Safety Cases: An Impending Crisis?

Ibrahim Habli, Rob Alexander, and Richard Hawkins

University of York, York, UK

ibrahim.habli@york.ac.uk, rob.alexander@york.ac.uk, richard.hawkins@york.ac.uk

Abstract *Safety cases have long been required by many safety standards and guidelines. Particularly in the UK, new systems in key sectors such as defence, nuclear and rail need a safety case before they can be certified and approved. Proponents of safety cases have justified this on the basis of some explicit theory (e.g. that of Toulmin) and by using a variety of plausible, common-sense arguments. There has, unsurprisingly, been a large amount of research on safety cases, on how to structure them, how to review them, and (increasingly) how to formalise them mathematically or generate them automatically. However, there has been very little research that evaluates safety case methods and practices as a whole. Do they "work", in terms of safety or other benefits? If so, when, how and why do they work? In particular, there has been almost no research on "the science of the safety case" — no systematic marshalling of all the claims used to promote safety case methods, and importantly no research programme to test each one. In this paper, we identify the key claims made by safety case proponents. We emphasise claims about mechanism — those about how and why safety cases provide benefits. From there, we spell out how to identify the most important research questions raised by those claims, and outline a community-wide research approach that could help answer those questions. To do this will require that we first understand how people currently develop safety cases and what their concerns, needs, constraints and problems are. There is little point in testing hypotheses and solutions that do not target real problems.*

1 Introduction

The adoption of safety cases is becoming widespread in many domains (Sujan 2016). There is little empirical evidence, however, as to what benefits safety cases actually deliver. On the one hand, proponents of safety case have made many claims about their benefits, such as those reported by (Rinehart 2017). On the other hand, several authors have made plausible objections to their use.

The objections to the use of safety cases take a number of forms. There are many anecdotal accounts of people questioning whether safety case activities are worth

the substantial cost involved. Some authors, e.g. (Leveson 2011), have gone further and suggested that safety cases may have negative effects even if cost is disregarded. Whilst some of these criticisms are easy to reject as being based upon false premises (such as that safety cases are only ever produced at the end of the development process), some of the criticisms are credible and deserve consideration (such as the risk of confirmation bias).

All safety activities could be subject to the same question — are they effective? And (if so) are they worth the cost? This is not a question that has been answered for most safety activities — no surprise, as it is very hard to measure (Rae 2020). Given that, the key question thus becomes not "do safety cases work?", but rather "is a safety case regime better than alternative approaches, and (if so) where, when, and how, exactly?".

To answer those questions, we will need to understand how people currently develop safety cases and what their concerns, needs, constraints and problems are. There is little point in testing hypotheses and solutions that do not target real problems.

This is all significant because without trustworthy and usable evidence of appropriateness, efficacy and cost-effectiveness, that supports actual needs, the vibrant safety case community risks a crisis of legitimacy. This is particularly the case with the significant uncertainties about the gap between the continuously emerging safety case methods and tools and actual safety case practices. Such practices rarely fit the abstract and simplified characterisation of industry problems and contexts as reported in research papers.

We consider three aspects:

1. **Establishing real-world safety assurance needs:** What are the actual needs that safety cases have the potential to meet? How do we establish these needs and understand the context within which they occur?
2. **Creating process theories that capture how safety case practice leads to meeting (or not meeting) needs:** How, under what circumstances, and to what extent do safety case methods meet these needs? What are the intermediate steps in that process? What does it look like when that process is going well; what about when it is going badly?
3. **Evaluating efficacy and efficiency results:** Which research methods and settings are appropriate and feasible to generate the efficacy and efficiency evidence? How can we appraise such evidence against the safety assurance needs?

In this paper, we will review existing work on addressing these questions for safety cases, and describe how we could make progress towards better answers.

2 A Very Brief History of Safety Case Research

In this section, we take a historic research-oriented view of safety case concepts and methods. Readers interested in the industrial roots of safety cases, notable accidents that led to wide usage of safety cases in certain industries and key standards that require or recommend their use are advised instead to consult the reviews by Haddon-Cave (Haddon-Cave 2009) and Sujan et. al. (Sujan 2016). Bloomfield and Bishop (Bloomfield 2010) also offer a practice-oriented review of safety cases, with a focus on software-based systems.

Although the explicit practice of developing safety cases has its roots in the nuclear industry in the 1950s (Arnold 2016), the conceptual basis was established in the 1990s, most notably through the seminal work of Kelly and McDermid at York (Wilson 1995, Kelly 1999) and Bishop and Bloomfield at Adelard (Bishop 2000). The research largely built on Toulmin's work on informal logic (Toulmin 1958) and laid the foundation for the notion of structured argumentation as an explicit and core part of safety cases. Further, influenced by the move towards more graphical models in software engineering (Rumbaugh 1999), combined with goal-based approaches to requirements definition (Dardenne 1993), two graphical notations for the representation of safety arguments were developed namely the Goal Structuring Notation (GSN) (ACWG 2018) and Claim Argument Evidence (CAE) (Bloomfield 1998). Realising the importance of the systematic construction of safety cases, detailed methods were created for the development, reuse and maintenance of safety arguments and for integrating the production of the safety case with the design and operation of the system (Kelly 1999, Bishop 2000).

