Exemplar papers for Reality-Based Safety Engineering

Rob Alexander*

** High-Integrity Systems (HISE) group, Department of Computer Science, University of York.*
rob.alexander@york.ac.uk

Technical Report, Department of Computer Science, University of York

Report number YCS-2023-506

Revision History

1.0.1	6 July 2023	First public release

Introduction

This document presents a set of *exemplar papers* for different kinds of safety engineering research. I chose these as exemplars because they embody good practice — they are worked examples of how to do good research. By studying these and adapting what they do, we can improve our own research practice.

You might ask — “Why not just point to the best research methods guides and tutorials”? Well, partly because there isn’t much of that for safety research — almost everything is about research in other domains which have different epistemic problems. But also because exemplars are more compelling and informative than methods tutorials are — social-science-type methods have big gaps of "and now you have to use informal judgement" in them, so method tutorials only take you so far. Finally, each exemplar implicitly says "someone else did this and got meaningful results, so maybe you could, too".

The document is organised as a series of categories — each category is a different “kind” of research, in that it has distinct goals and (therefore) methods. The category set is probably not unique to safety engineering, but it is specialised for such, in that I have emphasised types of studies that we particularly need to do because of the peculiarities of our subject.

The categories are, in turn, split into two “axes”:

1. Axis 1 categories contain studies that take a broad view, asking “How does the safety world work?”
2. Axis 2 categories contain studies of the narrower form “I (or someone else) made a specific method, process or tool… is it any good?”

This distinction is important because papers in different axes need to be methodologically different in ways that are often quite subtle. Confusing the two is an easy way for two people to argue at cross purposes.

Where possible, I’ve picked exemplars from safety engineering. These are thin on the ground, however, so I’ve often had to compromise and look further afield. In such cases, I’ve tried to give equal billing to papers studying *operational* safety and papers studying *non-safety* engineering, in the hope that between them they will be useful guides.

I’ve aimed for this paper to reach a broad audience of safety researchers, but I suspect it’s most valuable for junior researchers and for experienced researchers who are coming from outside safety.

A suggestion on how to use this document

My recommendation is that you read the exemplar papers in parallel with this document, paying particular attention to the properties they have that this document points out. Then discuss them with us and your (other) peers.

Exemplars for Axis 1 — Theories about how the safety world works

Category — Elicitation of how practitioners (or some other important group) understand the world

This category is for studies that ask “what does important practitioner group X believe about how safety happens?”.

These studies are important because practitioners have a wealth of knowledge that is otherwise inaccessible. Not everything every practitioner believes will be true, or even useful (especially if speaking outside their direct experience) but their beliefs provide some access to the reality of practice — especially for a researcher who does not have (recent, relevant) practical experience.

Crucially, practitioner beliefs and understandings provide access to the rationale (at least the ostensible rationale) behind many of their actions — something that is almost impossible to infer from observation. Much seemingly bizarre practitioner behaviour makes sense once the motives are known.

Current favourite exemplar

D. J. Provan, A. J. Rae, and S. W. A. Dekker, ‘An ethnography of the safety professional’s dilemma: Safety work or the safety of work?’ [1]

Why is that paper good:

- It’s in the safety operations domain, so the people studied have obvious parallels to the equivalent safety engineering roles. This is also a type of study where there is *probably* little difference needed between how you do it in operations and in engineering.
- It starts (first para sec 1) by citing a prior review of empirical studies that shows a worrying picture (there’s little evidence that having safety professionals in an org improves worker safety). That review was by someone else, which makes it a little more credible - it shows that someone else has independently come to the position they are starting from.
- It applies a previously developed theory (a model of types of safety work) (sec 1.2, sec 3, in particular the *headings* at the 3.x level). This theory is attractive because it is based on broader work outside the safety field (see first para p1.2). In the process it refines and extends that theory a little. See in particular sec 3.5 where it (a) uses the examples from the interviews to define a few subtypes of each type of safety work and (b) identifies two other activities that are important for safety professionals. It does a good job of explaining the theory, making it more understandable to readers/users, because each of the 62 interview “cases” is categorised and thus constitutes a concrete example of something that belongs in that category.
- Sec 2 describes their method, which others could adapt.

