The Deconstruction of Safety Arguments Through Adversarial Counter-Argument

James M. Armstrong[1] and Stephen E. Paynter[2]

[1] Centre for Software Reliability, School of Computing Science,
University of Newcastle Upon Tyne, United Kingdom.
J.M.Armstrong@newcastle.ac.uk
[2] MBDA UK Ltd, Filton, Bristol, United Kingdom.
stephen.paynter@mbda.co.uk

Abstract. The project Deconstructive Evaluation of Risk In Dependability Arguments and Safety Cases (DERIDASC) has recently experimented with techniques borrowed from literary theory as safety case analysis techniques. This paper introduces our high-level method for "deconstructing" safety arguments. Our approach is quite general and should be applicable to different types of safety argumentation framework. As one example, we outline how the approach would work in the context of the Goal Structure Notation (GSN).

1 Deconstruction in a Safety Context

French philosopher Jacques Derrida's concept of *deconstruction* rests upon the idea that, ironically enough, the meaning of an argument is a function of observations that it excludes as irrelevant and the perspectives that it opposes either implicitly or explicitly. On the one hand, if we recognise an opposing argument explicitly, we might be tempted to misrepresent it as weaker than we really feel it to be; but if this misrepresentation is detected, or if our own arguments do not convince, we may succeed only in perpetuating the opposing view. On the other hand, if we try to suppress our acknowledgment of credible doubt, we leave the reader mystified as to why we feel the need to argue our conclusion. To 'deconstruct' an argument is to try to detect such failures of "closure". Such failures need not necessarily lead one to an opposed conclusion (Armstrong & Paynter 2003, Armstrong 2003).

A deconstruction of an argument tries to show how the argument undercuts itself with acknowledgements of plausible doubts about its conclusion and betrays a nervous desire for the truth of assumptions and conclusions rather than unshakeable confidence. This perspective recognizes that deductive argument is unequal to the tasks of resolving contradictions and unifying the different explanatory narratives that underlie our debates. The deconstruction of a deductive argument has two stages. The *reversal* stage develops a counter-argument from clues offered within the original argument; the *displacement* stage compares the two arguments. In the safety assessment context we view reversal as an opportunity for the reassessment of the existing safety acceptance criteria.

A safety argument is required to be inferentially valid in some sense and its empirical premises must be *justified* in such a way that they seem plausible. Empirical

M. Heisel et al. (Eds.): SAFECOMP 2004, LNCS 3219, pp. 3–16, 2004.

claims can attain the status of knowledge only by means of supporting evidence of varying reliability. This is recognized in logics of justified belief that allow premises to be "warranted" to differing degrees; for example, Toulmin (1958). Starting with the *reversal* stage of safety argument deconstruction we ignore the warrantedness of the premises: instead, we try to produce a counter-argument that seems warrantable. Hence we provisionally assume that we could find sufficient evidence for justified belief in our counter-argument. In the *displacement* stage we deal with the relative strength of the warrants and backing evidence for both argument and counter-argument. Hopefully, after reversal we will be able to see that one argument (or both) is (are) unsatisfactory and act accordingly (either accept the system or require more risk reduction). However, there is a possibility that we get two opposing arguments that are "sufficiently" warranted. A deconstruction must explicitly recognize and analyze this particular failure of "closure". To question the "closure" of an argument is to try and find a possibility that has been excluded but which when re-introduced undermines faith in the argument by suggesting a plausible counter-argument. Thus the process of deconstruction is in the final analysis *adversarial.*

Section 2 of this paper presents a brief example of safety argument deconstruction using the Goal Structuring Notation (GSN). As yet we have no pragmatic justification (e.g. cost-benefit) for the use of safety argument deconstruction in safety processes. Therefore, in Section 3 we confine ourselves to a philosophical justification in terms of the lack of deductive closure in any non-absolute argument: we show that when safety decision makers act upon "sufficiently justified" beliefs – as they do when they accept or reject safety-critical systems – they are necessarily committing themselves to a variant of the 'lottery paradox'. We explain this using a *Warranted Deduction Schema* we have developed for the comparison of arguments and counter-arguments. Sections 4 examines political aspects of deconstruction in terms of the *Warranted Deduction Schema*. Section 5 outlines future issues in the pragmatic justification of safety argument deconstruction.