In short, the early research by York and Adelard emphasised the need for (1) an explicit, detailed and well-formed safety argument and (2) the integration of the safety case into the design and evolution of the system. This was clearly stated by Bishop and Bloomfield: "*the safety case life-cycle should be an integral part of the overall system development, and this should continue throughout the lifetime of the system*" (Bishop 2000).

Two notable extensions of the early research are worth highlighting. The first is safety argument patterns, building on the notion of design patterns in software engineering (Gamma 1995) and more fundamentally the pattern language that was defined by the architect and design theorist Christopher Alexander (Alexander 1977). Safety argument patterns provided means for documenting and reusing "successful" argument structures that relate to recurring safety assurance needs, requirements or strategies, e.g. arguments structured based on the concept that the risks should be as low as reasonably practicable (ALARP) or based on separating the product- and process-oriented assurance aspects of safety engineering (Kelly 1999). Unfortunately, very little reuse of patterns has been reported, particularly between organisations, mainly because the number of people who publish their experiences is very low.

Modularity in safety cases is the second major extension. Building on concepts and representations in software architectures, modular safety cases were created as

way to address inevitable changes and the scale and complexity in the system, environment and the safety case itself (Kelly 2001). Similar to argument patterns, the wider adoption of modular safety cases remains an open question.

Research in the last two decades has largely focused on implementing the aforementioned concepts. Key research themes include:

- *Notations*: extensions of existing notations such as Assurance Claims Points (ACPs) for GSN (Hawkins 2011) or for representing dialectic arguments (Yuan 2013)
- *Processes*: detailed processes for integrating safety cases into the overall safety, systems and software engineering processes (Denney 2015)
- *Model-based arguments*: providing a metamodel foundation and utilising model-based mechanisms such as model transformation, traceability and validation (Hawkins 2015, Denney 2014, Wei 2019)
- *Automation*: software tooling for representing and maintaining the argument and associated evidence (Denney 2012)
- *Formalism*: representing arguments in formal logic in order to "prove" certain assurance properties or support automated reasoning (Rushby 2010, Habli 2009)
- *System- or domain-specific arguments*: largely documented in patterns. e.g. for embedded automotive systems (Birch 2013), COTS (Ye 2005) or machine learning (Picardi 2019)
- *Assurance cases*: as a generalisation of safety cases in order to address other critical properties such as security (Rushby 2015)
- *Review*: detailed processes for evaluating different properties of safety cases (Chowdhury 2020), including known fallacies (Greenwell 2006)
- *Confidence*: assessment of different sources of uncertainty in safety cases, with a particular emphasis on quantification of confidence (Guiochet 2014, Ayoub 2012), most notably through Bayesian Networks (Littlewood 2007, Denney 2011)

Despite the scale of the above research areas, they have not advanced a new theoretical understanding of the discipline. For example, issues around automation and formalism were discussed in detail by Wilson and McDermid in 1995 (Wilson 1995). Around the same period, quantification of confidence was investigated through the Safety and Risk Evaluation using Bayesian Nets (SERENE) project (Bouissou 1999).

Finally, challenges and weaknesses in safety case concepts, methods and practices have been highlighted in published research. This is of course no bad thing — it is a necessary property of a healthy research community. The reported weaknesses relate to the complexity of the notion itself (e.g. safety case vs safety case report (Habli 2007)), the different interpretations of safety cases (Graydon 2017), practicalities particularly in regulatory contexts (e.g. integration into goal-based standards (Penny 2001)), the lack of sufficient skills and training (e.g. imbalance in skills between developers and assessors (Kelly 2008)), the need to

address cognitive biases (Leveson 2011), and the importance of stronger review and inspection mechanisms (e.g. similar to other prescriptive and standards measures in civil engineering (Wassyng 2010)). The lack of empirical evaluation has been highlighted by Sujan el. al. in a study funded by the Health Foundation (Sujan 2016) and more recently by (Graydon 2020) and (Reinhart 2017).

3 The Value of Safety Cases

A review of the literature reveals that there has been little serious research into whether safety cases are of net value for safety. By this we mean research that seeks to demonstrate that the money, effort etc. that is expended on developing a safety case — what (Woods 2015) calls "safety energy" — is better put to use on them than on other safety activities.

3.1 *The general research evaluation landscape*

The main work related to this was undertaken by (Reinhart 2017) that investigated "*assurance case practices*" with a focus on effectiveness. This work provides a starting point for understanding the value of safety cases. However, the results are based exclusively on literature review and self-reporting from interviews. The literature reviewed includes very little systematic empirical work.

Other work reports upon the success of big projects that have used safety cases (Eisner 2000). However these are little more than detailed anecdotes. In addition, it is very difficult to compare such examples as there is such variation in the nature of the systems considered and the ways in which the results are reported.

