# Chapter 12

# What It's All About

Why-Because Analysis represents a direct application of philosophical logic to what we believe is a pressing engineering problem – explaining the reasons for failures, or unwanted behavior of any sort. The importance of explaining failures has been emphasised in two notable books by Petroski [Pet85, Pet94] and the different nature of failure in complex systems was discussed by Perrow [Per84]. Existing engineering methods for explaining failure are, to our mind, largely incomplete [Lev95, Chs 3–5]. We wanted to devise an explanatory method whose product could be demonstrated to fulfil its aim.

## 12.1 The State of the Art

Sometimes things don't happen the way they should. If we spill a glass of water, that's probably no big deal, but if the aircraft in which we're flying augers in, we shall be a lot worse off. Certain incidents need investigating and explaining, because of their unusual social significance, like an airplane crash, but also because we wish to learn how to avoid 'similar' kinds of events in the future. So we must investigate the accident, and describe it in such a way that the similarities to other situations are apparent, and steps may be taken to change the behavior: the behavior of the machinery; of the humans involved, of the computers; maybe of the designers and overseers of the system. That's a lot of social weight to carry.

What is an incident or accident? First, observe that it contains a history: a collection of sequenced events and states in time that really happened. We can regard selecting and describing these events and states as an art, and we must apply this art in such a way as to elicit the similarities to other incidents or potential incidents. For example, a simple film of the occurrence, such as taken during the A320 crash in Mulhouse-Habsheim at an airshow in June 1988, does not suffice as a description, because we feel that we don't see *why* the aircraft behaved as it did. What are we in fact looking for?

Although every event sequence is unique, we may regard the similarities to

other events as real, objective components of the world as well. David Hume considered this problem 220 years ago:

> ....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 [Mil73a]. We shall clarify later what events and states may be.

So we're talking about causes. As David Lewis notes [Lew73a], there are in fact two definitions given by Hume, and over the course of the subsequent years, the consequences of these notions has been explored. The first, in terms of observable regularities that we have been briefly considering, leads to a psychological explanation of causality. The second is *counterfactual* – it talks of what might have been (but was not).

Do we go the psychological route? One wants to say: whatever for? An airplane banging into something is just about as objective an event as could happen, and it seems like a significant detour to wander through psychology in order to explain it. Consider what the U.S. Air Force says about accident explanations [Uni90]:

> **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. Richard Wood, the author of the USAF definition, illustrates the use of this definition on the following example in his manual on accident investigation [WS95]. The list of findings is:

1. The flight crew was properly certificated and qualified for this flight.
2. After the start of takeoff roll, the Number 2 engine oil pressure transmitter failed resulting in a cockpit indication of loss of all oil pressure to the engine.
3. Following flight manual procedures, the Captain rejected the takeoff at a speed of 120 knots, which was 10 knots below $V_1$.
4. The Captain reduced thrust on both engines to idle, deployed the speed brakes and applied full braking.
5. Almost immediately after full brake application, Number 2 tire on the right main landing gear failed at a point on the tire sidewall which showed evidence of previous damage.
6. Shortly after the failure of Number 2 tire, the Number 1 tire on the same axle failed due to overload.
7. The aircraft departed the right side of the runway at a speed of approximately 60 knots.
8. The left main gear failed when is struck the cement base of a runway distance marker which extended eight inches above the grade.
9. THe aircraft came to rest 120 feet from the edge of the runway. There was substantial damage to the aircraft, but no injuries to the crew or passengers.

Wood and Sweginnis comment

> Findings 2, 5, and 8 are each causal. Each had to occur in order to produce the final damage. Each was essential to the sequence of events. If any of those three do not occur, there is no accident.....

Thus appears the counterfactual test, which all three pass. According to the USAF definition, then, these three findings are causes. The USAF adheres to David Hume's second, counterfactual, definition of causality.

The problem which we address in this book is stated succinctly by Wood and Sweginnis [WS95]:

> The term "cause" (or "causes") is not well defined. The ICAO[1] definition (Annex 13) is, "Actions, omissions, events, conditions, or a combination thereof, which led to the accident or incident." There is apparently an assumption that everyone knows what cases are. Wrong! Everyone thinks they know what causes are, but there is little consensus.