2 An Example: The Goal Structuring Notation

The example deconstruction in this section is done in the context of the Goal Structuring Notation (GSN) and is adapted from Kelly (1998). The example argues a sufficiency of protection against a risk of catastrophic failure. In the source text, the example is only part of a larger GSN argument and thus some of the questions we put are answered there or are not relevant. We have taken the example out of its original context to illustrate the process of deconstruction. GSN is intended to make the structure of arguments clearer than in free text. Thus it provides a neutral and convenient format for the (de)construction of safety counter-arguments. GSN specifies:

- Goals (best expressed as predicates)
- Goal Decomposition (top down)
- Strategies (for explaining goal decompositions)
- Solutions (direct information sources)
- Justifications (for explaining rationale)
- Assumptions

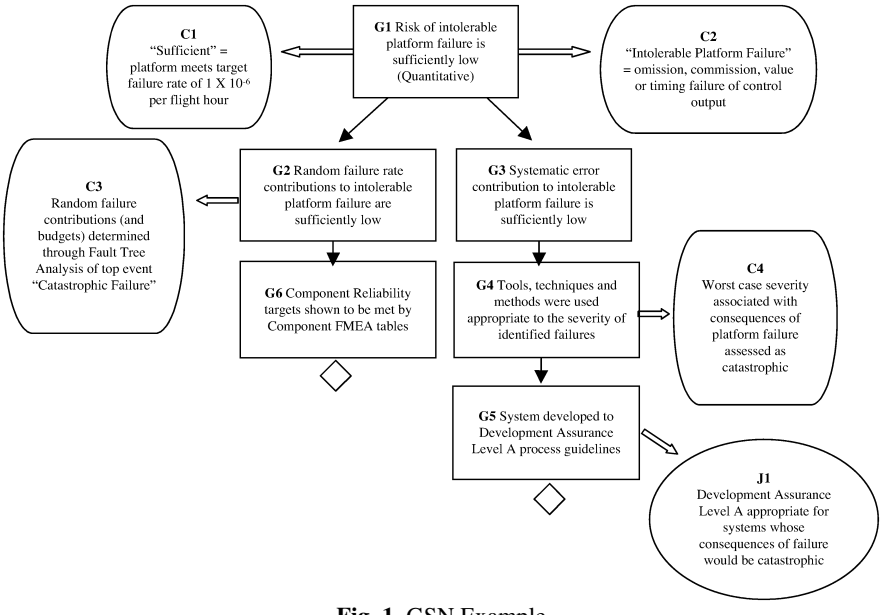

Fig. 1. GSN Example

There are also links to information and factors outside the scope of the argument itself: Contexts (for describing the circumstances of the argument), and links to Models of a system. Hence it is a simple matter to define a "shadow" GSN that provides a starting point for the construction of counter-arguments:

- *Anti-goals* (negations of the stated goals)
- *Anti-goal Deconstruction* (questioning the verifiability of a goal, the consistency between its anti-goal and stated subgoals, and the mutual independence of subgoals)
- *Tactics* (goal decompositions without an explicit strategy)
- *Questions* (to be placed against solutions)
- *Presuppositions* (unexplained rationale behind justifications)
- C*ounterassumptions* (negations of assumed facts)

In the reversal stage, GSN contexts and links to system models should be taken as givens. However, during displacement, if a counter-argument proves fruitful, the context in which it is stated may diverge in important ways from the original and this should be recorded in it.

2.1 Reversal

Given that the meanings of "sufficient" and "intolerable platform failure" are made clear in Fig 1, the negation of the top-level goal G1 to give an anti-goal is trivial. Looking at the decomposition of G1 we can see that the argument depends upon a distinction between random and systematic failure contributions, but this distinction is left unexplained. The deconstructor would hypothesize the absence of any explana-

tory strategy as an argumentative *tactic*. The way in which the two rates have been combined in the example is not clarified: for example, one can ask whether in order to get random failure rates sufficiently low the design has not used a complex scheme for redundancy that has made systematic errors more likely.

Furthermore, the distinction between "random" and "systematic" failures can be questioned. For example, "random" failure rates for hardware vary with intended operating conditions and it could be that Fault Tree Analysis (C3) has not taken account of this.

For "systematic" failure rates the chain of goals G3-G4-G5 could indicate flawed reasoning. For example, the negation of G3 is not in contradiction with G4. A presupposition behind J1 is that Development Assurance Level A and its associated tools, techniques and methods are "appropriate" for systems whose consequences of failure would be catastrophic. This most likely means that Level A development is required where failure is catastrophic; but it probably does not mean that adherence to Level A is considered sufficient to bound the predicted *rate* of systematic failure, or that the prediction must remain below any specific threshold of acceptability. Still less can a process be expected to bound the *measured* rate of catastrophic failure, as this is dependent upon the level of exposure to the system and its hazards that society chooses to accept.