Overall, the evidence base for the value of safety cases is poor. In addition, rival theories say that there are better ways to spend safety energy than through developing safety cases, e.g. by developing and following highly prescriptive and domain-specific compliance standards (Wassyng 2010).

A lack of rigorous empirical evidence for the effectiveness of safety cases coupled with the existence of credible rival approaches is sufficient grounds upon which to question the actual value of safety cases. However, it does not necessarily imply that safety cases are not an effective approach. It is important to acknowledge the difficulty in undertaking this kind of evaluation. This is true not just for safety cases, but for any safety approach. There are a number of impediments that exist here, including but not limited to: lack of public domain safety case examples, inability to share data, lack of control studies, lack of funding for empirical research in this area and labour (and time) -intensive nature of this research. There are also lots of good non-systematic-empirical reasons to think that safety cases might be good idea. For a recent illustration of this in another domain, see (Greenhalgh 2020)

on the need for thoughtful, interpretive critical reflection and narrative review when assessing the efficacy of face coverings in controlling the spread of COVID-19.

3.2 *Specific claims about the value of safety cases*

The literature discussed in the previous section makes some specific claims about the value of safety case approaches. (Reinhart 2017) extracts seven claims about them (albeit using the more generic term *assurance case*):

1. Assurance cases are successful where suitable.
2. Assurance cases are more comprehensive than conventional methods alone.
3. Assurance cases improve the allocation of responsibility over prior norms.
4. Assurance cases organize information more effectively than conventional methods.
5. Assurance cases address modern certification challenges.
6. Assurance cases offer an efficient certification path compared to other approaches.
7. Assurance cases provide a practical, robust way to establish due diligence.

Reinhart et al's work is a valuable step towards our goals. However, these claims are too high level and coarse-grained to be easily studied. (Graydon 2020), building on the review by Rinehart et al., derives more testable hypotheses. These are interesting, but do not form a coherent whole — Graydon proposes a large number of small variance claims that do not obviously fit together.

One thing that is revealed from considering the various claims made regarding safety cases is the sheer range of claims on what safety cases can do (Graydon 2017). This raises important questions such as:

- How compatible are the claims with each other?
- Do different "*schools of thought*" (Graydon 2017) suggest incompatible or awkwardly-compatible things?
- Can safety cases really do all those things? Can they be done at the same time?

4 Evaluating Safety Case Efficacy

4.1 *We need process, as well as variance, theories*

In the previous section we discussed some of the existing claims made regarding safety cases. These claims invariably seek to characterise the "what", in terms of what safety cases achieve. They thus constitute a *variance theory*, in that they

describe how one property (e.g. awareness of risk) varies with another (e.g. use of a particular approach to safety cases).

Such theories are necessary and valuable, but it is also important to establish *process theories* for safety cases — theories that explain *how* safety cases affect behaviour and outcomes (Van de Ven 2007). In a simple domain, where convincing experiments are easy to carry out, we maybe would not need process theories; we would just do large numbers of experiments to evaluate and tune our variance theories. But in the extremely complex sociotechnical domain that we are concerned with, we desperately need process theories so we can identify those few empirical studies that are practical and that will shed some light on what is happening.

We can use the work of Reinhart et al. and Graydon, but we need to do so in the context of building and evaluating process theories i.e. understanding the "how" claims. We need to model what a factor or property of interest does for safety and how it does it — we need to model the *mechanism* of its effect on safety. We need to turn the isolated claims into testable theories or, at least, into theories that imply a range of testable hypotheses.

4.2 *We need to compare multiple rival theories*

A straightforward way to make testable hypotheses is to define one theory we believe is important and set up studies to test it. In each study, we ask "are these results consistent with my theory, or are they just noise?" This is something like the default method in the natural sciences, and in much of social science too.

A null-hypothesis approach is unrealistic, however, for our concerns here — we know beyond reasonable doubt that adopting a safety case approach does *something* to an engineering process. If nothing else, it costs money — money which is likely to be taken away from some other safety activity. And given that many practitioners are often positive about them (see e.g. (Reinhart 2017)) we know that *in the very worst case* the act of producing safety cases gives engineers the *impression* that they are valuable; that the use of safety cases influences engineer perception of what good safety practice is.

In other words, it is unlikely that something as complex and far-reaching as safety case work does nothing at all in an engineering or operations process, unless the work is so siloed and isolated that the real development or operational personnel cannot even see it. At the very least, taking the most pessimistic position, safety cases will do nothing that is beneficial and some things that are not. They are extremely unlikely to do nothing at all.

We therefore need to define several rival process theories, and conduct studies to figure out which of them are the best, i.e. are most consistent with the observations we are able to make. For more on how to do that, see (Ralph 2019, Van de Ven 2007, and Yin 2009). Some of our theories could, and should, be quite negative about safety cases. For a given study, two rivals may well be enough; for a wider programme, more will be needed.