We take Hume's second, counterfactual, definition, as used by the USAF, and develop a rigorous form of reasoning concerning counterfactual causality. The result is a method for determining 'findings' and 'causes' which is objective, based on rigorous logical definitions, and in which all assumptions are explicit. This method allows rigorous formal proof that an explanation is correct and relatively sufficient, comparable with the rigor required for formal system verification. We call it '*Why...Because... Analysis*, WBA, and base it on a formal logic, *Explanatory Logic*, or EL.

David Lewis investigated Hume's counterfactual definition 20 years ago [Lew73a] and developed a semantics for counterfactual conditionals that he also axiomatised [Lew73b]. WBA is based on Lewis's counterfactual semantics, and EL is based on Lewis's axiomatisation as well as the Temporal Logic of Actions (TLA) of Lamport [Lam94c] for reasoning concerning sequences of events, human and system behavior. The counterfactual-conditional nature of singular causality (the kind in which we are interested, as opposed to general causal laws) was also considered favorably by John Mackie [Mac74, Ch. 2], who differs from Lewis on this issue primarily in how he explains the semantics of these counterfactuals. The proof of the running example in this book is supported as well by Mackie's view as by Lewis's, although we should not necessarily expect that to be true in general, since Lewis's axiomatisation is not known to be consistent with Mackie's proposed INUS-condition explanation.

So we have preliminary definitions of cause, and finding. But we haven't yet defined what we are trying to explain – accidents or incidents. A more abstract definition of accident, intended for most uses in engineering is given by Leveson [Lev95, Section 9.2]:

> An *accident* is an undesired and unplanned (but not necessaril unexpected) event that results in (at least) a specified level of loss.

A concrete definition for the case of *aircraft accident* is given by the U.S. Federal Aviation Regulations [U.Sc, Paragraph 830.2]:

> *Aircraft accident* means an occurrence associated with the operation of an aircraft which takes place between the time any person boards

[1]The International Civil Aviation Organisation, ICAO, is the UN body created in 1947 responsible for coordinating civil aviation procedures in 185 Contracting States on the basis of the 1944 Chicago Convention on International Civil Aviation.

> the aircraft with the intention of flight and all such persons have disembarked and in which any person suffers death or serious injury, or in which the aircraft receives substantial damage.

We shall be illustrating WBA using incidents from aviation. The principles, however, apply to any domain in which a complex system may have an accident – especially those subject to *normal accidents* or *system accidents* in the sense of Perrow [Per84].

## 12.2 Making the Reasoning Rigorous

*Why...Because... Analysis* (WBA) generates causal explanations of incidents, construed as behaviors of systems. We believe causal and explanatory judgements are objective: that is, that there is a fact of the matter about whether feature $A$ was a causal factor of event $B$, and that this fact of the matter is independent of personal preferences. First, a word on what we mean by '*objective*'.

Objective judgements can be contrasted with matters of taste, for example whether someone likes espresso coffee or not, in which the identity of the person making the judgement is a necessary input to determining the truth value of the predicate. Thus *John judges espresso to be tasty* is an objective judgement, whereas the truth of *espresso is tasty* cannot be determined unless it is understood to be, for example, a paraphrase of *I judge espresso to be tasty* referring to its utterer. So it might be said that there is a 'hidden subject' in statements of personal preference, which must be explicit to allow different people to come to the same judgement as to its truth. We believe that there is no such 'hidden subject' in causal attributions.

Objective judgements (such as that of causality) are also distinguished from subjective judgements (such as that of how tasty coffee is) by their ability to be mistaken. While it is very hard to see how John could genuinely hold a *mistaken* belief that coffee is tasty (if he *believes* it's tasty, isn't that enough for it to *be* tasty for John?), it is not hard to see how I could be mistaken in a belief that a particular lineup of mars and the moon entails that I'm going to get run over by a bus.

