes from open to closed.

Deriving Behavior Descriptions from State Comparisons Since we may compare two states to see what has changed, we may compare three, one after the other, to obtain a view of progressive change. Or four, or five, or a hundred. We may consider behavior to be a sequence of states such as this, specifying a series of changes. It is important to note that this sequence is *discrete*, that is, each state in it has a definite predecessor and a definite successor. Since we may need to chain together lots of these state comparisons, if we require very great detail or if the sequence goes on for a very long time, we consider that a state sequence may have a huge number or even an infinite number of member states. Thus we shall consider a *behavior* to be *an unending discrete sequence of states*. Figure 1.5 shows such a discrete sequence (at least, the first four states of one).

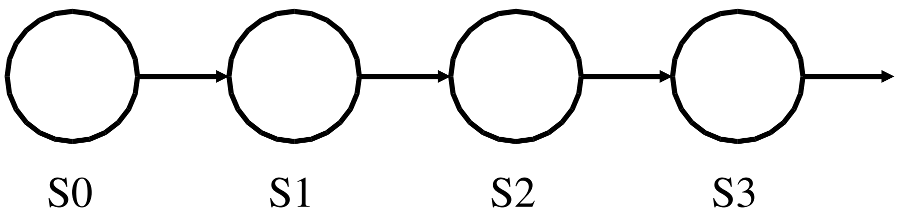


Figure 1.5: A Behavior

Another limitation: for technical reasons, we shall consider that a state sequence that forms a behavior shall always have a *first state*, that is, one that occurs before all others; that itself has no predecessor.

Near And Far State Changes We shall need to distinguish small from large state changes. We shall call small changes *near changes* and large changes *far changes*. These notions are intended to have their intuitive meanings. For example, Figure 1.6 shows a near change, in which the value of the fluent x changes from 2 to 3 and nothing else changes.

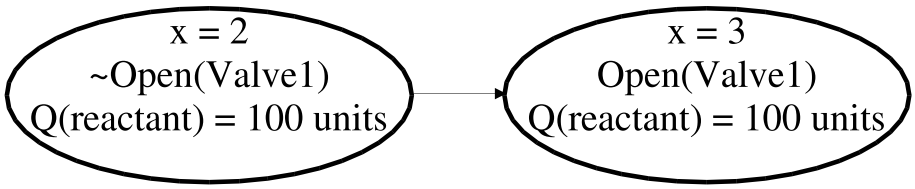


Figure 1.6: A Near Change

Figure 1.7 shows a far change, in which the value of the fluent x changes considerably to 54, and at the same time the quantity of reactant increases threefold.

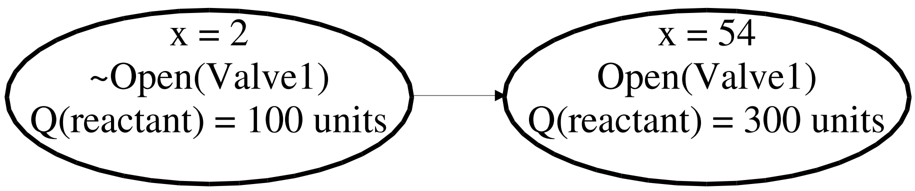


Figure 1.7: A Far Change

Near And Far Behaviors We shall also need the notion of near and far behaviors, which is a comparative notion. This notion is a comparison between three behaviors: a behavior B is *nearer to* a behavior A than a behavior C is to A. For example, the behavior in Figure 1.9 is nearer to the reference behavior in Figure 1.8 than is the behavior in Figure 1.10.

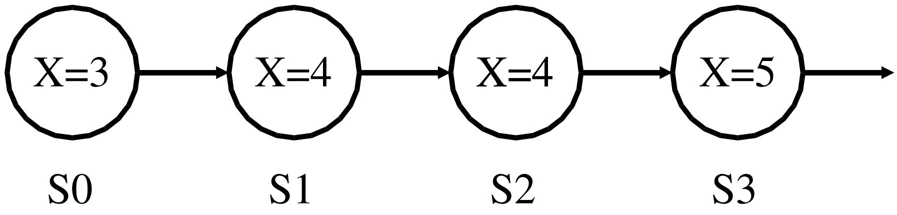


Figure 1.8: A Reference Behavior

The only difference between the reference behavior in Figure 1.8 and its near behavior in Figure 1.9 is that, in state $S1$ of the near behavior, the value of the fluent x is 5 rather than 4.

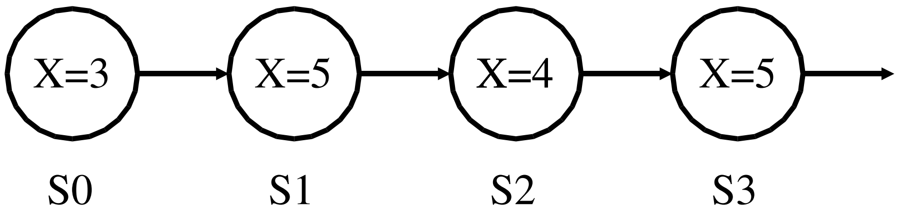


Figure 1.9: A Near Behavior to the Reference

In contrast, the values of x in states $S1$ and $S3$ of the far behavior are very different from the corresponding values of x in those states in the reference behavior, and the value of x in state $S2$ is also somewhat different.

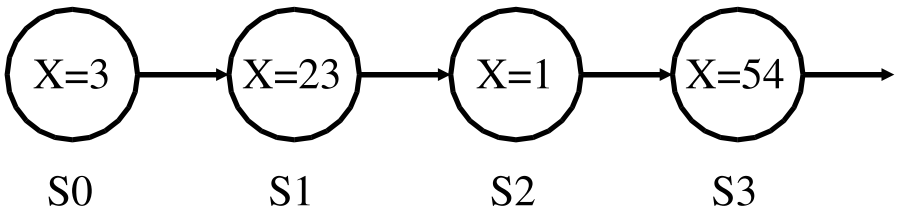


Figure 1.10: A Far Behavior to the Reference

The "Space" of Behaviors Suppose we were to represent behaviors as points in a Venn diagram. Then we could use the distance in the diagram to

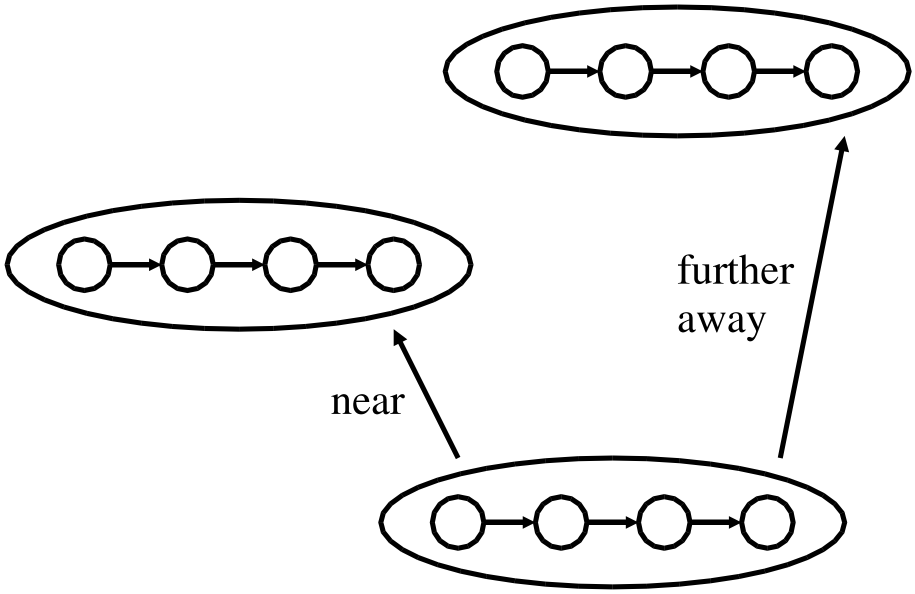


Figure 1.11: Nearer and Further-Away Behaviors

represent visually the nearness, respectively, farness, of the behaviors from each other, as in Figure 1.11. If we consider the "space" of all possible behaviors as a Venn diagram, and supposing we had a way of measuring nearness and farness on an *ordinal scale* [KLST71] (see Section 6.2 for an enumeration of the explicit properties meant), we could represent nearness and farness of all possible behaviors relative to a given behavior, the "*real world*", as in Figure 1.12.

The Purpose of The Definitions The point of this ontology is that

- there are provably complete forms of formal reasoning about these structures; and
- one can describe any "real world" situation adequately using these structures

Since we can describe any situation we may care about, and we can reason accurately and formally about that description, this ontology lends itself to rigorous reasoning about systems, which safety analysis requires.

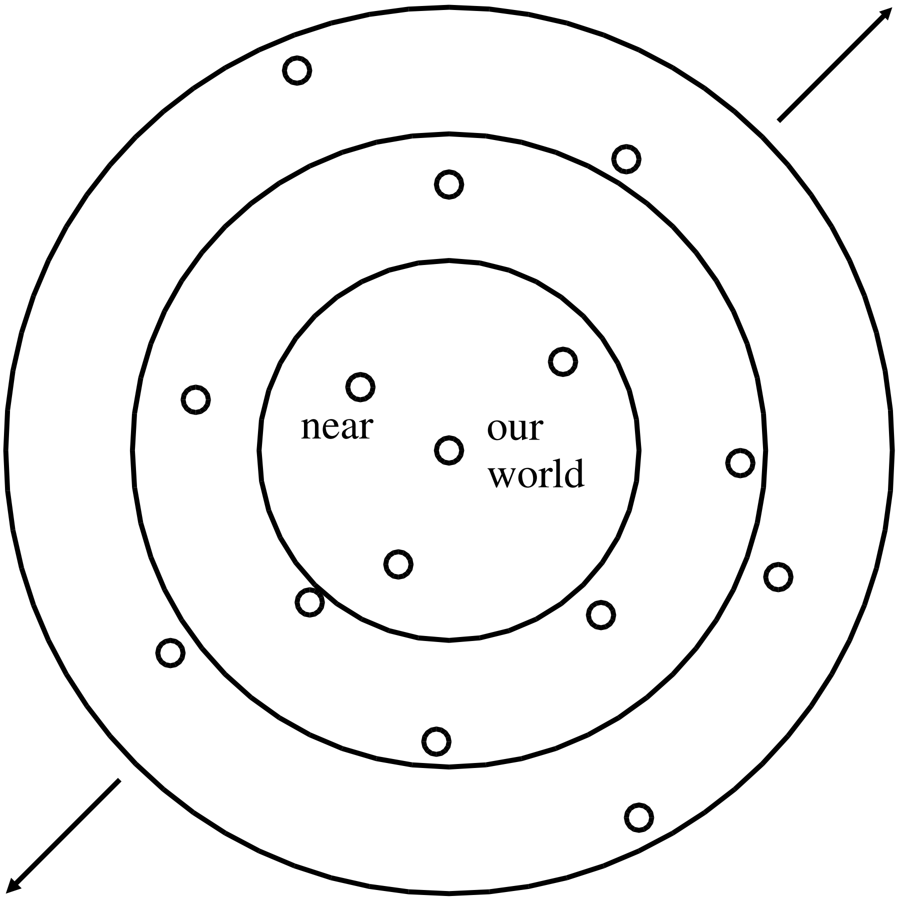


Figure 1.12: All Behaviors, Arranged in "Nearness" Circles

1.6 Objects, Parts and Failure Reasoning

Objects with Parts Consider a large chunk of computer code. Say, a program of a few thousand lines. This code *contains* procedures, and instructions. The procedures are *part of* the program; the instructions are *part of* the procedures.

Structural Parts The instructions are part of the procedures and the procedures are part of the program. But the progam, when looked at another way, is just a very long string of alphabetical symbols. Any contiguous substring is also part of the program, even though it may start in the middle of an instruction and end in the middle of another one. Probably any collection of contiguous substrings is part of the program also. We want to say that whole instructions and whole procedures are the meaningful parts of the program as far as the operation of the program is concerned, and that individual characters and strings of characters are not, except insofar as they constitute instructions and procedures. Compilers tacitly perform the distinction by having a preprocessor, a lexical analyser, which groups individual characters into *tokens*, which are regarded as the minimal meaningful objects as far as the operation of the program is concerned. We distinguish the cases by saying that instructions and procedural parts and the tokens identified by the lexical analyser (presuming it is correct) are *structural parts* of the program as far as its operation is concerned. Characters will be structural parts of the program if we are considering maybe its storage requirements, or if comparing it with the work of six monkeys sitting at typewriters.

Mereology and Fusion There is a logical science, *mereology*, concerning which parts of objects exist. One widely-accepted mereological operation is that of *fusion*, whereby from objects X and Y is formed the object $X \oplus Y$, the 'mereological sum', which has X and Y as parts, and such that any object with X and Y as parts has also $X \oplus Y$ as a part: the 'smallest' object one can make from X and Y in other words.

Parts and Failure If a system fails to perform its function, a subdivision into parts is often used in order to identify a part that failed to fulfil its function. Take a common example:

"This computational system failed. Its hardware didn't fail.
But the system is composed of hardware and software.
Therefore the software must have failed."

The surface logical form of the argument as presented is shown in Figure 1.13. Unfortunately, although we might want the conclusion to follow from the premises in this particular inference, the conclusion is false. There are examples in which the system failed, the hardware didn't fail, and the software did not fail to fulfil its designed function either. Ariane Flight 501 is an example. The situation, as in most failures of mission-critical or safety-critical systems, is that there was a misfit between the requirements for the design of the system, and the environment in which the system actually op-

$$\frac{\begin{array}{l} \mathit{Failed}(S) \\ \neg \mathit{Failed}(\mathit{hardware}(S)) \\ S = \mathit{hardware}(S) \oplus \mathit{software}(S) \end{array}}{\mathit{Failed}(\mathit{software}(S))}$$

Figure 1.13: Correct Failure Reasoning With A False Premiss

$$\frac{\begin{array}{l} \mathit{Failed}(S) \\ \neg \mathit{Failed}(\mathit{hardware}(S)) \\ S = \mathit{hardware}(S) \oplus \mathit{software}(S) \oplus \mathit{ReqSpec}(S) \end{array}}{\mathit{Either}\ \mathit{Failed}(\mathit{software}(S))\ \mathit{or}\ \mathit{Failed}(\mathit{ReqSpec}(S))}$$

(The term *ReqSpec* is used to denote the requirements specification)

Figure 1.14: Corrected Failure Reasoning

erated. A subroutine in the navigation hardware for the Ariane 5 had been reused from the Ariane 4. It needed to operate within certain bounds of its variables, which had been shown for the Ariane 4 not to be capable of overflowing during the flight environment. However, the initial trajectory of the Ariane 5 was different, and checks had not been made to see if the design assumptions for the Ariane 4 navigation routines were still valid for the Ariane 5. They weren't. A variable overflowed, causing a series of events which ended in loss of control and destruction of the vehicle.

If the first two premisses are true, and the conclusion is false, then either the third premiss is false or the reasoning is invalid. We may see from the Ariane example (and others that I haven't quoted) that the reasoning in Figure 1.14 is more appropriate.

The Role of Fusion in Failure Reasoning In the reasoning in Figure 1.13, the role of fusion is clearly indicated in the premiss:

But the system is composed of hardware and software.

The conclusion, that the software failed because the hardware didn't fail, was mistaken.

$$\frac{\begin{array}{l} \mathit{Faulty}(S) \\ S = \mathit{Part}_1(S) \oplus \mathit{Part}_2(S) \oplus \mathit{Part}_3(S) \\ \neg \mathit{Faulty}(\mathit{Part}_1(S)) \\ \neg \mathit{Faulty}(\mathit{Part}_2(S)) \end{array}}{\mathit{Faulty}(\mathit{Part}_3(S))}$$

Figure 1.15: Correct Failure Reasoning

In the otherwise similar premises in the reasoning in Figure 1.13, the role of fusion is indicated in the premiss:

But the system is composed of hardware and software and its requirements specification.

In contrast to the reasoning in Figure 1.13, the conclusion in Figure 1.14, that the software or the requirements failed because the hardware didn't fail, is correct.

The difference between the two cases is the premiss involving fusion, and the incorrectness, respectively correctness, of the conclusion. I conclude that getting the fusion premiss right is an important component of correct reasoning about failure.

It seems that we should like to be able to reason as in Figure 1.15. A little thought shows that this kind of reasoning goes into many failure analysis procedures (software people call this 'debugging'). However, the $software(S) \oplus hardware(S)$ example above shows that one must be very cautious in asserting that all the parts one thinks one has are all the significant parts of a system.

Is "Documentation" Part of the System? The difference between the reasoning with the false conclusion and that with the true conclusion is the constituents of the fusion in the premisses. In the Ariane example, the failure was actually in the specification of and determination of compliance with requirements. But our solution, including requirements in the fusion, entails the somewhat counterintuitive idea that the requirements specification, which includes or should include the limitations under which the system was designed to function, is actually part of the system itself. That is,

$$S = \mathit{software}_1(S) \oplus \mathit{hardware}_2(S) \oplus \mathit{requirements}_3(S)$$

It may indeed seem strange to include a *specification*, which is a piece of text, in with the physical components of a system. But it is not unprecedented: the system code is considered to be part of the system, and code is text too. How is a requirements specification different from, say, the system code? The system code can be considered a specification also; the system shall behave according to this-and-this instruction.

Adequate Decompositions One could define a decomposition of a system into parts as an *adequate decomposition*, when

(a) the system is the fusion of the proposed parts, and

(b) if the system fails, then one of the parts has failed also.

The purpose of an adequate decomposition is to enable reasoning about failure as in Figure 1.15. The example discussed, and others, show that most common engineering decompositions of systems into parts are not adequate decompositions.

System Accidents and the DEPOSE decomposition The accident sociologist Charles Perrow has argued in [Per84] that "*interactively complex*" systems which are "*tightly coupled*" suffer from a propensity to "*system accidents*", which are accidents caused by the system which cannot be put down to failures or misbehaviors of any of the parts.

If these accidents are considered to be failures of the system, then, according to the above definition, Perrow would be arguing that "interactively complex" and "tightly coupled" systems cannot have an adequate decomposition. There is, of course, no proof of this assertion, and it would be hard to see how there could be.

Instead, Perrow could be taken to be arguing that a humanly possible decomposition of an interactively complex and tightly coupled system is unlikely to be adequate.

He has himself proposed a scheme DEPOSE for classifying complex systems into types of components. DEPOSE stands for

- Design

- Equipment
- Procedures
- Operators
- Supplies and Materials
- Environment

While Perrow's classification emphasises essential features, such as the design of procedures and the training and behavior of human operators of the system, which have not traditionally been examined with the same care as the physical components, he does not provide any argument from which it can be concluded either that

- For any complex system T, DEPOSE provides an adequate decomposition, that is,

  $$T = D_T \oplus E_T \oplus P_T \oplus O_T \oplus S_T \oplus E_T$$

  or that

- DEPOSE enables a more thorough categorisation of failure categories than traditional investigative techniques special to each industry.

Nevertheless, Perrow's work has inspired significant contributions to the study of human error possibilities and procedure design in complex system engineering.

Common Decompositions into Component Types Adequate decompositions may exist for certain classes of systems. A discussion of some common or useful classifications of complex systems into component types may be found in [Lad99].

Chapter 2

Definitions for System Safety Analysis

2.1 Reliability and Safety

Reliability and Failure We have talked about failure, and inferring from failure of a system to failure of parts. But the failure of a system to fulfil its function, and the success of a system in filling its function, are not directly related to safety. If we install our LAN server in a fireproof room, and there are no essential functions of the company which depend on the computer functioning, then whether my LAN server fulfils its function most of the time or hardly at all is not a safety matter. *Reliability* is the property of a system whereby it fulfils its function. A firearm may reliably fire when the trigger is pulled; but if it's loaded and a child is playing with it, and there is no safety catch, it may reliably fire and kill someone.

Safety and Accidents The property of a system whereby it does not produce or encourage accidents is known as *safety.* An *accident* is taken to be any undesired or unwanted (but not necessarily unexpected) behavior. Definitions are taken from [Lev95]. This means that an accident can be almost anything you want it to be. Usually, we are concerned whether the operation of a system will kill or injure humans or other animals, but little in safety engineering techniques actually depends on whether this particular unwanted behavior is what one is considering to be an accident.

Reliability and Safety are Related However, situations such as just mentioned can be moderated by the introduction of safety mechanisms. For example, a trigger lock, which prevents the firearm being fired by anyone other than the keyholder. In order for the device to continue to function safely in these circumstances, the safety mechanism must be reliable. This is the most frequent connection between safety and reliability: safety is assured through the reliable operation of certain mechanisms.

Safety Mechanisms Safety is, roughly speaking, the *absence* of certain kinds of problems. Often, this absence is assured, or we attempt to assure it, through the *presence* of specific mechanisms, which are intended to inhibit rare but possible unsafe system behaviors. These systems must function reliably in order to ensure safety. But they are hardly ever used; just on the rare occasions when there would be a safety problem which triggers their operation. It is notoriously hard to ensure the reliable operation of a mechanism which is rarely used. Ensuring the reliability of safety mechanisms is often a much harder engineering problem than redesigning a system to avoid the potential safety problem without the use of specific mechanisms.

2.2 Definitions of Safety Concepts

Terminology Leveson notes that terminology in system safety has not always been used consistently [Lev95, p171]. She gives a series of definitions of such terms as *reliability*, *failure*, *error*, *accident*, *incident*, *hazard*, *risk* and *safety* [Lev95, Chapter 9: Terminology], which attempts to do the most justice to the engineering definitions, and is the result of considerable research into the engineering literature over a number of years. These definitions indeed seem to be amongst the most precise in the literature.

Reliability Leveson defines [Lev95, p172]:

> **Reliability** is the probability that a piece of equipment or component [of a system] will perform its intended function satisfactorily for a prescribed time and under stipulated environmental conditions.

Failure [Lev95, p172]:

> **Failure** is the nonperformance or inability of the system or component to perform its intended function for a specified time under specified environmental conditions.

Accidents and Safety [Lev95, pp172,181]:

> An **accident** is an undesired and unplanned (but not necessarily unexpected) event that results in (at least) a specified level of loss. [....] **Safety** is freedom from accidents or losses.

In order to use these definitions, one has to specify what one considers to be losses (and their levels). Such losses are often specified as numbers of deaths or injuries, financial losses to concerned parties, damage to the natural environment, and so forth. Typically, there is considerable agreement on what is to be considered a 'loss' (for example, deaths, injuries, money, damage), and how the levels are measured (mostly by numbers; more generally on ordinal or ratio scales [KLST71]). Leveson notes that this is stipulatory: it is up to us to specify what we consider a loss and what levels constitute an accident.

Accidents and the System Boundary There is nothing in the definition of accident concerning the system boundary; we may presume that many accidents involving both system and environment occur. Examples could be: the airplane crumples and dismembers, because the mountain rose through the cloud to smite it. When dealing with teleological systems, we may be presumed to be able to exercise more control over the constitution and behavior of the system than we may over the environment. We shall see that, depending on the openness of the system and various other factors, accidents may depend more or less on the interaction of the system with its environment.

System Contributions to an Accident The aircraft can be engineered to predict the looming presence of the mountain and fly above it; it is considerably harder to move the mountain out of the way of the encounter. Accordingly, we shall wish to speak about the part of the system that contributes to an accident, even though given favorable environmental conditions the accident will not occur: if the aircraft flies at or above a (true) 30,000ft (above mean sea level, MSL) altitude, there will be no mountain for it to encounter; if it flies through the Himalayas below 28,000ft MSL, there are some places it cannot fly without meeting an obstacle. Accordingly, we can distinguish airspace including an altitude of less than 28,000ft MSL over the Himalayas as hazardous, potentially leading to an controlled-flight-into-terrain (CFIT) accident, and other airspace as non-hazardous. The property of being hazardous or not has thereby been ascribed to the airspace, that is, part of the environment. However, there is a corresponding pair of properties of the aircraft, namely *being in/out of hazardous airspace*. One may wonder after considering this example whether hazards can be always be described either through environmental properties or through system properties, as desired. If so, there are reasons to classify system states and not environment states as hazards, namely that one brings them into the domain in which control and redesign can be exercise if necessary. But we shall see later that system and environmental hazard states are not always dual in this manner.

Hazard, Severity, and Risk The following definitions are said to be standard in U.S. System Safety engineering [Lev95, pp177-9]:

> *A **hazard** is a state or set of conditions of a system (or an object) that, together with other conditions in the environment of the system (or object), will lead inevitably to an accident (loss event). [....] A hazard is defined with respect to the environment of the system or component. [....] What constitutes a hazard depends upon where the boundaries of the system are drawn. [....] A hazard has two important characteristics: (1) **severity** (sometimes called **damage**) and (2) **likelihood** of occurrence. Hazard **severity** is defined as the worst possible accident that could result from the hazard given the environment in its most unfavorable state. [....] The combination of severity and likelihood of occurrence is often called the **hazard level**. [....] **Risk** is the hazard level combined with (1) the likelihood of the hazard leading to an accident (sometimes called **danger**) and (2) hazard exposure or duration (sometimes called **latency**).*

So a hazard, flying under 28,000ft MSL, in combination with other conditions in the environment (doing so in a particular direction in a particular geographical location, so that impact cannot be avoided) will inevitably lead to an accident (loss of airplane and death or injury of occupants) that may be more or less severe, depending on how many people on board there are, how expensive the aircraft is, what environmental damage is sustained, and so on. We shall later call this notion of hazard *Hazard-1*, to distinguish it from three other useful formulations of the concept.

The Concept of Hazard Partitions States It is important to note that this concept of hazard divides states of the system into two classes, consisting respectively of those states in which the aircraft is flying at an altitude greater than that of the obstructions in the vicinity; and of those in which the aircraft is flying at or below that altitude. The first category of states will not (because they cannot) lead to a CFIT accident, and states in the second category allow the potential for that kind of accident. Accordingly, the states in the second category are hazard states for CFIT, and those in the first category are not.

To take another example: an aircraft flying through cloud with the potential for embedded thunderstorms actually encounters one. The hazard consists in flying through cloud with embedded thunderstorms (rather than flying clear of such weather); the severity is loss of the aircraft and occupants; the 'most unfavorable state' of the environment is a thunderstorm of sufficient power to upset the aircraft and cause breakup under aerodynamic loads; the danger is how likely one is to fly through such a thunderstorm while flying through the stormclouds; and the duration is the length of time one flies through the stormclouds. One could presumably measure the relevant probabilities (likelihood and danger) by measuring the spatial distribution of thunderstorms in stormclouds of the given type, and the frequency of severe ones. All well and good. But do these concepts work generally?

Chapter 3

Problems Calculating Risk Via Hazard

We construct an example to show that the technique of calculating risk through hazard, as in the definitions in Section 2.2, does not give the intuitively correct answer, which in this case is separately calculable.

3.1 Five Notions of Hazard

3.1.1 The System Safety and Associated Notions

The "System Safety" Definition: Hazard-1 We denote by Hazard-1 the notion of hazard defined by [Lev95] and given in Section 2.2:

> A **Hazard-1** is a state of a system that, together with other conditions in the environment of the system, leads inevitably to an accident (loss event).

The Complementary Definition: Hazard-2 We have seen that it may make sense to have a term for a dangerous state of the environment that a system would like to avoid (an airplane avoiding thunderstorms, or mountains, or areas of dense traffic). Let us therefore define:

> A **Hazard-2** is a state of the environment of a system that, together with a particular reachable state or states of the system, leads inevitably to an accident (loss event).

The "Increased Likelihood" Definition: Hazard-3 Some safety engineers prefer to use a notion of hazard in which a hazard state is a system state in which there is a considerably increased likelihood of an accident happening. Accordingly, we define:

> A **Hazard-3** is a state of a system in which the likelihood of an accident is increased over the likelihood of an accident in precursor states.

The Enlarged System Safety Definition: Hazard-4 We have noted that sometimes one can use Hazard-1 effectively, and sometimes Hazard-2. It makes sense to consider whether one should define a hazard through a joint state of system and environment. We define

> A **Hazard-4** is a state of a system together with its environment that, together with other developments in the environment of the system, would lead inevitably to an accident (loss event).

3.1.2 The MIL-STD-882 Definition: Hazard-5

The MIL-STD-882 Definition of Hazard The MIL-STD-882 definition of hazard is *a condition that is prerequisite to a mishap* (wherein 'mishap' is essentially the same as 'accident' as we have considered it).

The State Predicate is Not Restricted This is a different notion of hazard to those three we have considered previously. First, observe that by "condition" is meant part of the state. Second, the state predicate is not restricted to be

- part of the system state (Hazard-1);
- part of the environment state (Hazard-2).

It is thus appropriate to consider any state predicate, which may contain elements of system state *and* environment state.