4.3 *We need to be selective about the theories we define and investigate*

We have assumed so far that we will build up our theories from the literature and from our own ideas. However there is a risk there that our prior assumptions may lead us into theories that are divorced from reality. So we should supplement our theory building with bottom-up descriptive work (Rae 2020). We need to prioritise what theories we develop and study. That is, we need to establish the assurance needs that potentially could be met by safety cases.

It is not hard for us to come up with theories, especially partial or fragmentary ones. Fleshing them out thoroughly and conducting basic stress testing (e.g. manual review by peers) is important but also expensive. Running studies to compare their explanatory power is extremely expensive. We can prioritise by a variety of criteria which include:

- Face plausibility
- Prominence in the literature, in interviews, or in other evidence of what people are believing
- Relevance to practice
- The degree to which they are embodied in existing safety standards and guidelines

If we want to change practitioners' behaviour based on empirical studies, we have to perform empirical studies directly about the beliefs that at least some of them hold. Otherwise, we are asking them to do too much work to figure out how research results might apply to them, and whether the research has any implications for their behaviour. After all, many practitioners are already predisposed to ignore academic research (Devanbu 2016).

Beyond that, if we can work out what theories are most "load-bearing" for stakeholders, i.e. which theories seem to most support the acceptance of safety cases as a practice, we can optimise our research to evaluate those. We should select rival theories in a similar way, in that we also evaluate those rivals that would significantly undermine trust in safety cases if people came to believe in them.

4.4 *Safety engineering theories need to be explicit about the effect of context*

As we have already emphasised, the starting point for a systematic evaluation process is to establish the safety assurance need or problems. Observational studies are useful instruments for identifying credible problems that safety cases have the potential to address and importantly for characterising the contexts within which

these problems occur, e.g. stakeholders, conventions and technical, social and financial constraints.

Observational studies are particularly relevant here for understanding contextual factors, as safety cases are complex interventions in that there is no concrete boundary between the safety case process and other safety activities and the wider technical and sociotechnical processes. This is often illustrated by the different interpretations of what a safety case is, how it is used and by whom (Graydon 2017, Rae 2020). For instance, are hazard-driven simulations part of the safety case or an input to the safety case?

Observational studies should also help appraise the nature, criticality and scale of the problem. For example, the fact that many safety case practices are manual or qualitative is often presented by researchers as a weakness. Observational studies could reveal, however, the extent to which safety cases work by stimulating engineers to understand their system better — something that might not have happen if the process had been automated.

Safety concerns are very context-sensitive, so it is not enough to just study variation in terms of the safety case technique used. For example, take the following form of hypothesis (Graydon 2020):

"form of argument" produces more/better "kind of value" for "actor" than "alternative"

That omits some key experimental variables — the type of system being developed and the context in which that is being done. We cannot control these variables — deploying one form of safety assurance is already an expensive exercise, so duplicating such effort is unlikely to be approved by senior management.

It follows that before we can study a safety intervention, we need to make a detailed description of that context. The key experimental variables lie in the system and the environment rather than just in the safety case solution, i.e. "form of argument". Proposing to control these variables can be neither feasible nor desirable. Deploying one form of safety assurance is already an expensive exercise. Duplicating such effort is unlikely to be approved by senior management. More importantly, in a complex intervention, we are interested in understanding how the proposed solution influences and is influenced by its dynamic environment. Constraining the environment is likely to invalidate the credibility of the results and could lead to the loss of valuable insights into the different contextual factors, e.g. impact of system and environmental complexity on the significance and efficacy of different safety argument formats.

A corollary of this is that when someone proposes a new safety case method, there is a lot that they can (and perhaps should) do to make it practical to evaluate it. They can support evaluation by defining an explicit, detailed theory of how it works, explaining exactly *how* the method produces outputs that meet the safety assurance need. If they do provide this hypothesised causal chain, then researchers

can evaluate it; if they don't, researchers will have to do considerable prior work to understand why this new method might work.

4.5 *A good result from a research study is a modification to a process theory or a change in our belief levels with respect to two rival theories*

The result of a research study of the form discussed above should be some combination of:

- A change to our relative levels of confidence as to which of the candidate theories is the best representation of what's going on
- A modification of a theory
  - E.g. by adding an applicability condition to the whole thing (e.g. "only when safety case developers perceived that a regulator was going to rigorously review their safety case")
  - E.g. by striking out a connection (e.g. "despite searching hard across multiple organisations, we have found no evidence that incidents or accidents ever modify the high-level architecture of a case")
  - E.g. by adding moderator to connection (e.g. "it appears that contradictions and ambiguities found by safety case developers can lead to changes in design engineer's view of how safe the system is, but this is very much moderated by the perceived competence and status of the safety case developers")

By necessity, we must progress by means of individual studies, but it is the *research programme*, not an individual study, by which substantial advances in knowledge can happen. Individual studies tend to be small in scope, and are invariably vulnerable to error. A research programme built around a process theory, or a set of rival process theories, can be cumulative over the time in that it accumulates the insight of many different studies.

