Safety Science xxx (2013) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Safety Science

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

Issues in safety science

Andrew Hopkins

Australian National University, School of Sociology, Canberra, ACT 0200, Australia

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Causation
Accident analysis
Theory
Reviewing

ABSTRACT

This paper deals with three issues. First, the question of the boundaries of safety science – what is in and what is out – is a practical question that journal editors and reviewers must respond to. I have suggested that there is no once-and-for-all answer. The boundaries are inherently negotiable, depending on the make-up of the safety science community.

The second issue is the problematic nature of some of the most widely referenced theories or theoretical perspective in our inter-disciplinary field, in particular, normal accident theory, the theory of high reliability organisations, and resilience engineering. Normal accident theory turns out to be a theory that fails to explain any real accident. HRO theory is about why HROs perform as well as they do, and yet it proves to be impossible to identify empirical examples of HROs for the purpose of either testing or refining the theory. Resilience engineering purports to be something new, but on examination it is hard to see where it goes beyond HRO theory.

The third issue concerns the paradox of major accident inquiries. The bodies that carry out these inquiries do so for the purpose of learning lessons and making recommendations about how to avoid such incidents in the future. The paradox is that the logic of accident causal analysis does not lead directly to recommendations for prevention. Strictly speaking recommendations for prevention depend on additional argument or evidence going beyond the confines of the particular accident.

1. Introduction

The call for papers provides me with an opportunity to reflect on the some of the issue that have been gnawing at me for years. These include:

- The boundaries of safety science.
- Problems with theories that are popular in our field.
- Accident analyses: causation and the problem of making recommendations.

These are all identified as foundational issues in the call for paper, and all raise very practical questions that we must grapple with as we go about our work. The reader is warned, therefore, that this is not an integrated paper, but deals with three discrete topics.

First, the boundaries of safety science is a pressing issue for journal reviewers who must decide whether articles are within scope. This paper takes what I imagine will be a controversial view, namely, that the discipline cannot be defined abstractly, but depends on the interests of the safety science community.

Second, normal accident theory, the theory of high reliability organisations and resilience engineering are all theories or theoretical perspectives that have been popular in our field. We cannot therefore ignore them; we must come to terms with them in some way. I argue here that each is defective in some way, raising questions about why they are so popular.

Third, the call for papers poses the question: "can we learn from past incidents and accidents in order to project useful predictions into the future?" I take this as meaning "in order to make useful recommendations." As the call notes, "on logical grounds, it is indeed impossible to justify prediction through observation of specific cases to be generalised." This paper discusses the logical difficulty of moving from accident analysis to recommendations and offers some pragmatic solutions. This is the most complex of the three topics, and for this reason, and not because it is any less important than the other two, it is reserved till last.

Finally in this introduction, a few comments on the style of this paper. One of the purposes of a special issue of the journal is to promote debate. I have taken this as a license to be provocative. As one reviewer put it, the paper provides the "first word" on the subjects touched on, "never anything close to the last". Moreover, I acknowledge that some of my criticisms are "particularly harsh". I leave it to the reader to judge whether that harshness is warranted.

2. The boundaries of safety science

As a reviewer for the journal, *Safety Science*, I frequently find myself asking: is the subject of this article really safety science? Is it suitable for this journal? For instance, I recently reviewed an

E-mail address: Andrew.hopkins@anu.edu.au

0925-7535/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ssci.2013.01.007

article entitled: "A one-piece coal mine mobile refuge chamber with safety structure and less risk of sealing under shock wave". I returned the paper to the editor saying:

> "My view is that this paper is pure engineering and therefore not appropriate for the journal. However this is really a matter of policy so you might like to think about making a policy decision."

The journal editor subsequently wrote to the author saying: "It seems to me not to be appropriate for publication in the journal".

Was this the right outcome? It depends on what we mean by safety science.

According to the editorial statement[1]:

- *Safety Science* serves as an international medium for research in the science and technology of human safety. It extends from safety of people at work to other spheres, such as transport, leisure and home, as well as every other field of man's hazardous activities.
- *Safety Science* is multidisciplinary. Its contributors and its audience range from psychologists to chemical engineers. The journal covers the physics and engineering of safety; its social, policy and organisational aspects; the management of risks; the effectiveness of control techniques for safety; standardization, legislation, inspection, insurance, costing aspects, human behaviour and safety and the like.

Taking this statement at face value, the decision referred to above was the wrong one.

But it is not as simple as this. *Safety Science* is a peer-reviewed journal. This can only work if there is some relevant community of peers. If it proves impossible to find people within the safety science community with the necessary expertise to act as reviewers, submissions cannot be assessed. Moreover, even if they can be assessed, they will not be read if they fall quite outside the areas of interest of this particular community. In other words the journal and its contents are inevitably and properly shaped by its readership and by its reviewers, not just by an abstract definition.

My judgment was that the article mentioned above lay outside the areas of expertise and interest of the current readership of *Safety Science* and that it would better sent to some journal of mining engineering, where editors will not have such difficulty finding peer reviewers and it is more likely to be read.

This position has far reaching implications. It involves the exercise of what has been called a "gate-keeping function". Moreover, influencing the content of the journal in this way necessarily influences the boundaries of safety science itself. The editorial statement above assumes that safety science can be defined independently of its practitioners. I believe, on the contrary, that the content of safety science must be inferred from the activities of its practitioners. This means that as the safety science community evolves, so too will the subject. For instance, climate change is a massive threat to human safety, and is theoretically encompassed by the editorial statement. But climate science is not currently part of safety science, although one can easily imagine the safety science community embracing aspects of climate science in the future, with the journal evolving accordingly.

All this raises the question of what is meant by the safety science community. Again, I think the answer is pragmatic rather than principled. The safety science community consists of people who are associated with self-identified schools of safety science, who go to safety conferences, who read each other's safety-related publications, and so on. This is a messy definition, but it is one that recognises the fluid and shifting nature of safety science. It would take a network analysis to identify the community with greater clarity.

As I write these words I see that *Safety Science* has recently accepted for publication an article entitled: "Effect of spark duration on explosion parameters of methane/air mixtures in closed vessels". I would have judged this to be outside the current community of interest. Clearly other reviewers and editors take a different view. The authors of the article come from the State Key Laboratory of Explosion Science and Technology, Beijing Institute of Technology. It would seem that Chinese researchers are testing the current boundaries and seeking to join what I have called the safety science community.