The example argument omits system and environment models from which systematic failure rate predictions must be derived. Instead, it argues that a Level A development process is commensurate with an acceptable systematic failure contribution. However, the best contribution that a development process can make to a systematic failure rate prediction is assurance that it provides the right context for the detection of unreliable systematic failure predictions: historically, it should have supported the derivation of reliable predictions, whatever those predictions might have been. Assuming this to be the case, the argument remains incomplete without the models that justify a specific predicted figure.

We can also speculate that justification J1 would be especially fallacious if the definition of Development Assurance Level A recognised that its tools and techniques – while appropriate for handling catastrophic hazards – were insufficient for the attainment of definite systematic failure rates. In such a case, the goal chain G3-G4-G5 would constitute a non-compliance with Level A and we might consider the argument to be what philosophers sometimes call a *performative self-contradiction* (a non-compliant assertion of compliance).

Such questions would lead to the counter-argument in Fig 2.

2.2 Displacement

The original GSN argument would be considerably improved by:

- The addition of a specific systematic failure contribution estimate
- Linking in system specification, test evidence, and hazard models as solution evidence for the systematic failure prediction G3
- Adding a strategy showing how the systematic failure rates were combined with the random failure estimate to give G1
- Stressing that goals G1 and G2 are only predicted failure rates in their text

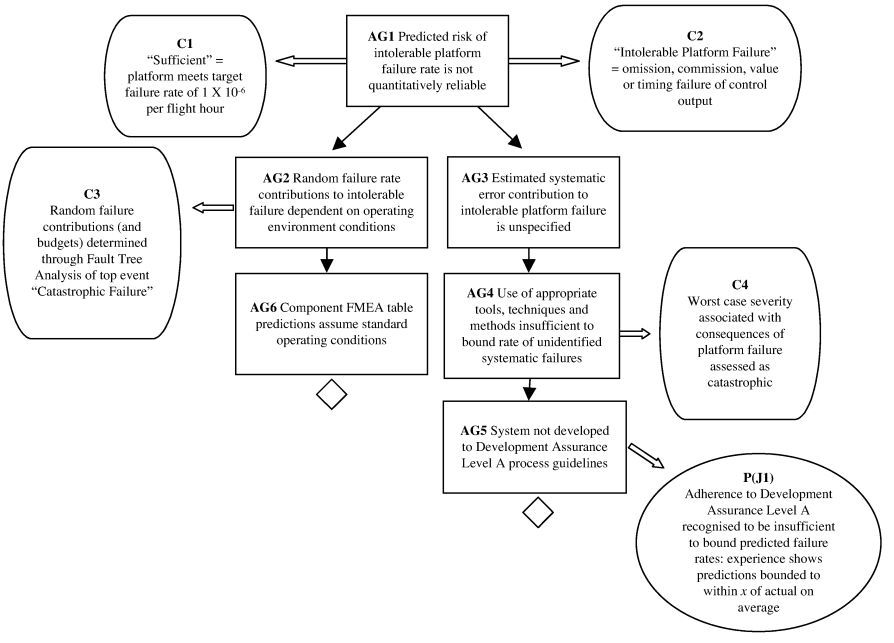

Fig. 2. Example GSN Counter-argument

- Making G5 part of the context of G1
- Combining G4 and J1 into a justification of the systematic failure rate prediction (by attachment to the modified G3)

A general conclusion that emerges from the deconstruction is that even when a top-level safety goal is clearly stated, the conditions of its verifiability might presuppose the acceptance that is supposedly being argued. Evidently no analysis can predetermine a predicted rate as equal to the measured rate; system acceptance is a condition for verifying such a prediction. Thus when a predicted catastrophic failure rate of 10^{-6} is set as a "goal", there is a risk that the safety argument focuses solely on finding a way of expressing the required prediction. The problem with our example argument is that only goals G6 and G4 can be reasonably considered verifiable *before* acceptance, whereas G1 and G2 and G3 have to be understood as predictions that could be verified only after a presupposed acceptance.

Such goals might be better thought of as conditions for continued system acceptability: for example, the system could be temporarily withdrawn for modification if the failure rate ever exceeded the predicted rate. However, it is sometimes the case that a system is withdrawn for modification as soon as it fails catastrophically *within* the predicted rate, or even after a "near-miss". There seems to be an implicit distinction between the acceptability of a predicted rate of catastrophic failure and the acceptability of a near-miss or an actual catastrophic failure, even where the rate prediction admits these possibilities. Put simply, we accept abstract and idealised dangers as they are predicted in safety arguments more readily than we accept the empirical prospect of danger or its actual consequences.