There is never likely to be a stopping rule here for the overall research programme. We are always going to be in a space of "should we study safety case efficacy more? if so, what specifically should we study?".

5 An Example Theory

Let us look at an example model, the McDermid Square (MoD, 2007), that theorises the necessary level of detail and rigour in safety cases against the novelty of the problem and solution. As depicted in Figure 1, the more novel the solution and the

problem are, the more extensive the argument and evidence need to be, including greater independence.

		Solution	
		Familiar	Unfamiliar
Problem	Familiar	Minimal argument and standard evidence from the domain, e.g. stability certificate.	Focused argument on reasons for novel solution, plus the appropriate evidence.
	Unfamiliar	Minimal argument and standard evidence from another domain, e.g. railway safety case.	Extensive argument and evidence, with substantial independent scrutiny.

Fig.1. McDermid Square: effect of problem and solution unfamiliarity on argument and evidence requirements (MoD, 2007)

Let us take this example further and use the McDermid Square as the basis for an evaluation of the level of rigor in safety cases for a specific industrial context. The evaluation is scoped in the form of a *logic model* (Figure 2)[1].

The specific logic model we define in Figure 2 focuses on an airborne software scenario. The goal is to justify that a *software system* performs its intended function with a level of confidence in safety.

The three forms of safety assurance approaches that we consider here are:

a) A software safety case with an explicit and structured argument, represented in GSN, that is managed in the commercially-supported software tool ASCE[2]
b) A tabular and detailed hazard log showing that all risks are low/tolerable
c) DO178C software lifecycle data, including Software Accomplishment Summary, in a free text format.

[1] A logic model is a research instrument that is used in social sciences for defining the causal links between the elements and activities of a complex intervention, including the hypothesised intermediate and long-term outcomes. Importantly, a logic model is used to explicitly define the contextual factors that are likely to influence the hypothesised outcomes.