"Inevitability" Is Not Predicated Furthermore, it does not contain within it the predicate of inevitability. If a condition is prerequisite to an accident, this means that the condition is *necessary* for an accident to occur. If an accident is inevitable, given the condition, this means that the condition is *sufficient* for an accident to occur. The system safety definition therefore requires the condition be necessary; the MIL-STD-882 definition that it be sufficient. These two criteria are very different!

This Distinguishes This Concept From That Of Hazard-4 Lack of the inevitability requirement distinguishes the MIL-STD-882 definition from that of Hazard-4.

The Definition We thus define:

> A **Hazard-5** is a state of a system together with its environment in which the likelihood of an accident is increased over the likelihood of an accident in precursor states.

3.2 Definition of the System S

The Objects There are three objects in the universe: x, y and z – let us call them 'atomic objects' – and thus also the objects $x \oplus y$, $x \oplus z$ and $y \oplus z$ and $x \oplus y \oplus z$.

Their Properties There are precisely three properties that may apply to any atomic object, which we shall write using standard formal notation, and we shall call *1*, *2*, and *3*. Furthermore, these properties hold exclusively of each object: if *1* holds of x, then *2* and *3* don't hold, and mutatis mutandis for *2*, *3* and y and z. And each object at any time has one of the properties; therefore, precisely one. The state of the universe may thus be described by specifying which property holds of which object.

Their Relations Let us suppose that there are no binary or ternary relations that are of significance.

The Assertions The collection of possible 'atomic' assertions is thus

1(x), 2(x), 3(x), 1(y), 2(y), 3(y), 1(z), 2(z), 3(z)

and, of these, precisely one involving a given object is true in any state.

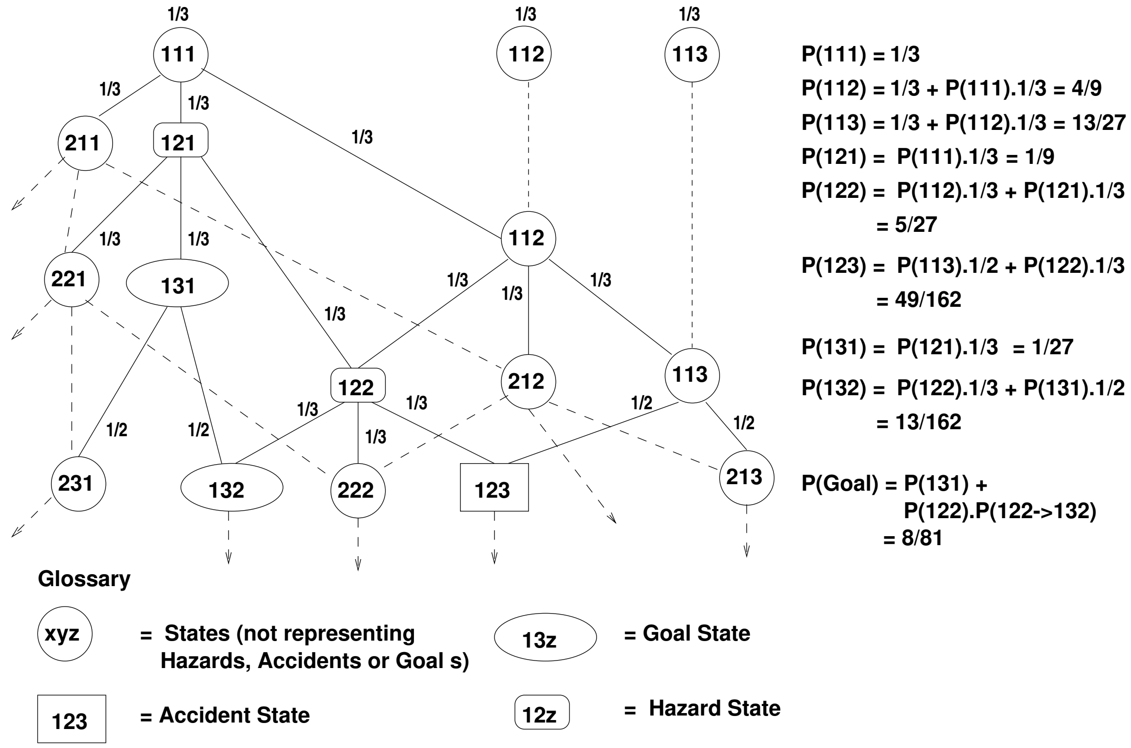
The States This collection of objects with their behavior will be called the 'universum'. The possible changes of the universum are simple: a change is possible from property *1* of any object to property *2* of that object; and from *2* to *3*; no other changes are possible. Let us also assume that changes are discrete: that no two changes happen simultaneously (this assumption is for convenience only; giving it up just complicates the arithmetic, as argued below).

Probability of Changes Assume that any possible change in state has an equal probability of happening. Thus in the state *112*, changes resulting in states *212*, *122* and *113* have each a probability 1/3 of happening; while in state *213*, changes resulting in states *313* and *223* each have probability of 1/2, because no change is possible to z. Let us also also assume that probabilities of transition are dependent only on what current state the universum is in: history is irrelevant.

The System and Its Environment We define a **system** S consisting of objects x and y; z constitutes the environment/rest of the universum. (This also means, if one so wishes, that S contains the object $x+y$; and that there are *mixed* objects, part system, part environment, namely $x+z$ and $y+z$. These ontological niceties need not concern us, since any properties of these objects may be defined logically from the properties of x, y and z.) The system is teleological: it starts in state *(11-)*, namely system state (*1(x) and 1(y)*), its goal state is *(13-)*, namely universum states *131*, *132* or *133*, and state *123* is a loss with a severity of unity (since it is the only loss). We assume there is an equal probability of S starting in any state of the environment; *111*, *112* and *113* are equiprobable universum states for the start of S, each with probability 1/3.

The Behavior of the System S works as follows. It starts in state *(11-)* and 'runs' (changes state) until no more actions are possible. As it changes, so does the environment. We suffer loss if the universum passes through state *123*, and we can consider S to have succeeded if it passes through state *13-* without me having suffered loss. We shall see that S is not very reliable (the probability of attaining its goal is about 1/10), and the chances of loss are quite high (about 1/3).

This system, indeed this universum, is just about as simple as could be. It is finite, with finitely many states and finite (terminating) behavior. We may see how the definitions given so far apply to this system. If we expect them to work in describing complex systems, we should be able to use them to describe such a simple system as S.



State-Action Diagram, with Probabilities, Hazards, Accidents and Goals

Figure 3.1: The Example System

The behavior of the universum is shown in Figure 3.1, along with the probabilities that the universum enters a given state. States of the universum are shown in circles, with system goal states shown as ovals; the loss state as a larger rectangle; and the various states important for various calculations of hazard are shown as boxes with rounded corners. One should observe that one can attain the goal state *133* only by passing through *132*, already a goal state, or through *123*, the 'specified level of loss' state. We may regard *131* and *132* as the two goal states that count, since in order to reach *133*, we would have either achieved our goal already or suffered an accident. But this is a point concerning reliability, not the safety definitions.

Initial States The initial system state is (11−), so the initial univerum states are *111*, *112* and *113*; each has probability 1/3.

Accidents An accident is defined as an event that *results in* a loss. The loss state is 123. Accordingly, there are precisely two sorts of accident events, namely the transitions 122 → 123 and 113 → 123.

3.3 Calculating Hazard-4 and Hazard-1 States

It will be easiest to calculate the hazard states for the various notions of hazard in a different order from that in which they were defined.

3.3.1 Identifying The Hazard-4 States

Hazard-4 states are universum states that are inevitable precursors of an accident. The two most obvious candidates are the preconditions of the two accidents *122* → *123* and *113* → *123*, namely *122* and *113*. Since *121* results in *122* without the system doing anything, *121* is a candidate also.

Candidates 121 **and** 122 There is no other place for the environment to go but to progress *1(z)* → *2(z)* → *3(z)*. Hence

- if the universum is in state 121 and the system does nothing, the universum will inevitably progress 121 → 122 → 123; an accident is inevitable;
- if the universum is in state 122 and the system does nothing, the universum will inevitably progress 122 → 123; an accident is inevitable.

Both 121 and 122 are therefore Hazard-4 states.

Candidate 113 An accident is not inevitable from the state 113, for the following reason. The environment cannot progress. If the system does nothing, the universum remains in state 113 for ever and no accident occurs. So 113 is not a Hazard-4 state.

Is 123 **a Hazard-4 State?** An accident is defined to be an event, and the two accidents are *122* $\rightarrow$ *123* and *113* $\rightarrow$ *123*. The loss state *123* is not a precursor of either of these two events, hence it is not a Hazard-4 state.

The Hazard-4 States The Hazard-4 states are thus 121 and 122.

3.3.2 Identifying the Hazard-1 States

The system state corresponding to the Hazard-4 universum states 121 and 122 is $(12-)$. The candidates for "most unfavorable" environmental state are thereby *1(z)* and *2(z)*. They are both inevitable precursors of an accident, as argued in Section 3.3.1. There doesn't appear to be much to choose between them. The progression of the environment makes either equally as bad as the other; one is the inevitable precursor of the other.

Is *(11-)* a Hazard-1 State? 113 $\rightarrow$ 123 is also an accident. We should ask ourselves whether system state *(11-)* is also a Hazard-1 state. It is not. The obvious candidate for "most unfavorable environmental state" is $3(z)$. Suppose the universum to be in state *113*. The environment cannot change any more, so if the system does nothing, no accident occurs. An accident is therefore not inevitable.

3.3.3 An Accident Without a Preceding Hazard

Note that an accident can be suffered without going through a hazard state: the change *113* to *123* is an accident, but *(11-)* is not a Hazard-1 state and *113* not a Hazard-4 state.

3.4 Calculating Probabilities

We cannot calculate the Hazard-3 and Hazard-5 states without estimating some likelihoods. We do so now.

Performing Probability Calculations The universum states corresponding to the system initial state *(11-)* have equal probability of occurring as the initial state in the system's behavior. Since transitions occur discretely, the probability of occurrence of a specific system behavior may be obtained by multiplying together the probabilities along the path of transitions that the system takes.

A Remark on Notation I use notation *P(xyz)* to denote the probability of occurrence of a state *xyz* in the 'run' of the system; since this is logically a temporal event (the system cannot be in this state forever, but only at certain times), this is really shorthand for *P(*$\Diamond$ *xyz)*, where *(*$\Diamond$ *xyz)* is to be read as '*eventually xyz*', that is, *in some future state, xyz*, as explained further in Section 5. *P(xyz* $\rightarrow$ *abc)* denotes the probability of occurrence of the event *(xyz* $\rightarrow$ *abc)* given that the system is in state *xyz*; using the standard notation for conditional probability, it is really a shorthand for *P(*$\Diamond$*(xyz* $\rightarrow$ *abc) / xyz)*. *P(xyz via abc)* is the probability that the system attains state *xyz* and passes through *abc* on the way; it is shorthand for *P(*$\Diamond$ *xyz and* $\Diamond$ *abc)*. Finally, I use the notation *P(11a init* $\rightarrow$ *abc* $\rightarrow$ *....* $\rightarrow$ *fgh)*, in which *(11z* $\rightarrow$ *abc* $\rightarrow$ *....* $\rightarrow$ *fgh)* is a path, or an initial segment of a path, commencing in the initial state, for the probability of occurrence of this path. We shall also need the probabilities that a path is followed *given that* we are already in the first state on the path. This is written *P(abc start* $\rightarrow$ *abc* $\rightarrow$ *....* $\rightarrow$ *fgh)*. Finally, $P(abc|efg)$ is the conditional probability that *abc* will be reached, given that the system is already in efg.

Calculation of Loss Probability Given Universum State We shall need the following probabilities of entering the loss state given certain universum states.

$$\begin{aligned}
P(123|111) &= P(111 start \ \rightarrow 121 \rightarrow 122 \rightarrow 123) \\
&\quad + P(111 start \ \rightarrow 112 \rightarrow 122 \rightarrow 123) \\
&\quad + P(111 start \ \rightarrow 112 \rightarrow 113 \rightarrow 123) \\
&= (1/3).(1/3).(1/3) + (1/3).(1/3).(1/3) + (1/3).(1/3).(1/2) \\
&= (1/27) + (1/27) + (1/18) \\
&= (7/54)
\end{aligned}$$

$$\begin{aligned}
P(123|112) &= P(112 start \ \rightarrow 122 \rightarrow 123) \\
&\quad + P(112 start \ \rightarrow 113 \rightarrow 123)
\end{aligned}$$

$$
\begin{aligned}
&= (1/3).(1/3) + (1/3).(1/2) \\
&= (1/9) + (1/6) \\
&= (5/18)
\end{aligned}
$$

$$P(123|113) = (1/2)$$

$$
\begin{aligned}
P(123|121) &= P(121 start \ \rightarrow 122 \rightarrow 123) \\
&= (1/3).(1/3) \\
&= (1/9)
\end{aligned}
$$

$$
\begin{aligned}
P(123|122) &= P(122 start \ \rightarrow 123) \\
&= (1/3)
\end{aligned}
$$

$$P(123|123) = 1$$

Calculation of Loss Probability Given System State We shall need the following calculations of entering the loss state, given certain system states. Notice that since the system starts in a state $(11-)$, we have

$$P(123|(11-)) = P(123)$$

Also, we have

$$P(123|(13-)) = P(123|(21-)) = P(123|(22-) = P(123|(23-) = 0$$

since a loss state is unreachable from these system states.

Calculation of $P(123|(12-))$ The calculation of $P(123|(12-))$ is a little tricky, because some of the accident occurrences $122 \rightarrow 123$ are counted already in $P(123|121)$ and we have to be careful not to count these again when assessing how likely it is that the accident $122 \rightarrow 123$ will happen when starting from state 122. We proceed by observing first that

- the ones we have already counted are those that come from $111 init \rightarrow 121 \rightarrow 122$.
- the ones we haven't already counted come via $111 init \rightarrow 112 \rightarrow 122$ plus those that come from $112 init \rightarrow 122$.

$$
\begin{aligned}
P(111init \to 121 \to 122) &= (1/3).(1/3).(1/3)\\
&= (1/27)
\end{aligned}
$$

$$
\begin{aligned}
P(111init \to 112 \to 122) &= (1/3).(1/3).(1/3)\\
&= (1/27)
\end{aligned}
$$

$$
\begin{aligned}
P(112init \to 122) &= (1/3).(1/3)\\
&= (1/9)
\end{aligned}
$$

It follows that

$$P(122 \textit{ via } 112) = (4/27)$$

$$P(122 \textit{ via } 121) = (1/27)$$

so we have counted one out of five accidents $122 \to 123$ in considering $P(123|121)$ and we need to consider the other four-fifths. It follows that:

$$
\begin{aligned}
P(123|(12-) &= P(121 \to 122 \to 123) + (4/5).P(123|122)\\
&= (1/3).(1/3) + (4/5).(1/3)\\
&= 17/45
\end{aligned}
$$

Likelihood of States Simpliciter We shall need the following state likelihoods.

$$
\begin{aligned}
P(112) &= P(111 \to 112)\\
&\quad +P(112init\ )\\
&= (1/3).(1/3) + (1/3)\\
&= (4/9)
\end{aligned}
$$

$$
\begin{aligned}
P(113) &= P(111 \to 112 \to 113)\\
&\quad +P(112init\ \to 113)\\
&\quad +P(113init\ )\\
&= (1/3).(1/3).(1/3) + (1/3).(1/3) + (1/3)\\
&= (13/27)
\end{aligned}
$$

$$
\begin{aligned}
P(121) &= P(111 \to 121)\\
&= (1/3).(1/3)\\
&= (1/9)
\end{aligned}
$$

$$P(122) = P(111 \to 121 \to 122)$$

$$
\begin{aligned}
&+P(111 \to 112 \to 122)\\
&+P(112 init \ \to 122)\\
&= (1/3).(1/3).(1/3) + (1/3).(1/3).(1/3) + (1/3).(1/3)\\
&= (1/27) + (1/27) + (1/9)\\
&= (5/27)
\end{aligned}
$$

$$
\begin{aligned}
P(123) &= P(111 \to 121 \to 122 \to 123)\\
&\quad +P(111 \to 112 \to 122 \to 123)\\
&\quad +P(111 \to 112 \to 113 \to 123)\\
&\quad +P(112 init \ \to 122 \to 123)\\
&= (1/3).(1/3).(1/3).(1/3) + (1/3).(1/3).(1/3).(1/3)\\
&= (1/3).(1/3).(1/3).(1/2) + (1/3).(1/3).(1/2)\\
&= (1/81) + (1/81) + (1/54) + (1/18)\\
&= (8/81)
\end{aligned}
$$

Calculation of Likelihood of System States We shall also need the following likelihoods of system states from which an accident is reachable:

$P((11-)) = 1$

$$
\begin{aligned}
P((12-)) &= P(111 \to 121).(P(121|121) + P(121 start \ \to 122))\\
&\quad +P(111 init \ \to 112 \to 122)\\
&\quad +P(112 init \ \to 122)\\
&= (1/3).(1/3).(1 + (1/3)) + (1/3).(1/3).(1/3) + (1/3).(1/3)\\
&= (8/27)
\end{aligned}
$$

There is little point to calculating the likelihood of other system states. We shall not need them, since an accident is unreachable from them.

3.5 Calculating Hazard-3 and Hazard-5 States

We are now in a position to determine the Hazard-3 and Hazard-5 states. We start as before with Hazard-5.

$P(123|(11-)) = P(123) = (8/81)$
$P(123|(12-)) = (17/45)$
$P(123|111) = (7/54)$
$P(123|112) = (5/18)$
$P(123|113) = (1/2)$
$P(123|121) = (1/9)$
$P(123|122) = (1/3)$
$P(123|123) = 1$
$P((11-)) = 1$
$P((12-)) = (8/27)$
$P(123) = (8/81)$
$P(121) = (1/9)$
$P(122) = (5/27)$
$P(112) = (4/9)$
$P(113) = (13/27)$

Figure 3.2: Summary of Calculations

3.5.1 Determining the Hazard-5 States

Hazard-5 states are those universum states in which the likelihood of an accident is increased over the predecessors. Looking at Figure 3.2 lets us read off as follows

Candidate 111 111 is an initial state, but an initial state has no precursor, so one cannot meaningfully speak of an increased likelihood over precursors. 111 is not a Hazard-5 state.

Candidate 112 112 has as sole precursor 111. It is itself an initial state, but an initial state has no precursor, so one cannot meaningfully speak of an increased likelihood over precursors when it occurs as an initial state. $P(123|112) = (5/18) > (7/54) = P(123|111)$. The likelihood of an accident is increased, therefore 112 is a Hazard-5 state.

Candidate 113 113 has as sole precursor 112. It is itself an initial state, but an initial state has no precursor, so one cannot meaningfully speak of an increased likelihood over precursors when it occurs as an initial state. $P(123|113) = (1/2) > (5/18) = P(123|112)$. The likelihood of an accident is increased, therefore 113 is a Hazard-5 state.

Candidate 121 121 has as sole precursor 111. $P(123|121) = (1/9) < (7/54) = P(123|111)$. The likelihood of an accident is decreased, therefore 121 is not a Hazard-5 state.

Candidate 122 122 has as precursors 121 and 112. $P(123|122) = (1/3) > (1/9) = P(123|121)$. The likelihood of an accident is increased. $P(123|122) = (1/3) > (5/18) = P(123|112)$. The likelihood of an accident is increased. Since the likelihood of an accident is increased over the likelihood of an accident in either of its precursor states, 122 is a Hazard-5 state.

The Hazard-5 States The Hazard-5 states are thus 112, 113, and 122.

3.5.2 Determining the Hazard-3 States

The two candidates are $(11-)$ and $(12-)$ as before, since the accident is unreachable from other system states.

Candidate (11-) $(11-)$ is the start state. It has no precursor, so so one cannot meaningfully speak of an increased likelihood over precursors.

Candidate (12-) $(12-)$ has as sole precursor $(11-)$. $P(123|(12-)) = (17/45) > (8/81) = P(123|(11-))$. The likelihood of an accident is increased. Since the likelihood of an accident is increased over the likelihood of an accident in its precursor state, $(12-)$ is a Hazard-3 state.

The Hazard-3 States The Hazard-3 state we have identified is $(12-)$. This makes calculations of risk identical in this case for Hazard-1 and Hazard-3.

Summary We summarise the Hazard states for each different notion of hazard in Figure 3.3.

Hazard-1 : (12−)
Hazard-2* : (12−)
Hazard-3 : (12−)
Hazard-4 : 121 and 122
Hazard-5 : 112, 113 and 122
* Recall that the Hazard-2 example is the "mirror", $S^\dagger$, of System S.

Figure 3.3: Hazard States For Each Notion of Hazard

3.6 The Calculation of Risk Via Hazard

Since severity is unity, the risk that we shall suffer loss is simply

$$1.P(123) = 1.(8/81) = (8/81)$$

The calculation of risk via hazard that we are supposed to perform is:

$$\mathit{Risk} \ = \sum_{\mathit{Hazard\ states}\ h} P(h).P(123|h)$$

and were these calculations to be accurate, we should obtain (8/81). Let us now perform these calculations, using the notation Risk_i for the notion Hazard-i. The numbers we use are summarised in Figure 3.2. The hazard states we use are summarised in Figure 3.3.

$$\mathit{Risk}_1 = P((12-)).P(123|(12-))$$

$$= (8/27).(17/45) \neq 8/81$$

$$\mathit{Risk}_3 = P((12-)).P(123|(12-))$$

$$= (8/27).(17/45) \neq 8/81$$

$$\mathit{Risk}_4 = P(121).P(123|121) + P(122).P(123|122)$$

$$= (1/9).(1/9) + (5/27).(1/3)$$
$$= (6/81) = (2/27) \neq (8/81)$$

$$
\begin{aligned}
Risk_5 &= P(112).P(123|112) + P(113).P(123|113) + P(122).P(123|122) \\
&= (4/9).(5/18) + (13/27).(1/2) + (5/27).(1/3) \\
&= (109/162) \neq (8/81)
\end{aligned}
$$

We have shown that the calculation of risk through combining hazard likelihood with likelihood of loss per hazard doesn't yield the appropriate figure, which is likelihood of loss *simpliciter*. The only exception is the calculation for Hazard-2, and for that we alter the example. For Hazard-2, we take the same example, but interchange system and environment. That is, the system becomes z and the environment x and y. Call this new example System $S^\dagger$. The loss state and its likelihood remains the same. The calculation of risk through Hazard-2 for System $S^\dagger$ is identical to the calculation of risk through Hazard-1 for System S, since we have just swapped system and environment, and thus system states for environment states and vice versa.

Conclusion The calculation of risk through hazard according to the definitions in [Lev95]do not work for any of the five notions of hazard we have considered.

3.7 The Problem

3.7.1 The Risk of Overcounting

The problems with calculating risk through hazard on Systems S and $S^\dagger$ come about partly through overcounting the paths. Namely,

- in the calculation of $Risk_1$ and $Risk_3$, both $P((12-))$ and $P(123|(12-))$ include a component assessing the likelihood of the transition $111 init \rightarrow 121 \rightarrow 122$. They are thus not independent.
- In the calculation of $Risk_4$, the term $P(121).P(123|121)$ counts some of the same paths as $P(122).P(123|122)$, again those that contain the transition $121 \rightarrow 122$.
- In the calculation of $Risk_5$, the term $P(112).P(123|112)$ counts some of the same paths as $P(113).P(123|113)$, namely those that contain the transition $112 \rightarrow 113$.

3.7.2 Not All Accidents Occur Through Hazards

Although the accident $122 \rightarrow 123$ starts in a hazard state for each of the different notions of hazard, the accident $113 \rightarrow 123$ attains a loss state *without passing through a Hazard-1, Hazard-3 or Hazard-4 state.* Thus the accident behavior $113init \rightarrow 123$ is omitted from the count that each of these risk assessments make. Hazard-5 is the only notion of hazard which includes 113 as a hazard state, but it suffers from overcounting problems as noted above.

3.7.3 Summary

We have used System S and System $S^{\dagger}$and their environments to demonstrate that there is no reasonable way via the notions of Hazard-1 through Hazard-5 to combine hazard probability with likelihood that a hazard state will result in an accident (along with severity of loss) to obtain an accurate estimate of risk, understood as the likelihood of loss (combined with severity). The concept of severity played no role in the argument; the problem lies in the attempt to combine hazard probabilities with likelihood that an accident will result. The problem lies partly in overcounting, and partly in undercounting accidents that occur in system behaviors that do not pass through a hazard state.

3.8 Trying To Fix It

Solution: Hazard-5 Plus Independence of Hazards? Although this example is combinatorially simple, intuition does not help a great deal in guessing its properties. It was deliberately constructed in order to demonstrate the risks of overcounting and undercounting. The risk of overcounting can perhaps be mitigated by trying to assure that all phenomena to be counted as hazards are independent of each other, in the sense of probability theory. It is because the situation of the universum entering state 121 is not independent of it entering 122 that we overcount. However, ensuring independence of hazard phenomena does not solve the undercounting problem, whereby accidents can happen without passing through a corresponding hazard state. But recall that Hazard-5 captured these states. This may suggest to some that a combination of

- employing the concept Hazard-5, and

- ensuring that identified hazards are independent

might be a useful solution to the problem of calculating risk through hazard. Note that both are needed: the Hazard-5s posed by 112 and 113 are not independent of each other. We will not go so far as to favor this solution. Rather, we prefer here simply to discuss the phenomenon.

Altering the Concept of Risk Another move would be to take the definition of risk as it is given; and conclude that the intuitive concept of risk as (in this case) likelihood of loss given unit severity is not the most appropriate concept of risk. But this would be a move to contradict intuition for the sake of otherwise unmotivated consistency. Besides, the problem remains in another form. We need to calculate the likelihood of loss, for example to calculate betting odds, and the problem is that the proposed calculation method cannot render this in all circumstances.

This Leaves One Open To Loss If I believe my risk is as in the definition, and I bet according to this, then I am betting according to some assessment of probability that is different from the actual probability that a loss state will occur. A bookmaker can thus construct a series of bets that I am prepared to accept according to my assessment of risk but which I am guaranteed to lose money on in the long run. This is of course only a way of phrasing the fact that if I incorrectly assess the probability of loss, I may make decisions which do not minimise my loss. This is not what one hopes for from a risk assessment.

3.9 Motivating The Conceptions of Hazard

Hazard-1 As we have seen, Hazard-1 is that used in System Safety engineering in the U.S. for some time, and that espoused for that reason by [Lev95]. This is reason enough for us to consider it. However, one should also note the rationale behind it, which is that in the safety assessment of a teleological system is that the the engineer has control over the system state but not over the environmental state. Any prophylactic measures can only be applied to circumstances and state components over which one has control. Therefore, hazard reduction must be applied to the system state.

Hazard-2: The Layman's Idea While driving my car, I am inclined to call a football bouncing into the road with a child running after it a hazard. I am also inclined to call a pothole in the road a hazard. Both of these are predicates of the environment in which I am driving, not of my car and its driver. If I consider my car with driver to be the system, these "hazards" are environmental predicates.

If I am driving my car in a more or less standard manner, these conditions could lead to - or would inevitably lead to - some sort of accident, depending on the system state. For example, if I am driving at 0.001 kph, the situations above would not lead to accidents, whereas they would if I am driving at the generally allowable 50kph.

Hazard-1 and Hazard-2 for Different Sorts of Systems? When considering a relatively closed system, such as a power plant or chemical plant, or electrical wiring in a building, it makes sense to conceive of hazard states as being system states. However, some complex systems are unavoidably open. An aircraft has to operate in weather and in terrain that is part of its environment. There is no system state which corresponds to a thunderstorm going from Level 2 to Level 4 within a matter of a few minutes, and it is wise to single out this area of the environment for special attention when it happens. This is the rationale for Hazard-2. The system safety definitions have no equivalent concept to that of Hazard-2, since there is no obvious way to reduce Hazard-2 to Hazard-1 in general, but it is hard to see how the use of Hazard-2 can be avoided in some cases.

Maybe Both At Once: Hazard-4 The point of considering and designing for system safety, however, is to avoid combinations of environmental conditions and system states leading to accidents. If one knows all such states, as in Hazard-4, then one can calculated Hazard-1 and Hazard-2 states from them, as we did for System S. Therefore Hazard-4 contains more information than either Hazard-1 or Hazard-2, and these latter are recoverable from it.

3.9.1 Weakening the Inevitability Requirement

Other definitions of hazard preserve its feature as a property of system states, but give up the insistence on inevitability. This led us to Hazard-3 and Hazard-5. Judging states by increased likelihood, not of accidents but of failure, is common in reliability engineering. Since safety may often depend on the reliability of safety-critical components, these are connected.