The preceding discussion is about the subject matter of safety science. It does not deal with the question of whether or to what extent safety science is truly a science. That question will no doubt be addressed by other contributors to this issue.

3. Popular theories

Certain theories have been popular in the safety science community in recent decades, in the sense that they have been widely cited. Anyone who is serious about safety science must therefore wrestle with them. There are three, in particular, with which I have wrestled: normal accident theory, high reliability theory and resilience engineering. The first of these is not mentioned in the call for papers but the latter two are. These theories have various defects, some fatal and others less so. Strangely, although these theories are often referred to in the literature, the difficulties that I shall identify are largely ignored.[2]

3.1. Normal accident theory

The theory of normal accidents is propounded by sociologist Charles Perrow (1999) in his book, *Normal Accidents*. It offers an explanation for why major accidents in many hazardous technical systems appear to be inevitable. He argues that where a system is characterised by both complexity and tight coupling, accident are inevitable, *no matter how well the system is managed* (Perrow, 2011:172). The paradigm case of a normal accident for Perrow is the Three Mile Island nuclear reactor near disaster in 1979. The terms complexity and tight coupling have a particular meaning for Perrow, but we do not need to define them for present purposes.

The question I want to ask is: how useful has this theory been in explaining the major accidents of our time? The answer is: not at all. Perrow (1994:218) himself acknowledges that few if any of the high profile accidents of recent decades are normal accidents. They were the result of poor management, cost pressures and the like, not the inevitable result of complexity and tight coupling. Most recently he conceded that the Gulf of Mexico blowout of 2010 was not a normal accident.[3]

In his book he devotes a chapter to analysing accidents in petrochemical plants, because this industry "provides some of the best examples or system (i.e. normal) accidents that we shall come across" (1984:101). Yet frequently in this chapter he undermines his argument. In one case he notes that "fairly gross negligence and incompetence seem to account for this accident", but he resists this explanation on the grounds that "a fair degree of negligence and incompetence is to be expected in human affairs, and under production pressures... we can expect forced errors" (1984:111). Else-

[1] http://www.journals.elsevier.com/safety-science/.

[2] One popular theory I shall not address here is Beck's "risk society" thesis (Beck, 1992). I offer a critique of this in Hopkins, 2005, chapter 13.

[3] http://theenergycollective.com/davidlevy/40008/deepwater-oil-too-risky posted July 19, 2010, accessed 18/8/2012.

Please cite this article in press as: Hopkins, A. Issues in safety science. Safety Sci. (2013), http://dx.doi.org/10.1016/j.ssci.2013.01.007

where he says about this same accident: "there was organisational ineptitude: they were knowingly short of engineering talent, and the chief engineer had left; there was a hasty decision on the bypass, a failure to get expert advice, and most probably, strong production pressures" (1984:112). The implication here is that complexity and tight couple did not make this accident inevitable: it could have been avoided with better management. In fact Perrow's descriptions support Turner's theory of sloppy management (1994) as the primary explanatory factor, not complexity and tight coupling.

These problems with Perrow's analysis led me to re-examine the paradigm case, Three Mile Island (Hopkins, 2001). I shall not reproduce the argument here. Suffice it to say that this accident, too, turned out to be a case of sloppy management, not an inevitable result of the technology. In short, the theory of normal accidents was inapplicable even to the accident that gave rise to the theory!

Given all this, the question that arises is: why has the theory of normal accidents proved so enduring? Perrow's political purpose is relevant here. He saw his theory as a way of combating the ubiquitous tendency to blame accidents on front line operators: if complexity and tight coupling were the real culprits then it was clearly inappropriate to blame the people who made mistakes on the day. That is a laudable purpose, but there are many other theories that do this, not the least of which is Turner's theory of sloppy management.

I suspect the fact is that while people continue to make reference to the theory, this is no more than lip service. We are dealing here with one of the more unfortunate aspects of academic practice. People refer to the works of others not necessarily because that work supports their arguments or are in any other way relevant to what is being said, but simply to establish that they are aware of the relevant literature. Such citations amount to little more than academic name dropping. I have myself been cited by people who seem unaware that my point is quite the reverse of theirs and that my work undermines their own conclusion, rather than supporting it. I suspect that this process of catch-all citation is part of the reason the theory of normal accidents continues to be cited.

3.2. The theory of high reliability organisations

The theory of high reliability organisations (HROs) is another widely known theory, popular both inside and outside academia. One prominent use outside of academia was in the analysis of the Columbia space shuttle accident. In the words of the Columbia Accident Investigation Board,

> "To develop a thorough understanding of accident causes and risk, and to better understand the chain of events that led to the Columbia accident, the Board turned to the contemporary social science literature on accidents and risk and sought insights from experts in High Reliability, Normal Accident, and Organisational Theory...Insight from each figured prominently in the Board's deliberations... The Board selected certain well–known traits from these models to use as a yardstick to assess the Space Shuttle Program, and found them particularly useful in shaping its views on whether NASA's current organisation... is appropriate" (CAIB, 2003, p. 180).

In fact the insights on which the Board ultimately relied came almost exclusively from HRO theory and it concluded that NASA fell a long way short of the way HROs operate.

Likewise, after the Buncefield explosion in the UK in 2005, which generated the largest peace-time fire ever known in Europe, the Incident Investigation Board recommended that organisations should be encouraged to behave as HROs (MIIB, 2005, Recommendation19).

So what is an HRO? Some years ago, this question became a very practical one for me. I had decided I wanted to learn more about HROs by studying one. How would I know if a prospective research site was an HRO? How would I know an HRO when I saw it?

The definitions offered in the literature offered little help. HROs were organisations that operated with hazardous technology in a "nearly accident-free" manner, or with many fewer accidents than might have been expected. But these formulations are much too imprecise to allow us to identify HROs and distinguish them from non-HROs. Importantly, the original HRO research that was done by a group of researchers at Berkeley in California in the 1980s was not based on organisations that had been identified as HROs. The three organisations originally studied... "were not so much selected, as offered to us by a conjunction of personal contacts and previous research. Although the selection process was far from 'objective'... the opportunity was not resistible" (Mannarelli et al., 1996:84). It was only after the work began that researchers coined the term HRO and began to think more carefully about how to define it, and ultimately to concede that "no truly objective measure is possible" of whether an organisation is or is not an HRO (Rochlin, 1993:17).