So if there is a fact of the matter about causal and explanatory connections, if it's possible to be mistaken, how do we ensure that our causal reasoning is correct? The aim of WBA is to derive a correct and relatively sufficient causal explanation ultimately from simple judgements of causal connection between details of states and events, and to prove rigorously that it is correct and relatively sufficient. This latter feature is, so far as we know, unique to WBA amongst current explanatory methods. The justification for requiring rigor is as follows. An explanation is a form of reasoning, and, in common with most real-world reasoning, the causal explanation of a real event, no matter how simple, is complex enough that reasoning mistakes can be made. A suitable reasoning method should ensure that

such mistakes are eliminated as far as possible. When the reasoning in causal explanations is explicit, then it can be more easily checked for accuracy. The basic tools for representing reasoning in a rigorous manner have existed for more than a century – formal logic. We have already motivated our basic choice of causal logic, namely the logic of counterfactuals of Lewis.

The Lewis semantics proceeds as follows. Suppose $A$ and $B$ actually happened (where $A$, $B$ are *acts, etc.*). Then $A$ is a causal factor for $B$ just in case, had $A$ not occurred, $B$ would not have occurred either. Thus Lewis appears to agree with the USAF, providing we interpret the USAF's 'cause' as Lewis's 'causal factor'. There is a logical subtlety here. The counterfactual definition (Lewis's 'causal factor') is that of a binary relation which is not transitive, that is, if $A$ is a causal factor of $B$ and $B$ is a causal factor of $C$, that doesn't mean that $A$ is a causal factor of $C$. However, 'cause' is usually taken to be transitive: if $A$ is a cause of $B$ and $B$ of $C$, then $A$ is a cause of $C$ also. This is a formal property: if 'cause' is counterfactual, it can't be transitive (the USAF assumes both, and in so doing is logically mistaken). However, it should be transitive and there is a solution. Lewis simply defines 'cause' as the *transitive closure* of the notion of 'causal factor' – the smallest transitive relation containing that of 'causal factor' (the notion is clarified further by the axioms and rules we use for it in EL).

The extra punch of Lewis's approach comes from his formal semantics and logic for counterfactuals [Lew73b], which is very much a main-stream semantics for a modal logic with some necessary bells and whistles. These two steps, the definition of *causal factor* in terms of counterfactuals, and the formal logic and semantics for counterfactuals, form the basis for our formal method WBA.

## 12.3 The Development of WBA

Our first major step in developing WBA was to apply the Lewis semantics for causal factor to two aircraft accident reports, those of Cali 1995 and Warsaw 1993 (see Chapters 23 and 24. We actually started development first on an example of Everett Palmer). We determined the causal relations amongst the significant events and states as listed in the bodies of the reports. We call this process the *WB-Graph method*. We took the nodes as being the *acts, etc.* listed somewhere in the report, and determined the binary relation of *causal factor* between them, using the Lewis definition.

In reworking the Cali and Warsaw reports, we found general agreement between our causal-factor criterion and the reports' judgements, and also some significant discrepancies. Some of these discrepancies related to the formal definition of the standard term *probable cause or contributory factor* that we used (in our definition, such a creature is simply a state or event which is a cause of the accident, but which is itself left causally unexplained). For example, where a report might state that lack of outside visual reference was a factor, or even

omit it altogether because procedures are intended to function also in conditions of poor or absent outside visual reference, our method might be unhelpfully more specific – citing lack of ground lighting and night conditions as separate contributory factors. However, other discrepancies concerned features of the event which were not judged by the report writers to be contributory factors, even though no other factors were adduced as their causes and they satisfied the USAF's definition – the accident would not have happened had they not pertained. Such discrepancies require individual explanation: how could states or events with those features possibly *not* be considered to be causal factors? But no such explanation was given in the reports, neither was there any indication that it might be needed. We concluded that discrepancies of this sort were simply reasoning mistakes. Since we had found such reasoning mistakes, and these mistakes were consequential for considering future prophylactic measures, we concluded that our method would provide an important service for accident investigators: that of rigorously checking the final reasoning.

However, two problems remained. Checking the final reasoning in an accident report is not all one needs to do in the way of reasoning. First, one wants to generate that reasoning, and also the list of *acts, etc.* which participate in it. And second, although the Lewis criterion is formal, we had applied it informally – we had no rigorous check, other than our own carefulness, which would enable us to conclude we had got it right. Perhaps we had made reasoning mistakes too.