Discrete Classification An example is commercial aviation. Lloyd and Tye [LT82] explain the various different likelihood categories used in civil aviation certification in the U.K. They note [LT82, Table 4-1] that both U.S. Federal and European Joint Aviation Regulations, in their parts 25, classify events as *probable* if their likelihood of happening lies between 10^{-5} and 1, *improbable* if between 10^{-9} and 10^{-5}, and *extremely improbable* if smaller than 10^{-9}; the JARs additionally classify probable events into *frequent* (between 10^{-3} and 1) and *reasonably probable* (between 10^{-5} and 10^{-3}), and improbable events into *remote* (between 10^{-7} and 10^{-5}) and *extremely remote* (between 10^{-9} and 10^{-7}).

The Purpose of the Discrete Classification The purpose is (was) to classify an event as *extremely improbable* if it was unlikely to arise during the life of a fleet of aircraft; *extremely remote* if it was likely to arise once during fleet life; *remote* if likely to arise once per aircraft life, and several times per fleet life; *reasonably probable* if likely to arise several times per aircraft life. Fleet sizes were assumed to be about 200 aircraft, with each aircraft flying 50,000 hours in its life (nowadays, we are seeing fleet sizes are of the order of 1,000 to 2,000, and aircraft flying more than 50,000 hours, altogether about a factor of 10 difference).

Classification of Effects Effects are also classified into *minor*, *major*, *hazardous* and *catastrophic*, according to damage, injuries and deaths.

The Certification Basis The certification basis is (was) to demonstrate that *major*, *hazardous* and *catastrophic* effects could occur at most with *remote*, *extremely remote* and *extremely improbable* frequencies respectively.

Reliability and Safety Conflated The certification basis attempted to assign probabilities to failures, which is a technique for reliability classification, but we have noted that reliability and safety are closely linked via reliability of safety-critical components. For example, multiple engine failures entail that the aircraft must land within a certain radius of its position, whether there is a suitable airport there or no. A fire on board that is not effectively extinguished will spread within a certain time, and be catastrophic unless the aircraft is on the ground at this time. Failure of various specific mechanical parts, or total failure of the flight control system, lead inevitably to an accident. However, reliability and safety may still be distinguished: a recognition light on the underbelly is a safety-critical item; a reading lamp in passenger class is not. The reliability of the latter is not a safety issue.

3.9.2 Avoidance Of The Problematic Notions

The IFIP WG 10.4 Definitions The series of definitions in [Lap92] concerned with dependability, which is taken by members of the IFIP WG 10.4 to include safety, does not include the concepts of hazard and risk at all.

3.9.3 Classifying Risk Through Statistics

An obvious way to avoid the problem is to have had the misfortune to have had sufficiently many accidents that one can calculate risks on a statistical basis from history. Let us briefly consider one plausible way in which this might be done, namely the U.S.A.F. mishap classification scheme.

U.S. Air Force Accidents as "Class A Mishaps" The most severe category of incident defined by the U.S. Air Force is a *Class A* mishap. This is a mishap resulting in loss of life or more than $ 1M in damage. This is similar to the U.S Federal Aviation Regulations definition, in which an accident is defined to be severe injury or loss of life, or "substantial damage" to an aircraft (the "or" is inclusive).

This definition may have unintuitive consequences Such accidents have occurred in military aviation in which both aircraft returned safely with more than $ 1M in damage [Gar98]. (This would be a case in which considerable damage was done, but no one died and the damage was repairable.)

The Distinction Between Events and Event Types With a slight change in parameter, say, a slightly different relative motion of the aircraft, then the event that occurred could have had catastrophic consequences. For example, loss of both aircraft with pilots. What does it mean to say that an event *could have had* other consequences? One way of interpreting this is to note that there is a class of incidents, *mid-air collisions* to which this event belongs. One can even go further and say: *mid-air collisions between two aircraft of type X in formation flying* or even more detailed: *mid-air collisions between two aircraft of type X in formation flying in clear weather, performing manoeuver Y*. These classifications define ever more precise and thus smaller *classes* or collections of events. These are *event types*.

Using the Distinction An individual event such as a mid-air collision with $1.01M damage but in which both aircraft and crew returned safety can be viewed either

1. as an individual event with specified damage; or
2. as a member of an event type whose average member is a catastrophic accident

How we view this event can have considerable influence on how we would treat it.

Different Classification Leads to Different Comparisons Consider the different reactions to the different classifications.

1. Suppose we treated the accident as an individual event with specified damage. Then we would be comparing with other events in which, say,
   - a ground service vehicle ran into a parked aircraft because it was travelling too fast for wet conditions and momentarily lost directional control; or
   - a misapplication of electrical power during routine service fried essential aircraft avionics and required thorough test and replacement
2. Suppose we considered the accident as an event belonging to the type *midair collision*. Then we would be comparing with other midair collisions, many with much more catastrophic consequences.

Different Comparisons Lead to Different Prophylactic Measures Let us for the moment consider non-injurious accidents. If a midair collision with minor consequences is classified together with midhandling of a ground vehicle or misapplication of electrical current during maintenance, we may be hard put to find similarities. The classification of these incidents into Class A mishaps uses a predicate which is

- primarily economic, an amount of money, and
- oriented towards consequences, the cost of management or replacement, rather than preconditions.

An incident classified in the according to the features, the state predicates, that were

- either necessary or sufficient precursors of the event, or else
- immediate postconditions

is of much more importance for the causal analysis of the event.

Different Views on Prophylaxis

- Management response to accidents is of the utmost importance. Data must be collected, resources allocated to response, trend data must be classified and analysed over time, and the results incorporated into institutional management procedures. All analysts agree uniformly that appropriate institutional treatment of accidents is crucial to safe systems operation.
- It should be evident that accidents themselves can only be avoided by mitigating their causes. The causes of an event can be regarded as a collection of individually necessary and jointly sufficient conditions for the event to have occurred [Mac74]. That they are individually necessariy means that if any one of these causal factors had not pertained, the event would not have happened. Identifying the causal factors identifies those factors which, were they to be avoided in the future, would avoid entirely future accidents with exactly those causal features.

Reconciling The Views An argument may thus be made that causal analysis is essential to, the *sine qua non* of, any prophylaxis. However, performing such a causal analysis requires allocation of resources and a decision must be made as to which incidents those resources should be allocated and to which not. The economic classification of mishaps is thus a practical guide to management, focusing resources on those mishaps for which there is a good economic argument to be made for avoidance, and thus encouraging political agreement with such decision. Care must be taken, however, not to confuse concepts which aid in causal analysis with concepts which aid in resource allocation.

Predicates That Matter, And Predicates That Don't Consider the event type of midair collisions. Each individual accident will have precisely locatable spatio-temporal features: such-and-such an aircraft part touched another part at a precise time in a precise time zone, in a precise altitude and geographical location (even if these precise coordinates are not so precisely determined). It is significant for the accident that there was spatio-temporal overlap of parts. The trajectories of the aircraft and their manoeuverability are regarded as causally relevant to this spatio-temporal overlap. The fact that it happened over the precise geographical point that is did and not 20 km, or even 20m, to the north, is usually regarded as less relevant, since nothing about the dynamics of the aircraft makes use of this information.

Allowing Revision of This Judgement It may be, however, that the exact geographical location is relevant; perhaps because of sun position and location of reflectors on the ground and position of the aircraft relative to the reflected image of the sun, one pilot was temporarily blinded and lost the precise dynamic control over his aircraft that formation flying requires. So revision of *a priori* judgements of causal relevance must remain an open possibility. However, in the example above, it should be clear that relative position (or angle) of sun, the presence of ground reflectors, and the requirement of perfect pilot vision for manoeuvering are pertinent causal factors, and the translation of the ground reflector and the aircraft 20m to one side of the (relative) coordinate frame does not affect causality.

Precursors to An Accident Must Be Causal Precursors Causal analysis is uniformly accepted as the predominant accident analysis technique.

A significant justification for this acceptance is that in order to avoid repeat accidents, it is both necessary and sufficient that a necessary part of a sufficient causal condition for the accident be absent from future behavior of the same or similar systems in their environment.

Why Not Correlates? Factors may have high correlation with accidents. However, correlation does not mean that there is a specific causal relation, for three reasons:

- Correlation (or, as Mill called it, *concomitant variation* [Mil73]) is a symmetric relation (if A correlates with B, then B correlates with A), whereas being a casual factor is an asymmetric relation (if A is a causal factor of B, then B cannot be a causal factor of A);
- if A and B vary concomitantly, then that may be because they have a (maybe unidentified) causal factor in common;
- the possibility remains that it could just be chance.

Correlation Focuses the Hunt For Causality Identifying correlation between factors helps to focus attention in the hunt for causality. Causal factors will be correlated, so identifying correlations narrows the potential relationships to be considered in identifying causal relations, without excluding any.

This is Not a Universal Method However, in order to identify statistical correlations, one needs a sufficient number of sufficiently similar incidents or accidents, or a sufficient number of observations of subsystem behavior. These may not always be available. One circumstance in which these would be available are in a case in which safety is correlated with reliability of a system component, and sufficient analysis of this component has been performed to be able to identify a statistical reliability. Such components are most likely “safety mechanisms”, on whose reliability the safety of the system operation is predicated.

3.10 Summary

We have seen that, even for a simple case such as System S and System $S^\dagger$, there are serious problems with the notions of hazard and risk as used about systems. This despite probabilities in S and $S^\dagger$ being determinate, with probabilities of change independent of history, and assuming trivial severity and ignoring duration.

There are three components to the argument as presented:

1. Risk in the case of unit severity is likelihood of loss;

2. Calculating risk through hazard likelihood combined with likelihood that an accident will result overcounts some paths in the case in which one hazard state inevitably leads to another,

3. Calculating risk through hazard likelihood combined with likelihood that an accident will result omits to count the likelihood of accidents which occur without passing through a hazard state.

Chapter 4

More Theory: Types of Predicates

Asserting State Predicates We have accepted that the world state consists of objects which have properties, and relations between them. This suggests that the vocabulary of first-order logic is appropriate for talking about state.

Types of Objects We have classified objects into

- System objects
- Environment objects
- Neither (objects belonging to "the world")

Because objects also have parts, and we allow ourselves the operation of fusion $\oplus$, it is possible that we may talk about objects which are part system and part environment: let O_s be a system object and O_e belong to the environment. Then $O_s \oplus O_e$ is part-system, part-environment. However, in the absence of a specific reason for doing so, and in view of the fact that we are mostly concerned with *artifacts*, with systems that we ourselves design and build, it seems wise to attempt to reduce confusion by avoiding talk of such composite objects as far as we can.

Types of Properties Relative to the classification of objects, properties of objects that are relevant to system operation thus can have the following types:

- properties that only system objects can have
- properties that only environment objects can have
- properties that both system objects and environment objects can have

Types of Relations Relative to the classification of objects, relations amongst objects that are relevant to system operation can have the following types

System Predicates relations between system objects alone

Environment Predicates relations amongst environment objects alone

Hybrid Predicates of Type 1 relations that may pertain between system objects and environment objects

Hybrid Predicates of Type 2 relations that may be between system objects, or between system and environment objects

Hybrid Predicates of Type 3 relations that may be between environment objects, or between system and environment objects

We Only Need Hybrid Predicates of Type 1 We may consider hybrid predicates of type 2 to be the union of a hybrid predicates of type 1 with a system predicate, and a hybrid predicate of type 3 to be the union of a hybrid predicate of type 1 with an environment predicate. Their interdefinition is possible in any language which contains the predicates "*Object x belongs to the system*", that is, $BelongsToSystem(x)$, and "*Object x belongs to the environment*", that is, $BelongsToEnvironment(x)$, as follows. Suppose $A(x, y)$ is a hybrid predicate of type 2. Then

$$BelongsToSystem(x) \ \& \ BelongsToSystem(y) \ \& \ A(x, y)$$

is a system predicate,

$$BelongsToSystem(x) \ \& \ BelongsToEnvironment(y) \ \& \ A(x, y)$$

is a hybrid predicate of type 1, and

$$A(x, y)$$

$$\textit{if and only if}$$
$$(\textit{BelongsToSystem}(x)\ \&\ \textit{BelongsToSystem}(y)\ \&\ A(x,y)$$
$$\textit{or}$$
$$\textit{BelongsToSystem}(x)\ \&\ \textit{BelongsToEnvironment}(y)\ \&\ A(x,y))$$

Henceforth, we will consider only hybrid predicates of type 1, and omit the type.

Limiting the Types of Relations by Fiat We may assume that if a relation important to the system operation pertains between environment and world objects, that all the objects that can be in the relation to each other (the so-called *domain* of the relation) should be considered to be part of the environment. By this means, we rule out the need to consider relations involving "world" objects.

Discrimination of State Predicates Predicates of the world can have as arguments

- just system parameters, in which case we call them *system predicates*;
- just environment parameters, in which case we call them *environment predicates*; or
- some system parameters and some environment parameters, in which case we call them *hybrid predicates*.

Every predication is precisely one of the three types (a), (b) or (c).

System State and Environment State

- The *system state* consists of all true system predicates;
- The *environment state* consists of all true environment predicates;
- the collection of all true hybrid predications is the *hybrid state*;
- The *world state* consists of the union of the system state with the environment state with the hybrid state.

Relations Between the Types of States Note that a hybrid predication is related to certain system predicates and environment predicates by quantification. For example, let x be a system parameter, t be an environment parameter, and A a hybrid predicate. Then $A(x,t)$ is a hybrid predication, related to the system predicate $(Existsn.A(x,n))$ and the environment predicate $(Existsm.A(m,t))$.

Hybrid Predications Are Essential Information Although from every hybrid predication one can obtain a system predication, respectively an environment predication, the reverse is not necessarily the case. Suppose

- any state containing $A(x,t)$ would be an accident state, but that
- there are no such states.

Suppose further that

- t is the only instance of n in which $Existsn.A(x,n)$ occurs in an accident state;
- that x is the only instance of m in which $Existsm.A(m,t)$ occurs in an accident state.

Suppose also that

- there are plenty of states in which there is an n such that $A(x,n)$ and
- plenty in which there is an m such that $A(m,t)$, and
- plenty in which there are both such n and m.

Now, observe that

- a predication involving A, x and t is crucial to determining certain accidents;
- no predication involving A and x alone is going to help you to determine a predication involving A, x and t;
- no predication involving A and t alone is going to help you determine a predication involving A, x and t.

Therefore, no observation of system states and environment states is going to help you with analysing the chance of this accident. Ergo, hybrid predications are essential information.

Chapter 5

An Example: Playing Golf

It has been suggested by colleagues [Mel00, Lev00] that safety concepts can be applied to a simple system such as a golf game. Recall that all that is needed to have a system is objects and behavior.

Intuitive Interpretation The idea is that

- an accident is some event such as playing too many strokes (and thereby losing the game);
- Landing in a bunker entails with high probability that one increases the number of strokes required to play the hole, and thus to play the game
- Thus landing in a bunker, or being in a bunker, represents a hazard
- The risk associated with the hazard (with the bunker) is the expected number of extra strokes one must take to get out of the bunker

5.1 The Basics: Objects, Predicates, Accident

We provide here a reconstruction of the example according to the concepts we have already introduced.

Objects The system and environment are defined by objects. These are

- A golf ball: b
- A player: p
- A course: C
- A bunker: B

Let us for simplicity assume that there is only one bunker. Notice that the bunker is *part of* the course: $B \prec C$.

Predications There really isn't a whole lot that can be said yet in the way of predications. Intuitively,

- the ball can be on the course: $loc(b, C)$
- and when on the course it can be in a bunker: $loc(b, B)$
- the player can be on the course: $loc(p, C)$
- and the player can be in the bunker: $loc(p, B)$

Accident An accident is too high a stroke count at the end. It seems we need a fluent $TotalStrokes$ for the total number of strokes played. We have simplified by having only one player, so the player must be playing against a set total, N. An accident would be a total stroke count greater than this limit: $TotalStrokes > N$.

5.2 The System And Behavior

We only have four main objects, so there are only a few choices.

- the ball b alone belongs to the system; the other objects to the environment. This would mean that the predications above are all hybrid or environment predicates:
  - $loc(b, C)$ and $loc(b, B)$ are hybrid;
  - $loc(p, C)$ and $loc(p, B)$ are environment.

System predicates would be those obtained from the hybrid predicates by quantification. There is only one, namely

  – $ExistsX.loc(b, X)$

And if we assume that the ball remains on the course the entire time, this sole system predicate turns out to be always true.

- The ball b and player p belong to the system. It follows that all four predicates above are hybrid, and the system predicates are $ExistsX.loc(b, X)$ and $ExistsX.loc(p, X)$ If we assume that the player remains on the course with the ball, both of these are always true.

- The course C belongs to the system, player and ball to the environment. Because B is part of C, $B \prec C$, B must belong to the system also. The system predicates would be $Existsx.loc(x, C)$ and $Existsx.loc(x, B)$: there is a player or ball on the course or in the bunker. Again, the former is always true; the latter is what we intuitively have called a hazard.

- The bunker B belongs to the system; the course C to the environment. Whether ball and player belong to system or environment, one may classify the predicates similarly to above.

- Everything belongs to the system. In this case, all predicates are system predicates.

Additional Objects In order to say what we meant by an accident, we introduced the fluent $TotalStrokes$ and the natural numbers (or at least one of them, N). If we include the natural numbers, we have more to say. In principle, we should inquire whether the fluent $TotalStrokes$ and the natural numbers belong to the system or not. In practice, it doesn't matter. In general, expressive artifacts we introduce in order to be able to talk about a system will not need to be classified with the system, but in certain circumstances they may.

Greater Expressive Capability Accumulating strokes by landing in a bunker raises the chances of $TotalStrokes$ exceeding the target N, because of the increased chance of extra strokes being needed. But one extra stroke in a bunker may be neutralised by a reduced number of strokes (due to luck or skill) later. So it's not inevitable that an accident will occur if one lands in a bunker.

Behavior We must say what kind of behavior the system can engage in (whatever we take the system to be). In order to specify behavior, we have to say what changes can occur, of system and environment. The ball and player are always on the course, so there can be no change in this predication. However, both ball and player can be in or out of the bunker. This gives the possibility of four changes. In the formalism below, I use the prime symbol "$'$" on a predicate to indicate that this is true *after* the change, and any predicate without a prime is asserted to be true *before* the change.

- Player in bunker, then player out:

$$loc(p, B) \ \& \ \neg loc'(p, B)$$

- Player out of bunker, then in:

$$\neg loc(p, B) \ \& \ loc'(p, B)$$

- Ball in bunker, then out:

$$loc(b, B) \ \& \ \neg loc'(b, B)$$

- Ball out of bunker, then in:

$$\neg loc(b, B) \ \& \ loc'(b, B)$$

Furthermore, strokes are being accumulated, so an additional change can occur to the fluent $TotalStrokes$:

$$TotalStrokes' = TotalStrokes + 1$$

I use the prime notation here to indicate that the value of $TotalStrokes$ after the change is 1 greater than the value before.

5.3 Expressing Constraints on Behavior

Every Change Means (At Least) A Stroke Intuitively, each change in location of the ball must be caused by a stroke. However, if the ball is be hit from place to place on the course without landing in the bunker, we cannot express that change in the language we have. So the stroke count can increase without a change in (expressible) location. We might suppose that the condition on location change is

$$loc(b, B) \ \& \ \neg loc'(b, B) \Rightarrow TotalStrokes' = TotalStrokes + 1$$

$$\&$$

$$\neg loc(b, B) \ \& \ loc'(b, B) \Rightarrow TotalStrokes' = TotalStrokes + 1$$

but we would be wrong. We are measuring change by comparing two states. But these two states may not represent consecutive hits of the ball. We may be comparing two states relatively far apart, say 4 or 5 strokes apart. Thus the correct condition is

$$loc(b, B) \ \& \ \neg loc'(b, B) \Rightarrow TotalStrokes' > TotalStrokes$$

$$\&$$

$$\neg loc(b, B) \ \& \ loc'(b, B) \Rightarrow TotalStrokes' > TotalStrokes$$

Expressing the Bunker Constraint We want to say that if the ball lands in the bunker, this is likely to increase the final score, but not definitely because of the chance of a birdie, as noted above. One way of saying it is that landing in the bunker increases the expected final stroke count. We need a new fluent $ExpectedScore$. The condition that landing in the bunker increases one's expected score by 1 is expressed by

$$\neg loc(b, B) \ \& \ loc'(b, B) \ \& \ TotalStrokes' = TotalStrokes + 1$$

$$\Rightarrow ExpectedScore' = ExpectedScore + 1$$

Here again, the extra conjunct in the antecedent ensures that we are comparing a change due to one stroke, not to many.

The Game Must Stop We have been assuming that $ExpectedScore$ is a positive integer. There are formal ways of expressing all these conditions such that all such assumptions are explicit - for example TLA [Lam, Lam94, Lad97], from which this notation is lifted. However, let us continue without worrying about these technical details for the moment. One important constraint on behavior is that the game stops. That means that at some point in the future, the stroke count $TotalStrokes$ maintains a fixed value for ever.

Using Tense-Logical Operators Let us phrase this in terms of behavior and state. We use the tense-logical operator $\Diamond$: asserting $\Diamond A$ is to assert that *at some state in the future, A will hold.* We want to say that at some point in the future, the stroke count does not change for ever more. The "*dual*" of $\Diamond$ is the tense-logical operator $\Box$: $\Box A$ asserts that *at all states in the future, A will hold.* It is straightforward to check that

$$\Diamond A \Leftrightarrow \neg\Box\neg A$$

and

$$\Box A \Leftrightarrow \neg\Diamond\neg A$$

by referring to the behavior diagram in Figure 1.5, and imagining it going on for ever. The two equivalences can be expressed in natural language as follows. If A is true in one of the states (one can write it in to make it clear), then it cannot be the case that A is false in all the states; that is to say, that $\neg A$ is true in all the states. Similarly, if A is true in all of the states, then it cannot be the case that A is false in one of them; that is to say, it cannot be the case that $\neg A$ is true in one of them.

Expressing the Stopping Constraint To say that, at some state in the future, it will be the case from then on that the stroke count will not change, can be expressed as follows. $Box(TotalStrokes' = Totalstrokes)$ says that the stroke count remains the same for ever more. We need to say that this assertion becomes true at some state in the future. This means it needs to be prepended with $\Diamond$. So $\Diamond\Box(TotalStrokes' = TotalStrokes)$ expresses the stopping constraint. Notice that we haven't said *when* the game stops. We have not built a stopping criterion into the example yet, and won't do so.

Expressing the Accident We can now say that an accident will happen: $\Diamond TotalStrokes > N$.

Constraints Are Expressible Independent of the System Notice that we have been able to express the constraints unambiguously without deciding exactly in what the system consists. This phenomenon is more or less general: one can express constraints on system and environment without necessarily needing to distinguish them.

5.4 Hazard Definitions and Consequences

Hazard The reason that landing in the bunker is a hazardous situation is that the expected number of strokes is thereby increased. If one has already birdied five times, then landing in a bunker and expecting to lose a stroke is not particularly problematic, or hazardous, because one then expects merely to finish four under instead of five under. I suggest that landing in the bunker becomes hazardous when the expected number of strokes increases to above the limit N. In fact, bunker or no bunker, this is a hazard, and remains a hazard until one birdies to get it down again. So we might try to express being in the hazardous situation by the state predicate $ExpectedScore > N$.

Hazard and World State There are only two objects (one fluent and one integer) mentioned in the state predicate that expresses a hazard. Neither of these objects belongs to the system as we have so far conceived it, or the environment as we have so far conceived it. There are the following possibilities:

- If integers and the fluent belong to the system, then the hazard definition is a system predicate; thus a Hazard-1-type definition.
- If integers and the fluent belong to the environment, then the hazard definition is an environment predicate; thus a Hazard-2-type definition.
- If one belongs to the system and the other to the environment, then the hazard definition is a hybrid predicate; thus a Hazard-4-type definition.

Is An Accident Is Inevitable? From the state predicate $ExpectedScore > N$ it does not follow that an accident is inevitable. However, we may imagine that $ExpectedScore$ is really the *expected score*, and if we get to within one stroke of finishing and the expected score is still greater than N, then the total we have already, $TotalScore$, must be already greater than or equal to

N. So if the expected score rises to above N, and remains there until the game finishes, then an accident is inevitable. That is,

$$ExpectedScore > N \;\&\; \Box(ExpectedScore > N)$$

$$\Rightarrow \Diamond(TotalScore > N)$$

That is, an accident is inevitable if the predicate

$$ExpectedScore > N \;\&\; \Box(ExpectedScore > N)$$

ever becomes true. A technical point: in tense logic as used in engineering, the statement $\Box A \Rightarrow A$ is taken to be an axiom, for any statement A. Hence we may write the statement

$$ExpectedScore > N \;\&\; \Box(ExpectedScore > N)$$

as its tense-logical equivalent

$$\Box(ExpectedScore > N)$$

without loss of expressiveness.

Summary of the Hazard Definitions We may summarise the situation with regard to the use of hazard definitions as follows.

- if we use the hazard definition $\Box(ExpectedScore > N)$ then
  - We may use a Hazard-1 definition only if both $ExpectedScore$ and N belong to the system;
  - We may use a Hazard-2 definition only if both $ExpectedScore$ and N belong to the environment;
  - If one of $ExpectedScore$ and N belongs to the system and the other to the environment, we may use Hazard-4;
- Consider the hazard definition $(ExpectedScore > N)$. An accident is not inevitable from a state satisfying $(ExpectedScore > N)$ unless it also satisfies $\Box(ExpectedScore > N)$. To use a Hazard-1 or Hazard-2 definition, one of these predicates would have to be a system predicate and the other an environment predicate (recall that according to

these definitions, and accident is inevitable if a hazard state (system or environment respectively) *coupled with* a complimentary state (of environment or system, respectively) leads inevitably to an accident. Since both predicates mention the same objects, they are either both system or both environment, and cannot be one and the other. Hence a Hazard-1 or Hazard-2 definition cannot be used. A Hazard-4 definition also requires inevitability. Hazard-3 is the only definition which requires just increased likelihood. Hence if the hazard definition is taken to be $(ExpectedScore > N)$, a Hazard-3 definition must be used.

Hazard Definition Type Depends on What's In The System We may conclude that which type of hazard definition is used depends on what one considers to be part of the system and what not.

- If one insists on using a Hazard-1 definition, then one *must* include $ExpectedScore$ and integers (at least, N) as part of the system.
- If one insists on using a Hazard-2 definition, then one *must* include $ExpectedScore$ and integers (at least, N) as part of the environment.
- One cannot use a Hazard-4 definition.
- A Hazard-3 definition can be used which is logically simpler (it does not include the tense-logical operator $\Box$), no matter where $ExpectedScore$ and N are chosen to belong.

Chapter 6

Some More Conceptual Machinery

We have already addressed the need for rigor in system analysis, and seen that a rigorous approach to describing systems demonstrates subtleties in the definition of hazard, and difficulties in the proposed analysis, whereby one attempts to calculate overall risk as a function of hazard, severity, and likelihood that a hazard will lead to an accident. We now introduce some further notions which will help in the analysis of systems.

6.1 System Properties in the Large

Causality in Aviation It is required by international treaty (the 1948 Chicago Convention, setting up the International Civil Aviation Organisation) that accidents to commercial aircraft be investigated, and a *probable cause* and *contributing (causal) factors* for the accident determined. Commercial aviation represents a significantly complex system, involving complex systems such as air traffic control (considered by Perrow [Per84]) and invididual commercial airliners as parts. We may conclude that the causal influences of complex system parts on each other and on the environment, and vice versa, is an important and significant feature of such systems.

Commercial Aircraft As Complex Systems Commercial aircraft themselves are highly complex systems, with functioning parts that are mechanical (engines, control surfaces), parts that are electrical (lighting, control systems and control system signalling), digital electonics (avionics) and human (pilots).

What's a Part? We include, say, pilots as parts of the aircraft, because the aircraft's behavior is presaged on the (formal) behavior of the pilots; and the behavior of the pilots is specified as part of the aircraft's operation. In general, one can consider an object O with behavior to be a part of a system insofar as the physical behavior of the system is coupled to the behavior of the object O. So we would consider pilots to be part of the aircraft system, because they physically manipulate objects in the cockpit which have a direct effect on the behavior of the aircraft. We do not consider air traffic control to be part of the aircraft because ATC behavior is mediated through pilot understanding and compliance, and they have no direct influence, as do pilots, on the behavior of the aircraft. ATC is an aviation-system part which communicates with aircraft-system parts. We are free to draw the boundary of a system where we like (that is, to include or exclude certain objects) and criteria which we may use include

- tradition
- the intensity of interoperation
- the mode of interoperation

Interactive Complexity and Tight Coupling: Perrow We have identified *complexity* as a feature of systems. It is an intuitive notion that most people understand (although they may judge differently depending on a number of factors, including the intellectual tools they use for understanding). Other large-scale features of systems are important, including two singled out by Perrow:

- *Interactive complexity.* This feature represents the degree to which system parts need to communicate and interact with each other during normal system operation.
- *Coupling.* Coupling is further classified as *tight coupling* and its contrary *loose coupling.* This feature represents the degree to which a change in state or behavior of a system part causes changes in state or behavior in other system parts, and the degree of such changes. In a

> tightly coupled system, many other parts of a system will be sensitive to small changes in a single system part, and the associated state or behavioral changes will be significant.