It was left to Karl Weick and his colleague Kathleen Sutcliffe to organise the empirical findings into what can best be described as a model. For them an HRO, which they now called a "mindful organisation", had five characteristics:

> "HROs manage the unexpected through five processes:
> (1) preoccupation with failures rather than successes,
> (2) reluctance to simplify interpretations,
> (3) sensitivity to operations,
> (4) **commitment to resilience** and
> (5) deference to expertise, as exhibited by encouragement of a fluid decision-making system.
> Together these five processes produce a collective state of *mindfulness*"

(Weick and Sutcliffe, 2001, v. Bold for later reference; italics in original)

This is a model, a yardstick, an ideal, against with real world organisations can be measured. Implicitly, Weick and Sutcliffe have abandoned the idea that any real world organisation can measure up to this ideal. In principle they are not even committed to the view that the organisations originally studied were HROs in all respects.

It is indeed prudent to let go of the claim that any particular organisation is an HRO. HRO researchers write about the organisations they have studied with great fondness and enormous enthusiasm. At times they seem breathless with excitement. This is not a state of mind conducive to objectivity and one wonders about the organisational flaws that may have escaped the researchers' notice. The literature on regulation talks about the capture of the regulator by the organisations they are supposed to be regulating. There is a similar risk of researchers being captured by the organisations they study. Perrow (1994) describes the problem well:

> "It is easy to be awed by these behemoths and the intense level of activity in some of them, such as flight exercises at night on the rolling deck of an aircraft carrier. I cannot even walk through a nuclear power plant without being awed by the immensity of controlled power and danger. The reflection that the plant that I am in has been safe for 15 years and might be for ten more, erodes the critical scrutiny I should maintain. I remember being very impressed by one such visit and chastened to later read that it was closed down for gross violations after a series of near misses".

Since the initial investigations by the Berkeley group in the 1980s, there has been a growing body of work that situates itself in the HRO literature. For instance, various studies of hospital intensive care units and of US wildfire fighting organisations cite this literature. However the researchers do not claim to be studying organisations that have been previously identified as HROs. Indeed some of the organisations studied are found *not* to be performing as HROs (Weike and Sutcliffe, 2007, 3–18; 126–137). For these researchers, as for Weick and Sutcliffe, HRO functions as model against which real organisations can be measured. (A nice recent example of this approach is Lekka and Sugden, 2011.)

Treating the high reliability organisation as an ideal in this way resolves one of the difficulties in which the original HRO researchers found themselves. In 1989, three years after the Challenger space shuttle disaster, they described NASA as an HRO (Roberts and Rousseau, 1989:133,137), while two years later, in 2001, they concluded that it did not exhibit the characteristics of an HRO (Roberts and Bea, 2001:179). While it is conceivable that NASA had reverted from HRO to non-HRO status in this two year period, it is far more likely that NASA at no stage measured up completely to what was in fact an unattainable ideal and that the researchers had in mind different aspects of this real world organisation at the times they were writing about it.[4]

This leaves HRO theory in strange kind of limbo. It purports to provide us with a theory about what organisations need to do to achieve safe operation. But it is an untestable theory. If HROs cannot be identified in some *a priori* way then there is no way that we can establish whether HROs indeed exhibit the characteristics identified by Weick and Sutcliffe. Nor can we can demonstrate that HROs are safer than non-HROs. Finally, there is no way new instances of HROs can be identified for further study. These are major drawbacks for any theory. The propositions of HRO theory are inherently persuasive: the theory identifies mechanisms that should in principle make organisations safer. That may be enough reason to retain the theory. But the lack of any real empirical foundation for the theory is, to say the least, unsettling.[5] To summarise, my quest to identify a real live HRO that I could study was doomed to failure. HROs are very elusive creatures that inhabit the realm of theory rather than the real world.

3.3. Resilience engineering

A third perspective that has become popular in recent years is resilience engineering. Its proponents do not describe it as theory. Rather, they talk about "concepts and precepts" (Hollnagel et al., 2006). Since they do not claim it as an explanatory theory it is not appropriate to critique it in the way that I have critiqued HRO theory. Nevertheless there is a fundamental issue that concerns me about resilience engineering: it offers itself as something new, when in fact it is hard to see in what way its "precepts and concepts" depart from those of HRO theory. It is this that I wish to explore here. I shall do this by reference to three books of readings on resilience engineering that have appeared in recent years. (Hollnagel et al., 2006, 2008; Hollnagel et al., 2011).

The issue was raised in 2006 by Hale and Heijer in the very first book of readings on resilience engineering. They said:

> "we would, however, ask whether we do not have other terms already for that phenomenon [resilience], such as high reliability organisations, or organisations with an excellent safety culture"(2006:40)

I raised this issue again in the pages of *Safety Science* in 2009 when I wrote:

A resilient organisation . . ., seems indistinguishable from a high reliability organisation. . . . I hope that resilience theorists will someday explain the difference, if there is any, between these two ideas. (Hopkins, 2009:510).

Hollnagel claims that resilience engineering is indeed something new (2008:xi), but some practicing safety scientists have treated the two ideas as interchangeable (Le Coze and Dupre). I want to demonstrate here just how similar the resilience engineering approach is to HRO theory.

Resilience is one of the five cardinal features of HROs identified by Weick and Sutcliffe, in the quotation above (see bold font). According to Weick and Sutcliffe (2001:14), resilient organisations are not disabled by errors or crises but mobilise themselves in special ways when these events occur, so as to be able to deal with them. A commitment to resilience is actually a commitment to learn from error.

> "To learn from error (as opposed to avoiding error altogether) and to implement that learning through fast negative feedback, which dampens oscillations, are at the forefront of operating resiliently" (2001:69, brackets in original)

Hollnagel says something very similar when he formally defines resilience as "the ability of a system or an organisation to react to and recover from disturbances at an early stage with minimal effect on dynamic stability" (Hollnagel, 2006:16). It is striking that, in providing this definition, he makes no reference to HRO theory.

Hollnagel provides a more elaborate definition of resilience in the second book of readings (2008:xii–xiii), which he repeats in the third book (2011:xxxvii). In this definition he identifies four essential feature of resilience, described as "the four cornerstones". I consider them in turn.