We were left with the following questions. Given a rough history of an accident, of the sort normally found in preliminary reports or resumés (for examples, see Chapter 13), how does one go about rigorously constructing a causal explanation? How does one determine what other possible *acts, etc.* to look for? And once one has determined the causal relations amongst these *acts, etc.*, how does one determine that this particular collection of *acts, etc.* provides a sufficient causal explanation of the incident? And how does one determine that one's reasoning is in fact rigorously correct? Finally, how does one put all this together into a uniform method that will work in the large? For it is a notorious property of solutions-in-the-small that they work wonderfully on the smaller problems for which they are chosen, but when scaled up to larger problems in wider environments, one can find that their requirements on specifying and reasoning can begin to conflict with each other.

The full WBA method explained in this book, using the logic EL and the hierarchical method of organising formal proofs, answers these questions.

## 12.4 Some Properties of WBA

An application of the WBA generates a sound and relatively sufficient causal explanation of any incident, based on certain basic judgements of the truth of counterfactuals (and as few as possible of these), along with a formal proof that

this explanation is indeed sound and relatively sufficient. A causal explanation is *sound* if all the causal-factor relations asserted in it were indeed causal factors in the incident. A causal explanation is *relatively sufficient* if, for each state and event adduced in the explanation, the causal factors in the explanation form a sufficient set of causal factors for the state or event indeed to have occurred. Note that relative sufficiency does not say that those sufficient causal factors were all the causal factors that there in fact were: it just says that they are all that one needs to explain why the state or event occurred.

An explanation generated by the WBA is by no means unique. It depends on judgements by the investigator as to the depth of detail to which she wishes to go; conversely, the level of generality she wishes to remain at. For example, some investigators may be content with a single attribution of '*pilot error*' as an explanation of an event; others may wish to analyse this factor further into whether it was a perception, attention, reasoning, decision, intention or action error (according to the PARDIA model we use for classifying human systems behavior – see Chapter 18). WBA expects the user to provide her own guidance on the depth of explanation and as far as we are aware is consistent with most such choices.

WBA is rigorous, based on a formal logic EL and includes the module capabilities of the temporal logic specification language TLA+ [Lam] for stating system specifications, norms and temporal dependencies, extended by the logic *VCU* of Lewis, and admits of rigorous formal proofs of the relative sufficiency of causal explanations. The basic causal primitive is the binary operator of *counterfactual conditional*, a formal logic of which is based on the rules and axioms for counterfactuals in [Lew73b]. From this operator we define the notion of *causal factor* and *cause*, with the semantics of [Lew73a]. We need other operators: an alethic modality for the concept of *necessary implication*, to express that a certain *act, etc.* event follows from a specification and the presence of its preconditional events and states; and a deontic modality to express that operating procedures *should* be followed and specifications *should* be conformed to, even though they may not be in a particular incident. Finally, we need a notion of *sufficient causal explanation* of an *act, etc.*

It turns out that the alethic modality is definable from the counterfactual conditional, and the 'standard' deontic modality is definable from the alethic modality with the introduction of a new propositional constant standing for '*violation*'. We wrote down many of the EL inference rules for these modalities before we were aware of these definitions. Many of the EL inference rules we had written down turned into derived rules under these definitions.

One important operator, that of *sufficient cause*, $\Box\!\!\Rightarrow$, we chose not to define using the counterfactual. We were tempted to do so for technical reasons, namely to reduce all the modal operators, and by implication their inference rules, to the counterfactual conditional, and to inherit the completeness results of Lewis [Lew73b] also for $\Box\!\!\Rightarrow$. However, defining $\Box\!\!\Rightarrow$ in terms of the coun-

terfactual would impose unnecessarily strict conditions on a correctness proof – we would have had to search for a *minimally* sufficient set of causal factors in each explanation we gave, and this seems to be intuitively unreasonable. For example, if a machine gets in a particular state $B$ because it was in a previous state $A$ and performed a particular action $F$, and the specification to which it conformed implies that the state which follows $A$ when $F$ is performed is $B$, then we consider that it is a sufficient causal explanation to note that the machine conformed to its specification, that the machine was in state $A$ and performed $F$, and that the specification entails that state $B$ results when $F$ is performed in state $A$. However, this is usually not a *minimal* causal explanation, because the specification also entails statements about other transitions that have no obvious relevance to the case of $A$, $B$ and $F$.