Perrow claimed that tightly-coupled, interactively complex systems were unusually subject to what he called *system accidents*, which is an accident which one cannot attribute causally to failure of any one precise system part. All parts seemed to function as they were designed to do, but nevertheless an accident occurred. Perrow went so far as to claim that when systems are highly tightly-coupled and interactively complex, system accidents were virtually inevitable. He provided a number of studies in [Per84] to support his claim. A significant in-depth study of one highly complex system, the U.S. military's system of nuclear weaponry, was investigated by Sagan [Sag93], who came to conclusions supporting Perrow's contention.

System Decomposition We have already noted that one should not conflate reliability (the property of a system to continue to perform its intended function, or not) and safety (the avoidance of accidents). It follows that a system accident in Perrow's sense may not represent a system failure (a failure to perform its desired function), unless avoiding accidents was an explicit function of the system. It may not be.

According to the failure reasoning in Figure 1.15, a failure in the system as a whole may be put down to a failure in some system component, provided that the components form an adequate decomposition. Assuming that an accident represents a system failure (as it should if the required safety properties of the system are included in the system requirements specification), an adequate decomposition will determine in which part of the system a failure is located. Perrow suggests the DEPOSE composition; his contention that there exist system accidents suggests that DEPOSE is not an adequate decomposition. But he provides no reasoning to suggest that adequate decompositions do not exist. Intuitively, they do. Features of systems are found which contribute to accidents. These features are part of some decomposition of a system. No accident has ever occurred in which investigators have simply given up, and said that although they know everything there is to know about how the accident occurred, nevertheless they cannot say anything about any system part which contributed.

Heterogeneity I call a system *heterogeneous* if it includes parts of widely disparate types: for example, mechanical, electrodigital, human, procedural. An aircraft includes electromechanical, electrodigital, and human parts and its correct operation requires procedural parts also. An air traffic control system has similar parts in different proportions; minimal ATC systems involve radios and recorders as the sole electromechanical parts, have no electrodigital parts, and are humanly and procedurally intensive. The importance of heterogeneity lies in the different operational and failure modes of the different types of parts. For example

- electrodigital systems are functionally reliable, do not adapt to situations they are not explicitly designed to handle, and fail in unpredictable ways;

- humans are functionally unreliable (relatively speaking), adapt to situations they were not explicitly trained to handle, and fail in predictable ways

- procedures do not have behavior, they specify behavior, hence the notion of functional reliability does not apply; they do not adapt to situations they are not explicitly designed to handle, and they fail in unpredictable ways.

Openness An *open system* is a system whose constitution or behavior is comparatively affected by the environment. In constrast, a *closed system* is one whose constitution or behavior is relatively unaffected by the environment. For example

- Computer communication subsystems may be closed or open:
  - Communication in a computer network connected by appropriately shielded cables is relatively unaffected by the location of other objects in the space, by temperature, by light, by radio signals and by electromagnetic fields.
  - Communication in a network connected by infrared sensors is affected by the location of other objects (it is "line-of-sight"), and by the presence of other infrared radiation such as that generated by spotlights.

  - Communication in a network connected by radio is relatively unaffected by the location of other objects except for building structures which are relatively radio-opaque, and is highly affected by the presence of other radio signals, of which there are many.

  A communication subsystem connected by cabling is therefore relativel closed; one connected by radio or infrared is relatively open

- A pressure tank subsystem of a chemical plant may be affected by the ambient temperature, but this may also be suitable controlled by a cooler which belongs to the system. Else, it is affected mainly by the inflows and outflows, which are part of the overall system, and which may be regulated within certain specified limits. The pressure tank may be adversely influenced by bombs, nuclear explosions, earthquakes (to some extent) and large-scale plant fires, and little else. It is a relatively closed system.

- An aircraft in flight is significantly affected by the motions of the air mass it is flying through, and by the presence or absence of terrain and other non-gaseous physical objects in its flight path. It is a relatively open system.

A Proposal Connecting Hazard Definition And System Properties
It is worth considering if the appropriate definition of hazard for a safety analysis of a system can partly be determined from system properties. For example, Hazard-1 was developed in the commercial nuclear and chemical industries, both of which deal with relatively closed systems. In a relatively closed system, it makes sense to focus on the system state as the major component of a risky situation, since the system state is the major factor affecting subsequent behavior. In contrast, many aviation and other safety analysts seem to prefer Hazard-2, in which properties of the environment are singled out as the significant contributors to a hazardous situation. In a case in which system behavior is largely affected by the environment state, this focus seems to make sense.

6.2 Causality

Formal System Descriptions Suffice to Define Causality We have construed systems as consisting of objects with behavior, and have described in Section 1 how we may consider these formally:

- system state is described through state predicates
- state predicates can be written in, say, a first-order logical language
- behavior is construed via a discrete unending sequence of states
- temporal operators such as $\Box$ and $\Diamond$ are used to make assertions about future states in behaviors
- states can be considered to be *near* and *far* from each other
- altenatives behaviors can be considered to be *near* and *far* from each other

Construing systems in this way allows us precisely to define causality. We shall now see how.

6.2.1 Hume

Hume's Second Definition David Hume gave two definitions of causality over 200 years ago. Here is his second.

> ....we may define a cause to be *an object, followed by another, and where all the objects similar to the first are followed by objects similar to the second.* Or, in other words *where, if the first object had not been, the second never had existed.*

[Hum75, Section VII, Part II, paragraph 60].

We may consider the word '*object*' to refer also to events, maybe states, as noted in the work of John Stuart Mill [Mil73].

David Lewis notes [Lew73a] that there are two definitions given by Hume, and over the course of the subsequent couple of hundred years, the consequences of these notions has been explored. The first definition, in terms of observable regularities, leads to a psychological explanation of causality and is of less interest for our purposes. The second definition, above, is *counterfactual* – it talks of what might have been but was not.

6.2.2 The U.S. Air Force

This is what the U.S. Air Force says about accident explanations [Uni94]:

> **3-11. Findings, Causes, and Recommendations.** The most important part of mishap investigation is developing findings, causes and recommendations. The goal is to decide on the best preventive actions to preclude mishap recurrence. To accomplish this purpose, the investigator must list the significant events and circumstances of the mishap sequence (findings). Then they must select from among these the events and conditions that were causal (causes). Finally, they suggest courses of action to prevent recurrence (recommendations).
>
> **3-12. Findings:**
>
> a. Definition. The findings ..... are statements of significant events of conditions leading to the mishap. They are arranged in the order in which they occurred. Though each finding is an essential step in the mishap sequence, each is not necessaily a cause factor......
>
> **3-13. Causes:**
>
> a. Definition. Causes are those findings which, singly or in combination with other causes, resulted in the damage or injury that occurred. A cause is a deficiency the correction, elimination, or avoidance of which would likely have prevented or mitigated the mishap damage or significant injuries. A cause is an act, an omission, a condition, or a circumstance, and it either starts or sustains the mishap sequence.....

In the paragraph defining causes, the counterfactual definition is used.

6.2.3 Lewis

Lewis's Formal Definition of Causal Factor Suppose A and B are state predicates or state changes (which we shall call *events* from now on). David Lewis's definition of causal factor proceeds as follows. *A is a (necessary) causal factor of B* just in case, had A not occurred, B would not have occurred either. This definition is counterfactual. Before we explain the formal meaning of counterfactual expressions, also due to Lewis [Lew73b], we illustrate the definition of causal factor. Consider a system in which there is a programmable digital component which contains a bit, stored in a variable named X. With systematic ambiguity, we shall refer to this bit as X. Suppose the electronics is wired such that, when X is set, a mechanism (say, an interlock) is thereby set in motion. Suppose the interlock has been well enough designed so that it can only be set in motion by setting X. Then X is a causal factor in any setting in motion of the interlock according to the Lewis definition: *had X not been set, the interlock would not have moved.* Furthermore, let us suppose that the digital component is well-designed, so that X can only be set by a specific operation O of a processor to set it, and that this operation is performed by executing a specific program instruction I. Then,

- *had the operation O not been performed, X would not have been set*, and
- *had the instruction I not been executed, the operation O would not have been performed.*

It follows that

- Performance of O is a necessary causal factor in setting X, and
- Executing I is a necessary causal factor in performing O

The Meaning of A Counterfactual Lewis also gives a formal meaning to a counterfactual. The counterfactual *had A not occurred, B would not have occurred* is interpreted as follows [Lew73b]. We have construed the real world as a behavior, and we have a relation of nearness amongst behaviors. Now, in the real world, B occurred, as did A. But we can consider the *nearest behaviors* to the real world in which A did not occur. The counterfactual *had A not occurred, B would not have occurred* is defined to be true (in the real world) just in case, in all these nearest behaviors in which A did not occur, B did not occur either.

The Semantics Applied to the Example We can consider behaviors near enough to the real world such that I was not executed. We're focusing on system predicates and environment predicates of this system, so we may presume that the more of them that are the same, the nearer the states of the alternative behavior are to the real world. It follows that in the nearest behaviors the design and intended operation of the system can be assumed to be identical to its design and intended operation in the real world. For these behaviors, then, in which I was not executed, O was not performed. And in these behaviors in which O was not performed, X was not set. And in these behaviors in which X was not set, the interlock was not set in motion. So consideration of the nearest behaviors shows that the counterfactuals are to be evaluated as true. Consequently, the assertions of causality (or, rather, *causalfactorality*) are true.

A Comment on the Relation of Nearness The relation of nearness between behaviors is ternary, and comparative: behavior B is nearer than behavior C to behavior A. For reasons that we shall not go into here, Lewis's formal semantics for counterfactuals requires that the nearness relation have a certain form. Fix A, then the relation *B is nearer than C to A* is a binary relation between B and C. Lewis's requirement is that this binary relation must be *ordinal*: it must define an order relation. That is:

Comparability every two worlds B and C are comparable: either B is nearer to A than C, or vice versa, or they are both equally near.

Asymmetry if B is nearer to A than C, then it cannot be the case that C is also nearer to A than B

Irreflexivity B is not nearer than itself to A

Transitivity if B is nearer than C to A, and C is nearer than D to A, then B is nearer than D to A

There are two further conditions on the order relation, that it be closed under arbitrary upper bounds and lower bounds, that need not concern us.

The Notion of Causal Factor is Not Transitive Lewis points out that his notion of causal factor is not transitive, that is

- If A is a causal factor of B, and B is a causal factor of C, this does not necessarily mean that A is a causal factor of C.

Since the intuitive idea of a cause is something that propagates through a "chain" of causal factors, Lewis proposes to define "cause" as the "transitive closure" of the relation of causal factor. The *transitive closure* of a relation R is the smallest (or "tightest", most narrowly defined) relation R^* which

- is transitive, that is, if aR^*b and bR^*c, then aR^*c, and
- contains R, that is, if aRb then aR^*b.

Another way of defining the transitive closure is by a recursive definition; it is demonstrable that the two definitions are equivalent for any relation R. The recursive definition is as follows. aR^*b if and only if

1. aRb, or
2. there is a c such that aRc and cR^*b, and
3. aR^*b only if this can be shown by (repeated) application of Rules 1 and 2 above.

We won't concern ourselves further with the notion of transitive closure. It suffices to know that

- there is a purely formal way of obtaining a unique transitive relation from a given binary relation, called the transitive closure, and
- the intuitive notion of "cause" appears to be transitive, so
- we may rigorously define "cause" as the transitive closure of "causal factor".

We shall need the notion of cause when it comes to discussing the "probable cause" of an aviation accident, in Section 8.

6.2.4 Aside: Causality and Computers

Relation Between Instruction and Execution is Causal This example also illustrates that, according to the formal definition, the design of a digital system ensures that the relation between the form of an instruction and and its execution is causal. The instruction I says to increment regiester R. I is executed; R is incremented. Had the instruction not been to increment register R, then R would not have been incremented. Therefore, the form of I, that I is an instruction to increment R, is a causal factor in incrementing R when the instruction is executed.

Debugging is Causal Analysis This observation entails that debugging computer programs is a form of causal analysis. We shall use this observation later to motivate a method, *Why-Because Analysis*, of causal analysis of complex system failures. One can consider it akin to 'debugging' complex systems. Not only by analogy, but formally.

Chapter 7

Causal Analysis of a Pressure Tank

We have described the ontology of objects and behaviors, states and state predicates, which we use to describe systems and their behavior. We have also introduced intuitively the notion of nearness of behaviors and states, and a formal notion of causality, which can be used with this ontology and the notion of nearness. We claim that the notion of causality is central to system analysis and we demonstrate how by means of an example of a pressure tank in this section.

7.1 Basic Concepts: Object, Properties, Relations

The Pressure Tank The simple pressure tank is shown in Figure 7.1. It contains three input streams, for steam, hydrocarbon and catalyst, on the left. Each stream is controlled by a valve. The tank itself has a pressure sensor, shown above the tank, not currently connected to anything. It contains three output streams, one for the normal output of the product and two vents.

The Accident The accident for this analysis is defined to be an explosion of the tank due to overpressure.

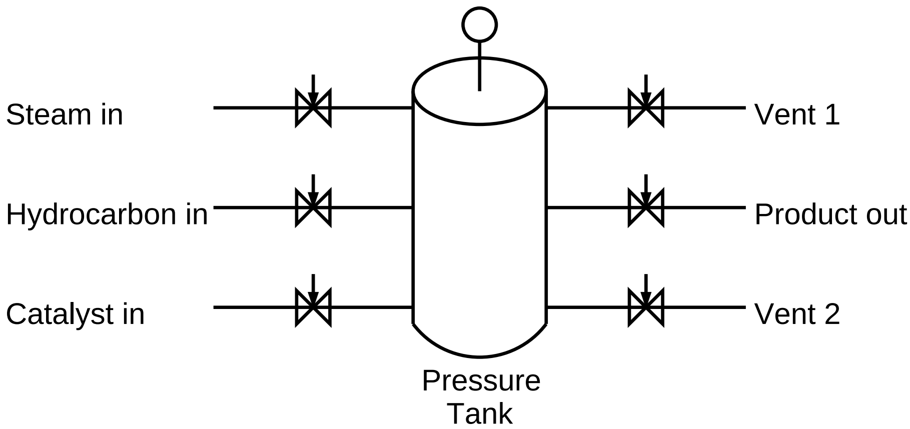


Figure 7.1: The Pressure Tank Without Safety Mechanisms

Safety Analysis Levels On the level on which the design has been given to us, we can specify certain properties and predicates amongst the components of the system. Examples include the quantity, temperature and pressure of steam, hydrocarbon, catalyst, and product; the open/closed states of the valves and maybe even which components (tubes, tank, valves) are fulfilling their specification (which we are not given) and which not. There are other components, such as joints, screws, surface coatings, controlled climate, and so on, which we are not given. We cannot therefore assess the state or behavior of these components, although this might be a significant factor in any real accident behavior. This is why we speak of *levels of analysis*. One cannot infer anything about things one is not given, or properties one is not made aware of. It is essential, however, to determine precisely what one can know and what one cannot know, and assign the latter to a different stage of analysis.

Objects We have quite a few objects, even for so simple an example.

- *Tank*
- *SteamPipe*
- *HCPipe*
- *CatalystPipe*

- *ProductOutPipe*
- *VentPipe1*
- *VentPipe2*
- *TankPressureSensor*
- *SteamPipeValve*
- *HCPipeValve*
- *CatalystPipeValve*
- *ProductOutPipeValve*
- *VentPipe1Valve*
- *VentPipe2Valve*
- *Steam*
- *HC*
- *Catalyst*
- *Product*

Properties The following properties pertain to certain objects.

- *Intact* and its contrary *Ruptured*, to *Tank*, *SteamPipe*, *HCPipe*, *CatalystPipe*, *ProductOutPipe*, *VentPipe1*, *VentPipe2*;
- *Open*, *Closed* and *Partopen*, to *SteamPipeValve HCPipeValve CatalystPipeValve ProductOutPipeValve VentPipe1Valve VentPipe2Valve*;
- *Temperature*, *Pressure*, *Quantity*, to *Steam HC Catalyst Product.* Although we have called these properties, in fact they are fluents, taking values; different values at different times.

7.2 Causal System Analysis (CSA)

Formal Definition of Accident The accident may then be defined as

$$Ruptured(Tank)$$

What Can Cause The Accident? In this case, we are lucky that the causal antecedents to the accident at this level of analysis are fairly resticted. Indeed, it is a goal of, and a criterion for, a hierarchical division into safety analysis levels that each analysis level allows one to delimit the causal antecedents to events and system states.

A rupture in the tank can only occur if the tank is breached from outside, or if there is a sustained overpressure in the tank above a certain level. This is a causal statement. If we rule out the breach, and an accident occurs, then *the accident would not have happened had there not been overpressure for a particular time in the tank.*

We use the symbol "$\Rrightarrow$" to denote "*is a causal factor of*". The causal relation between accident and condition we can thus write:

$$Pressure(Tank) > N\,Units\ over\ time\ T \Rrightarrow Rupture(Tank)$$

In fact, it is much more reasonable to consider the certainty of occurrence of the accident to be a function, not of simple overpressure for fixed time, but as some function of overpressure and time that is monotonic in both arguments. There is probably some overpressure value N under which the tank would rupture instantaneously, but much more likely is a sustained smaller overpressure. Nevertheless, in order to indicate how a condition may depend on time, without complicating the argument, we consider overpressure above a fixed value over a fixed time interval. The reader should keep in mind, however, that this is a simplification.

Hazard Condition We can thereby intuitively designate $Pressure(Tank) > N$ to be a hazard condition. An accident is not inevitable provided that the pressure is reduced inside a particular time. But that the pressure has been greater than N for some time increases the chances that an accident will occur. The argument is as follows. It rests on certain assumptions, called *stasis* and *temporal strengthening*, which are debatable and by no means universally true without conditions.

Suppose $T = t + s$, that both t and s are non-zero, and the pressure has already been greater than N for time s. Call this the precondition.

- Then the chances that the pressure will continue to be greater than N are equal to or greater than if the the pressure had not been greater than N in previous time. This we call *stasis*.

- The condition for an accident to occur, given the precondition, is that $Pressure(Tank) > N for\ time\ t$. Since t is less than T it follows from temporal logic that $Pressure(Tank) > N for\ time\ T$ tense-logically implies $Pressure(Tank) > N for\ time\ t$ but not vice versa. If A tense-logically implies B but not vice versa, we assume that the a priori probability of B is higher than that of A. This assumption is called *temporal strengthening*.

The chances that an accident will occur given the precondition are thus at least the a priori probability that $Pressure(Tank) > N for\ time\ t$ (and possibly raised by stasis). Temporal strengthening says that this is greater than the a priori probability that $Pressure(Tank) > N for\ time\ T$, which is the a priori probability that an accident will occur tout court, given the causal dependence of the accident on this condition.

It follows that the accident is more likely to occur, given the precondition. Therefore the precondition is a Hazard-3.

A simple argument that the precondition or some transformation of it fulfils one of the other hazard conditions seems difficult to obtain. But the assumptions of stasis and temporal strengthening are crucial even to the argument that the overpressure condition is a Hazard-3.

Causal Factors of the Hazard The hazard condition is unusual in that there is just one condition which leads to an accident. We now inquire about the causal factors of the hazard condition.

Knowledge of the gas laws tells us that the pressure in the tank is a monotone increasing function of the quantity of the product $Quantity(Product)$ and the temperature of the product $Temperature(Product)$. "*Monotone increasing*" means that the value increases with each increase in each argument. Let us make the further assumption (which must be justified through chemical knowledge), that the pressure of the product rises as the hydrocarbon and steam convert into desired product. Thus the pressure of the product for given inputs and temperature is itself an increasing function of time:

$$Pressure(Tank) = F(Quantity(Product), Temperature(Product), time)$$

We are not concerned with the exact form of F, just in knowing that it is monotone increasing with its arguments. We may summarise this causally as

$$Quantity(Product) \Rightarrow\!\!\succ^{+,t} Pressure(Tank)$$

$$Temperature(Product) \Rightarrow\!\!\succ^{+,t} Pressure(Tank)$$

The superscript indicates the monotonic increasing dependency of values, as well as the *hysteresis*, the lag in time of the effect.

Discrete Factors and Value-Influence Factors The simple counterfactual definition of "$\Rightarrow\!\!\succ$" talks about the presence or absence of factors. We call such factors *discrete factors*, for which it makes sense to talk about their presence or absence *simpliciter* in a behavior.

We have moved from a simple counterfactual definition of causality to describing a causal tendency:

- not only that one extensively-measurable state predicate (or fluent, as we have called it) is a causal factor in another extensively-measureable state predicate, but
- that the measurements depend upon each other in a certain way: namely monotonically increasing or decreasing, or threshold-triggered, or time-triggered.

We call such causal factors *value-influence factors*. We assert here without further argument:

- that these specific four features may be brought within the counterfactual definition in a straightforward way, for example
- we have shown by example how time-triggering may be handled in our discussion of the condition $Pressure(Tank) > N for\ time\ T$ above, and
- these qualitative features of quantitative causal regularities are (with maybe some others) all that is needed for an adequate causal analysis for safety purposes.

This last point can be taken to suggest that so-called Qualitative Physics, as studied for example under "Common-Sense Physics" by AI researchers, can have a role to play in the future in adequate causal analyses for safety. This field is still quite young, however.

Following Causality Backwards We now consider the causal factors of the fluent $Quantity(Product)$. Through simple chemistry, these are $Quantity(Steam)$ and $Quantity(HC)$. Furthermore, $Quantity(Product)$ is monotonic increasing in these values. $Quantity(Catalyst)$ remains unchanged and does not contribute – this is the property of a catalyst. Thus

$$Quantity(Steam) \Rrightarrow^{+,t} Quantity(Product)$$

$$Quantity(HC) \Rrightarrow^{+,t} Quantity(Product)$$

From now on, we shall say that a quantity is *positively causally dependent* on another if the first is causally dependent on the second, and if this causal dependency is monotonically increasing. Similarly, we shall say that a quantity is *negatively causally dependent* on another if the first is causally dependent on the second, and if this causal dependency is monotonically decreasing.

Boyle's Law of gases tells us that, for fixed volume, such as contained in the inside of a pressure vessel, the pressure rises with the temperature. If the chemical reaction is *exothermic*, the temperature of the product is positively causally dependent on the quantity of reactants (steam and hydrocarbon). If the reaction is endothermic, the causal dependency is negative. Let us assume the reaction is exothermic. Then we have

$$Quantity(Steam) \Rrightarrow^{+,t} Temperature(Product)$$

$$Quantity(HC) \Rrightarrow^{+,t} Temperature(Product)$$

and of course what goes in must come out, so the temperatures also show a positive causal dependency, but without hysteresis:

$$Temperature(Steam) \Rrightarrow^{+} Temperature(Product)$$

$$Temperature(HC) \Rrightarrow^{+} Temperature(Product)$$

$$Temperature(Catalyst) \Rrightarrow^{+} Temperature(Product)$$

7.3 The Causal Influence Diagram

The Causal Influence Diagram (CID) We can represent the causal influences we have derived so far as a graph, which we call a *Causal Influence Diagram* (CID).

7.3.1 Generating the CID from CI-Script

Software cid2dot We have software *cid2dot* which automatically builds a CID from a specification in *CI-Script*, using the *dot* graph-layout tool from AT&T Research. Figure 7.2 shows the CI-Script for the analysis we have performed so far.

```
[0] /* Ruptured(Tank) */
    [1] /* Pressure(Tank) > N // +, TIME */
[1] /\ [-.1] /* Quantity(Product) // +, TIME */
    /\ [-.2] /* Temperature(Product)  // +, TIME */
    /\ [-.3] /* Fixed Volume V units */

  [1.1] /\ [-.1] /* Quantity(Steam) // +, TIME */
        /\ [-.2] /* Quantity(HC) // +, TIME */

  [1.2] [-.1] /* Quantity(Steam) // +, TIME */
  [1.2] [-.2] /* Quantity(HC) // +, TIME */
  [1.2] [-.3] /* Temperature(Steam) // + */
  [1.2] [-.4] /* Temperature(HC) // + */
  [1.2] [-.5] /* Temperature(Catalyst) // + */
```

Figure 7.2: The CI-Script for the Pressure Tank CID

The resulting CID is shown in Figure 7.6. Because the labels are somewhat obscured (we have not finished modifying the code from its use for generating WB-Graphs yet), we include another version generated by *dot* from hand-prepared input. This version, which is intended to be identical, but with the labels drawn felicitously on the causal relations (arros) instead of obscurely in the nodes, is shown in Figure 7.7.

7.3.2 Analysing the CID

Conditions Derived From the Meaning of Causal Factor The CID shows the causal influences on the processes in the pressure tank at this System Level which lead to an accident. There are two consequences of the fact that the causal conditions are all necessary conditions, demonstrable from the meaning of "$\Rrightarrow$":

discrete factors removing any one of them will lead to avoidance of an accident;

value-influence factors decreasing any one of the monotone-increasing influences in sufficient quantity will lead to amelioration of the conditions causing the accident

Removing a single discrete factor will avoid the accident. However, it is not enough simply to reduce the value of a value-influence factor by itself to avoid the accident, because the lowest value to which one can reduce the factor *simpliciter* may not be enough to avoid the accident by itself, given the unaltered values (of value-influence factors) or presence (of discrete factors) of other factors. In this case, one may have to consider reducing the value of multiple value-influence factors in order to avoid the accident.

How To Proceed We work backwards from the accident through the graph in the reverse direction of the causal arrows. The motivation for this process is that seeing how one can ameliorate the immediate causal factors of an accident is the most direct form of avoiding the accident that presents itself.

The Obvious Top Condition We proceed therefore by considering whether we can ameliorate $Pressure(Tank) > N$ *simpliciter*. We cannot, because it is a value-influence factor, hence we have to look at its causal determinants. These are

- *Fixed Volume V units*
- *Quantity(Product)*
- *Temperature(Product)*

We observe that *Fixed Volume V units* is a discrete factor. We can remove it by changing the value of V. But by Boyle's Law, volume is a value-influence factor of pressure, so we cannot ameliorate the accident simply by picking any old value of V.

Changing Volume According to Boyle's Law, the volume V is a negative value-influence factor. Accordingly, we can consider increasing V appropriately. We can do this, for example, by opening either *Vent1* or *Vent2*. Let us build in a mechanism to do this:

- we put *Vent1* under computer control from the pressure sensor in the tank top;
- we put *Vent2* under human operator control; inform the human operator of the pressure via a warning signal (a discrete overpressure warning, or simply a pressure reading dial); and put procedures in place for the operator to open *Vent2* under suitable states of the indicators.
- we ensure that this measure *by itself* is sufficient to increase the volume enough to remove the factor $Pressure(Tank) > N$.

We have then designed the system in Figure 7.3.

7.3.3 Analysing The Modified System

The CID The CI-Script for the modified system is shown in Figure 7.4, and the CID thereby generated in Figure 7.8. Again, a *dot* version from hand input is shown in Figure 7.9.

Ameliorating the Factors Reconsidered We have introduced two new discrete factors into the CID, namely *Closed(Vent1)* and *Closed(Vent2)*. So, concentrating on the discrete factors leads us to remove these factors as a way of ameliorating the hazard condition. This will work, but leads us to a new accident analysis.

Causal Analysis of the Valves We have modified the accident scenario, but have not yet performed a full causal influence analysis of the new system. The analysis is specified in CI-Script in Figure 7.5, and the CID generated is

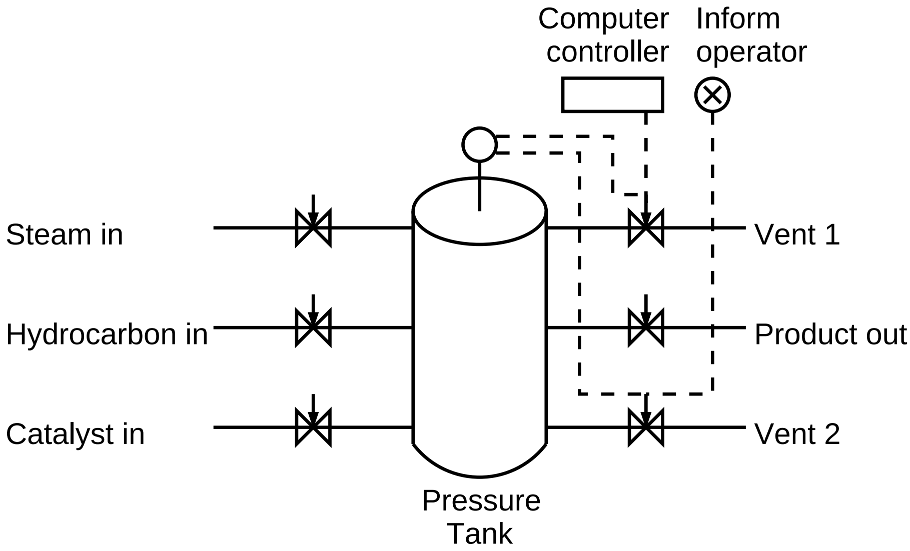


Figure 7.3: The Modified Pressure Tank

shown in Figure 7.10. A hand-prepared-input *dot* version is shown in Figure 7.11.