> 1. "Knowing. . . how to respond to regular and irregular disruptions and disturbances either by implementing a prepared set of responses or by adjusting normal functioning. . ."

This is precisely what HRO theory is about. "HROs differentiate between normal times, high-tempo times, and emergencies and clearly signal which mode they are operating in" (Weick and Sutcliffe, 2001:17). Moreover they are adept at managing the unexpected. This is in fact the title of the book by Weick and Sutcliffe and the quotation from Weick and Sutcliffe in the preceding section details exactly how this is to be achieved.

> 2. "Knowing . . . how to monitor that which is or can become a threat in the near term. . ."

Here is what Weick and Sutcliffe have to say about monitoring.

> "The key difference between HROs and other organisations in managing the unexpected often occurs at the earliest stages, when the unexpected may give off only weak signals of trouble. The overwhelming tendency is to respond to weak signals with a weak response. Mindfulness preserves the capability to see the significant meaning of weak signals and to give strong responses to weak signals. This counterintuitive act holds the key to managing the unexpected" (2001:3–4)

Perhaps there is more here than mere monitoring, but the statement certainly implies that HROs will be very diligent about monitoring possible future threats.

[4] That said, it is possible that NASA approximated the HRO ideal more at one time than another. According to the Columbia Accident Investigation Board, NASA's Apollo-era research and development culture and its prized deference to the technical expertise of its working engineers was overridden in the space shuttle era by 'bureaucratic accountability' (CAIB, 2003:198, quoted in Buljan and Shapira, 2005:149). It is possible to read this as a statement that NASA had operated closer to the HRO ideal during the Apollo era than it did during the shuttle era.

[5] A reviewer points out that the early work treats high reliability as a nominal variable. (organisations are, or are not, HROs). If we could identify a suitable *continuous* variable we would be able to locate any and all organisation along a single dimension from high to low reliability. This would facilitate empirical analysis.

Please cite this article in press as: Hopkins, A. Issues in safety science. Safety Sci. (2013), http://dx.doi.org/10.1016/j.ssci.2013.01.007

3. "Knowing…how to anticipate developments, threats and opportunities further into the future…"

HRO theory does not appear to distinguish between near term and longer term threats as is done in 2 and 3 above. The preceding commentary about response to weak signals covers both.

4. "Knowing… how to learn from experience…"

HRO theory emphasises this aspect of mindfulness. HROs are learning organisations par excellence. This point has already been made above.

In summary, Hollnagel's four cornerstones are central features of HRO theory.

The preceding analysis is all rather abstract. Let me move briefly to the detailed empirical level. One of the lead figures in resilience engineering provides a description of the hospital emergency department stretching to cope with high workload. "These local adaptations are provided by people and groups as they actively adjust strategies and recruit resources so that the system can continue to stretch" (Woods, 2009:500). This description of what happens in the emergency department is powerfully reminiscent of the description in the HRO literature of how air traffic control centres mobilize additional resources at times of heaviest workload (La Porte and Consolini, 1991:38), providing "extras pairs of eyes" to ensure that nothing is missed. The very same phenomenon is being described in both cases.

Given that resilience writers are making the same points as have been made by HRO theorists, one would have expected that they either acknowledge this explicitly and conceptualise their work as a contribution to the HRO literature or, on the other hand, explicitly distinguish what they are doing from that literature. But they do neither. They write with almost no reference to HRO theory.[6] It is as if that body of work hardly exists for them. HRO is not mentioned in the indexes of any of the three resilience engineering books reviewed for this article.

What is going on here? In the absence of any explanation from the resilience engineers we can only speculate. In the social sciences generally, generations of theorists and theories succeed each other, without there being necessarily any real advances in our knowledge or understanding. Beck's theory of *risk society* generated great excitement when it appeared twenty years ago, and the concept of risk subsequently dominated social theory for a period. In recent years the concept of resilience has become fashionable (Sheffi, 2005) – although the resilience engineers do not locate themselves in this literature – and "resilience" is now beginning to vie with "risk' as a dominant organising idea. It is almost as if every new generation seeks to make its mark by developing new theories, or at least new language with which others must then come to terms. One wonders whether safety science may be exhibiting something of this pattern.

This can be put in more sociological terms. As described above, the HRO school developed in the early 1980s among a network of researchers based at Berkley in California. The research came first and the HRO concept came later, almost as an afterthought. It served as a label for their work as well as serving to unify the group. From this point of view it is understandable that the theory of HROs has always been somewhat problematic.

The banner of resilience engineering was raised almost a generation later. A group of like-minded researchers gathered together for a symposium in Sweden in 2004. As one of the organisers notes, the symposium was held because "most of the people we had in mind were able and willing to interrupt their otherwise busy schedules to attend the symposium" (Hollnagel, 2006:xii). This group, plus and the book of proceeding that followed, provided a nucleus around which a new community of scholars could crystallise. The term resilience engineering served as a label for their work and as well serving to unify the group.

If this analysis is correct, we come to a disturbing conclusion. The emergence of new concepts must be understood in terms of the social functions they perform for their proponents, rather than the intellectual work they do. The failure of the safety science community to look more critically at the theories it embraces is what is most disturbing about this. Without this critical scrutiny, theory is merely a matter of fashion.

4. Major accident analysis

The final topic I want to address is the problem of drawing policy recommendations from major accident investigations. Every such investigation is carried out with the express purpose of learning lessons and making recommendations to prevent a recurrence. The assumption is that the recommendations can be read off from the accident analysis in a straight forward way. They cannot. This is the problem I want to discuss here.

This is a personally pressing issue. In my recent analysis of the Gulf of Mexico blowout (Hopkins, 2012), I highlighted two particular recommendations: the first is that companies need to change their reporting lines so that engineers (and more generally safety specialists) do not report to low level line managers, but rather to higher level engineering (or safety) managers. The second was that bonus arrangements for senior executives should include measures of how well the company is managing major hazard risk. The question I am sometimes asked is: how can I be sure that adopting these recommendations will make a difference?