[2] ASCE Software: https://www.adelard.com/asce/choosing-asce/index/

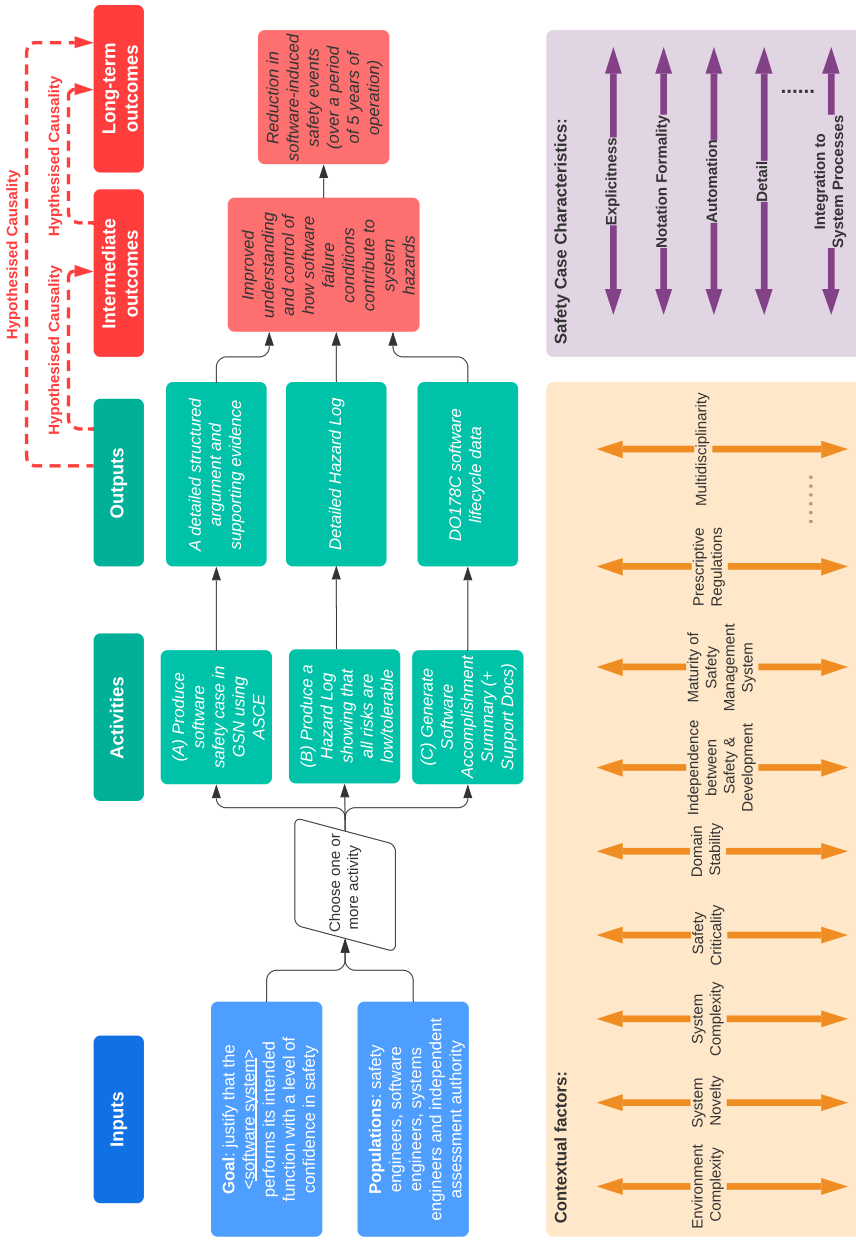

Fig.2. Scope of an Example Safety Case Research Study

We can note that in (b) and (a), the safety argument is implicit. They are thus not safety case techniques per se, but they are rival safety techniques which may lead to the same outcomes. They are thus credible candidates for comparison.

The intermediate outcomes that we specify are improvements in the understanding and control of how software failure conditions contribute to system hazards. These outcomes could be gauged via surveys, interviews and observational studies. The hypothesised long-term outcomes take the form of reduction in software-induced safety events, over a period of 5 years of operation. This could be established by reviewing software safety reports issued over that period.

The *software system* here is a key variable. We can illustrate by considering two scenarios:

1. A rule-based aircraft wheel braking software system, developed in MISRA C[3], in a human piloted commercial aircraft
2. An autonomous taxiing software system in an unmanned aircraft.

For each scenario it is essential to detail the key contextual factors, such as the ones we list in Figure 2, and hypothesise their influence on the outputs and outcomes. Two plausible hypotheses would be:

1. Increased rigor and clarity in the argument representation for the aircraft wheel braking software has very little effect on the stated outcomes.
   but conversely
2. Increased rigor and clarity in the argument representation for the autonomous taxiing system ultimately reduces the level of the software-induced safety events

Hypothesis (2) could also be further qualified by stating that *too much rigor* in the representation of the argument, e.g. through formal mathematical notations and automation, is counterproductive in that such rigor complicates the communication between the different stakeholders and masks the sources of uncertainty that are hard to formalise. This could be analysed when gauging the intermediate outcomes.

We realise that systematically and rigorously conducting the above study would be expensive. It is unclear whether a case could be made for similar evaluation studies to be funded by industry. What is clear however is that such studies are infeasible to conduct without the direct engagement of practising engineers, users and regulators and without access to the real development and operational settings (while of course maintaining the confidentiality of sensitive information). What is feasible, but unfortunately underutilised, is case study research — in the model describe by (Yin 2009) — in which one safety case approach could be evaluated in its natural industrial context. Experimental, hypothesis-driven studies are preferred

[3] https://www.misra.org.uk/Publications/tabid/57/Default.aspx

but case studies, particularly those utilising observational designs, are more practical and in some cases offer more insights into the complex contextual factors.

6 Conclusions

Empirical evidence for the value of safety cases is weak. This does not mean we should assume they *don't* have value, but it does mean that improvements identified and proposed by the research and wider engineering community are not adopted in practice by organisations. This will mean ineffective use of safety cases and missed opportunities for safety improvements. Conversely, if developments in safety cases are adopted by industry without empirical evidence of effectiveness, this has the potential to not just fail to improve industrial practice, but in the worst case to make it worse.

What is needed is a way to theorise well about how safety cases are effective, and a way to compare them to rival approaches. There are several credible theories and it looks like they can be usefully described. With the research approach we have discussed here, we think that it is possible and practicable to design and conduct studies that are likely to shed some light on key issues and reveal important distinctions.

Beyond safety cases, the research methods outlined here may be applicable to many topics in the safety domain. Those topics all involve questions about social behaviour in organisations that is long term and hard to observe, and are all concerned with outcomes that are extremely rare.

In short, our final messages are as follows:

- Progress can only be credibly made through close collaboration between safety engineers who can identify the assurance needs and provide access to credible industrial settings, and researchers who can design and help execute empirical research studies.
- Experimental, hypothesis-driven studies are preferable but case studies, particularly observational ones, are more likely to be practicable.
- In order to have a genuinely cumulative research programme in safety cases, we need to develop and revise explicit process theories for how safety cases work. Having such theories makes research (a) practical, in that they allow you to identify those intermediate-result studies that *can* be done and (b) cumulative, in that experimental results can contribute to changing the theory, thus making the prevailing theor(ies) a record of our best current understanding of safety cases and their effects.
- Researchers wishing to develop new approaches to safety cases, e.g. new notations, formalisms and tools, should explain and justify the safety assurance

need, the process theory they are assuming, and the evaluation methods. If they use a case study, they should explain whether it is conducted in its natural industrial context or merely used for illustrative purposes. The latter has explanatory value, but no external validity.

- When practising engineers report their experiences with safety case approaches, they should clarify whether these experiences correspond to their personal engineering judgments and opinions or the results of empirical research conducted in the relevant and clearly defined industrial settings (i.e. they should distinguish between opinion-based and evidence-based practice (Hampton 2002)).