The vent-subsystem causal analysis shown in the Causal Influence Diagram is not a causal analysis of an accident, as the other diagrams were. It shows the normal causal operation of the vent subsystem, which is a safety subsystem.

The Safety Subsystem Function Fulfils Its Purpose It may be seen directly by comparing the CIDs in Figures 7.9 and 7.11 that the vent subsystem fulfils its intended safety function. Place similarly-labelled nodes on top of each other, namely the $Pressure(Tank) > N$ and $Volume$ nodes, and look at the precursors of the volume nodes. Both of them, in both diagrams, involve the objects $Vent1$ and $Vent2$. However, the predicates in the accident CID are contraries of those in the vent-subsystem CID ("contrary" means that it is not possible for them both to be true at the same time of the same object). That means simply that the intended operation of the vent subsystem precludes the situation described in the accident CID; they are incompatible. Since the relevant state predicates are discrete predicates, their falsity ensures that the accident cannot happen, as explained above.

```
[0] /* Ruptured(Tank) */
    [1] /* Pressure(Tank) > N // +, TIME */
[1] /\ [-.1] /* Quantity(Product) // +, TIME */
    /\ [-.2] /* Temperature(Product)  // +, TIME */
    /\ [-.3] /* Fixed Volume V units */

  [1.1] /\ [-.1] /* Quantity(Steam) // +, TIME */
        /\ [-.2] /* Quantity(HC) // +, TIME */

  [1.2] [-.1] /* Quantity(Steam) // +, TIME */
  [1.2] [-.2] /* Quantity(HC) // +, TIME */
  [1.2] [-.3] /* Temperature(Steam) // + */
  [1.2] [-.4] /* Temperature(HC) // + */
  [1.2] [-.5] /* Temperature(Catalyst) // + */

  [1.3] [-.1] /* Closed(Vent1) */
  [1.3] [-.2] /* Closed(Vent2) */
```

Figure 7.4: The CI-Script for the Modified Pressure Tank

Hence the vent-subsystem CID demonstrates visually and directly that the discrete state predicates of the vents, required for an accident to happen, do not pertain when the vent subsystem operates as designs. Ergo, the accident cannot happen.

7.3.4 Causal System Analysis of the Vent Subsystem

From Normal Operation to Failure We have not yet identified improper operation of the vent subsystem. The vent-subsystem CID is a CID of normal operation. The system does not function properly, it fails, precisely when one of the causal arrows is "broken", that is, it is not there in the case of a discrete factor, or it has null or opposite influence if it is a value-influence factor. These may be considered one at a time from the CID, and their causal influence traced, as follows.

- remove the chosen causal link;
- remove all successors of that link up to the point at which another

```
[0] /* Volume */
    /\ [1] /* Open(Vent1) // +, TIME */
    /\ [2] /* Open(Vent1) // +, TIME */

[1] /\ [-.1] /* Command(Open(Vent1)) */

  [1.1] /\ [-.1] /* On(Sensor)  */

    [1.1.1] /\ [-.1] /* Pressure(Tank) > N  */

         [1.1.1.1] [0]

[2] /\ [-.1] /* Operator commands Open(Vent2)  */

  [2.1] /\ [-.1] /* Operator perceives On(WarnLight)  */

    [2.1.1] /\ [-.1] /* On(WarnLight)  */

      [2.1.1.1]  [1.1.1]
```

Figure 7.5: The CI-Script for the CID of the Vents

path combines (i.e., up to the first point at which there are two or more in-arrows to a node;

- place the resulting CID "over" the accident CID as before and see if they are consistent;
- if they are not consistent, the failure does not result in an accident; if they are consistent, this failure allows the accident

For example, if the arrow between node `[2.1.1.1] On(WarnLight)` and node `[2.1.1]: Operator perceives On(WarnLight)` is "broken", then the chain from here forwards to the next joint with another chain, at the `Volume` node, must be removed. This chain is indicated by the dashed lines in Figure 7.12. After removal, the CID is shown in Figure 7.13. Note that the other chain remains: `Vent1` will still open, volume will be increased, pressure reduced. When this modified CID in Figure 7.13 is placed "over" the accident CID, the nodes `Open(Vent1)` and `Closed(Vent1)` still contradict. The causal link we removed represents the case in which an operator did not perceive the warning light. Heshe did not thereby act to open `Vent2`.

It is easy to see that removing any single arrow from `Volume` backwards renders the vent-subsystem CID still incompatible with the accident CID. Hence the modified pressure tank system is immune to single-point failures of the vent subsystem.

One Must Consider Multiple "Breaks" The previous operation only dealt exhaustively with single failures of the vent subsystem. One must remove arrows two at a time, three at a time, and so forth in general to obtain a complete analysis. However, from the form of the graph, it is easy to see what those consequences will be. Any pair of arrows removed, one from each parallel chain, will remove both `Open(Vent1)` and `Open(Vent2)` and the resulting diagram will be compatible with the accident CID.

It is easy to see that a pair of arrows removed from both chains is both necessary and sufficient to render the vent subsystem compatible with the accident CID, by the "placing over" test.

The safety analysis thereby explicitly produces a general condition both necesary and sufficient for the vent subsystem to be compatible with an accident. One cannot always expect such an analysis to be so clean - this is an example, after all. But certain features stand out:

- it is easy to see how to perform an exhaustive analysis, even though the combinatorics might not always be so felicitous;
- it is easy to check that one's analysis has been exhaustive; since this is merely a graph-theoretic counting exercise;
- it is visually much easier to check one's reasoning than, say, to check a fault tree.

A Comparison With Fault Tree Analysis For comparison, to substantiate especially the last point above, a fault tree from [Lev95, Figure 14.5, p331] for this system is shown in Figure 7.14.

Fault Tree Analysis (FTA) is not based upon a formal notion of cause. So there is no means of checking its correctness except through informal inspection by experienced practioners. The advantage of FTA seems to be threefold:

- to perform an FTA, one has to inspect a system and its components thoroughly. Any thorough inspection is bound to help highlight inadequacies in system design, including safety inadequancies.
- FTA has developed graphical methods of handling system decompositions into components, which enables one to combine FTAs performed independently over an adequate decomposition.
- There is long experience with FTA, and its strengths and weaknesses are known, as well as a "library" of individual ways of handling specific cases which can be drawn on by other users

These advantages are not to be sneezed at. However, the advantages of basing an analysis method on a formal notion of causal factor are also important:

- one has a formal criterion for correctness;
- it is in principle possible to develop criteria for completeness;
- although CIDs have to be constructed by hand by analysts, in principle checking them against each other can be automated, since it is based on logical consistency methods

That is, the use of the formal notion of cause means that correctness checking and analysis of safety mechanisms such as we have described can be automated. This cannot be done with FTA.