4.1. Accident causation

Major accident inquiries are implicitly or explicitly inquiries into cause. We must therefore begin with some observations about causation. For present purposes, I distinguish two distinct meanings of cause. The first is *sufficient* cause, meaning a factor or set of factors that is sufficient to produce the outcome. This is a strong sense of causation. It is not however the most useful meaning of cause, because to identify the sufficient cause of an accident, that is, the entire set of factors that went into the producing the accident, is impossible, practically speaking.[7]

[6] I have found just one reference -Hollnagel et al., 2006: xi.

[7] Some authors (e.g. Ladkin, 2001) argue that it is possible to identify a discrete set of factors that together provide a sufficient cause. However their analysis assumes that all other factors that might affect the outcome remain unchanged. This is a crucial limitation. It may be that the best way to prevent recurrence is not to focus on a discrete set of causes but to identify some background factor that, if changed, would prevent a recurrence. For example, suppose the analyst identifies a particular set of factors that together are sufficient to cause a house to burn down. A potentially infinite set of background factors is ignored in this way. One that we might want to highlight is that the house was not fitted with a back-to-base alarm. Had it been, the fire brigade might have been on the scene sufficiently quickly to save the house. That in turn suggests a possible policy response – all houses should be fitted with back-to-base fire alarms. Whether this is a sensible proposal is not the point here. The point is that this is a recommendation that would not emerge from an inquiry that was seeking to establish sufficient cause in Ladkin's sense. More generally, the philosopher, Bertrand Russell (1872–1970), among others, makes the point that no finite set of causes can be regarded as sufficient. Putting a penny in the slot machine might be seen as a sufficient to produce the ticket. "But [, says Russell, suppose that] before I can draw out my ticket there is an earthquake which upsets the machine and my calculations. In order to be sure of the expected effect, we must know that there is nothing in the environment to interfere with it. But this means that the supposed cause is not by itself adequate to insure the effect." (Quoted in McIver: 1964:44-5). We do not need to resort to *force majeure* in this way to make the point. Slot machines may jam if not properly maintained. In other words, putting the penny in the slot is not in and of itself sufficient to produce the ticket. The machine must also be properly maintained. Interestingly, the expression – "the penny dropped" – may have originated as a description of a jammed slot machine that finally operated, perhaps after being nudged or thumped (www.phrases.org.uk/meanings/280900.html, accessed 1/8/12). In such a case, a thump is a necessary element in the set of sufficient factors.

The second meaning of cause is a factor that was *necessary* for the outcome to occur. Such a factor can be called a *but-for* cause, in the following sense: but for this factor (had it been otherwise), the accident would not have happened. Most accident analyses implicitly adopt this second meaning. They aim to identify a relatively small set of necessary causes, in the absence of any one of which, the accident would not have happened.

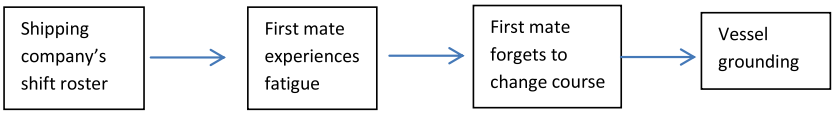
This is basis of the ubiquitous Swiss cheese model of accident causation: an accident occurs because the holes in the barriers all line up. Had any one of the barriers operated as intended, the accident would not have happened. Each of the barrier failures was thus necessary for the accident to occur. Each, in this sense, was a cause.

Similarly, the accimap (accident map) method of accident analysis developed by Rasmussen (1997) and Rasmussen and Svedung (2000) implicitly uses this but-for logic. Accimaps graphically depict an array of factors ranging from the most immediate through to aspects of general organisational functioning and even to failures of the regulatory regime. The accimap arranges these factors so as to demonstrate their inter-relationships – how one thing contributed to the next. The accimap logic has been well described in Branford et al. (2009) and in a publication by the Australian Transport Safety Bureau (ATSB, 2008) which uses the accimap methodology for its investigations of aircraft crashes and other major transport accidents. This publication is a sophisticated and thoughtful discussion of the philosophical underpinnings of the ATSB's accident investigation strategy and I shall draw on it more than once in what follows.

The reasoning involved in identifying necessary factors is counterfactual – making an argument about what would have happened had this factor been otherwise. This raises the question: how can we know what would have happened had this particular factor been otherwise?

In some situations the but-for argument is a matter of logic. So, in the case of the Swiss cheese model, it is a matter of logic that had a barrier not failed, the accident would not have occurred. This kind of reasoning is often possible where technical causes are involved. For example, had the pressure relief valve lifted as it was intended to do, the pressure vessel would not have exploded.

For more remote, organisational causes, it becomes a matter of expert judgement, and the causal connections become probabilistic statements rather than logical deductions. For example, ATSB provides the following example, where the arrows can be read as "contributed to".



In this situation the analyst cannot be certain that a better shift roster would have prevented the accident. Perhaps the first mate was dealing with some other issue that distracted him and, even in the absence of fatigue, this would have caused him to forget to change course. The best the analyst can do is make a judgement, based on assessment of the all facts of the case, that a better roster would *probably* have prevented the accident.

As this example makes clear, the more remote the presumed organisational cause, the less certain we can be that, had it been otherwise, the accident would not have occurred, that is, the less certain we can be that it was indeed a but-for cause. However, this disadvantage brings with it a corresponding advantage – the more far-reaching the organisational cause that we identify, the larger is the class of potential accidents that may be avoided (assuming we can convert the but-for cause into a recommendation – to be discussed below). In this case, designing shift rosters in such a way as to reduce fatigue will reduce the number of fatigue-induced errors and hence reduce the risk of a broad range of fatigue-induced accidents.

The preceding paragraphs have identified two bases for drawing conclusions about but-for cause: logical argument and expert judgement. There is sometimes also an empirical way to demonstrate that more remote organisational factors are but-for causes. If we can identify a comparable situation to the one under consideration where the causal factor of interest was different, and if the outcome was also different, then we can infer that the factor of interest was a but-for cause. This is the method of difference, made famous by the nineteenth century philosopher, John Stuart Mill, and long used by sociologists as a method of establish causal connections. For example, Max Weber (1930) noted that capitalism developed initially in Western Europe and nowhere else. He noted further that Protestantism developed just prior to the emergence of capitalism in Western Europe and nowhere else. He then identified certain fundamental similarities between Protestantism and capitalism – in a nutshell, the work ethic. Finally, he concluded that Protestantism was a but-for cause of capitalism – had Protestantism not emerged when and where it did, neither would capitalism have emerged when and where it did.