So we are left with probably incomplete inference rules for $\Box\!\!\Rightarrow$, but no definition of it. This also restricts the scope of the relative completeness results we can assert, which follow directly from Lamport's completeness result for TLA and Lewis's completeness result for his logic *VCU*. $\Box\!\!\Rightarrow$ is not a part of the language for which the completeness result is valid, because it is not apparently definable in the form in which we need to use it either through TLA or through *VCU*.

As well as the EL logic, WBA currently uses the PARDIA human agency classification system, which we also developed for this task. But whereas WBA is dependent on EL, it is not dependent on PARDIA – any other desired human agency model could be used, provided only that some axioms are written for it. We use PARDIA because we found that the classification was necessary for identifying where a particular problem occurred in a chain of human interaction with machine. It is necessary to identify this location for deciding on the domain in which to determine appropriate prophylactic measures, as shown in Figure 18.3. The domains are some combination of human-machine interface design, training, cognitive psychology, operating manual and training manual design, navigation avionics and paper chart design. The PARDIA classification helps identify which sorts of general measures could help avoid repeat problems. It does *not* attempt to explain the human behavior by giving a general model of human agency, neither are we convinced that such a model is necessary for the purposes for which accident investigation is undertaken, although of course it would be likely to help if there were such a generally-established model. We do not necessarily believe that there is likely to be such a model in the near future, and are reluctant to place bets on the future of promising but complex models when a simpler classification seems to suit the purpose of accident analysis.

PARDIA is an extended so-called '*information processing*' classification of human behavior, similar to those used in some aviation human factors research institutions. We anticipate that there will be many such models, both crude and sophisticated, which will be used for different tasks in the future, so WBA is not committed to PARDIA. Choosing such a model is not a matter for logic, but

a matter for enlightened human-factors expertise. We believe we have ensured that WBA remains consistent with the great variety of (if not all) such models, including those that have not been developed and thus do not exist yet. The properties of the human-factors model should be included as a series of modules in the specification language TLA+, divided into axioms (which are intended to capture the meaning of the concepts used and are therefore inviolable during use of WBA) and norms (which describe usual or intended behavior and thus may be violated in incidents – and it may be part of the explanation of the incident that they are so violated).

In order to classify human or system errors, there needs to be some standard against which the behavior of either human or system is compared. For systems, this would be the system specification. Through experience with it, we believe TLA+ (and the associated logic TLA) suffices to specify systems in the manner we require. But what about the human side? There are also norms and specifications of how the human agent should act. In aviation these norms are Standard Operating Procedures (SOPs) of various sorts, and regulations (for example, the U.S. Federal Aviation Regulations [U.Sc]). We specify SOPs and regulations in TLA+ in a similar manner. We write TLA+ modules which specify operating procedures, including aviation regulations and guidance ([U.Sc, U.Sa], air carrier Standard Operating Procedures, Flight Crew Operating Manual Procedures, Air Traffic Control procedures, and so forth), divided into *axioms* and *norms*. The modules which we include here for the PARDIA classification are not intended to be complete or to be the final word on PARDIA, they are intended to suffice for analysing the running example which we use to explain WBA.

While using the full WBA is not for the technically squeamish, it is usable with sufficient training and familiarity, as we demonstrate on the running example. We also claim it is not too complex to preclude use on even the most complex of incidents, as we argue by contrasting the complexity of our fully-worked running example, considering the formal proof to be linear in the size of the WB-graph, against the size of the WB-graph for the most complex aviation accident reports we know (specifically, the 1994 Nagoya accident). Our worked example is roughly a quarter the size of the Nagoya WB-graph (Chapter 25), and although one can expect that a formal WBA reworking of the Nagoya accident may produce more nodes, we regard it as unlikely that it would produce an order of magnitude more nodes. We consider an effort an order of magnitude greater than that we spent on our running example in this book to be eminently feasible. Therefore we propose that WBA, as it stands, is a practical, mathematically rigorous method for failure analysis. We hope the reader will agree with us.