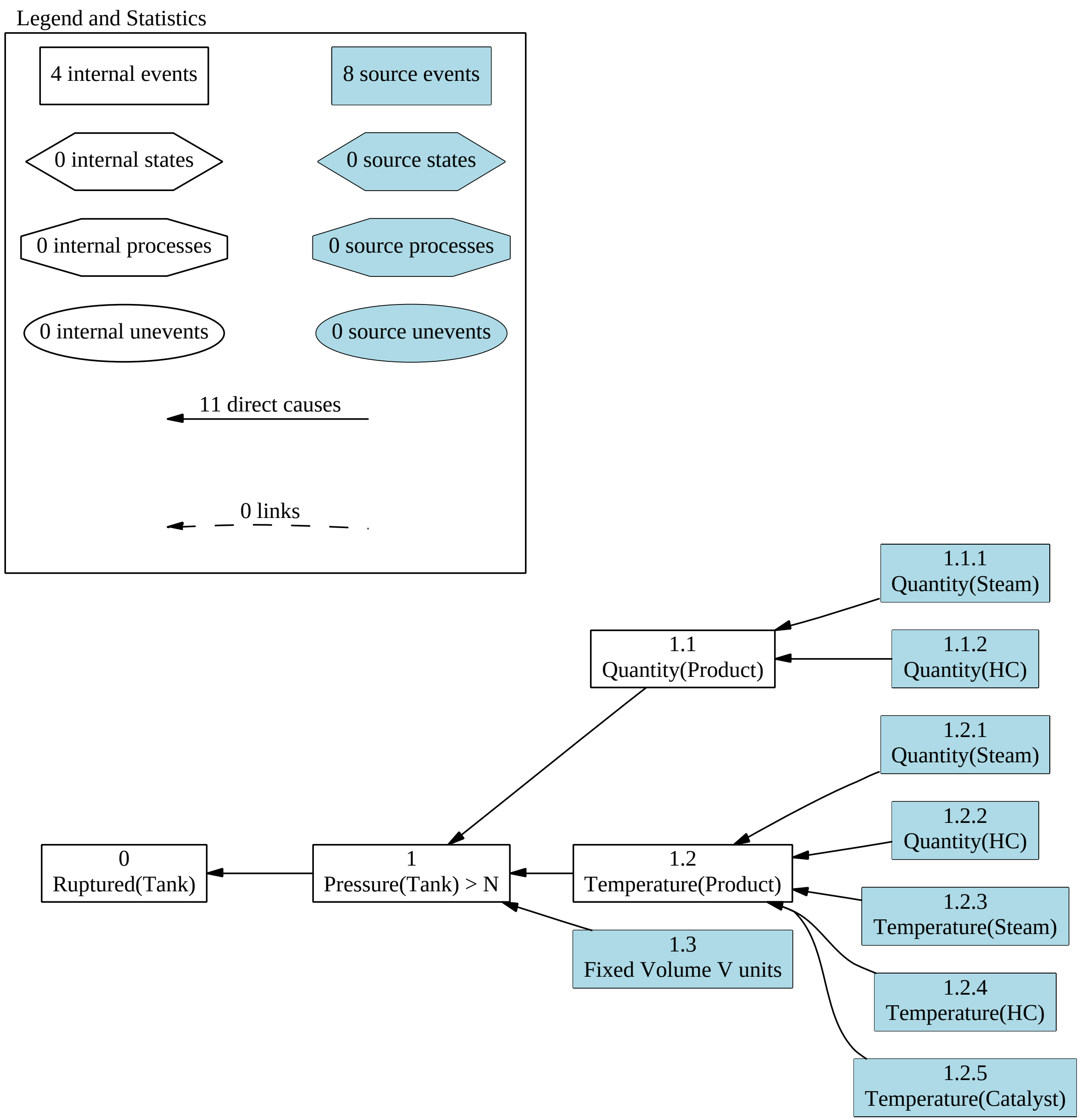


Figure 7.6: The CID for the Pressure Tank

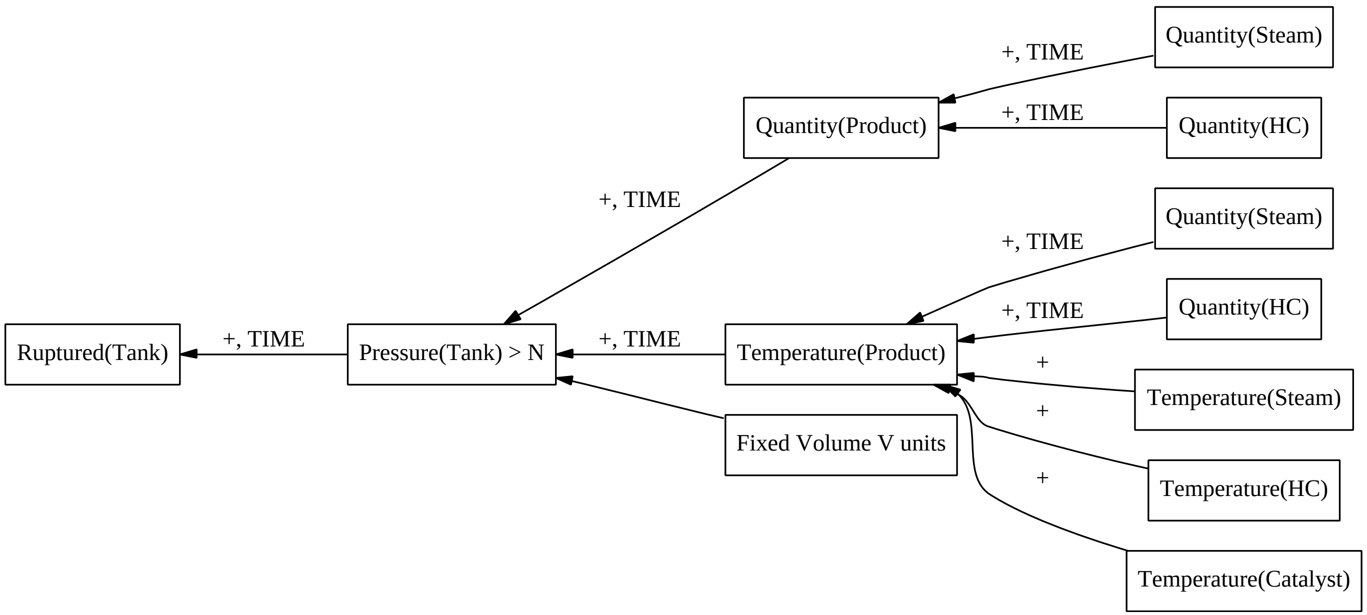


Figure 7.7: The CID for the Pressure Tank: Version 2

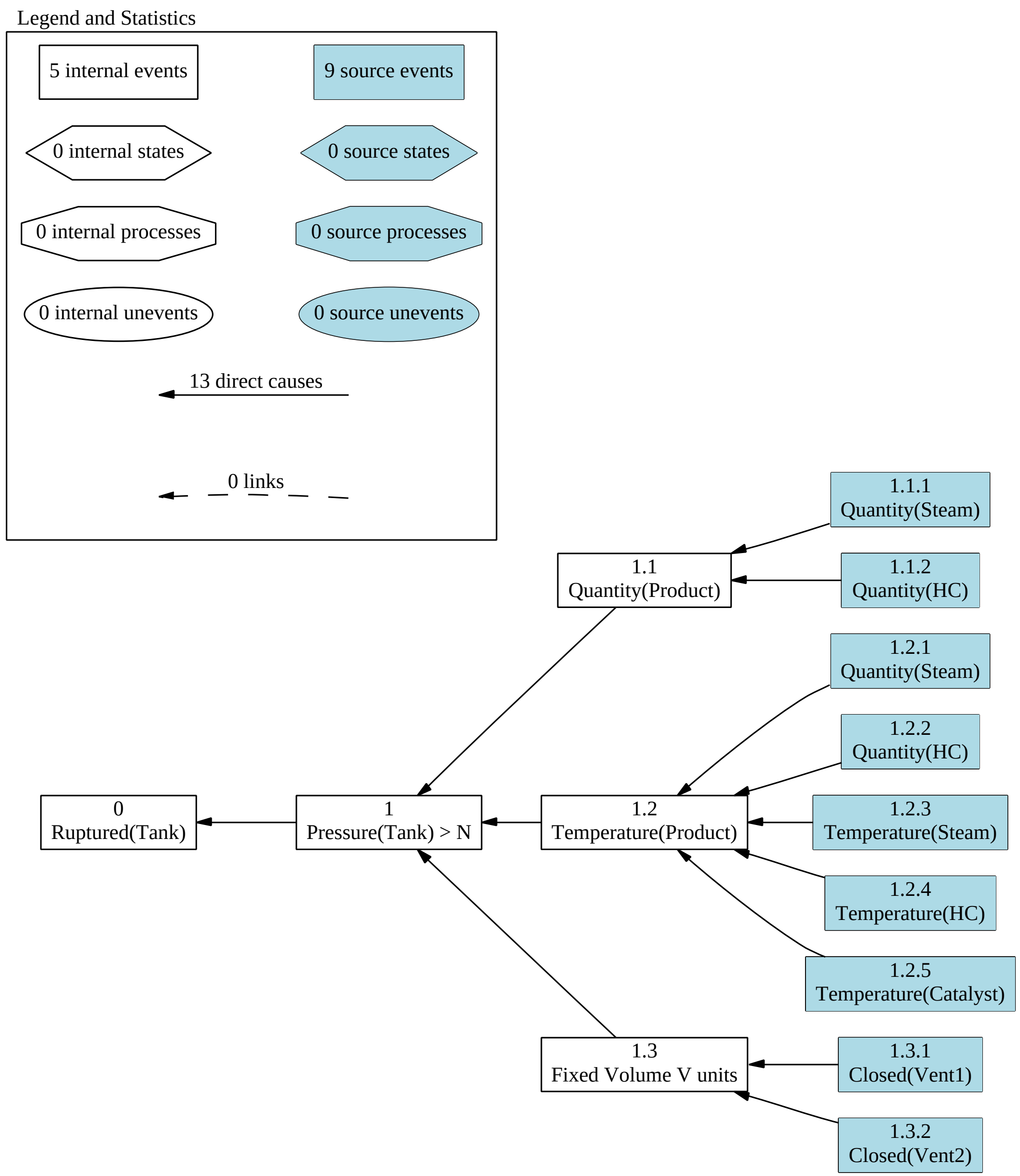


Figure 7.8: The CID for the Modified Pressure Tank

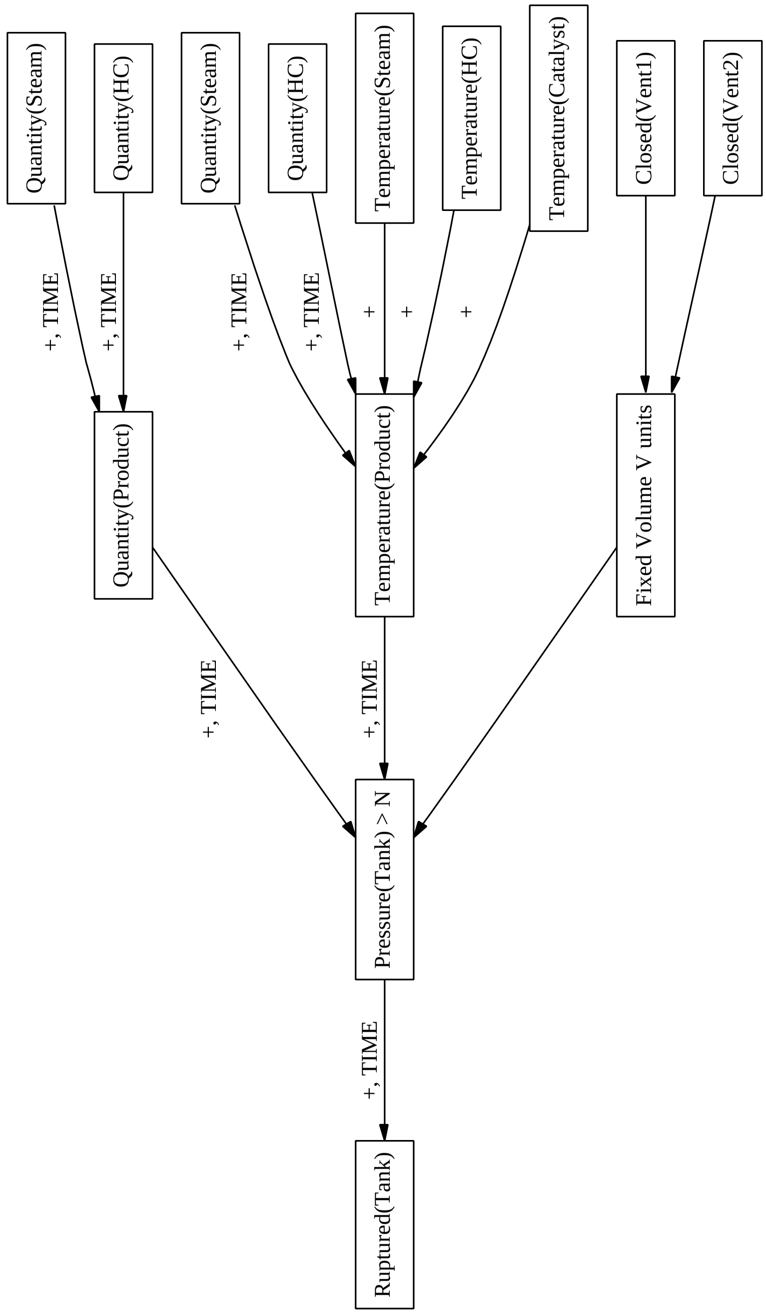


Figure 7.9: The CID for the Modified Pressure Tank: Version 2

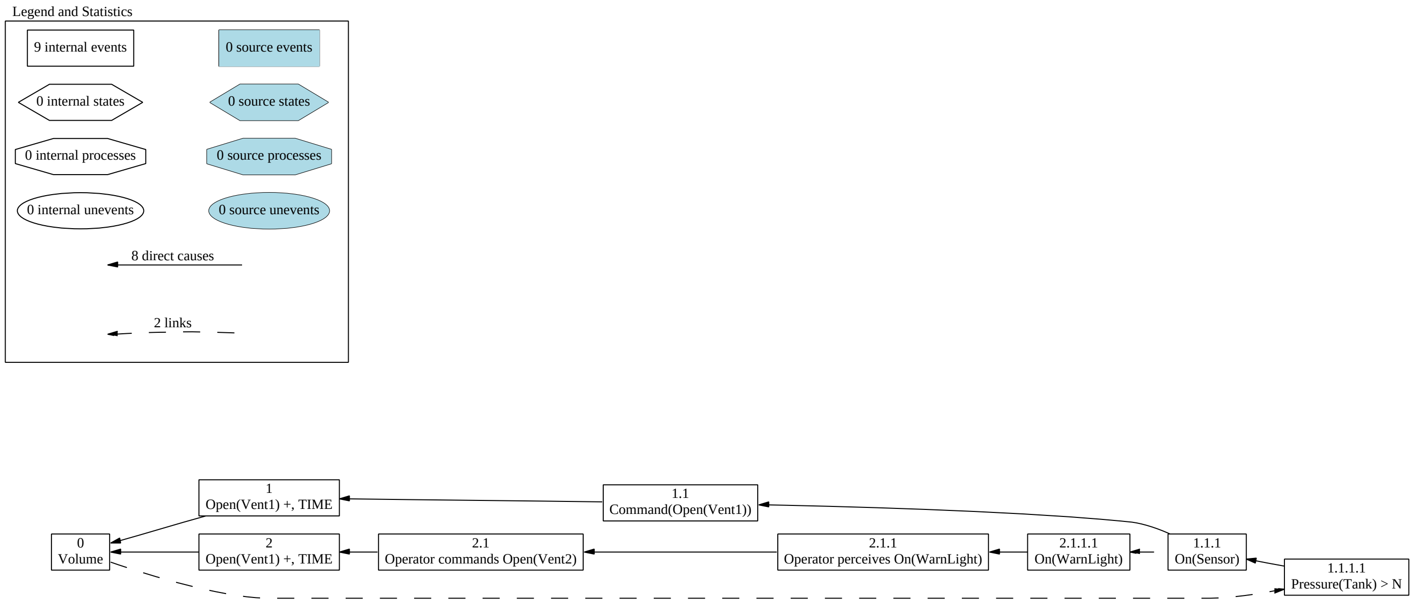


Figure 7.10: The CID for the Vents

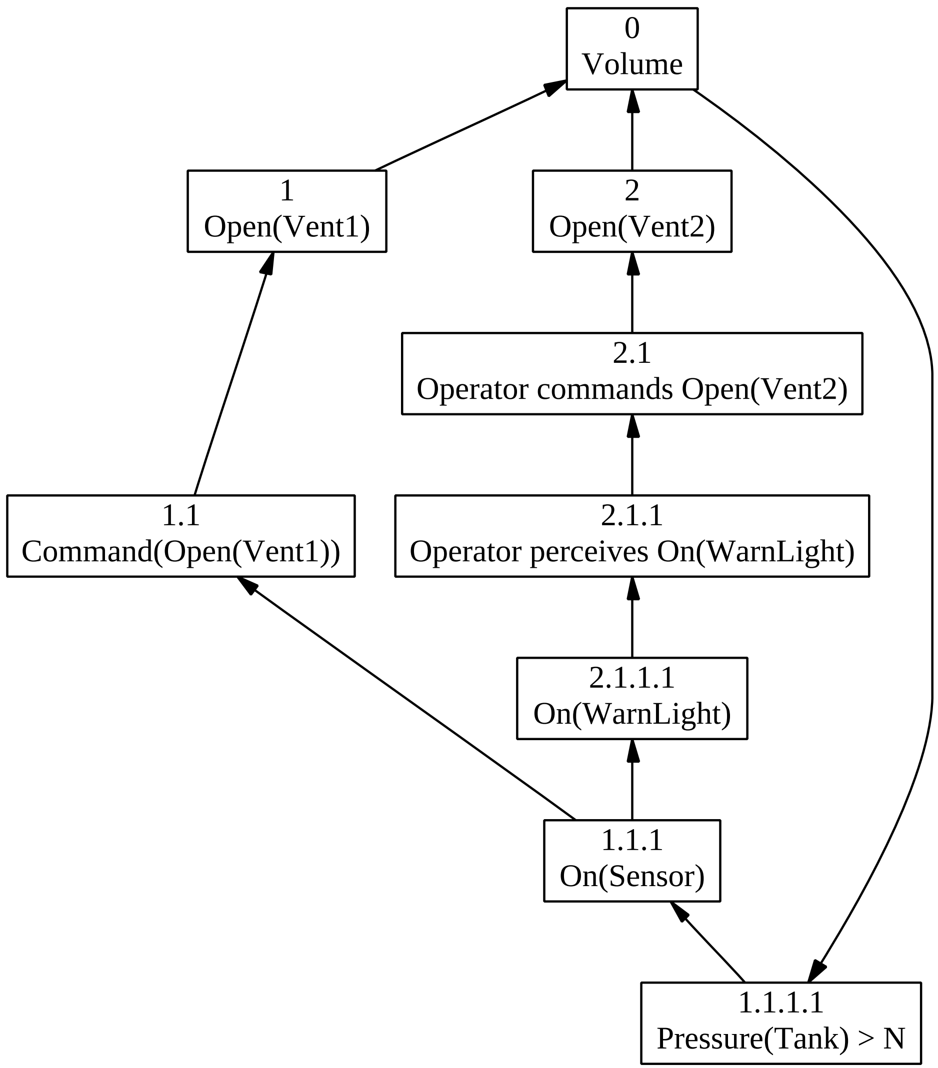


Figure 7.11: The CID for the Vents: Version 2

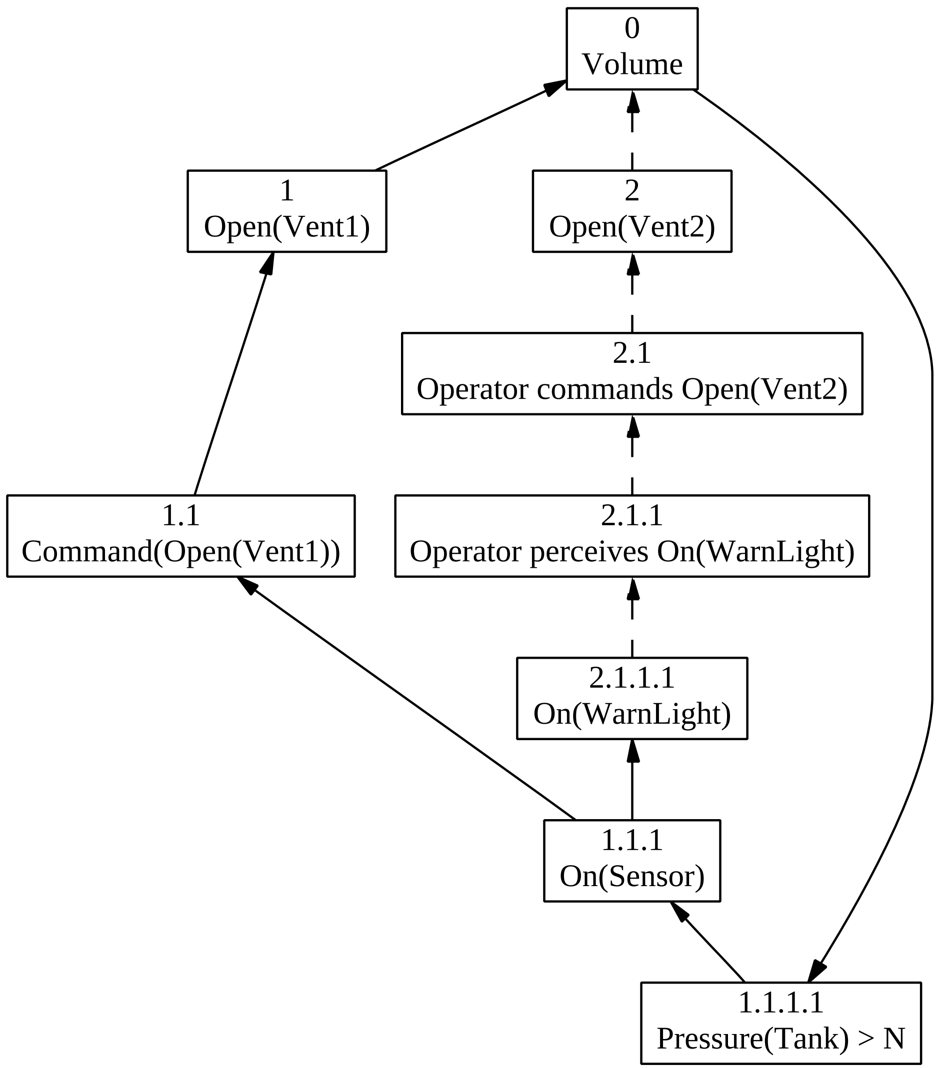


Figure 7.12: Removing a Causal Chain After Breaking a Link

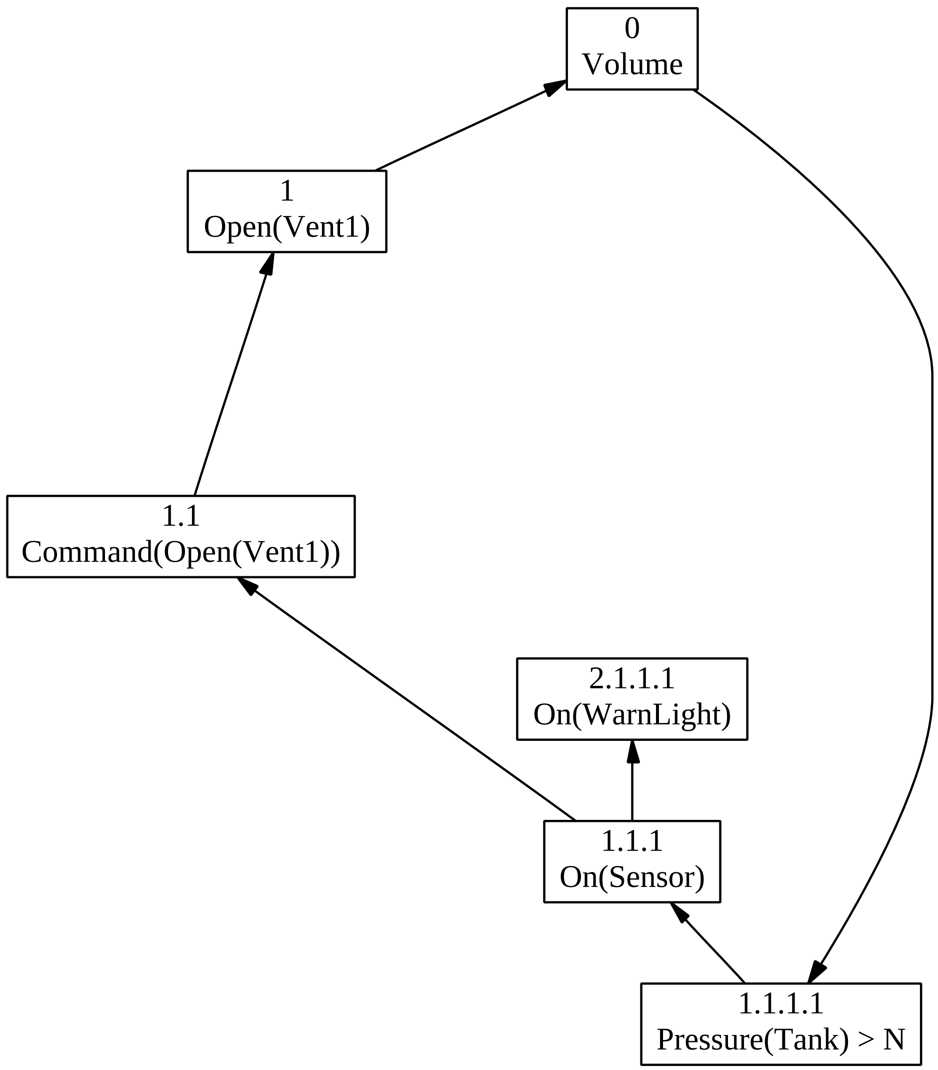


Figure 7.13: The CID of the Vent Subsystem After Breaking a Link

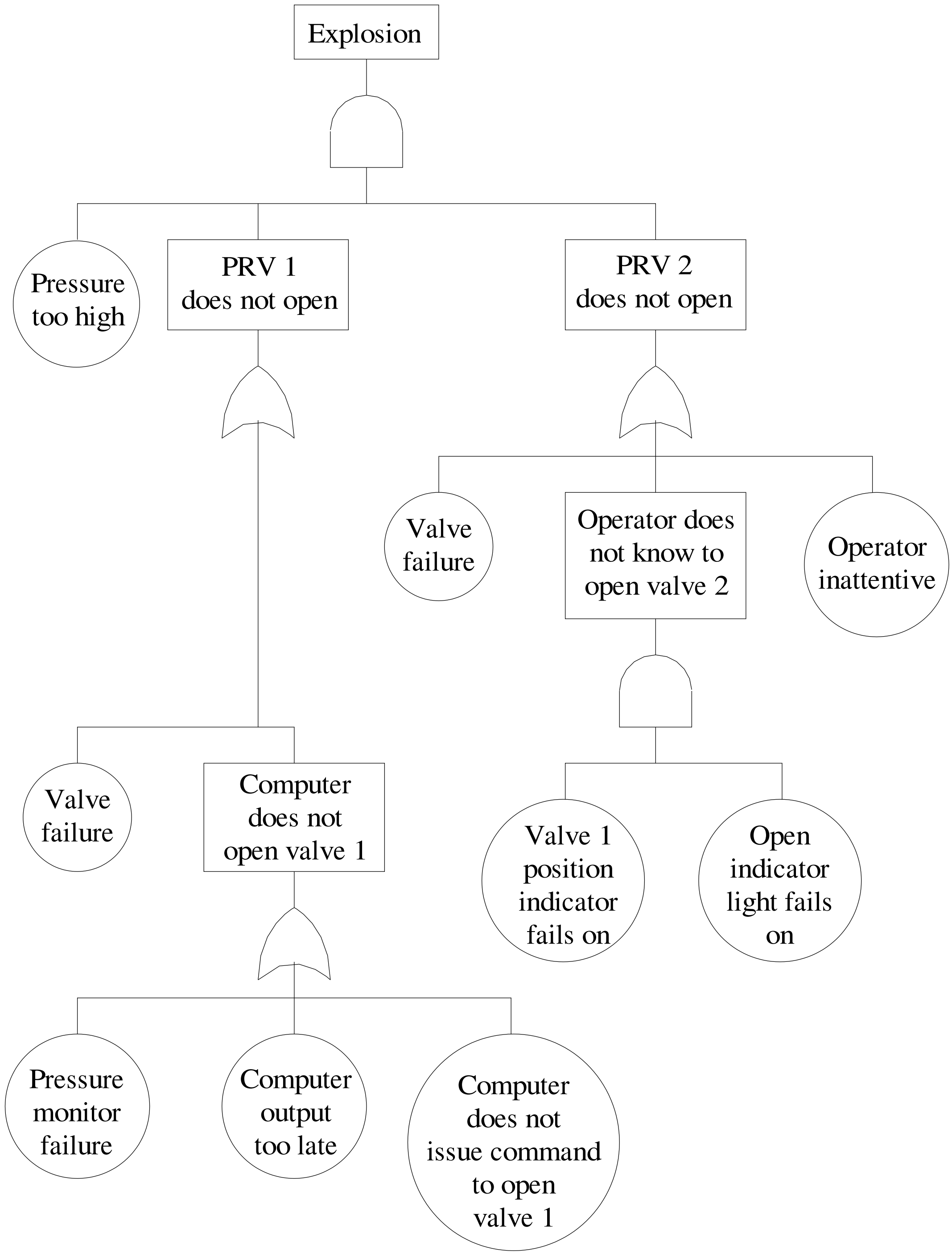


Figure 7.14: The Fault Tree for the Pressure Tank

Chapter 8

Accident Analysis: Why-Because Analysis

Because Why-Because Analysis (WBA) has been described elsewhere [LL98], we will skim briefly over its features and advantages here.

WBA is Causal Influence Analysis With Discrete Factors An accident has already happened; there is thus a history to discover and analyse. All fluents took on particular values, and some of these values and trends had causal effects on others. But the history is given: there are no alternative behaviors to consider. Therefore, all factors in the CID, here called a Why-Because Graph or WBG, are discrete factors.

Prophylactic Measures Become Easier to Determine Because all factors are discrete, prophylaxis consists therefore choosing a factor or factors and deciding what action(s) to take to ensure that these factors are omitted in all future similar cases.

A WB-Analysis of the 1993 Warsaw A320 Accident On 14 September 1993, the Lufthansa Airbus A320 "Kulmbach" landed in Warsaw during a heavy rainstorm. It overran the runway, hit and overran an earth bank, and burned. Two people died, one pilot from trauma and a passenger, asphyxiated while unconscious. The accident report may be read in [Lad]. The points mentioned in this section are drawn from a longer paper [HL98]. The report cited probable cause as follows:

> Cause of the accident were incorrect decisions and actions of the flight crew taken in situation when the information about windshear at the approach to the runway was received. Windshear was produced by the front just passing the aerodrome; the front was accompanied by intensive variation of wind parameters as well as by heavy rain on the aerodrome itself.
>
> Actions of the flight crew were also affected by design features of the aircraft which limited the feasibility of applying available braking systems as well as by insufficient information in the aircraft operations manual (AOM) relating to the increase of the landing distance.

Making a WBG We arranged (with Michael Höhl) the states and events described in the official accident report into a WB-Script (a CI-Script for WB-Graphs). The WBG that resulted is shown in Figure 8.1.

Focusing In on Factors We can focus on the upper portion of the graph, where it narrows down to one node. This portion is shown in Figure 8.2. It is rare that a WBA of an accident results in a graph with a width of one. What is this single node?

```
AC hits earth bank
```

Take away this node, and you've avoided the accident. What are its immediate precursors?

```
AC overruns RWY
Earth bank in overrun path
```

The report's attribution of probable cause focused entirely on causal factors contributing to the first of these two events. What about the second? Why was there an earth bank in the overrun path? Because

```
Bank built by airport authority for radio equipment
```

Prophylaxis: Don't Overrun Or Don't Build So there is clearly something to consider. Don't build earth banks for radio equipment at the ends of runways in the overrun area. Or don't overrun runways. Well, measures are taken to minimise cases of the latter, but most authorities consider that no matter what one does, aircraft will still overrun runways once in a while. So if you want to prevent or minimise such catastrophic overrun accidents, one had better take the other option and not build in the overrun area.

Leaving Clear Overruns is Just Good Practice In fact, leaving a clear overrun area at the end of runways is regarded not only as good practice but as essential practice by most Western European and US authorities and by practically all pilots.

This Was Omitted from the Report's Conclusions The report's conclusions about probable cause and contributing factors said nothing about building earth banks in overrun areas.

This is Demonstrably A Mistake in Causal Reasoning The WBA of the accident shows clearly that this omission is a mistake in causal reasoning that the report made. The information necessary to infer it was a contributing cause was contained in the body of the report - that is where we obtained the factors in the WB-Graph in Figurefig:Warsaw-WBG. The WBA shows it to be a causal factor.

Rigorous Causal Reasoning Helps This is not the only causal reasoning mistake in the Warsaw report, neither is it the only report in which significant causal reasoning mistakes may be demonstrated by WBA. Another, the report on the 1995 American Airlines B757 accident on approach to Cali, Colombia is one, which also omits demonstrably causal factors in its statement of probable cause. The omitted factors in that report were, however, taken into account by the U.S. National Transportation Safety Board in their letter to the U.S. Federal Aviation Administration containing their safety recommendations based on their analysis.

Using rigorous methods of causal reasoning such as WBA would thus help considerably in ensuring correctness of these important reports. Prophylactic measures are based on the reports' analyses. It is important to reduce future accidents that resources be pointed in the appropriate directions, and one can only do this if a report's reasoning is correct.

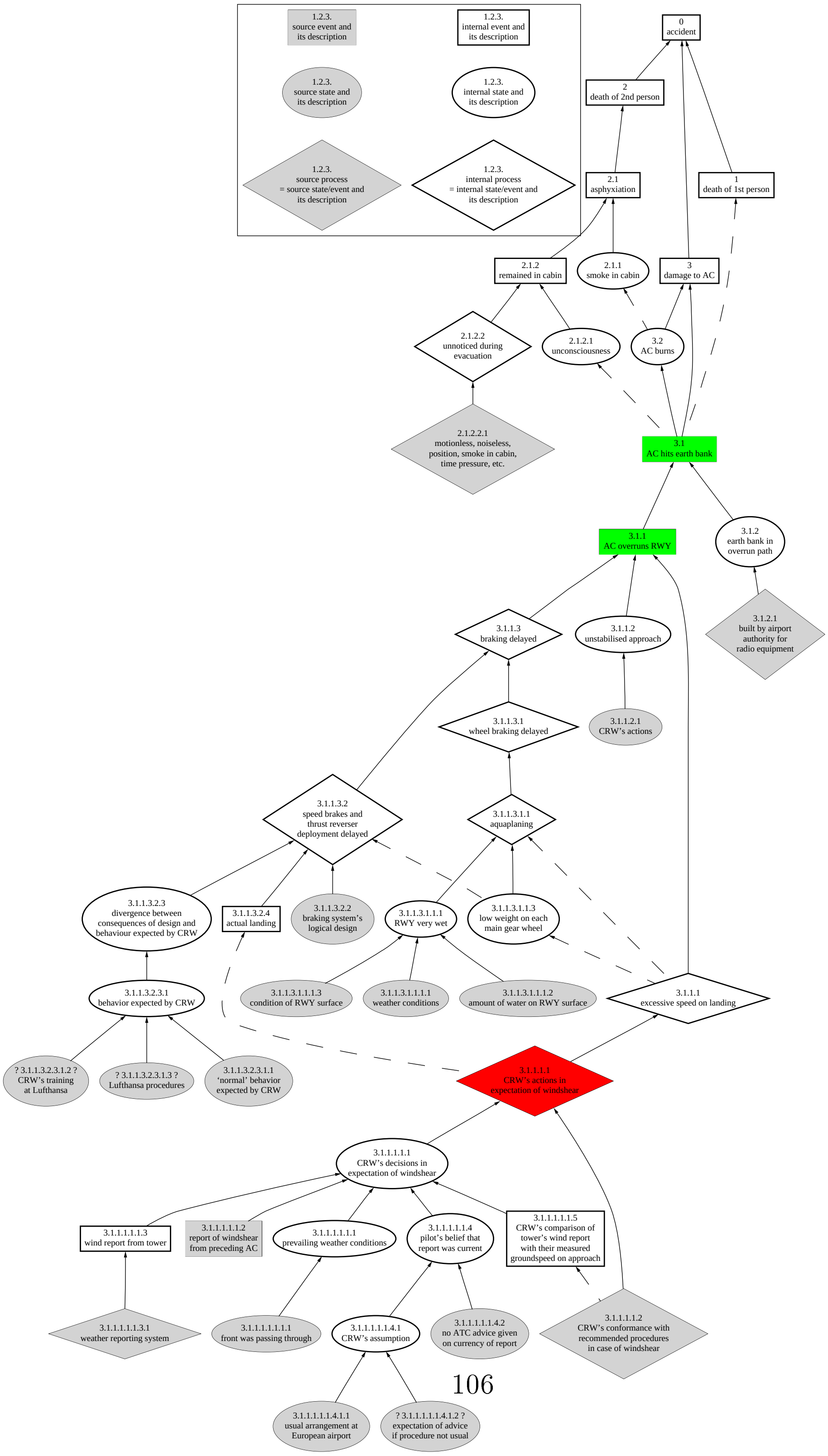


Figure 8.1: The Warsaw WB Graph

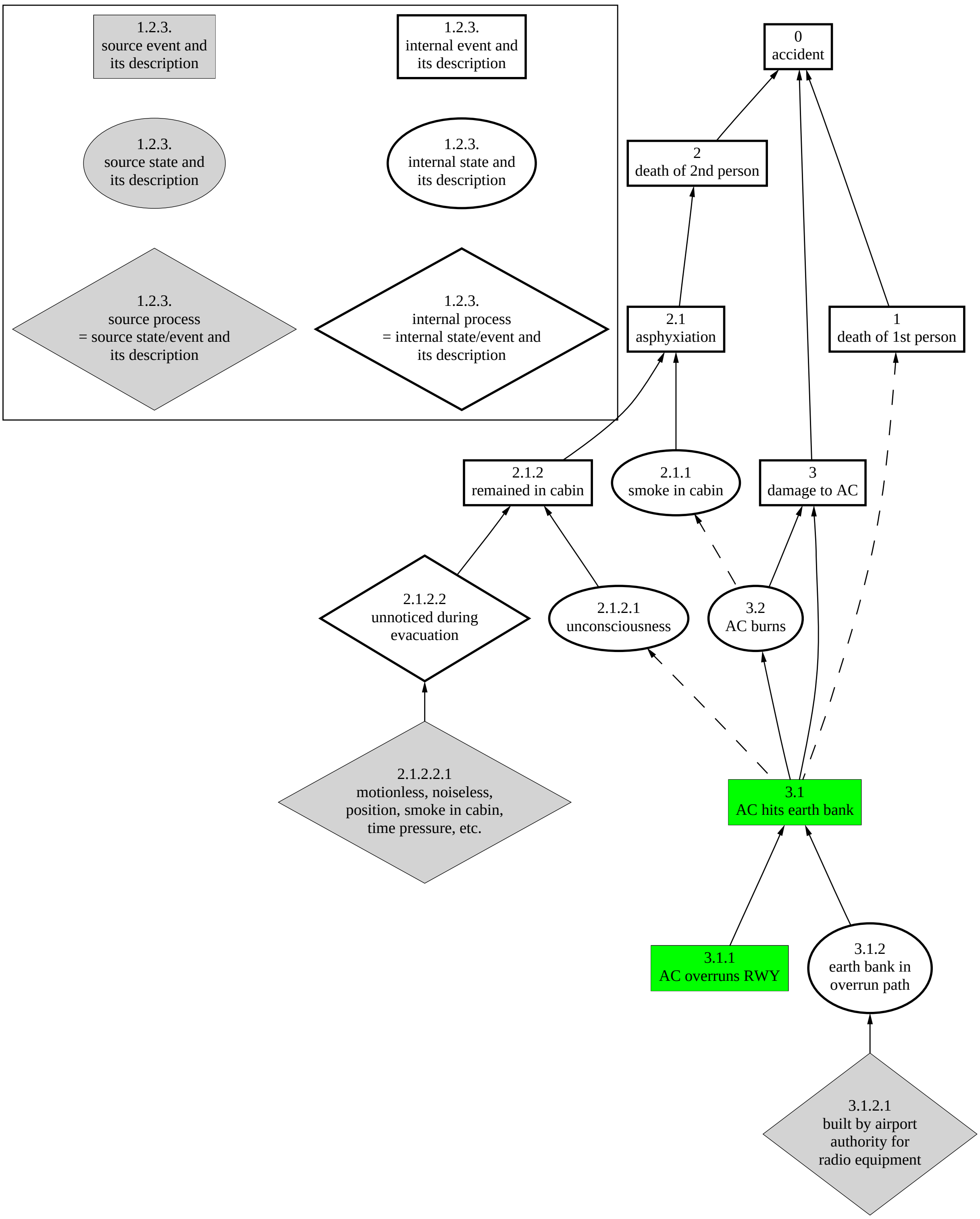


Figure 8.2: The Warsaw WB Graph, Upper Part

Chapter 9

The Social Background to Technological Risk

9.1 What Is Risk?

Risk As An Everyday Notion In everyday speech, one *risks* something, or one *takes a risk*, if a goal outcome is not certain to be attained by a course of action, or if a course of action has a certain likelihood of engendering deleterious consequences in the course of attempting to reach the goal.

9.2 Risk And Teleological Systems

Risk As Associated With Purposive Behavior Notice first that we are speaking of *purposive* or *intentional behavior*. One has a goal of sorts, although sometimes this goal may be just to execute the particular course of action. Wandering around aimlessly can be considered as much a purposive behavior as running for the bus.

The Risk Background of Teleological Systems When we build teleological systems, the background includes amongst other things

- the operation of the system to achieve the goal,
- the reasons for wanting to achieve the goal,
- the consequences of achieving the goal,

- effects of the course of action implemented to achieve the goal,
- the consequences of failure to achieve the goal,
- effects of the failed course of action utilised
- the range of accidents possible through achieving the goal
- the range of accidents possible through trying and failing to achieve the goal
- the range of accidents possible through trying to achieve the goal through the particular course of action embodied in the system

Typical System Risk Analysis A typical system analysis considers only

- the operation of the system to achieve the goal,
- the range of accidents possible through trying to achieve the goal through the particular course of action embodied in the system,

which is a somewhat restricted subset of the total background questions on safety.

9.2.1 Risk Analysis As Profession

Risk Assessment and Management as Profession It has been mooted [FLS$^+$81, KH99] that risk assessment is a professional skill by itself. There are a number of reasons for this.

- It involves techniques that one does not learn in the technical professions involved in designing the teleological systems.
- There is a variety of well-developed techniques which may be brought to bear on any of the issues involved in risk assessment, that are significantly more effective than naive approaches to the issues
- The techniques involved in risk assessment in a variety of fields have much in common with each other; more than they have in common with the techniques of teleological system building within the field itself

While [KH99] is primarily concerned with calculation techniques and [FLS+81] with management, [MH90] deals with all the techniques they have found useful as specialists in handling uncertainty in assessment situations. Further to [FLS+81]'s view that risk assessment is a decision problem, the practical techniques of so-called Decision Science, that is, rational decision making in uncertain situations, are described in [KKS93]. The fundamentals of rational choice, along with expositions of significant items of theoretical interest such as Arrow's Impossibility Theorem, may be found in [Res87]. Those interested in the foundations of inductive reasoning and the fundamentals of Bayesian decision theory may be interested in [Sky99].

General Difficulties with Risk Assessment Speaking against the acceptance of risk assessment as a profession are the following phenomena.

- The problems often defy precise formulation

- Although there are well-developed techniques, these techniques are often best employed not as decision methods, but in concert with other "competing" techniques as decision aids only.

- A large number of techniques involve subdisciplines of already-established fields such as social-psychological interviewing techniques, elicitation of expert judgements, and statistical evaluation.

- Some effective techniques, although relatively few, such as Fault Tree Analysis and Hazop, are specifically bound to technical disciplines already.

- Because of the nature of the problems, in which many of the data are unknown or very uncertain, the very best and most careful analyses can still lead only to rough, approximate answers; a situation which is theoretically and often socially unsatisfactory.

9.3 Risk Assessment

9.3.1 Two Principles: Know And Consult

Two principles which one finds uniformly in writings on risk assessment are

Know Everything One should inform oneself as thoroughly as possible what the facts of the matter are;

Ask Everyone All "stakeholders" should have an appropriate degree of involvement in deciding whether to implement a system with attendant risks.

9.3.2 Fact And Value

One Principle, Two Views There is one principle which one finds asserted by one group of professionals and denied by another.

Fact and Value Should Be Separated Technical experts in safety analysis should present "decision makers" with the facts concerning the level of risk inherent in a system design; the decision makers should solicit the decision through social processes. For an example of this point of view, see Section 9.4.4.

Fact and Value Cannot Be Separated Implicit in any technical assessment of a system are a series of assumptions about what matters and what does not matter. These assumptions are value judgements and should enter explicitly into an assessment of the values involved in a decision about risk.

A Value-"Fact" It is widely accepted by almost everyone (except certain U.S. Congresspeople) that there is no such thing as *Zero Risk*. But what does this mean? Another way of putting it is that no course of action is risk-free.

9.3.3 "Acceptable Risk": A Confused Concept?

A Confused Concept The idea of "*Acceptable Risk*" has been proposed as an alternative to the concept of zero risk. Various definitions of "acceptable risk" have been proposed.

- A chance of less than 1 in 10^6 of an untimely death during a lifetime [Lew90, pp95,105], see also [Ato76] as quoted in [FLS$^+$81, p85]

- A chance of less than 1 that a catastrophic accident will happen to any device (aircraft) during the lifetime of the fleet [LT82, p37]

- No significant increase in risk over "background" levels without the technology [Wei79] as quoted in [FLS+81, p87];
- Nothing scandalous about my behavior making it into the news during my time in office [various heads and former heads of state of Western countries, 1998-9]

A Less Confused Concept Rather than talk about levels of risk being "acceptable" or not, one may prefer to talk about risky "*options*" being acceptable or not:

> Strictly speaking, one does not accept risks. One accepts options that entail some level of risk among their consequences.
> [FLS+81, p3]

By "options" is meant, for example,

- a course of action (including doing nothing);
- design of a teleological system;
- forms of use of a teleological system.

The following definition from [FLS+81, p2] suggests exactly this:

> Acceptable-risk problems are decision problems; that is, they require a choice among alternative courses of action. What distinguishes an acceptable-risk problem from other decision problems is that at least one alternative option includes a threat to life or health among its consequences. We shall define *risk* as the existence of such threats.

Notice that the definition of "risk" is more narrow than the technical definition we proposed in [Lad00]. That definition suggested that we could define "loss" whatever way we wanted, and not necessarily as a threat to life or health.

9.3.4 Risk As Decision

The Decision Problem The decision problem identified by [FLS+81] consists of (quote)

> ......the following five interdependent steps
>
> 1. Specifying the objectives by which to measure the desirability of consequences
> 2. Defining the possible options, which may include "do nothing"
> 3. Identifying the possible consequences of each option and their likelihood of occurring should that option be adopted, including, but not restricted to, risky consequences
> 4. Specifying the desirability of the various consequences
> 5. Analyzing the options and selecting the best one
>
> [FLS+81, p2]

They point out that no known techniques allow each of these processes to be conducted optimally during an analysis. They evaluate the known techniques against this "wish list".

9.4 Alternative Conceptions of Risk

9.4.1 Risk as Interplay of Knowledge and Consent

[DW82] suggest that

> Risk should be seen as a joint product of *knowledge* about the future and *consent* about the most desired properties.
> [DW82, p5]

They pose the problem of assessing risk in the form of a table, shown in Table 9.1.

Complete Consent	Certain Knowledge Problem: Technical Solution: Calculation	Uncertain Knowledge Problem: Information Solution:Research
Contested Consent	Problem: (dis)Agreement Solution: Coercion or Discussion	Problem: Knowledge and Consent Solution: ??

Table 9.1: Douglas and Wildavsky's Problem Table

Some Components Are Missing It should be clear from a few moments thought that the explanation of [DW82] was not intended as a definition. The notion of *loss* is absent, and it is hard to see how this could be explained in terms of knowledge and consent. Supposing I were to wish to steal a sheep, know that I can do so, and have perfect knowledge of the likelihood that I would be caught. I would presumably be as unlikely to consent to the consequence that I would pay a fine of $ 10,000 as I would to consent to the consequence that my hand would be chopped off. However, the latter would be regarded by most as a more *severe* consequence and as a greater loss should it come about. (It may well be that notions of loss and of severity of consequence are interdefinable.) This preference cannot be explained by the notions of knowledge and consent, since by hypothesis these are the same in the two cases. However, it can obviously be explained by the notions of loss or severity, as the very phrasing of the example shows.

The Subjects Are Missing Knowledge doesn't exist in a vacuum. People have knowledge, and different people can have very different knowledge about outcomes. Consent exists even less in a vacuum – although one can reasonably speak of accumulated knowledge without supposed this accumulation is realised by any one person, it is hard to speak of consent without asking whose consent. One may assume that consent of a stakeholder is meant; the further question becomes how to identify "stakeholders".

A Tricky Example Consider the Jonestown massacre in Guyana, in which a nominally Christian cult which originated in San Francisco committed apparently willing mass suicide under the direction of the so-called "Reverend" Jim Jones. The participants are presumed to know what they were doing, to know that their actions of drinking the poisoned liquid would result in their deaths, and to have consented to this action. This would put them in the top left box in Table 9.1, which suggests that all they would need to do to find out their "risk" is to calculate. This cannot be so simple.