I used this method of differences in my analysis of the causes of the blowout of the Macondo well the Gulf of Mexico blowout. There were several BP drilling teams in the Gulf of Mexico and the team drilling the Macondo well was the only team in which engineers reported directly to line managers rather than more senior engineers. They were thus subordinated to the schedule and cost reduction pressures imposed on them by line managers which could only undermine the quality of their engineering decisions. The fact that the accident occurred with this drilling team and no other, supports the argument that had the Macondo engineers reported to higher level engineering managers, the accident would not have happened. The conclusion is not inevitable because there were certain other differences between the Macondo team and the other drilling teams that might have played a part. The ideal would be to compare two situations which are similar in all respects, except for the causal factor of interest and the outcome. In such circumstances, the causal argument would follow as a matter of logic. Given that we can at best approach this ideal and never reach it, the method of difference is persuasive rather than conclusive. But it remains a powerful empirical basis on which to mount counterfactual arguments.[8]

Notice that the preceding argument is about necessary conditions, not sufficient conditions. The Macondo drilling engineers had been subordinated to low level line managers for years without a major accident. The subordination of the engineers was therefore not a *sufficient* condition for the accident. The method

[8] I also used the method of difference to demonstrate that that the inappropriate process safety indicators were a but-for cause of the Macondo blowout.

of difference, as used here, establishes merely that it was *necessary* for the outcome to occur.

4.2. Making recommendations

Major accident investigations are essentially case studies. They therefore suffer from many of the limitations of case studies, in particular, the difficulties involved in trying to generalise. Some of these limitations are more perceived than real. Flyvbjerg (2001: chapter 6) has shown that there are various ways that conclusions from case studies can go beyond the particular case. But given that a major accident investigation is a case study, we need to think carefully about how such an investigation can generate recommendations going beyond the particular case.

The problem is this. Accident analyses are backward looking. Making recommendations is forward looking. In making a recommendation we want to be able say something like this: the evidence is that if you follow this course of action you will reduce the risk of a major accident. Unfortunately, a but-for analysis only provides evidence of how we may avoid an *identical* accident in the future – that is the but-for logic. But no future set of circumstances will ever be identical with those that gave rise to the accident that has been analysed, and the more dissimilar the future circumstances the less certain we can be that changes based on a previous but-for analysis will have any preventive effect.

We cannot even be sure that recommendations based on a but-for analysis will reduce the *likelihood* of future accidents appreciably. To see why, consider the Macondo case again. The causal analysis showed that the reporting arrangements for engineers were a but-for cause of the accident. It follows, for what it is worth, that having engineers report to more senior engineers would prevent identical accidents in the future. More importantly, we can argue that having engineers report to more senior engineers is likely to prevent other accidents *in which poor engineering is a but-for cause*. However major accidents can occur for reasons that have nothing to do with the quality of engineering practices. If we think in bowtie terms for a moment (Bice and Hayes, 2009), engineering excellence may be a barrier on one of the causal pathways but not on others. In order to know whether improving the quality of engineering is likely to reduce significantly the frequency of major accidents we need some additional information about the frequency with which poor engineering is one of the but-for causes of such accidents.[9] In other words, the evidentiary basis for any recommendation to strengthen engineering reporting lines lies beyond the data provided by a single accident investigation.

There is one way in which a much firmer foundation could, in principle, be provided for such a recommendation. That would be to do additional empirical work to demonstrate that organisations with stronger engineering reporting lines have fewer major accidents. Unfortunately such research would be enormously difficult to do. Here are two reasons why. First, major accidents are relatively rare so it would be difficult to assemble sufficient data to test the proposition. Second, reporting lines are so variable and fluid that it would difficult to array organisations along a single dimension for the purposes of the correlational analysis.[10]

Even if we could produce such data it would be subject to a very significant qualification. The results would be of this form: organisations with longer engineering reporting lines have on average fewer major accidents. In other words, longer reporting lines are associated with a reduced risk of major accident.[11] Moreover, given that simple correlations like this leave many factors uncontrolled, the demonstrated reduction in risk is likely to be quite slight. In short, even if these results could be demonstrated, there would be no guarantee that any particular organisation with longer reporting lines for engineers would not experience a major accident.

In the absence of systematic correlational evidence we may nevertheless fall back on "expert judgment". Expert judgement is based on evidence of a more qualitative kind – experience. In this sense expert judgement draws on data beyond the case in question and provides a sounder basis for action than may be provided by that case alone. It is noteworthy that following the Macondo accident, BP took the view that poor engineering practice was a sufficiently significant causal factor that it made major organisational changes in order to promote engineering excellence. Clearly it was BP's expert judgement that this would significantly reduce the risk of major accidents.[12]

There are however circumstances in which it is reasonable to make recommendations, even in the absence of evidence going beyond the particular case. A good investigation will highlight various but-for causes that contributed to accident, the full significance of which had not been previously recognised. Once recognised, the very logic of how they operate may be sufficiently persuasive to justify making certain recommendations.

For instance, it is well known that the safety of pressure vessels depends on the having reliable pressure relief valves that open automatically if the pressure rises to high. It may not be well known that one the reasons they can fail to function as designed is that operators may deliberately defeat them, for what they regard as their own good reasons.[13] Where this is identified as a but-for cause of an explosion (had operators not defeated the safety system the explosion would not have happened), the obvious recom-

[9] This reasoning rules out the possibility of using the method of differences to derive firm, empirically-based recommendations from the Macondo case. Interestingly, the method of differences has sometimes been used to go beyond the particular case. In *Tokagawa Religion*, Bellah (1957) noted that western style capitalism developed independently in Japan, and his explanation was that the Tokagawa religion had many of the features of Protestantism. In making this argument he implicitly converts Weber's discussion, from a backward looking, but-for explanation, into a forward looking prediction: wherever we find a religion with the features of Protestantism, we can expect the independent emergence of capitalism. In so doing he is treating the causal connection identified by Weber as a sufficient cause, which takes a whole lot of other things for granted, as noted in an earlier footnote.

[10] Similar problems have bedevilled any attempt to show that the enactment of safety case regimes has had a significant effect on major accident rates (Fenning and Boath, 2006).