In this case, predictions of catastrophic failure rate are best seen as a basis for assessing and controlling exposure to danger, so that "over-egging" the predictive power of a development process is to be avoided. Since no development process could be shown to bound *measured* systematic failure rates without reference to a specification (which determines which behaviours are considered as failures) and since the process arguably *limits* the reliability of predicted failure rates, to do so might breed false confidence on all sides.

3 A Philosophical Justification for Safety Argument Deconstruction

To have any force, a safety argument must really consist of an argument and a "meta-argument". The argument says that from a certain set of premises $P_1,\ldots, P_n$, the conclusion C follows. The conclusion might consist of a number of claims $C_1,\ldots, C_n$ but we will assume a single conclusion for simplicity. We will use *deductive validity* as the interpretation of "follows" for now, but we do not mean this to be taken too literally: we expect that any kind of inference rule could be used – including rules that allow exceptions, such as Toulmin's (1958) "warrants" with their backing evidence and qualifiers. Thus our safety arguments will be of the form:

$$P_1,\ldots, P_n \mid\!- C$$

Since deductive validity does not guarantee that the inference is *sound*, the argument must also justify belief in each premise. For now we will not worry about the strength of this justification. Thus the argument being offered is equivalent to:

SA $\quad justified(P_1),\ldots, justified(P_n) \mid\!- justified(C)$

To have the force that it has, the argument needs to claim that because the premises are justified and the conclusion follows from them, then the conclusion is also justified. There is usually a constraint on the strength of the claim. We shall consider this matter presently.

The principle relied upon to make inferences of warranted beliefs from warranted premises is called the Deductive Closure Principle (***DCP***). As Olin (2003, p.83) expresses it:

Deductive Closure Principle (DCP). If you are justified in believing $P_1,\ldots, P_n$ and $P_1,\ldots, P_n$ jointly imply Q, and you see this, then you are justified in believing Q.

Olin gives several reasons why this principle is more problematic than it first appears. Her observations relate to the "lottery paradox". In this paradox, we hold a ticket in a thousand-ticket lottery. We know one ticket will win. We assess the probability against winning as 999/1000 and decide that this level of probability justifies us in believing we will not win: but our ticket is just like all the others, and our inference is therefore equally justified for all other tickets; which would mean that no ticket would win. So in making the inference that we will not win we are implicitly accepting contradictory propositions (Olin 2003). We have drawn analogous conclusions from a consideration of the underlying logic of acting upon evidentially justified beliefs that is the basis of the safety process.

Consideration of ***DCP*** leads to a variant of the lottery paradox that applies to the notion of a "sufficiency" of confidence in a defeatable statement, such as a safety claim. Note the words "and you see *this*" in ***DCP***: Derrida's deconstruction is concerned with what happens when one sees "this" but does not see what is opposed to "this". It suggests that we ask what the opposite of "justified" could be. The answer "unjustified" yields a problem: "unjustified" does not mean the same as "unjustifiable". All "justified" means is that a good justification for a statement has been constructed. All "unjustified" means is that no good justification has been offered so far. Ironically, a claim of "unjustifiability" would be itself *unjustified* in the context we are considering: in denying the possibility of empirical justification to a given statement, "unjustifiability" makes an implicit appeal to the absolute truth of the negation of that statement. We are trying to minimize reliance on such appeals for empirical premises; yet we cannot assert the "unjustifiability" of an assertion where, to avoid the charge of scepticism, *all* empirical reasoning has been put on non-absolute and evidential grounds. However, we will continue (at least provisionally) to adhere to the view that non-empirical (i.e. logical) contradictions are unjustifiable.

We have found a number of problems with ***DCP*** as Olin (2003) formulates it. Firstly, it is not clear whether "implies" is a material implication and if it is, whether we are to believe that the inference $P_1,\ldots, P_n$ *implies* Q is itself justified. Secondly, since *false* implies every statement, ***DCP*** can be used to deduce belief in contradictions from contradictory premises. We fix these problems by requiring that the set of premises $P_1,\ldots, P_n$ be *satisfiable* (their conjunction is logically consistent) and that there must be a deductive argument from them to the conclusion in question:

Strengthened Deductive Closure Principle (SDCP). Given:

a) a set of premises $P_1,\ldots, P_n$
b) a justification for each premise in $P_1,\ldots, P_n$
c) an argument that $P_1,\ldots, P_n$ is a *satisfiable* set of premises
d) that $P_1,\ldots, P_n \vdash Q$ is a *deductively valid* inference
e) and one sees this,

then one is justified in believing Q.