- Determining the consequences for the participants themselves might be straightforward, but
- determining the social consequences of their actions upon their relatives would not be so straightforward,
- and upon the social effectiveness in their roles of former religious colleagues of Jones in San Francisco,
- and upon the social tolerance of cults in formerly tolerant California, would be even less straightforward

This hangs on the assumption that the stakeholders in the action were also relatives, former colleagues, and members of Northern California society, and not just the participants themselves.

The Example Doesn't Fit the Proposed Paradigm Douglas and Wildavsky might reply by noting that I am saying that knowledge of the consequences was incomplete, and therefore this example fits rather in the upper right (we presume there is no dispute over consent). But their "solution" there involves "research". I doubt that research in advance of the mass suicide would have revealed the social consequences upon San Francisco and California society. So their prescriptions in Table 9.1 appear too facile.

An Extended Metaphor One might view Douglas and Wildavsky's contribution in [DW82] as an extended metaphor, intended to persuade readers of the essential dependence of risk on cultural norms, and emphasising the primacy of risk perception as a full-fledged component of risk. This view entails that fact and value cannot be separated.

9.4.2 The Royal Society's View

The Douglas-Wildavsky proposal stands in stark contrast to the "scientific" or "engineering" view, which tries to separate fact from value and which regards risk as the topic of calculations, such as in a Risk Cost-Benefit Analysis (RCBA), and perceptions of risk to be illusory in so far as they depart from the conclusions of such an RCBA. Compare the Royal Society's 1983 definition [Roy83], quoted in [Ada95, p8]:

> The Study Group views "risk" as the probability that a particular adverse event occurs diring a stated period of time, or results from a particular challenge. As a probability in the sense of statistical theory, risk obeys all the formal laws of combining probabilities.
>
> [.....]
>
> [*Detriment* is] a numerical measure of the expected harm or loss associated with an adverse event .... it is generally the integrated product of risk and harm and is often expressed in terms such as costs in £s, loss in expected years of life or loss of productivity, and is needed for numerical exercises such as cost-benefit analysis or risk-benefit analysis.

But Even Here Things Change Thanks to the arguments concerning separation of fact and value, the Society's view had changed by 1992 [Roy92]. The Study Group's terms of reference, quoted in [Ada95, p9]:

> [to] consider and help to bridge the gap between what is stated to be scientific and capable of being measured, and the way in which public opinion gauges risks and makes decisions.

The Study Group concluded, amongst other things, that

> the view that a separation can be maintained between "objective" risk and "subjective" or perceived risk has come under increasing attack, to the extent that it is no longer a mainstream position [Roy92], quoted in [Ada95, p9].

9.4.3 The National Research Council's View

Risk Characterisaction As Process The National Research Council Committee On Risk Characterization published seven principles for "implementing the [risk characterisation] process" [Nat96, p2]:

1. Risk characterisation should be a *decision-driven activity*, directed toward informing choices and solving problems [Nat96, p2].

2. Coping with a risk situation requires a *broad understanding* of the relevant losses, harms, or consequences to the interested and affected parties [Nat96, p2].

3. Risk characterization is the outcome of an *analytic-deliberative process*. Its success depends critically on systematic analysis that is appropriate to the problem, responds to the needs of the interested and affected parties, and treats uncertainties of importance to the decision problem in a comprehensible way. Success also depends on deliberations that formulate the decision problem, guide analysis to improve decision participants' understanding, seek the meaning of analytic findings and uncertainties, and improve the ability of interested and affected parties to participate effectively in the risk decision process. The process must have an appropriately diverse participation or representation of the spectrum of interested and affected parties, of decision makers, and of specialists in risk analysis, at each step [Nat96, p3].

4. The analytic-deliberative process leading to a risk characterisation should include early and explicit attention to *problem formulation*; representation of the spectrum of interested and affected parties at this early stage is imperative [Nat96, p6].

5. The analytic-deliberative process should be *mutual and recursive*. Analysis and deliberation are complementary and must be integrated throughout the process leading to risk characterization: deliberation frames analysis, analysis informs deliberation, and the process benefits from feedback between the two [Nat96, p6].

6. Those responsible for a risk characterization should begin by developing a provisional *diagnosis of the decision situation* so that they can better match the analytic-deliberative process leading to the characterization to the needs of the decision, particularly in terms of level and intensity of effort and representation of parties [Nat96, p7].

7. Each organisation responsible for making risk decisions should work to *build organizational capability* to conform to the principles of sound risk characterization. At a minimum, it should pay attention to organisational changes and staff training efforts that may be required, to ways of improving practice by learning from experience, and to both costs

> and benefits in terms of the organization's mission and budget [Nat96, p8].

If the NRC Committee thought that risk could be characterised as a likelihood coupled with a severity, it is unlikely they would need to suggest such a complicated recursive, mutual, analytic-deliberative, feedback-oriented, stakeholder-intensive, proactive, organizationally-structurally-supported process. A few engineers with calculators coupled with a few stakeholders to tell how much it hurts should have sufficed. One can conclude that the Committee accepts the inevitable intertwining of fact and value in risk characterisation, and thereby the necessity of a political process to elicit those values and the consent of stakeholders. One only wishes they could have expressed it somewhat more succinctly.

9.4.4 A Software Safety Expert's View

Leveson writes that

> Making decisions such as how safe is safe enough involves addressing moral, ethical, philosophical, and political questions that cannot be answered fully by algebraic equations or probabilistic evaluations [Lev95, p17]. ...... We must also realise that decisions about safety will cause legitimate disagreements that cannot be resolved by simple utilitarian arguments [Lev95, p18].

Leveson quotes Alvin Weinberg, former head of Oak Ridge National Laboratory, as suggesting that it is the scientist's duty to

> ...inject some order into this often chaotic debate by distinguishing scientific from trans-scientific problems [Lev95, p18].

Rather than attempt to characterise reasoning as "moral", "ethical", "philosophical", "political" or "scientific", I would prefer to say that probabilistic reasoning is one form of reasoning that one can expect to use when reasoning about risk, and one should use it where appropriate. If there are reasons why one should clearly delimit where is is appropriate to use such forms of reasoning, as Weinberg suggests is the "scientist's duty", then presumably those reasons are good reasons for distinguishing "political" from "moral" or other forms. I doubt whether such a line can be drawn; utilitarian calculations belong as much to moral reasoning as they do to "scientific" calculations.

Furthermore, reasoning is reasoning, and valid reasoning takes the same form in any subject matter. The question in any domain is more one of hidden assumptions than it is of any distinction in the notion of valid reasoning between "moral", "ethical", "philosophical" or "scientific" deliberations.

9.4.5 Risk Decisions As A Feedback System

Adams [Ada95] proposes to characterise risk decisions as a personal feedback system, in which a balance is sought between the rewards of risk-taking, one's personal propensity to take risks, the perceived danger, and knowledge of related accidents. Risk-taking decisions are the result of balancing these competing factors. He emphasises risk compensation (below) as an important factor in the system which is discounted in many risk analyses, and provides strong evidence for its existence.

A Quick Example of Risk Compensation I have just obtained a reclining bicycle, which I delight in riding. I wear a tie to work, and noticed that when I mounted the bicycle to ride to work, as I leaned over, the tie came perilously close to the oily chain, which is at thigh-level on the bicycle and only partly protected. I was very careful in mounting.

I then bought myself tie clips, to keep my ties attached to my shirt. I am much less careful about mounting the bicycle on my way to work. One day, if my tie clip fails to perform its function, because I have not attached it securely, I will suffer an oily tie.

It has been speculated that risk compensation may be a strong factor in the behavior of cyclists with helmets [Ada95, pp144-151], [Hil93].

I think we may conclude that risk compensation behavior is apparent. The question is, how significant a factor is it in assessing behavior while making decisions under risk?

9.4.6 Perception is an Irreducible Component of Risk

An Example Consider an example from [Ada95, p9]. Slipping and falling on ice is a game for young children, but potentially fatal for old people. The probability of such an event is influenced directly by the perception of its probability: old people see the risk of slipping on an icy road to be "high"; they take avoiding action, thereby reducing the probability for their group. The young people take minimal action, or even encourage it. Furthermore, older people share experiences and perception of the risk; so do young children in their peer group. Behavior is different; perception is different; consequences are different; probably even the mechanics are different. The role that such an event plays in the lives of these two groups is thoroughly different. This is sufficient grounds to speak of a *cultural difference.*

Intertwining of Perception and Risk: Another Example Adams quotes the author Roald Dahl, relating how he excitedly rode his new tricycle to school each day:

> All this, you must realise, was in the good old days when the sight of a motor car on the street was an event, and it was quite safe for tiny children to go tricycling and whooping their way to school in the center of the highway [Dah86], quoted in [Ada95, p11].

To put Dahl's feeling that it was "quite safe" in context, Adams notes that between 1922, the period about which Dahl was writing, and 1986, the number of children under the age of 15 killed annually on the roads in England and Wales *fell* from 736 to 358, although the amount of motorised traffic increased by a factor of 25. The child road death rate is now about half what it was then; per motor vehicle it has fallen 50-fold.

Changes in Exposure Before one puts this risk assessment down to a perception, one might query whether this reduction in number is not because the roads have become "objectively" safer, but primarily because the exposure of children to traffic is much reduced. Some figures suggest this: 80% of children made their way unaccompanied to school in 1981, for example, compared with only 9% in 1990. The difference, according to surveyed parents, was mostly their worries about the danger of traffic.

Changes in Behavior; Vigilance Adams also wonders how much may be due to changes in behavior: playing *alongside* the street rather than *in* it. The *vigilance* of motorists towards children may have changed, also the children's reaction to the speed, volume and variability of traffic. Measuring changes in exposure effectively presents all but insurmountable problems [Ada95, p13]. The general problem

> ...for those who seek to devise objective measures of risk is that people to varying degrees modify both their levels of vigilance and their exposure to danger in response to their *subjective* perceptions of risk [Ada95, p13].

9.4.7 Risk Compensation

Purpose of Characterising Risk Is To Manage It According to Adams, the Royal Society's purpose in devising risk assessment procedures is to manage the risk. When people respond to their perception of risk by altering their behavior, then management of risk does not operate against a static background that can be measured, but against people's adjustment to a newer risky situation. This adjustment is termed *risk compensation.*

A Model of Risk Compensation Adams proposes a model of risk compensation that he attributes to Wilde in 1976 [Ada95, pp14–16]. This model is based on the following propositions

- Everyone has a propensity to take risks
- This propensity varies between individuals
- The propensity is influenced by the rewards of risk-taking
- Perceptions of risk are influenced by the experiences of self and others with accident losses
- Individual risk-taking decisions represent a balance between the perception of risk and the propensity to take risks
- accident losses are a consequence of taking risks.

Because of the feedback from consequences to perception and the mixing with propensity, it follows that managing risk is an interactive phenomenon. Adams illustrates this idea with what he calls the "*risk thermostat*", Figure 9.1. Just how complicated matters can be to assess when two risk managers, one riding a bicycle and the other driving a truck, meet on a wet curve in the road can be seen in Figure 9.2. One can imagine how complicated this gets with many "risk managers" all at once. With this, Adams hopes to illustrate how simplistic the current assessment methods are in comparison with reality.

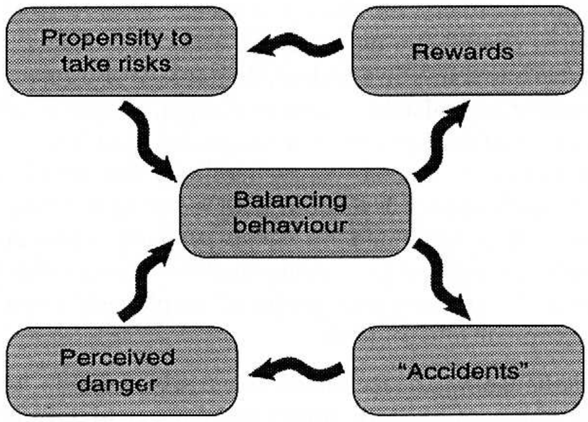


Figure 9.1: Adams's "Risk Thermostat"

9.4.8 Summary: Risk As Cultural Artifact

We may take it, as the Royal Society suggested in 1993, that there is nowadays significant weight given to the two theses that

- assessment of risk is culturally dependent, and
- risk perception is an irreducible and inseparable component of risk itself.

9.5 Cultural Theory

9.5.1 Attitudes to Nature and Risk

Myths of Nature Adams [Ada95] identifies four anthropomorphic attitudes to nature, which stem from the observations of Holling [Hol79, Hol86] concerning different management strategies for managed ecosystems that appeared to be explicable in terms of the managers beliefs about nature. He identified three belief styles, extended to four by Schwarz and Thompson [ST90] and developed into so-called *cultural theory* of risk in [TEW90]. These myths are

nature benign: nature is stable, robust and forgiving of human insult. In the technical vocabulary of dynamics, the state of nature is a *stable*

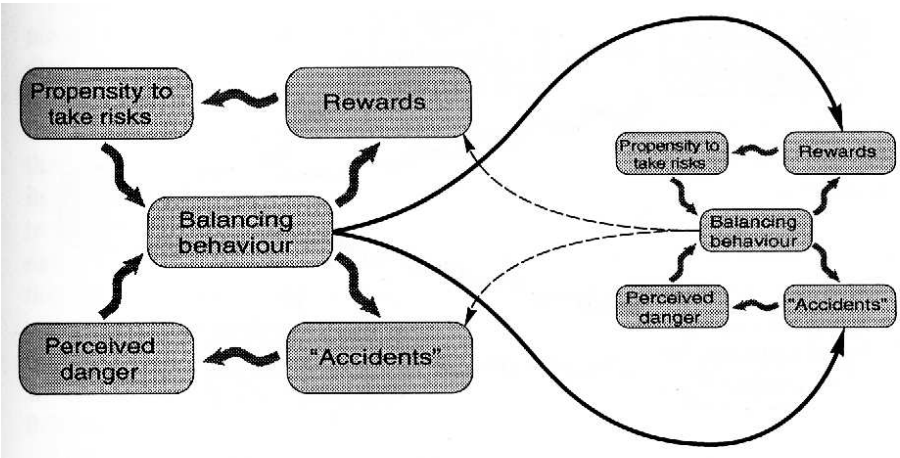


Figure 9.2: Risk Thermostats Interacting

equilibrium. The appropriate management style is *laissez-faire.*

nature ephemeral: nature is fragile, precarious, unforgiving. We must tread carefully on the earth. The state of nature is an *unstable equilibrium.* The appropriate management style is *precautionary.*

nature perverse/tolerant: Within limits, nature can be relied upon, but care must be taken not to exceed those limits. The state of nature is a *local stable equilibrium that is not global.* The appropriate management style is *interventionist*

nature capricious: Nature is unpredictable. The state of nature is that there are *no equilibria.* The appropriate management style is *resignation: do nothing.*

It should be clear that the four models refer to various features of so-called dynamical systems, a field of mathematics which uses the analysis of differential equations to study predator-prey and other ecological systems. Adams suggests this with his diagram illustrating the four myths, reproduced in Figure 9.3. It is clear to those engaged in such modelling that dynamical systems include and are included in other dynamical systems, and is perfectly mathematically in order to consider the union of all such natural dynamical systems. One can identify this with "nature", and plausibly ask about its

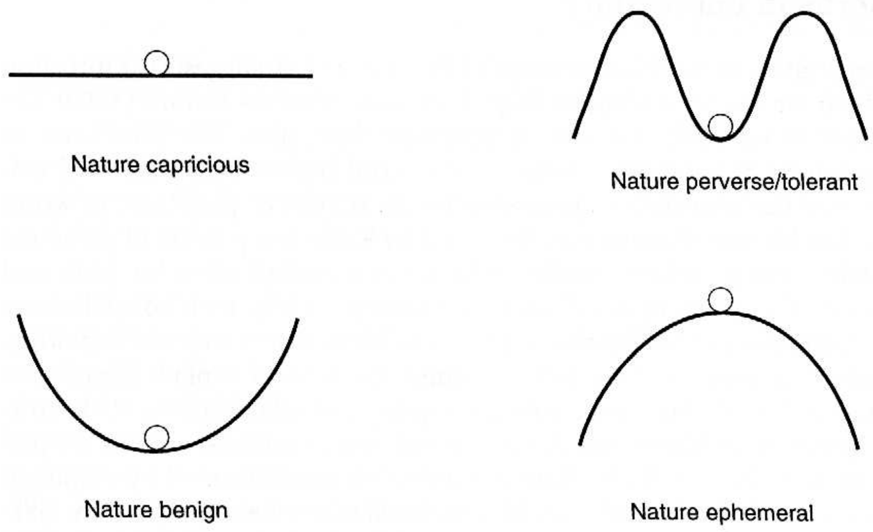


Figure 9.3: The Four "Myths" of Nature

equilibrium properties, as one can for any dynamical system. But because the system is so complex, one cannot hope to answer this question definitively. Hence the various "myths of nature" correspond to the variety of possible beliefs about the global equilibrium properties of "nature". They need not stem from anthropomorphic roots after all. All very reasonable so far.

Applied to Risk-Taking If we regard risk-taking as the management of uncertainty, and this uncertainty concerns the way the "world" is, one can plausibly identify "the world" with "nature", as long as nature includes human activity also. Thus can the four views of nature be adapted as the background to risk-taking decisions, as in [ST90, TEW90].

- If nature is in stable equilibrium, then I can take risks as I like, strive to exert control over my environment and people in it, and the "world" will accomodate. I am an *individualist* about risk.

- If nature is in unstable equilibrium, then I must manage my risk by consensus to ensure uniformity of action by all and avoid disturbing the equilibrium. I work by consensus under strong group cohesion; leaders arise through force of personality and persuasion; rules from outside the group don't apply. I am an *egalitarian.*

- If nature has local but not global stable equilibria, then I may act inside defined boundaries, and outside these boundaries others must have the say. I construct management structures; I am a *hierarchist.*

- If nature has no equilibria, then it is not possible for me to manage my risk, that is, to affect the "world" in such a way as to get it to respond more favorably to my wishes. Management is impossible; I am a *fatalist.*

It is clear that, although the "myths" of nature are bound up with speculation about the global nature of a well-defined model, the attitudes to risk just enumerated are indeed myths. They are socially constructed parables about how the world behaves which lead to management paradigms.

Grounding These Paradigms The paradigms may, however, be grounded in abstract social views. Consider the two dimensions, denoted by their extrema, of

- Individualist - collectivist. This dimension describes one aspect of the nature of the human animal. Eagles are individualist, rabbits are collectivist. We use the acronyms *I/C*

- Prescribed/unequal - prescribing/equal. This dimension describes organisation. At one end, social choices are constrained, prescribed, by a superior authority and social and economic transactions are characterised by inequality. At the other end, transactions are negotiated by participants as equals, without externally prescribed constraints on choice. We use the acronyms *U/E*

Given this typology of social organisation, we can categorise the four views of risk as

- Individualists are IE.

- Egalitarians are CE.

- Hierarchists are CU.

- Fatalists are IU.

This placement is illustrated in Figure 9.4 This provides what amounts to a theory of how risk attitudes arise. But does it work to explain risk behavior?

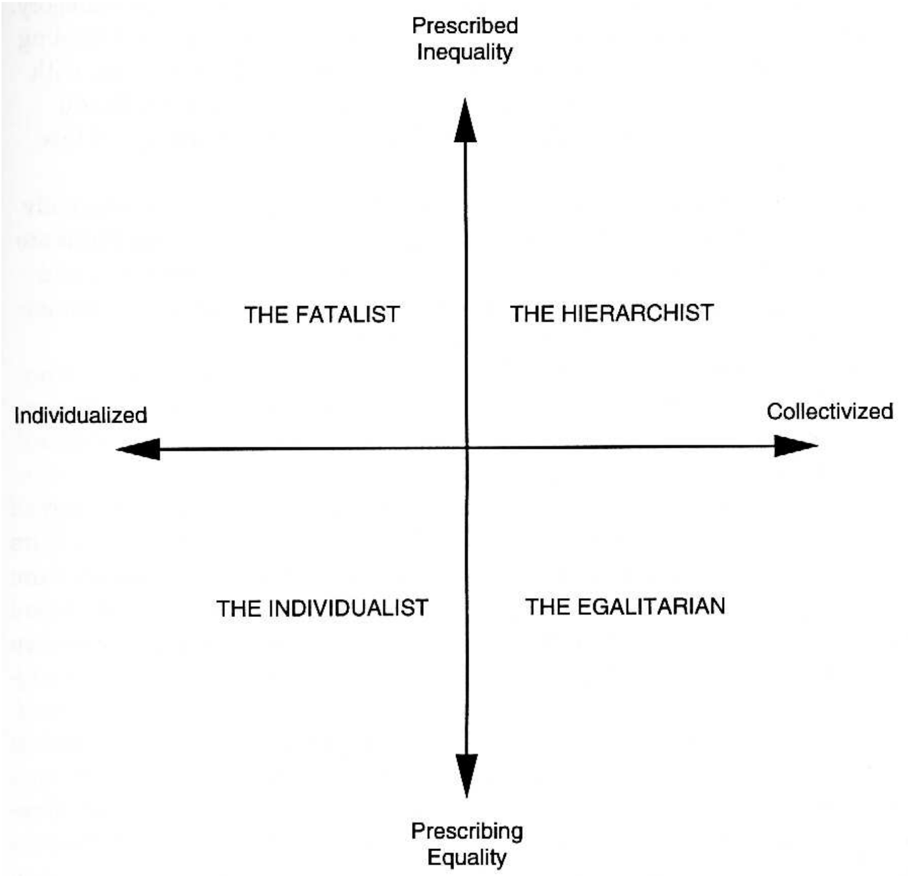


Figure 9.4: The Myths As Four Quadrants in Two Dimensions

Empirical Evidence is Lacking It has proven hard to identify these attitudes with groups of risk-takers. Adams reports [Ada95, p64] that [Dak91] had "limited success" in trying to substantiate the hypothesis that social concerns are predictable given people's cultural biases. Recent studies assessing the fruitfulness of the four cultural categories in predicting risk attitudes have also doubted whether the categories best describe individuals, and have suggested that a given individual may well have a mix of attitudes to different risk situations. Marris and colleagues [MLO98] distributed psychometric-style questionnaires to residents of Norwich, England, and found that psychometrics explained a “far greater proportion” of risk variance than cultural biases as explained by cultural theory. However, they discovered a “key point” that cultural biases were associated with concern about distinct types of risk, and that the pattern of responses in these cases was compatible with that predicted by cultural theory. Furthermore, the psychometric questionnaire could only allocate individual respondents unequivocally to a unique cultural category in 32% of the cases. Brenot and colleagues used a version of a questionnaire developed by Dake [Dak91, Dak92] to test the correlation between cultural bias and 20 social and environmental risks. They found a “weak positive” correlation: cultural bias explained just 6% of the risk variance. They compared with other studies in other countries, and concluded that “new methods, more qualitative and contextual”, are needed to investigate cultural perceptions of risk.

“More Studies Are Needed” So the jury is out on cultural theory. The model is based more firmly on structure and less on parable than some proponents have credited it with. Maybe one can usefully compare the situation with that of individual political views versus party systems.

The “Party System” Analogy A cultural category is like a party manifesto. But I can still have strong political views on a number of issues without following the party line. Party X believes as person A does that that untrammeled libertarianism is the best model for social welfare economics; quite in distinction to party Y, which believes in putting jobless and homeless people up in the local 5-star hotel at tazpayers’ expense as compensation for not having a home or salary. However, party Y also believes that the new autobahn should not be built at all, let alone in front of A’s house, which is where party X decided to put it. Assigning a party affiliation to A on the basis of this information would be unwise, I propose. Maybe we can pursue an analogy with risk characterisation.

Distinct Attitudes for Distinct Risks Although I may ride my bicycle with relatively great care, I do believe that whether I am assaulted by an auto is largely due to chance rather than under my control, and the residual variance I can affect is limited. By contrast, it is apparently the case that many car drivers believe themselves to have greater control than they actually do – most drivers believe themselves to be "above average" [NS75, Sve81] quoted in [KST82, p469], which is a collective contradiction. However, I may well believe that my career is largely under my control, through my performance, although of course it is significantly affected by my age, where I choose to work, what choices I have that suit my capabilities, and what potential colleagues think of my personal presentation, as well as what political role I play for them. Further, I tend to think that paper acceptance at academic conferences is largely a matter of chance, whereas paper acceptance by journals is much more determined by the relative quality of the contents. Also, while I talk with my colleagues and plan action about matters of mutual concern, I don't necessarily believe that what they agree to and what they do are perfectly correlated. I am thus an egalitarian about bicycle riding, largely an individualist about my career and about journal papers, and largely a fatalist about conference papers and my neighbors' neighborliness quotient. As a whole, my attitudes to risk management appear to be diverse. I would be surprised to find that I were atypical. The research results of Marris and colleagues are unsurprising.

9.6 Perception Heuristics

9.6.1 Problem Presentation Affects Choice

In a series of classic experiments, Daniel Kahneman and Amos Tversky, amongst others, have investigated the probability reasoning of laypeople (i.e., those who are not probability theorists or statisticians, but who might be aware of probability calculations, such as students). One experiment asked practicing physicians to answer the following [TK81]:

1. Imagine that the United States is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the consequences of the programs are as follows:

    - If program A is adopted, 200 people will be saved.

- If program B is adopted, there is one-third probability that 600 people will be saved, and two-thirds probability that none will be saved

Which of the two programs would you favor?

2. Imagine that the United States is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the consequences of the programs are as follows:

   - If program C is adopted, 400 people will die.
   - If program D is adopted, there is one-third probability that nobody will die, and two-thirds probability that 600 people will die.

   Which of the two programs would you favor?

Most physicians preferred A over B, and D over C. Note that the two problems are formally equivalent. Programs A and C save 200 and let 400 die; programs B and D give one-third chance that all will be saved and none will die, and two-thirds chance that none will be saved and all will die. Those paying attention to the actual outcomes, if they prefer A over B, should also prefer C (identical with A) over D (identical with B). But they don't, despite being au fait with the numbers. That seems to be a simple contradiction, and would render the majority choices irrational. Furthermore, the expected number of deaths is identical for all four programs. Someone choosing strictly according to expected number of deaths (one common measure of risk in co-operative situations, when a qualitative measure is called for) would have no preference amongst the four choices. One might exhibit a preference for certainty, or on the contrary for the chance of a jackpot, but this would also entail consistent choice, which the majority do not exhibit.

The outcome of this experiment is reproducible, with different populations, in different formulations, and roughly in the proportions of respondents preferring which alternatives. It is a *result* of social cognitive psychology, if anything is. It appears to demonstrate a certain kind of irrationality in choices under uncertainty.

One conclusion is clear.

- The mode of presentation of an uncertain choice, a risk, affects the choice. And how.

9.6.2 Prospect Theory

Exactly how this presentation affects choice is explained by *Prospect Theory* [ST95, pp80–81]. First observe that "saving people" is a *gain* and "people dying" is a *loss*. A preference for a risky outcome over a "sure thing" with the same expected value is termed *risk seeking*; a preference for a "sure thing" over a risky outcome with the same expected value is termed *risk aversion*. The experiment demonstrated that risk aversion holds for gains and risk seeking for losses. This is true in general, except for choices involving very small probabilities. Prospect theory posits the following three phenomena:

Diminishing sensitivity: I am more sensitive to a difference in expected outcome varying between, say, \$50 and \$150 than I would be to a difference in expected outcome varying between, say, \$8,050 and \$8,150.

Relative Value: I am sensitive to gains and losses rather than to total wealth.

Loss Aversion: I am more sensitive to losses than I am to gains of equal magnitude.

It turns out that prospect theory can explain many of the apparently irrational, but reproducible, preferences expressed in choice problems. Various other phenomena to complement those explained by prospect theory have been identified.

9.6.3 Other Heuristics

[SFL82] describe other heuristics of risk perception.

Availability: People using this heuristic judge an event as likely or frequent if instances of it are easy to imagine or recall. Aircraft accidents, shark attacks (after *Jaws*), atomic powerplant accidents (after Brown's Ferry, Three Mile Island and Chernobyl). It surprises people to realise that twice as many people were killed on the roads in Northern Ireland than were killed in the sectarian violence over the last quarter century [Ada95, p62]. Since rare events tend to get reported and discussed, in constrast to relatively common events, one might expect people to overestimate the frequency of rare events and underestimate the frequency of common events. Such a result can be seen in Figure 9.5, in which participants were asked to estimate the frequency of various causes of death in the U.S.