[11] But-for causation is about the cause of individual events, while correlational analysis is about causal relationships between variables. In the latter context, the analyst first postulates a causal model – that is, a set of relationships between variables. Given particular values for the independent variables, the model then predicts the mean value of the dependent variable over a number of cases, but not its value in any particular case. There is always an error term in the model that expresses this uncertainty in any particular case (Blalock, 1964:16). In causal models of this nature, the departure of the dependent variable from the mean, in any particular case, is assumed to be a result of unknown variables not included in the model. The implication is that if we were able to introduce all relevant variables into the model we would be able to predict outcomes in individual cases. In other words we would have identified a sufficient cause of the event in question. However, as Blalock (1964:16) observes, "in real-life situations it will be impossible to take account of all relevant variables or to obtain perfect measurements", so in practice we can never identify the sufficient cause of a particular event. Hence, although the causal modelling approach is guided by a deterministic vision of causation, in practice it provides only a statistical or probabilistic account. However, this way of thinking does mean, as Blalock again points out, that "whenever we find a high degree of unexplained variation, we immediately look for other variables that have not been included in the causal system, expecting that we can successively reduce this variation by adding further variables".

[12] As a reviewer points out, this assumes that the experts are in agreement. If there is a real difference of opinion among the experts, it may be impossible to all back on expert opinion in this way.

[13] As was discovered in the inquiry into the 1999 Kaiser aluminum smelter explosion at Gramercy, Louisiana. http://www.msha.gov/disasterhistory/gramercy/report/reportdept.htm#desc accessed 17/8/12.

mendation is that organisations should redouble their efforts to ensure that this is not occurring at their own sites. In this situation the but-for analysis serves as a warning about what can happen and the steps that need to be taken to avoid such happenings.

A second example of this nature came to light in the analysis of the Texas City refinery disaster (Hopkins, 2008). It was discovered that managers had performance agreements that provided incentives for them to minimise cost, but no incentives to attend to major hazard risks. Subsequent inquiries recommended that the remuneration for senior managers include some component based on how well they were managing major hazards.

It would be almost impossible to demonstrate empirically that altering remuneration systems in this way will reduce the number of major accidents. But the recommendation does not depend on such evidence. It is based on a theory of human motivation, viz, that people are likely behave in ways that please their boss, especially when their material interests are dependent on pleasing the boss. The strength of the recommendation depends on the strength of this theory of human motivation. For many people the theory will be regarded as self-evident, without need for further evidence. The logic here is that the particular accident analysis has shown how a certain causal factor *can* work and the theory of human motivation has provided the basis for generalisation and for the belief that if this recommendation were implemented the risk of major accident would be reduced. It would of course be desirable to have research evidence that directly supported the recommendation, but those who have responsibility for controlling major hazards cannot afford to wait until such evidence is available. They must make judgements on the basis of whatever information is currently available, and the information generated by the Texas City inquiries is arguably quite sufficient to justify the recommendation about remuneration.

In summary, the conclusions about cause that emerge from major accident inquires do not automatically give rise to recommendations that can be applied more widely. That step of generalisation depends of some additional evidence, or least some additional argument that is, in and of itself, persuasive, independently of the but-for causes identified in the particular accident analysis.

4.3. Separating cause and recommendation

Few accident investigating bodies recognise the logical gap that exists between identifying causes and making recommendations. One agency that does is the Australian Transport Safety Bureau. It distinguishes between but-for causes and "safety issues" that become obvious during the course of an inquiry. For example, while it may be impossible to demonstrate that the shift roster in the ATSB example above was a but-for cause, it may well be judged a "safety issue" worth addressing and worth making recommendations about. Quite explicitly, then, the recommendations of the ATSB do not necessarily depend on the detailed causal analysis that it undertakes in its accident investigations. According to the ATSB, "it is important that safety investigation reports discuss the safety issues identified during an investigation, regardless of whether they contributed" (2008:21).[14]

That being the case, why bother with the causal analysis in the first place? The ATSB's answer is quite pragmatic (2008:21–2).

- Stakeholders require an investigation into causes.
- Some organisations will only appreciate the significance of a safety issue if it can be plausibly linked to the accident in question.
- The concept of causal contribution to a particular accident provides a central organising principle and serves to limit the scope of the inquiry that could otherwise develop into a boundless inquiry into safety.

To this pragmatic list we might add the following reasons.

- The causal analysis of a major accident identifies accident-producing mechanisms of which we previously may have been only dimly aware.
- Companies cannot afford to wait until there is conclusive evidence of the efficacy of some safety measure. They must act on the basis of available evidence. The evidence from detailed causal analyses may be the best available.
- The causal analysis of a major accident amounts to a powerful story from which others can learn.
- When the findings of a series of major accident inquiries are put together, patterns may emerge from which, in turn, policy recommendations follow quite persuasively.

5. Conclusion

The three topics I have discussed here are quite disparate, but each goes to the heart of safety science, in its own way. Each is something that members of the safety science community must wrestle with.

First, the question of the boundaries of safety science – what is in and what is out – is a practical question that journal editors and reviewers must respond to. I have suggested that there is no once-and-for-all answer. The boundaries are inherently negotiable, depending on the composition of the safety science community.

The second issue is the problematic nature of some of the most widely referenced theories or theoretical perspective in our interdisciplinary field, in particular, normal accident theory, the theory of high reliability organisations, and resilience engineering. Normal accident theory turns out to be a theory that fails to explain any real accident. HRO theory is about why HROs perform as well as they do, and yet it proves to be impossible to identify empirical examples of HROs (beyond those originally studied) for the purpose of either testing or refining the theory. Resilience engineering purports to be something new, yet on examination it is hard to see where it goes beyond HRO theory.

The question then is why these theories have enjoyed such popularity. The answer, I believe, has something to do with the functions they perform for the theorists themselves.

The third issue concerns the paradox of major accident inquiries. The bodies that carry out these inquiries do so for the purpose of learning lessons and making recommendations about how to avoid such incidents in the future. The paradox is that the logic of causal analysis does not lead directly to recommendations for prevention. Strictly speaking recommendations for prevention depend of additional argument or evidence going beyond the confines of the particular accident. There is in fact a fundamental disconnect between the causal analysis of major accidents and the recommendations that often emerge from those analyses. This is a troubling conclusion for those of us who are interested in the causes of major accidents, since it calls into question the whole rationale for doing this kind of analysis. It seems that the value of accident analysis lies not so much in the conclusions about the causes of the accident investigated, but in the way the whole investigation process draws attention to safety issues that can usefully be made the subject of recommendations.