## 12.5 Failure Analysis as Formal 'Debugging'

Debugging computer programs is an art, and is very hard. It constitutes the analysis of failure in a closed system whose only parts relevant to performance, for the most part, are the programs themselves and the hardware they run on. In contrast, *open* systems are those whose behavior and performance are significantly affected by the environment, over which they are supposed to have little or no control. *Heterogeneous* systems are those which have working parts of many different sorts: digital parts, mechanical or other 'physical' parts (a combination of these two is often called '*hybrid*'), human parts, other biological parts. Heterogeneous systems are, by nature, distributed (a human agent is by nature an individual processor, although her role in the system may not be so conceived) and real-time (that is, they must respond to certain hard time deadlines). A *distributed system* of processes is often distinguished from a parallel system by the relative unreliability of communication between its parts. In systems with human components, communication between human and machine can be complex. Because of these qualities, '*debugging*' open, heterogeneous systems appears to be much harder than debugging computer programs.

Open heterogeneous real-time systems (OHRS's) are very often complex, although simple ones can exist. Because of their open, heterogeneous nature, specifying them and understanding their behavior is of itself a complex activity, especially when they fail. Rather than speak of 'debugging', we use the more distinguished term *failure analysis*. Failure analysis involves mostly trying to explain a failure. As we have pointed out in previous sections, explanation is a reasoning activity and it's appropriate to expect failure analysis to have a large reasoning component, which should be explicitly described in order to be objective, thorough, accurate and powerful; hence a formal logic. The idea of formulating appropriate reasoning methods for a problem domain, the 'right logic for the job' so to speak, is elegantly defended in [Haa78] and [Haa96].

We give a brief technical overview of WBA, summarising the features previously remarked. A formal logic involves a formal language, with axioms and formal inference rules showing which forms of formal sentences may be 'derived' from which others, all-in-all representing a formal encoding of the assumptions and legitimate reasoning steps that (the investigators hope) lead to the desired conclusions. A formal analysis method involves a little more than that, namely

- a logic (a set of *well-formed formulas* (wffs), axioms and inference rules amongst the wffs);
- a semantics, an intended meaning or interpretation, for the wffs;
- formal guidelines for using the logic (meta-axioms) in analysis of the problem domain.

WBA is designed for, and we believe a suitable failure analysis method for, open heterogeneous systems and OHRS's. The associated logic EL is a multi-modal logic, involving counterfactual conditionals, causal logic, tense logic, deontic logic and alethic modal logic. A significant part of EL is formulated using Lewis's axioms for the counterfactual conditional primitive, and definitions of the causal, alethic and deontic modalities from this counterfactual; furthermore tense-logical wffs occur strictly 'inside' the other modalities, so that one can treat the tense-logical formulas as propositional primitives for the counterfactual-causal-alethic-deontic part. Soundness and (the reasonable) relative completeness results for EL follow.

The formal method WBA is supported in the following ways:

- use of a formal semantics for causality (specifically for an event or state being a '*causal factor*' in the occurrence of another event or state) [Lew73a], which has been shown to be fruitful in analyses of the complete causal reasoning in aircraft accident reports [GLL97a, HL97];

- the use of a formal specification method for system subparts (for example, aircraft subsystems) and procedures (for example, ATC handoff communications) which is supported by a formal tense logic (TLA [Lam94c]) and has proved itself fruitful in the specification and verification by hand of non-trivial system components and algorithms;

- use of an analogous specification method for 'human' components: SOPs, regulations and so forth;

- the use of rigorous, completely formal proofs of correctness and relative sufficiency for explanations using Lamport's hierarchical proof method [Lam95b], which guarantees a rigorous proof with no informal reasoning steps.