Acknowledgements. This work is partly funded by the Assuring Autonomy International Programme https://www.york.ac.uk/assuring-autonomy.

References

C. Alexander, A pattern language: towns, buildings, construction. Oxford university press, August 1977.

L. Arnold, Windscale 1957: anatomy of a nuclear accident. Springer, 2016.

Assurance Case Working Group, "GSN community standard version 2," Available on-line: https://scsc.uk/scsc-141B, 2018.

A. Ayoub, B. Kim, I. Lee, and O. Sokolsky, "A systematic approach to justifying sufficient confidence in software safety arguments," in International Conference on Computer Safety, Reliability, and Security. Springer, 2012, pp. 305–316.

J. Birch, R. Rivett, I. Habli, B. Bradshaw, J. Botham, D. Higham, P. Jesty, H. Monkhouse, and R. Palin, "Safety cases and their role in iso 26262 functional safety assessment," in International Conference on Computer Safety, Reliability, and Security. Springer, 2013, pp. 154–165.

P. Bishop and R. Bloomfield, "A methodology for safety case development," in Safety and Reliability, vol. 20, no. 1. Taylor & Francis, 2000, pp. 34–42.

R. Bloomfield and P. Bishop, "Safety and assurance cases: Past, present and possible future–an Adelard perspective," in Making Systems Safer. Springer, 2010, pp. 51–67.

R. Bloomfield, P. Bishop, C. Jones, and P. Froome, "Ascad—Adelard safety case development manual," Adelard, vol. 5, 1998.

M. Bouissou, F. Martin, and A. Ourghanlian, "Assessment of a safety-critical system including software: a bayesian belief network for evidence sources," in Annual Reliability and Maintainability. Symposium. 1999 Proceedings (Cat. No. 99CH36283). IEEE, 1999, pp. 142–150.

T. Chowdhury, A. Wassyng, R. F. Paige, and M. Lawford, "Systematic evaluation of (safety) assurance cases," in International Conference on Computer Safety, Reliability, and Security. Springer, 2020, pp. 18–33.

A. Dardenne, A. Van Lamsweerde, and S. Fickas, "Goal-directed requirements acquisition," Science of computer programming, vol. 20, no. 1-2, pp. 3–50, 1993.

E. Denney and G. Pai, "A methodology for the development of assurance arguments for unmanned aircraft systems," in 33rd International System Safety Conference (ISSC 2015), 2015.

E. Denney and G. Pai, "Automating the assembly of aviation safety cases," IEEE Transactions on Reliability, vol. 63, no. 4, pp. 830–849, 2014.

E. Denney, G. Pai, and J. Pohl, "Advocate: An assurance case automation toolset," in International Conference on Computer Safety, Reliability, and Security. Springer, 2012, pp. 8–21.

E. Denney, G. Pai, and I. Habli, "Towards measurement of confidence in safety cases," in 2011 International Symposium on Empirical Software Engineering and Measurement. IEEE, 2011, pp. 380–383.

P. Devanbu, T. Zimmermann, and C. Bird, "Belief & evidence in empirical software engineering," in 2016 IEEE/ACM 38th International Conference on Software Engineering (ICSE). IEEE, 2016, pp. 108–119.

H. Eisner, "The channel tunnel safety authority," Safety science, vol. 36, no. 1, pp. 1–18, 2000.

E. Gamma, Design patterns: elements of reusable object-oriented software. Pearson Education India, 1995.

M. S. Graydon, "Towards efficacy hypotheses for safety cases," in 2020 16th European Dependable Computing Conference (EDCC). IEEE, 2020, pp. 51–58.

P. J. Graydon, "The many conflicting visions of 'safety case'," in 2017 47th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W). IEEE, 2017, pp. 103–104.

P. J. Graydon, "The safety argumentation schools of thought," in 3rd AAA 2017 International Workshop on Argument for Agreement and Assurance, Nov 2017. [Online]. Available: https://ntrs.nasa.gov/search.jsp?R=20180000378

T. Greenhalgh, "Face coverings for the public: Laying straw men to rest," Journal of Evaluation in Clinical Practice, p. e13415, 2020.

W. S. Greenwell, J. C. Knight, C. M. Holloway, and J. J. Pease, "A taxonomy of fallacies in system safety arguments," 2006.

J. Guiochet, Q. A. Do Hoang, and M. Kaaniche, "A model for safety case confidence assessment," in International Conference on Computer Safety, Reliability, and Security. Springer, 2014, pp. 313–327.

I. Habli and T. Kelly, "A generic goal-based certification argument for the justification of formal analysis," Electronic Notes in Theoretical Computer Science, vol. 238, no. 4, pp. 27–39, 2009.

I. Habli and T. Kelly, "Safety case depictions vs. safety cases–would the real safety case please stand up?" in 2nd Institution of Engineering and Technology International Conference on System Safety. IET, 2007, pp. 245–248.