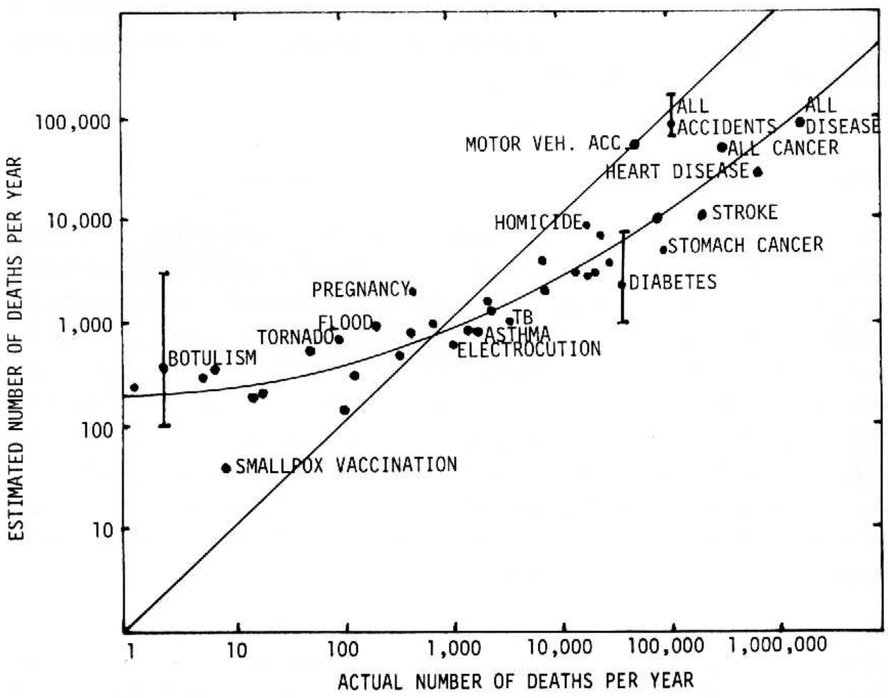


Figure 9.5: Estimates of Event Frequency Plotted Against True Frequency

Overconfidence: People typically have greater confidence in judgements under uncertainty than warranted. One notable result was remarked by Hynes and Vanmarcke [HV76], who asked seven "internationally known" geotechnical engineers the height at which an earth embankment would cause the clay foundation to fail, and to specify "confidence bounds" around this value that were wide enough to have a 50% chance of enclosing the true failure height. In other words, they were asked to guess and hedge their guess to 50% likelihood. None of the intervals from any of the seven experts enclosed the true failure height. The results from this experiment are illustrated in Figure 9.6

Anchoring: Judgements are "anchored" to initially presented values. For example, individuals were asked to estimate the frequency of death in the U.S. from 40 different causes. When told initially that total annual driving deaths were about 50,000, people tended to give higher estimates of fatalities for all causes than if told initially that 1,000 people die annually in the U.S. from electrocution [SFL82, p481].

That experts as well as "laypersons" are subject to the same heuristics and biases (i.e., these are cognitive phenomena) makes the elicitation of expert opinion under uncertainty far from the objective assessment that one hopes it might be. Various tricks, or "elicitation protocols" have been devised to obtain estimates as free from the effects of heuristics and biases as possible [MH90, Chapter 7].

9.7 Difficulties With the Numbers

Biases make it harder to obtain what one would like to believe were roughly accurate judgements about event frequencies and likelihoods. Further problems are

- often a dearth of available statistics from which desired conclusions could be drawn, and
- the wide variance in calculated values, even given apparently sufficient statistics.

9.7.1 An Example: The Value of a Life

Various attempts have been made to compare how much has been spent overall, across many different industries and social themes, to save how many lives. If an estimate can be made of how much has been spent on safety measures, and how many lives have been saved, one can divide the one number by the other and call it, somewhat crudely, the "value of a life saved". It represents the marginal average cost that society has been willing to spend in the past not to forgo a life.

One of the earliest estimates came up with $200,000 per life saved [TR76]. An estimate a few years later came up with $2m [Rap81]. This is an order of magnitude higher. Nearly twice as high again is the estimate of [Mar92], quoted in [Ada95, p103] of £2m–3m. Thes are hardly figures on which one can place much faith.

9.7.2 Example: Cigarette Smoking Deaths

In a well-known comparison of various ways to increase one's chances of dying prematurely by 1 in 10^6, 1 in a million, [Wil79] (quoted in [FLS$^+$81, p81]) includes smoking 1.4 cigarettes.

Another estimate can be obtained as follows. Let us take the average male lifetime to be 75 years. According to [CL79] (quoted in [FLS$^+$81, p82]), cigarette smoking will reduce a life expectancy by 2,250 days, which is roughly 6 years. So let us take the average lifetime of a cigarette smoker to be 70 years. Assume he started smoking at age 15 years, giving a smoking duration of 55 years. The average German smokes 5 cigarettes a day, and since about one-third of the population claim to smoke, we may obtain an estimate of 15 cigarettes a day per smoker, which is about 5,000 cigarettes per year, and thus 275,000 per lifetime. Every second smoker may expect to die from smoking-related causes, so the average number of cigarettes per death is 550,000. This is approximately one-third the figure given by [Wil79]. While not an order of magnitude difference as the the "value of a life saved", this is still a notable difference.

Don't Forget: Probability Allows Anything One should not forget that a probability estimate is compatible with most individual outcomes. If there is an infinitesimal likelihood that I will receive a dose of pigeon dropping on the head today, that does not rule out that I'll be hit by a pigeon every day of the next year. As Chauncey Starr is reported to have said concerning Three Mile Island [Ada95, p51]:

> On the technical side, this accident, while no one wanted it, has a statistical probability that falls within the predicted probability of this type of accident.

9.8 Excessive Prudence Is Disadvantageous

One may wonder if safety problems arise primarily because of a lack of will to fix them. In fact, there can be considerable disadvantage to excessive prudence. [Ada95, p55] lists some.

- People may spend more money on insurance, needlessly

- Motorists may drive more slowly and with more space between vehicles if they believe that there is "black ice" on road, hindering traffic flow
- The construction industry may waste money and resources on "over-building", for example, building to earthquake safety standards in regions which have little or no earthquake risk
- On the railways in Britain, excessive expenditure on safety measures raises ticket prices and encourages people to use even less safe modes of transportation such as cars instead.
- An inordinate fear of physical attack leads some women and elderly people not to venture outdoors as often as they would prefer to.

9.9 How Biases May Affect Assessments

9.9.1 Cultural Biases

One way in which the four cultural types may be seen to affect assessments follows from the types of error they may make. Suppose one is attempting to evaluate a hypothesis such as

Hypothesis: CO_2 emissions threaten a runaway greenhouse effect

- a *Type 1 error* is made when a hypothesis is accepted that should be rejected;
- a *Type 2 error* is made when a hypothesis is rejected that should be accepted.

The four types distribute themselves amongst the error categories thus:

- Egalitarians are at high risk of a Type 1 error and low risk of a Type 2 error
- Individualists are at high risk of a Type 2 error and low risk of a Type 1 error
- Hierarchists would reject the statement of the hypothesis as unspecific on critical limits
- Fatalists would ignore the hypothesis and not attempt to determine its truth or falsity

9.9.2 Evaluation Biases

A common elicitation technique used to attempt to set a uniform value (usually a monetary value) on factors in a risk problem is to ask the value of compensation. This can take two forms:

- What is one *willing to pay* (WTP) for a certain advantage that one does not have.
- What is one *willing to accept* (WTA) in compensation for loss of a resource or capability that one values.

These quantities are used in an attempt to achieve an equitable distribution of risk or of consequences of a course of action. The quantities are not dual. In a successful transaction, the range of amounts that one party is willing to pay overlaps the range of amounts that the other party is willing to accept, but not all risk and compensation problems are of this type. For example, consider asking a fatally ill person what could compensate him for loss of his life. The answer might well be that no amount of money would suffice for him to consider himself suitably compensated. However, the amount he would be willing to pay to have his life saved is rigorously limited by his assets.

In general, WTPs can be very much less that WTAs, and this leads to bias in distribution [Ada95, p99].

9.9.3 An Example: Negotiating a Smoke

Consider two rules for smoking in a compartment of a railway carriage [Ada95, p99].

- Under the *permissive rule*, one may smoke. In this case
  - A smoker may consider a WTA for giving up smoking for his journey
  - A non-smoker may consider a WTP for experiencing a smoke-free journey
- Under the *restrictive rule*, one may not smoke, unless all parties are agreed to it. In this case
  - A smoker may consider a WTP for smoking during his journey

- A non-smoker may consider a WTA for suffering smoke during his journey

The consequences for smoker and non-smoker alike of the preexisting rules are different, given that WTPs are less than WTAs. Whoever has the right of WTA is likely to prevail. Thus the permissive rule favors the smoker and the restrictive rule favors the non-smoker.

A Personal Comment I am a non-smoker who strongly does not like to breathe air polluted with cigarette smoke, and avoid it wherever possible. But I do believe the decision to smoke or not is a personal one, so have nothing against smoking per se. The difference between the number of smokers in the U.S. and in Germany is tiny. About a quarter of Americans say they smoke and about a third of Germans. The difference is one-twelve - about 8% of the population. However, in the U.S. I have no trouble avoiding smoke when I wish. Restaurants are completely non-smoking or have adequately ventilated non-smoking areas; offices are mostly or entirely smoke-free. This is supported by Federal and State regulations. However, in Germany, there are few regulations. Restaurants are to me so unpleasantly smoky that I do not go to eat in restaurants any more, although that was a hobby when in the U.S. and I went out most nights. My office, although nominally in a non-smoking corridor, is invaded by the strong odor of cigarette smoke many times daily, and by the end of a working day I have noticeable physical effects from it. At bus stops or in train stations with more than two or three people waiting, there will be cigarette smoke. People and exchange students who visit from Great Britain, the U.S. or Ireland have also remarked on the comparative pervasiveness of smoke here. The difference between the biases implicit in the permissive and restrictive rules is real and palpable.

9.10 Professional Attitudes To Risk Management

9.10.1 Engineering Codes of Ethics and Their Consequences

Engineers have to manage risk, whether they are familiar with the tools of technical risk management or not. [Ada95, pp186–189] reports on a conference [Fel90] of engineers concerning "preventable disasters". The Rules of Conduct for Chartered Engineers require engineers to pay due regard to

- the safety of the public,
- the interests of their client or employer,
- the reputation of other engineers,
- the standing of the profession.

Adams notes that only the first of these has anything to do with safety. The other three are political or social group interests. It is well possible, of course, that failing to pay due regard to safety could ultimately influence the standing of the profession. The point here is, as has been well-documented in, for example, the history of DC-10 carog-door failures [FB92] and the decision to launch the Challenger space shuttle in low temperatures with a known temperature-affected weakness, that engineers who warn of problems are very often ignored by management or have their concerns submerged in the flow of the organisations involved. That is, very often the criteria above may lead to contradictory choices of action. Client or employer *versus* safety, for example. Safety *versus* (sometimes undeserved) public trust in engineering capability (that may, for example in the case of very large and complex software systems, not even exist in appropriate measure).

9.10.2 An Example of What Counts: The Therac-25

But what do practicing engineers expect from the development of safety-critical devices? The history of a series of accidents caused by the Therac-25 radiation therapy machine in the mid-1980's in [Lev95, Appendix A] indicates what the authors, as well as their readers presumably, single out as problems.

One Possible Attitude

A thought experiment. Suppose the makers of the Therac-25 had said, OK, our machine is killing people because of the way it has been used. However, it has saved many more lives than that. So on balance (using, if you like, an RCBA), there are benefits. Let's leave everything as it is.

This fits with "standard" evaluation techniques, an RCBA, by hypothesis. What exactly, if anything, would be wrong with such an attitude?

One might say, not all the interests of all the stakeholders are taken into account [Nat96, FLS+81]. The interests of the people who were accidentally irradiated much more than they should have been were not taken into account. But they could have been, consistently with this attitude. Suppose each of them was informed of the chances of successful radiation and of overexposure beforehand, and they had consented (as we suppose many of them still would have). Then their interests had been taken into account.

Other stakeholders include insurance companies, hospitals, regulators, the company itself, and users of the machine, amongst others. I think it is fair to say that the patients are the primary stakeholders, however. So the claim, that stakeholders' interests have not adequately been taken into account, could well fail. The question remains: what exactly is wrong with the utilitarian argument for doing nothing?

The Implicit Critique

Leveson and Turner analyse the design of the machine. The machine made much more extensive use of software control than its purported predecessors [Lev95, p516].

Turntable Positioning

The authors noted that positioning of the turntable holding the patient was crucial, and that protection against inappropriate positioning, or inappropriate activation of the device with the turntable in a disadvantageous position, was traditionally provided by mechanical interlocks. In the Therac-25, software checks were substituted for many of the hardware interlocks [Lev95, pp517–518].

Operator Interface

The design of the operator interface, displayed on a 25 line by 80 character computer screen, left a lot to be desired. Error messages were "cryptic", containing codes (numbers 1 through 64) for various types of malfunction. The codes were not explained in the operator's manual. Apparently malfunction messages were commonplace and did not usually involve patient safety [Lev95, pp591–520].

Hazard Analysis

A hazard analysis was performed by the manufacturer. The analysis excluded the software. Three assumptions were explicit:

- Programming errors had been reduced by testing on simulator hardware and under field conditions. Any residual software errors were not included in the analysis.
- Software does not degrade due to wear, fatigue, or the reproduction process.
- Computer execution errors are caused by faulty hardware or by "soft" random errors caused by alpha particles or electromagnetic noise.

Information-Gathering About the First Accidents

The first accident led to a lawsuit from the patient involved. It was not officially investigated. The company claimed the first it had heard was when a lawsuit was filed against it by the patient about 9 months later. Others claim that the company was officially notified of a lawsuit about five months after the accident. The accident was not reported to the U.S. Food and Drug Administration, the reponsible government authority, until adter further accidents in the next year.

After the second accident, the company was informed and sent a service engineer to investigate. Regulators and users were informed that there was a problem, although users claimed not to have been told of an accident.

Company Response

The company investigated, made some hardware and software modifications, and informed users that the hazard rate of the new system offered a *five-orders-of-magnitude* improvement over the old system. They had, however, been unable to reproduce the reported hardware behavior in their investigations.

Further Accidents and Response

There were further accidents with the machine. Eleven machines had been installed altogether, six in the U.S. and five in Canada. Altogether, there were six accidents at four different sites. Over a third of the installed sites suffered an accident.

The Software Bugs

Two different software bugs were found in response to two different accident scenarios. Both involved so-called "race conditions", conditions in which a particular instruction execution sequence is required for correct operation, but in which the instructions could and did execute in a different order. The first involved a hardware operation taking about eight seconds. The operator was able to change certain settings, and these changes were reflected on the terminal screen, but the machine could not correctly attend to the desired changes until after the eight-second hardware cycle. A missequenced series of operations followed [Lev95, pp534–537].

The second bug also required an operator action, which triggered the missequencing of operations when it occurred simultaneously with an internal software operation [Lev95, pp542–544].

The Causal Factors

The authors invoke the following causal factors:

- Overconfidence in Software
- Confusing Reliability with Safety
- Lack of Defensive Design
- Failure to Eliminate Root Causes
- Complacency
- Unrealistic Risk Assessments
- Inadequate Investigation or Followup on Accident Reports
- Inadequate Software Engineering Practice

One remarks first that all these causal-factor statements are value judgements or negative human attributes: *overconfidence, confusion. failure, complacency, unrealism, inadequate practice.* This suggests

- that there is a standard, or many standards, which this particular machine and its development did not meet

- that many of the causal factors were human failures which need not have occurred

There is thus a strongly moral tone to this assessment. No one is saying "well, this machine is really complicated and we know it fails but no one knows how to do better than this". On the contrary, the authors are saying "best engineering practice was not followed in this and this and this respect".

Conclusion: What Counts To Engineers

At the heart of the reports of many accident investigations lie similar attributions. Overall, they can be summarised thus.

We know how to do better and we could have done better in this case.

This is strictly a moral judgement. I believe it may distinguish engineering safety concerns from other technological areas in which risk and uncertainty assessments need to be taken into account.

On a final, not completely satisfactory, note, I remark that the injunction to

- Use best practice and perform as well as possible

is not generally part of codes of engineering ethics, for example in Section 9.10.1 above. It is, however, becoming enshrined in increasingly many standards governing certification of teleological systems.

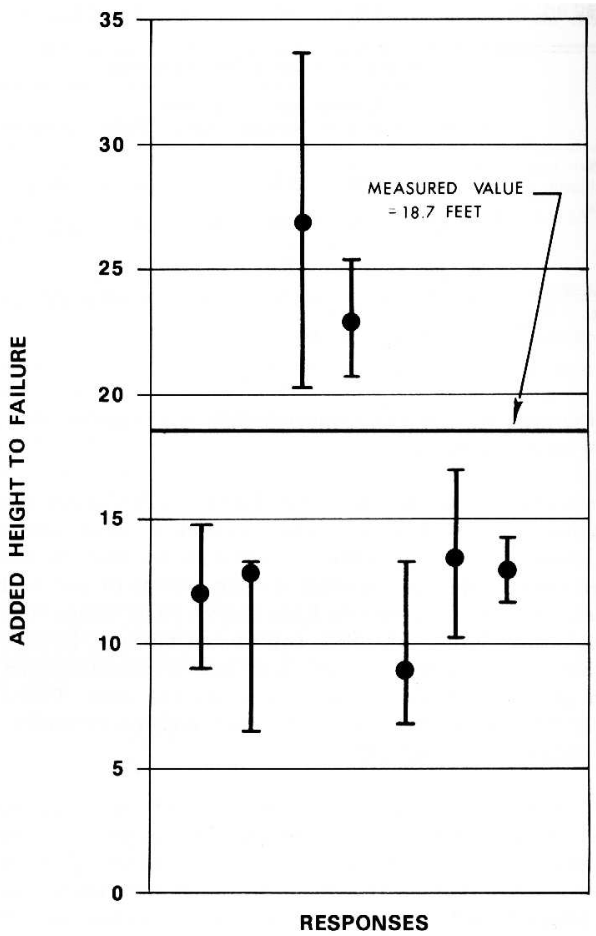


Figure 9.6: Estimates of Dam Failure Height With 50% Confidence Levels

Bibliography

[Ada95] John Adams. *Risk*. UCL Press, London, 1995.

[Ato76] Atomic Industrial Forum. Committee on reactor licensing and safety statement on licensing reform. Technical report, Author, New York, 1976.

[CL79] B. Cohen and I. S. Lee. A catalog of risks. *Health Physics*, 36:707–722, 1979.

[Dah86] Roald Dahl. *Boy*. Penguin Books, London, 1986.

[Dak91] Karl Dake. Orienting dispositions in the perception of risk: an analysis of contemporary worldviews and cultural biases. *Journal of Cross-Cultural Psychology*, 22(1):61–82, 1991.

[Dak92] Karl Dake. Myths of nature: Culture and the social construction of risk. *Journal of Social Issues*, 48:21–37, 1992.

[DW82] Mary Douglas and Aaron Wildavsky. *Risk and Culture: An Essay on the Selection of Technological and Environmental Dangers*. University of California Press, Berkeley, Los Angeles and London, 1982.

[FB92] John H. Fielder and Douglas Birsch, editors. *The DC-10 Case*. State University of New York Press, Albany, New York, 1992.

[Fel90] Fellowship of Engineering. Warnings of preventable disasters conference. Author, London, 6 September 1990.

[FLS+81] Baruch Fischhoff, Sarah Lichtenstein, Paul Slovic, Stephen L. Derby, and Ralph L. Keeney. *Acceptable Risk*. Cambridge University Press, Cambridge, U.K., 1981.

[Gar98] Ken E. Garlington. Personal communication. June 1998.

[Hil93] Mayer Hillman. *Cycle Helmets: the case for and against.* Policy Studies Institute, London, 1993.

[HL98] Michael Höhl and Peter B. Ladkin. Analysing the 1993 Warsaw Accident with a WB-Graph. Technical Report RVS-Occ-97-09, Networks and distributed Systems Group, Faculty of Technology, Bielefeld University, Bielefeld, Germany, 1998.

[Hol79] C. S. Holling. Myths of ecological stability. In G. Smart and W. Stanbury, editors, *Studies in Crisis Management.* Butterworth, Montreal, 1979.

[Hol86] C. S. Holling. The resilience of terrestrial ecosystems. In W. Clark and R. Munn, editors, *Sustainable development of the biosphere.* Cambridge University Press, Cambridge, U.K., 1986.

[Hum75] David Hume. *An Enquiry Concerning Human Understanding.* Oxford University Press, third edition, 1777/1975. Ed. L. A. Selby-Bigge and P. H. Nidditch.

[HV76] M. Hynes and E. Vanmarcke. Reliability of embankment performance predictions. In *Proceedings of the ASME Engineering Mechanics Division Speciality Conference*, Waterloo, Canada, 1976. ASME, University of Waterloo Press.

[Jer97] Robert Jervis. *System Effects: Complexity in Political and Social Life.* Princeton University Press, New Jersey, 1997.

[KH99] Daniel M. Kammen and David M. Hassenzahl. *Should We Risk It?: Exploring Environmental, Health, and Technological Problem Solving.* Princeton University Press, Princeton, N.J., 1999.

[KKS93] Paul R. Kleindorfer, Howard C. Kunreuther, and Paul J. H. Shoemaker. *Decision Sciences: An Integration Perspective.* Cambridge University Press, Cambridge, U.K., 1993.

[KLST71] David H. Krantz, R. Duncan Luce, Patrick Suppes, and Amos Tversky. *Foundations of Measurement, Volume 1: Additive and Polynomial Representations.* Academic Press, New York, London, 1971.

[KST82] Daniel Kahneman, Paul Slovic, and Amos Tversky, editors. *Judegement under uncertainty: Heuristics and biases.* Cambridge University Press, Cambridge, U.K., 1982.

[Lad] Peter B. Ladkin et al. Computer-related incidents with commercial aircraft. Technical Report RVS-Comp-01, RVS Group, Faculty of Technology, University of Bielefeld. Compendium of digitised accident reports and commentary, available through [LR].

[Lad97] Peter Ladkin. Using the Temporal Logic of Actions: A Tutorial on TLA Verification. Technical Report RVS-RR-97-08, Networks and distributed Systems Group, Faculty of Technology, Bielefeld University, Bielefeld, Germany, 1997. Invited Tutorial on TLA, Second International Conference on Temporal Logic, Manchester, England, 14-18 July, 1997, also available through [LR].

[Lad99] Peter B. Ladkin. On classification of factors in failures and accidents. Technical Report RVS-Occ-99-04, RVS Group, Faculty of Technology, University of Bielefeld, July 1999. Available through [LR].

[Lad00] Peter B. Ladkin. Notes on the foundations of system safety analysis. Unpublished Manuscript, May 2000.

[Lam] Leslie Lamport. The Temporal Logic of Actions (TLA) Page. `http://www.research.digital.com/ tla/`.

[Lam94] Leslie Lamport. The temporal logic of actions. *ACM Transactions on Programming Languages and Systems*, 16(3):872–923, May 1994.

[Lap92] J.-C. Laprie, editor. *Dependability: Basic Concepts and Terminology, in English, French, German, Italian and Japanese*, volume 5 of *Dependable Computing and Fault Tolerance.* Springer-Verlag, Wien, New York, 1992. Prepared by IFIP Working Group 10.4 on *Dependable Computing and Fault Tolerance.*

[Lev95] Nancy G. Leveson. *Safeware: System Safety and Computers.* Addison-Wesley, 1995.

[Lev00] Nancy G. Leveson. Personal communication. February 2000.

[Lew73a] David Lewis. Causation. *Journal of Philosophy*, 70:556–567, 1973. Also in [Lew86, ST93].

[Lew73b] David Lewis. *Counterfactuals*. Oxford University Press, Inc., Blackwell, 1973.

[Lew86] David Lewis. *Philosophical papers, Vol.II*. Oxford University Press, Inc., 200 Maddison Avenue, New York, New York 10016, 1986.

[Lew90] H. W. Lewis. *Technological Risk*. Norton, New York and London, 1990.

[LL98] Peter B. Ladkin and Karsten Loer. Why-Because Analysis: Formal Reasoning About Incidents. Technical Report RVS-Bk-98-01, Networks and distributed Systems Group, Faculty of Technology, Bielefeld University, Bielefeld, Germany, 1998. Draft book manuscript available through [LR].

[LR] Peter Ladkin and RVS Group. RVS Group Publications. RVS Group, Technische Fakultät, Universität Bielefeld. Available through `http://www.rvs.uni-bielefeld.de`.

[LT82] E. Lloyd and W. Tye. *Systematic Safety: Safety Assessment of Aircraft Systems*. Civil Aviation Authority, London, 1982.

[Luh91] Niklas Luhmann. *Soziologie des Risikos*. Walter de Gruyter, Berlin, New York, 1991.

[Mac74] J. L. Mackie. *The Cement of the Universe: A Study of Causation*. Clarendon Press, Oxford, 1974.

[Mar92] A. Marin. Costs and benefits of risk reduction. Appendix to [Roy92], 1992.

[Mel00] Peter Mellor. Personal communication. February 2000.

[MH90] M. Granger Morgan and Max Henrion. *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. Cambridge University Press, Cambridge, U.K., 1990.

[Mil73] John Stuart Mill. *A System of Logic, Books I-III*, volume VII of *Collected Works*. University of Toronto Press, London: Routledge & Kegan Paul, 1973.

[MLO98] Claire Marris, Ian H. Langford, and Timothey O'Riordan. A quantitative test of the cultural theory of risk perceptions: Comparison with the psychometric paradigm. *Risk Analysis*, 18(5):635–647, October 1998.

[Nat96] National Research Council Committee on Risk Characterization. *Understanding Risk*. National Academy Press, Washington, D.C., 1996.

[NS75] R. Näätänen and H. Summala. *Road-user behavior and traffic accidents*. North-Holland, Amsterdam, 1975.

[Per84] Charles Perrow. *Normal Accidents: Living with High-Risk Technologies*. New York: Basic Books, 1984.

[Qui64] Willard Van Orman Quine. *From a Logical Point of View*. Harvard University Press, second edition, 1964. Revised.

[Rap81] E. Rappoport. *Unpublished doctoral dissertation*. PhD thesis, University of California, Los Angeles, 1981.

[Res87] Michael D. Resnick. *Choices: An Introduction to Decision Theory*. University of Minnesota Press, Minneapolis, 1987.

[Roy83] Royal Society. *Risk assessment: a study group report*. Author, London, 1983.

[Roy92] Royal Society. *Risk: analysis, perception and management*. Author, London, 1992.

[Sag93] Scott D. Sagan. *The Limits of Safety: Organisations, Accidents and Nuclear Weapons*. Princeton University Press, New Jersey, 1993.

[SFL82] Paul Slovic, Baruch Fischhoff, and Sarah Lichtenstein. Facts versus fears: Understanding perceived risk. In Daniel Kahneman, Paul Slovic, and Amos Tversky, editors, *Judgement Under uncertainty: Heuristics and biases*. Cambridge University Press, Cambridge, U.K., 1982.

[Sky99] Brian Skyrms. *Choice and Chance: An Introduction to Inductive Logic*. International Thompson Publishing, fourth edition, 1999.

[ST90] M. Schwarz and M. Thompson. *Divided We Stand: Redefining politics, technology and social choice*. Harvester Wheatsheaf, Hemel Hempstead, 1990.

[ST93] Ernest Sosa and Michael Tooley, editors. *Causation*. Oxford Readings in Philosophy. Oxford University Press, Oxford, 1993.

[ST95] Eldar Shafir and Amos Tversky. Decision making. In Daniel N. Osherson, editor, *An Invitation to Cognitive Science, Volume 3: Thinking*, chapter 3. MIT Press, Cambridge, Mass., second edition, 1995.

[Sve81] O. Svenson. Are we all less risky and more skillful than our fellow drivers? *Acta Psychologica*, 47:143–148, 1981.

[TEW90] M. Thompson, R. Ellis, and A. Wildavsky. *Cultural Theory*. Westview Press, Boulder, Colorado, 1990.

[TK81] Amos Tversky and Daniel Kahneman. The framing of decisions and the psychology of choice. *Science*, 211:453–458, 1981.

[TR76] R. Thaler and S. Rosen. The value of saving a life: Evidence from the labor market. In N. Terleckyj, editor, *Household production and compsumption*. Columbia University Press, New York, 1976.

[Uni94] United States Air Force. *Air Force Instruction 91-204*. Author, July 1994.

[Wei79] N. D. Weinstein. Seeking reassuring or threatening information about environmental cancer. *Journal of Behavioral Medecine*, 16:220–224, 1979.

[Wil79] R. Wilson. Analyzing the daily risks of life. *Technology Review*, 81(4):40–46, 1979.