We work in subsequent chapters through a running example, an aviation incident we wish to explain. The end result is a causal analysis, complete with formally-described uncertainties, along with a formal proof that the analysis is correct and relatively sufficient (sufficient to provide a causal explanation at the level of detail to which it was chosen to take the analysis).

We make no claim that WBA and EL as presented here suffice to provide all possible causal explanations for all possible events. First, logic only goes so far, and we're interested in explaining real events and states that happened. We do not require that the events themselves be described in a formal language; were we to do so, we would have to solve the problem of knowing when a formal description suffices, and to solve the problem of when events may be individuated through their descriptions alone (there is significant philosophical justification to remain agnostic – see [Dav80, Chapters 6-9]). Thus considerable weight lies in WBA on expert judgements of the truth of certain counterfactual conditionals. However,

the framework we build around these judgements is rigorous, and therefore suffices to concentrate the expert attention where it is required to justify the causal analysis through the assessment of the counterfactuals. In the formal proof, these expert judgements will appear as assumptions, from which the analysis follows by pure logic alone. The assumptions are explicitly listed, and it is therefore a theorem of logic that the conjunction of sentences which represents the WBA analysis follows from the conjunction of these assumptions.

We developed many inference rules for EL, to discover later that they were in fact derived rules of Lewis's logic *VCU* [Lew73b], which served to suggest to us that we were on the right track. We use strict implication and deontics in conjunction with the Difference Condition (a criterion due originally to Mill) in order to describe the consequences of procedures that were followed as well as the consequences of them *not* being followed. It was intuitive confirmation that we were on the right track that the inference rules originally devised for EL sufficed to complete the proof of the example (with the possible exception that we originally forgot one major rule, but rapidly perceived the need for it). Also, the domain axioms specifying the procedures were added to, but not essentially changed. This confirms our judgement that a stable set of axioms and rules can be found to specify the domain. In the example, the domain consisted of two facets: landing procedures, and ATC handoff procedures. We engaged in moderately 'lazy' specification, defining facets as we needed them, rather than the 'greedy' approach of attempting to axiomatise a facet before knowing what was needed.

The proofs of correctness and relative sufficiency are formulated in EL using the hierarchical proof method suggested by Lamport [Lam95b]. In a hierarchical proof, one starts with the conclusion, and enumerates premisses from which this conclusion follows by an inference rule; then one does the same for the premisses; and so on until one needs to go no further. Because there are many premisses for a given conclusion, this process branches. *Axioms* are inference rules with no premisses, so of course an axiom terminates a branch of this process. A branch may also be terminated by an explicit *assumption*. The assumptions we make are generally of two forms:

- formal interpretations of natural-language sentences
- counterfactual conditionals involving events and states

The former are justified by appeal to inspection. The latter are usually justified by an informal argument showing that they are true under the Lewis semantics. Both are included in the list of assumptions. The correctness and relative sufficiency of the analysis follows logically therefore from the assumptions. It is of course possible that one may dissent from the truth of certain of the assumptions (else they wouldn't be assumptions) but our aim is to reduce the explanation to assumptions of which the plausibility will be generally agreed.

WBA proceeds as follows:

- we establish a rough history (temporal succession) of the incident with a few universally-accepted facts (a few putative 'findings' in USAF terminology);
- we attempt to convert the temporal succession into a causal chain;
- we specify relevant procedures and subsystems rigorously, at a level of detail appropriate to analysis of the incident (deeper where things go wrong, high-level where no fault is suspected);
- we handle faulty procedures and faulty devices (devices that do not satisfy their specifications) through the use of a deontic operator, adding '*non-events*' where something was required to happen by procedure but didn't[2];
- we use Predicate-Action Diagrams (PADs) [Lam95a] to describe indeterminate situations with precision;
- we prove formally the correctness and explanatory sufficiency at the chosen level of detail of the causal explanations uncovered;

The formal proof is necessary. We found that even using our judgement to determine what was a causal factor of what, even using the Lewis semantics informally, we needed to correct some steps that after a closer look during the formal proof we realised could not be proved in EL. This experience is consistent with that of using formal proofs in system verification. Informal proofs just do not suffice for 100% accuracy – people make mistakes.

[2]First suggested in [LP96].