Below we develop a "deconstructive" schema for dialectical argument that allows different levels of confidence to be assigned to a statement. We follow Toulmin (1958) in allowing that what he calls a "warrant" – an inference rule – itself needs to be justified. It can be strengthened by backing evidence or weakened by data about exceptions. Thus confidence in a justification is a matter of degree. Rather than introduce backing as a separate term, we define the level of confidence in a justification as a function ω that maps the claimed statement to a value n where $0 \leq n \leq 1$. We refer to this as the *warrantedness* of the statement. Warrantedness can be absolute or zero, so that we can include total certainties and unwarranted statements should they be claimed.

We also need to be able to record how far the deductive argument itself is warranted. For example, we might use a proof tool to do a deduction, and have some worries about its reliability; or if the derivation is long and complex, we might have doubts about our own capability to check it. If we accept such possibilities, we have to adopt a logically weakened version of ***SDCP***:

Warranted Deduction Schema (WDS). Suppose we have:

a) a set of premises $P_1,\ldots, P_n$
b) a degree of warrant ω, where $0 \leq \omega \leq 1$, for each of $P_1,\ldots, P_n$
c) an argument that $sat(P_1,\ldots, P_n)$, i.e. we have a *satisfiable* set of premises
d) a deductively valid inference $P_1,\ldots, P_n \mid\!- Q$
e) a degree of warrant ω for the argument in (c)
f) a degree of warrant ω for the deduction in (d)

then we are justified in believing $\omega(Q)$, where this is defined as:

$$\omega(P_i) \times \omega(sat(P_1,\ldots, P_n)) \times \omega(P_1,\ldots, P_n \mid\!- Q_i)$$

where $\omega(P_i)$ is the minimum of the warrants in b): $min(\{ x \mid x = \omega(P_j)$ for all j in $1 .. n\})$

Note that a consequence of ***WDS*** is that a zero degree of warrant for any statement or deduction immediately nullifies the degree of warrant for the conclusion. The derivation of a certainty would require all statements and deductions to be certain. In bounding confidence in our argument by the least warranted premise, we have adopted a conservative approach. We do not currently allow the mutual consistency of premises to increase confidence in the set as a whole.

A theory of inductive justification requires the idea of a *sufficient* degree of warrant that justifies belief in a defeatable statement. This is analogous to a basic presupposition of probabilistic reasoning that the lottery paradox puts into question (see Olin 2003, p. 79):

Principle of Sufficient Warrant (PSW).
There exists a degree of warrant ω, such that $0.5 < \Omega \leq 1$, and such that if statement P has warrant $\omega(P) \geq \Omega$, then we are justified in believing that P.

In what follows, where a statement P has the sufficient degree of warrant Ω, we will simply write $\Omega(P)$. This principle leads to a contradiction analogous to the lottery paradox as we shall show.

Consider the notion of a *least sufficiently warranted argument*. This is an argument in which Ω is attained exactly for all premises and inferences:

LSWA1

$\Omega(P_1), \ldots, \Omega(P_n)$	(the premises are all warranted)
$\Omega(sat(P_1,\ldots, P_n))$	(the satisfiability of the premises is warranted)
$\Omega(P_1,\ldots, P_n \mid\!- Q)$	(the deduction is warranted)

Therefore, by ***WDS*** we have $\Omega(Q)$.

However, suppose we have *not seen* the following counter-argument:

LSWA2

$\Omega(A_1), \ldots, \Omega(A_n)$	(the premises are all warranted)
$\Omega(sat(A_1,\ldots, A_n))$	(the satisfiability of the premises is warranted)
$\Omega(A_1,\ldots, A_n \mid\!-$ *not* $Q)$	(the deduction is warranted)

Therefore, by ***WDS*** we have Ω (*not* Q).

It may be that we could not find such an argument even if we looked for it. However, the possibility of an argument of ***LSWA2***'s form is not denied by the mere fact that we found ***LSWA1*** first: perhaps if we had set out to prove *not Q* we would have found ***LSWA2*** first and missed ***LSWA1***. What justifies the "blindness" when we see that ***LSWA1*** is valid and claim we are therefore justified in believing *Q*?

For example, a dishonest attempt at persuasion might avoid drawing attention to a sufficiently warranted counter-argument that has already been made. The case where one can sense the possibility of ***LSWA2***, but cannot pursue the matter further is a very difficult one and not remote from everyday life. We may feel a particular conclusion is forced upon us by the circumstances we are in. Thus force of circumstances can defeat the requirement for sufficiency of warrant before action. A deconstruction of a safety argument will look for implicit clues to uncontrolled factors, but will also try to understand the nature of the "force" of circumstances where *force majeure* is explicitly claimed: for example, one can ask how far the force of circumstances was a result of previous freely-taken decisions.