[14] Sidney Dekker (2011:66) makes a similar point: "...there is a difference between those things that can explain why a particular event happened, and those things we should focus our attention onto make sure that similar things do not happen again. In complex systems, separating these out probably makes great sense..."

References

ATSB [Australian Transport Safety Bureau], (2008). Analysis, Causality and Proof in Safety Investigations. ATSB, Canberra.
Beck, U., 1992. Risk Society. Sage, London.
Bellah, R. 1957. Tokagawa Religion New York, The Free Press.
Bice, M., Hayes, J., 2009. Risk management: from hazard logs to bow ties. In: Hopkins, A. (Ed.), Learning from High Reliability Organisations. CCH, Sydney (Chapter 4)
Blalock, H., 1964. Causal Inferences in Nonexperimental Research. Univ of North Carolina Press, Chapel Hill.
Branford, K., Naikar, N., Hopkins, A. 2009. Guidelines for accimpap analysis. In: Hopkins, A. (Ed), Learning from High Reliability Organisations. CCH, Sydney(Chapter 10).
Buljan, A., Shapiro, Z. 2005. Attention to production schedule and safety as determinants of risk-taking in NASA's decision to launch the Columbia shuttle. In: Starbuck, W. Farjoun, M. (Eds.), Organization at the Limit: Lessons from the Columbia Disaster. Blackwell, Oxford pp. 140–156.
CAIB (Columbia Accident Investigation Board), 2003. Report, vol. 1. NASA, Washington.
Dekker, S., 2011. Drift into Failure. Ashgate, Aldershot.
Fenning, N., Boath, M., 2006. Impact Evaluation of the Control of Major Accident Hazards (COMAH) Regulations 1999. UK Health and Safety Executive, Research Report 343.
Flyvbjerg, B. 2001. Making Social Science Matter. Cambridge Univ Press.
Hale, A., Heijer, T. 2006. "Defining resilience". In: Hofnagel, E., Wood, D., Leveson, N.(Eds.), Resilience Engineering: Concepts and Precepts. Ashgate, Aldershot, pp. 35–40.
Hollnagel, E., Woods, D., Leveson, N. (Eds.), 2006. Resilience Engineering: Concepts and Precepts. Ashgate, Aldershot.
Hollnagel, E., Nemeth, C., Dekker, S. (Eds.). 2008. Remaining sensitive to the possibility of failure: resilience engineering perspectives, vol. 1. Aldershot, Ashgate.
Hollnagel, E., Paries, J., Woods, D., Wreathall, J., (Eds.), 2011. Resilience Engineering in Practice. Aldershot, Ashgate.
Hopkins, A., 2001. Was Three Mile Island a normal accident? Journal of Contingencies and Crisis Management 9 (2), 65–72.
Hopkins, A., 2005. Safety, Culture and Risk. CCH, Sydney.
Hopkins, A., 2008. Failure to Learn: the BP Texas City Refinery Disaster. CCH, Sydney.
Hopkins, A., 2009. Reply to comments. Safety Science 47, 508–510.
Hopkins, A., 2012. Disastrous Decisions: the Human and Organisational Causes of the Gulf of Mexico Blowout. CCH, Sydney.
Ladkin, P. 2001. Causal System Analysis, self-published http://www.rvs.uni-bielefeld.de/publications/books/CausalSystemAnalysis/index.html.
La Porte, T., Consolini, P., 1991. Working in practice but not in theory: theoretical challenges of high reliability organisations. Journal of Public Administration Research and Theory 1 (1), 19–47.
Le Coze, J., Dupre, M., (In preparation). How to prevent a normal accident in a high reliable organisation? The art of resilience, a case study in the chemical industry.
Lekka, C., Sugden, C., 2011. The successes and challenges of implementing high reliability principles: a case study of a UK oil refinery. Process Safety and Environmental Protection 89, 443–451.
MIIB (Major Incident Investigation Board), 2005. The Buncefield Incident, The final report, vol. 2. 11 December 2005.
McIver, R.M., 1964. Social Causation. Harper and Row, New York.
Mannarelli, T., Roberts, K., Bea, R., 1996. Learning how organisations mitigate risk. Journal of Contingencies and Crisis Management 4 (2), 83–92.
Perrow, C., 1994. The limits of safety: the enhancement of a theory of accidents. Journal of Contingencies and Crisis Management 2 (4), 212–220.
Perrow, C., 1999. Normal Accidents. Princeton University Press, Princeton.
Perrow, C., 2011. The next catastrophe: reducing our vulnerabilities to natural, industrial and terrorist disasters. Princeton University Press, Princeton, New Jersey.
Rasmussen, J., 1997. Risk management in a dynamic society: a modelling problem. Safety Science 27 (2–3), 183–213.
Rasmussen, J., Svedung, I., 2000. Proactive risk management in a dynamic society. Swedish Rescue Services Agency, Karlstad.
Roberts, K., Bea, R., 2001. When systems fail. Organisational Dynamics 29 (3), 179–191.
Roberts, K., Rousseau, D., 1989. Research in nearly failure-free, high-reliability organisations: having the bubble. IEEE Transactions of Engineering Management 36 (2), 132–139.
Rochlin, G., 1993. Defining "high reliability" organisations in practice. a taxonomic prologue. In: Roberts, K. (Ed.), New Challenges to Understanding Organisations. Macmillan, New York.
Sheffi, Y., 2005. The Resilient Enterprise: Overcoming Vulnerability for Competitive Advantage. MIT Press, Cambridge.
Weick, K., Sutcliffe, K.M., 2001. Managing the unexpected: assuring high performance in an age of complexity. Jossey Bass, San Francisco.
Weike, K., Sutcliffe, K., 2007. Managing the unexpected, second ed. Jossey Bass, San Francisco.
Woods, D., 2009. Escaping failures of foresight. Safety Science 47, 498–501.
Weber, M. 1930. [1905] The Protestant Ethic and the Spirit of Capitalism Unwin Hyman, London.

Please cite this article in press as: Hopkins, A. Issues in safety science. Safety Sci. (2013), http://dx.doi.org/10.1016/j.ssci.2013.01.007