C. Haddon-Cave, The Nimrod Review: an independent review into the broader issues surrounding the loss of the RAF Nimrod MR2 aircraft XV230 in Afghanistan in 2006. London: The Stationery Office, 2009, vol. 1025.

J. R. Hampton, "Evidence-based medicine, opinion-based medicine, and real-world medicine," Perspectives in Biology and Medicine, vol. 45, no. 4, pp. 549–568, 2002.

R. Hawkins, T. Kelly, J. Knight, and P. Graydon, "A new approach to creating clear safety arguments," in Advances in systems safety. Springer, 2011, pp. 3–23.

R. Hawkins, I. Habli, D. Kolovos, R. Paige, and T. Kelly, "Weaving an assurance case from design: a model-based approach," in 2015 IEEE 16th International Symposium on High Assurance Systems Engineering. IEEE, 2015, pp. 110–117.

T. P. Kelly, "Arguing safety-a systematic approach to safety case management," DPhil Thesis York University, Department of Computer Science Report YCST, 1999.

T. Kelly, "Concepts and principles of compositional safety case construction," Contract Research Report for QinetiQ COMSA/2001/1/1, vol. 34, 2001.

Kelly, T., 2008. Are safety cases working. Safety Critical Systems Club Newsletter, 17(2), pp.31-33.

N. G. Leveson, "The use of safety cases in certification and regulation," 2011.

C. Picardi, R. Hawkins, C. Paterson, and I. Habli, "A pattern for arguing the assurance of machine learning in medical diagnosis systems," in International Conference on Computer Safety, Reliability, and Security. Springer, 2019, pp. 165–179. 31.

B. Littlewood and D. Wright, “The use of multilegged arguments to increase confidence in safety claims for software-based systems: A study based on a bbn analysis of an idealized example,” IEEE Transactions on Software Engineering, vol. 33, no. 5, pp. 347–365, 2007.

Ministry of Defence (MoD), “Standard 00-56 on safety management requirements for defence systems,” Ministry of Defence, Directorate of Standardisation, Kentigern House, vol. 65, 2007.

J. Penny, A. Eaton, P. Bishop, and R. Bloomfield, “The practicalities of goal-based safety regulation,” in Aspects of Safety Management. Springer, 2001, pp. 35–48.

P. Ralph, “Toward methodological guidelines for process theories and taxonomies in software engineering,” IEEE Transactions on Software Engineering, vol. 45, no. 7, p. 712–735, Jul 2019.

A. Rae, D. Provan, H. Aboelssaad, and R. Alexander, “A manifesto for reality-based safety science,” Safety science, vol. 126, p. 104654, 2020.

D. J. Rinehart, J. C. Knight, and J. Rowanhill, Understanding what it Means for Assurance Cases to “work”. National Aeronautics and Space Administration, Langley Research Center, 2017.

J. Rumbaugh, I. Jacobson, and G. Booch, “The unified modeling language,” Reference manual, 1999.

J. Rushby, “Formalism in safety cases,” in Making Systems Safer. Springer, 2010, pp. 3–17.

J. Rushby, “The interpretation and evaluation of assurance cases,” Tech. Rep. SRI-CSL-15-01. Comp. Science Laboratory, SRI International, 2015.

M. A. Sujan, I. Habli, T. P. Kelly, S. Pozzi, and C. W. Johnson, “Should healthcare providers do safety cases? lessons from a cross-industry review of safety case practices,” Safety science, vol. 84, pp. 181–189, 2016.

S. Toulmin, “The uses of argument cambridge university press,” Cambridge, UK, 1958.

A. H. Van de Ven et al., Engaged scholarship: A guide for organizational and social research. Oxford University Press on Demand, 2007.

A. Wassyng, T. Maibaum, M. Lawford, and H. Bherer, “Software certification: Is there a case against safety cases?” in Monterey Workshop. Springer, 2010, pp. 206–227.

R. Wei, T. P. Kelly, X. Dai, S. Zhao, and R. Hawkins, “Model based system assurance using the structured assurance case metamodel,” Journal of Systems and Software, vol. 154, pp. 211–233, 2019.

S. Wilson, S. P. and McDermid J. A., “Integrated Analysis of Complex Safety Critical Systems”. Computer Journal, vol. 38, no. 10, pp. 765-776, 1995.

D. Woods, M. Branlat, I. Herrera, and R. Woltjer, “Where is the organization looking in order to be proactive about safety? a framework for revealing whether it is mostly looking back, also looking forward or simply looking away,” Journal of Contingencies and Crisis Management, vol. 23, no. 2, pp. 97–105, 2015.

F. Ye, “Justifying the use of cots components within safety critical applications,” Ph.D. dissertation, Citeseer, 2005.

R. Yin, Case Study Research: Design and Methods, 4th ed. SAGE, 2009.

T. Yuan, T. Kelly, T. Xu, H. Wang, and L. Zhao, “A dialogue based safety argument review tool,” in Proceedings of the 1st international workshop on argument for agreement and assurance (AAA-2013), Kanagawa, Japan, 2013.