Our difficulty derives from the fact that nothing in ***WDS*** or ***PSW*** allows us to claim that an argument of the form ***LSWA2*** cannot exist: all inductive reasoning is *defeatable* in the light of new information, which might make possible an argument of the form ***LSWA2***. What we call "twenty-twenty hindsight" sometimes reveals just such an argument. Whilst acting on ***LSWA1***, one might claim that the existence of an argument of the form ***LSWA2*** is highly improbable. However, ***LSWA2*** is only a schema, so there could be an infinite number of such arguments, or none; in practice, we can at best estimate the amount of effort spent on trying to find a counter-argument and take our assurance from how hard it is to find one. This is why we propose the formulation of the best possible counter-argument to a safety argument before system acceptance.

We can express the dilemma of choosing between equally strong but opposing arguments more explicitly as the decision to accept or reject the following conjecture:

Sufficient Warrant Conjecture (SWC): *Ω(P) implies not Ω(not P).*

So long as we have ***LSWA1*** and noone has found any ***LSWA2***, we can substitute Ω(*P*) into ***SWH*** in order to state that the opposite conclusion is not sufficiently warranted: *not* Ω (*not P*). But of course this does not mean it is *unwarrantable*. If someone does find an argument of form ***LSWA2*** and *we have not seen it*, they can then also use ***SWH*** to argue *not* Ω(*P*) and we have a contradiction with Ω(*P*).

So perhaps we decide to deny ***SWC***, so that we believe: *not*(Ω(*P*) *implies not* Ω(*not P*)). In that case, one can ask what interpretation should be attached to the following statements:

Insufficiency of Warrant Conjectures:

1. Ω (*P*) *and* Ω (*not P*) – a statement and its opposite can be sufficiently warranted
2. Ω(*P*) *and not* Ω(*P*) – a statement can be both sufficiently warranted and insufficiently warranted
3. *not* Ω(*P*) *and not* Ω(*not P*) – no sufficient warrant exists for either alternative

4. $\Omega(P \text{ and not } P)$ – warranted belief in contradictions (radical *dialethism*)
5. *not* $\Omega(P \text{ and not } P)$ – non-belief in contradictions (the classical principle)

We do not argue for (4) here. Interestingly however, logician Graham Priest (2002) does make a relatively strong case for dialethism; see Olin (2003, Chapter 2) for a critique.

We accept (5) provisionally, but show that doing so leads us to a "meta-problem" in choosing whether to accept (2). The denial of ***SWC*** makes (1) consistent: we have to accept the possibility that a statement and its negation can both be warranted to a sufficient degree; but suppose that we have derived $\Omega(P)$, and act as if P. The action is an implicit appeal to ***SWC*** (which we have denied) in order to deny (2). Do we not act *as if* we believe there is no equally warranted argument for $\Omega(not(P))$? Certainly, we can only act confidently in the belief that the sufficiency of our warrant is not defeated even as it is asserted.

The contradiction involved in our thinking comes into sharper focus in a situation where we can actually *see* an adequate counter-argument to P, and have to decide to act as if P or *not* P. In such a case, we are forced to recognise (2), as it follows from (1) and (5): the "sufficient" warrant of the one argument defeats that of the other. Since we cannot tell which argument is at fault, we have a case where (2) seems to be true for each argument.

We interpret this possibility as meaning that where we have equal 'sufficient' belief in arguments for P and for *not* P we have *no* justification for believing either (3). Unfortunately, if we were to capture this principle in an inference rule, we would create an unsound logic that allows (2):

defeatability_of_warrant(dow) $\Omega(P)$, $\Omega(not\ P) \mid\!\!-\ not\ \Omega(P),\ not\ \Omega(not\ P)$

So far we have assumed that we have arguments for both P and *not* P that meet the requirements for sufficiency exactly. In the more usual case where one claim seems to have a higher degree of warrant than the other, one usually chooses to act according to the more warranted argument; but on what grounds? If the "sufficient" degree of warrantedness really is sufficient for belief, then there is no advantage to be gained from *more* than "sufficient" confidence; the opposed arguments should still disqualify one another. Where we have sufficiently warranted arguments for both P and *not* P, as (1) allows, then we should in practice have *no* confidence in either statement.

If P and *not* P are opposite outcomes of some type of trial, and we decide to test which argument is correct by direct observation, then from our viewpoint and in our circumstances, the outcome is a matter of sheer chance. If we try to justify the undertaking of the trial by means of only one of the arguments, then the warrant for the argument we act against is not only simultaneously sufficient and defeat*able*, it is simultaneously sufficient and defeat*ed*.

For example, say we have equal warrant for *heads* and *tails* in a coin toss – the argument that $\mathbf{P} = 0.5$ for both outcomes; then naturally, we can have no justified confidence that either outcome is more likely. However, say we ended up with arguments of $\mathbf{P} = 0.6$ for *heads* and $P = 0.7$ for *tails*, where we needed $\mathbf{P} = 0.51$ for sufficient belief. Since this is a contradiction according to probability theory and we accept (5), the situation is no different; both arguments must be insufficient for confidence. If we leave the matter at that, we have *no* warrant for predicting what the outcome will be at all.

In practice, other courses of action than trial might be available to us. We could appeal to factors not covered in either argument, effectively trying to find a third which is more warranted than either of the first two; or we could appeal for more work to be done to test one or the other argument, or both. Nonetheless, these courses of action imply that we have invalidated the degree of sufficiency for belief Ω. Indeed, we must invalidate it, for not to do so is to accept the justifiability of belief in contradictory statements (4).

Thus, for any predictive argument at a particular time, if the discovery of an equally good or better counter-argument is possible at that time, then we do not really *know* the degree of sufficient warrant for the argument we have: in acting on it, we merely *assert* belief in it. At best we can claim that we have expended as much effort as possible on trying to find counter-arguments against our prediction and can see no reason why it should go wrong; if noone else can either, that is the most assurance we are ever going to get. Thus our schema illustrates that the "sufficient" level of belief we attach to a statement retains a potential for destabilisation and is itself something that can be renegotiated in the light of experience. This goes some way to explicating the problem behind the oft-asked question "how safe is safe enough?" and why a definitive and context-free answer cannot be given.

In a dilemma such as this, force of circumstances is often offered as a justification for following a certain course of action. Such assertions need to be considered carefully. We can ask the following questions:

a) How does a party (perhaps ourselves) represent the circumstances they are in to themselves?

b) Is the representation accurate, e.g. does it symbolise hidden value systems, *emphasise* certain interests and *de-emphasise* or *exclude* others?

c) When they act, does a party use political power to *change* the circumstances whilst arguing that they are subject to them?

These are key themes of postmodern philosophy and questions that might help safety assessors understand the "safety culture" of an organisation that puts a safety-critical system up for acceptance.

However, they are also questions that assessors should ask of themselves, since to ask a), b) and c) at all presupposes a viewpoint that differs from the viewpoint being assessed.

4 The Politics of Safety Argument Displacement

The safety process sometimes involves 'meta-arguments' about the acceptability of prearranged acceptance criteria as well the adherence of a system to them. In an adversarial approach, the worst-case scenario is formally "warranted" incompatibles. Such an outcome requires that displacement specify how the acceptance criteria need to be evolved and improved. As uncommon as it might be, this scenario is disorienting for all concerned. Differing viewpoints, competing interests, and changes in circumstances only complicate the problem. Common agreement might evade concerned parties: for example, the failure to find a good counter-argument might not be total.

Furthermore, even in the best case, the question of why submission to the test of experience was accepted precisely when it was may arise later if safety problems do occur.

Furthermore, if "warrant" is relative to how much justificatory work is undertaken then, since one could theoretically work on the warrant for a particular statement forever, a politics of "creative inertia" becomes possible: the supplier of a safety argument ***SA***, being initially intrigued by a counter-argument ***CA***, might agree that ***CA*** seems strong; but they might then argue that the warrant of one of its statements – say R_j – needs more backing. The supplier of ***CA*** might agree; but they could also object that one of the premises of ***SA*** – say P_i – also needs more work, and so forth.

Our suggestion has been that should a reversal succeed well enough to cause a deadlock situation then *neither* argument should be considered valid. Otherwise, the only way to break a deadlock in the dialectic process is through an action that implicitly subordinates one argument to the other. Where there are equally plausible arguments for opposite outcomes, involved parties sometimes cannot see any other option but to make the test of experience.

However, in so doing they assert their cultural values. Thus safety processes depend upon cost-benefit analysis to resolve political deadlocks. Nonetheless, it is not unusual to encounter decisions with benefits to some (e.g. increased profits) that would be costs to others (e.g. increased dangers). The 'resolution' of these dilemmas is often forced by the application of principles that are little more than surreptitious assertions of power. This can been illustrated by the difference between "willingness to pay" (an amount that would *prevent* a loss) with "willingness to accept" (an amount that makes the loss *acceptable*) compensation approaches (Adams 1995, p. 98).

To describe deadlock situations we need to consider the various arguments in the light of implicit assumptions about the urgency of a decision. This suggests that for any outcome *C* we consider:

1. an argument ***SA*** for doing action *a* because it will probably have the positive outcome (*C*)
2. an argument ***CA*** against doing *a* because the outcome will probably be *negative* (*not C*)
3. a proposition that ***SA*** should be accepted *now*, i.e. we should test *C or not C* by doing *a*
4. a proposition that ***CA*** should be accepted *for now*, i.e. we should not do *a* and not test *C or not C*

Implicit propositions like (3) and (4) are apparent in any "battle of wills": assumption (3) might be made explicit in order to defeat ***CA*** as a matter of exigency, thus attracting no criticism; but (3) could also be *enforced* by one party on the other. In a case where (3) gives ***SA*** priority over ***CA***, ***CA*** is in effect given *no* priority, whatever steps have been taken against the failure of *C*. Likewise, (4) can be enforced by the party with the more political power and resources: (4) need consist only of a plea for more evidence (short of testing *C*) used as a delaying tactic to defeat (eventually) ***SA*** by "putting off the evil day" until the proposer of ***SA*** either loses interest or runs out of resources to do *a*.

Unsurprisingly, the deadlock situation brings underlying power struggles to the surface: but we can only make sense of the situation through attempts to understand viewpoints that differ from our own, and exposure of our own viewpoint to analysis and criticism. To make sense of the political controversy and hopefully avoid wasteful argument, the displacement stage must consider what factors *in addition* to their belief in their proffered arguments parties might have for whichever of proposition (3) or (4) they favour. Indeed, such factors *must* be operative, since in the absence of unexpressed considerations, the justifications for *both* courses of action would be entirely circular, as follows: to do action *a* is to commit to (3), which presupposes ***SA***; to "do" *not a* and commit to (4) presupposes ***CA***. *Both* parties must be acting according to preferences and interests not made explicit in their arguments. Trying to make these new criteria explicit, should it prove necessary, will probably be the most difficult and protracted part of safety argument displacement.

5 The Pragmatics of Safety Argument Deconstruction

The DERIDASC project did not set out to assess the advantages and disadvantages of our approach in industrial practice: we felt that experimentation with an immature method might prove obstructive. However, our experimental applications of the Warranted Deduction Schema to example safety arguments suggested the following benefits:

- an approach to safety assessment that is more visibly adversarial, leading to the construction of better safety arguments
- more reliable and unambiguous rejection of unsatisfactory safety arguments
- the ability to monitor the effect of new information and knowledge on accepted safety arguments
- a "ready made" assessment approach for different safety argument notations (through the definition of accompanying "shadow" notations)
- a method by which regulators can explicitly manage the incorporation, comparison, and assessment of different viewpoints on the safety of a system, including arguments addressed to the lay public from differing viewpoints
- a way of explaining the evolution of safety acceptance criteria to the public

The issues yet to be addressed concern practical safety argument deconstruction in an industrial context. These issues are:

- are "in-house" counter-arguments an effective way for suppliers to identify and remedy objections before regulatory assessment takes place?
- what resources need to be set aside for the production of counter-arguments?
- would through-life counter-argument maintenance be cost-effective?
- is public trust enhanced by the explicitly adversarial nature of the approach?

A key question about our adversarial approach is whether it will really prove resistant to the production pressures, unimaginative complacency, and excessive bureaucracy that are generally alleged as the root causes of safety failure. A fascinating and perhaps morally necessary deconstructive exercise would be to apply our strategy to itself, that is, to our own justification for it, in collaboration with independent colleagues.

References

Adams J (1995). *Risk*, Routledge ISBN 1-85728-068-7.
Armstrong (2003). *Danger:Derrida at Work*, Interdisciplinary Science Reviews, Vol. 28, No. 2, June 2003, pp. 83–94.
Armstrong J & Paynter S (2003). *Safe Systems: Construction, Destruction, and Deconstruction*, In: *Proceedings of the Eleventh Safety Critical Systems Symposium*, Bristol UK, Edited by Redmill F and Anderson T, Springer, ISBN 1-85233-696-X, pp. 62–76.
Kelly TP (1998). *Arguing Safety: A Systematic Approach To Managing Safety Cases*. DPhil Thesis, Department of Computer Science, University of York. Available from the author's homepage: http://www-users.cs.york.ac.uk/~tpk/
Olin D (2003). *Paradoxes*. Central Problems of Philosophy Series Editor John Strand, Acumen Publishing, ISBN 1-902683-82-X.
Priest G (2002). *Beyond The Limits of Thought*, Oxford University Press, ISBN 0-19-925405-2.
Toulmin S (1958). *The Uses of Argument*. Cambridge University Press, ISBN 0-521-09230-2.