- It combines interview and observation techniques, something that's often advised by methods tutorials.
- The numerous quotes from participants, along with the third-person accounts of participant experience, provide a vivid and evocative window into working reality (sec 3.1–3.4).
- It notes conflicts between its observations and both (a) the way things "should" work according to a prominent safety theory/philosophy/school (sec 3.6 para *"Resilience engineering, safety II, …"*) and (b) a general model of what needs to be done to achieve certain aims (sec 3.8, especially the second and third para).
- It makes a proposal for change that follows from the results of the study and the (wide range of) literature it surveys (sec 3.7 last para), and it does so without presenting that proposal as something elaborate and complex, or indeed as the main contribution of the paper.
- It makes an interesting observation on the (non-) impact of research that was revealed by the study (sec 3.9 especially first para).
- It draws non-obvious conclusions (sec 4). In particular, note its conclusion about how the increasing alignment of safety professionals with line management, and the increasing move towards social, demonstrated and administrative safety work, look to have become a problem, leading to high "safety" spend but no improvement in operational safety.

Key caveats on that paper:

- It's not about safety *engineering* — focus is on safety operations
- The organising theory used and refined was developed by the same authors. This isn't bad per se, but a healthy subfield won't *only* have this kind of relationship - you need people applying each other's theories, too.
- Their method description (sec 2) is not that detailed
- It's not really clear how the observation work was combined with the interview results

Category — Observation and description of how engineering work happens in practice

This category is for studies that ask “what actually happens during engineering work X” and tries to answer it by observing engineers at work.

Studies in this category are important because it complements “Elicitation of how practitioners understand the world …” by observing in the field those activities that the elicitation merely asked about, thereby avoiding the “talk vs walk” problem (i.e. response bias in surveys and interviews). Of course, interpreting observations of knowledge work is difficult, so it invariably involves some degree of asking questions, but usually quite narrow ones and in the context of live evidence. (E.g. an engineer might believe they do some kind of check “regularly”, but their own log book could give the lie to that).

These papers are distinct from the Axis 2 category “Evaluation of an existing method, process, or tool” because it asks *how* rather than *how well*. I.e. rather than asking *“how well does FMEA work?”* we’re asking *“what do people actually do when they are ‘doing FMEA’”*.

(If you’re familiar with the distinction between *variance* and *process* theories [2], this category is mostly about process ones — it’s part of discovering the mechanism by which variance of outcome can happen.)

Current favourite exemplar

J. Rooksby, M. Rouncefield, and I. Sommerville, ‘Testing in the Wild: The Social and Organisational Dimensions of Real World Practice’ [3]

Why is that paper good:

- It’s an ethnographic study of people doing engineering work over an extended period of time.
- It studies four separate development projects (albeit in the same organisation), thus reducing the risk that it’s reporting on one very unusual case.
- Its results, though perhaps not surprising to anyone with first-hand experience of testing in practice, are not *obvious*, certainly not to the majority of researchers in the software testing space. The paper makes this easier to see by making explicit (a) what that understanding by the majority of researchers seems to be and (b) how these results challenge that understanding

Key caveats on that paper:

- Not safety — This is very much “medium-integrity” software, in a minimally-regulated setting.
- “[...] The data is under-analysed and under-theorised. A little bit more analysis and they could have come up with good questions. [...] [e.g.] They hint at time mattering in ways more interesting than just being short on time, but they don't explore it. [...]” (Drew Rae,

personal communication)

Category — Theory specification, in a way that supports empirical study and comparison with rivals

This category is about papers that state, specify, or refine theories about how the world works, specifically about how safety happens (or doesn't happen).

The main distinction from "Elicitation of how practitioners understand the world" papers, and to some extent "Observation and description of how engineering work happens in practice" ones, is that "Theory specification…" papers put forward claims by academics in their own voice. They make claims about how the world works *in a form suitable for empirical tests*, or at least that make significant progress towards that.

The *"in a way..."* clause in the category name is to rule out the grand-scope but loosely-specified theories of e.g. Leveson, Hollnagel etc. These grand theories have a role to play in how we think about safety, but it's not reasonable for the vast majority of researchers to ever propose one of their own.[1]

Current favourite exemplar

A. Rae and D. Provan, 'Safety work versus the safety of work' [5]

Why is that paper good:

- It presents a theory that is potentially testable. A crude summary of that theory would be — "the proximal purpose of most operational safety actors, or at least the effect of the actions of most such actors, is not to achieve operational safety — their purpose is social, administrative, or demonstrative"). It might be possible to test that versus a rival of roughly "actually, they really do (aim to) achieve operational safety". Sec 5.2 in the paper does suggest some directions for testing it, and we can note that their later paper *'An ethnography of the safety professional's dilemma: Safety work or the safety of work?'*[2] carries out something they suggest (*"In particular, [we should study] how do safety practitioners […] explain how and why they perform safety work"*).

Key caveats on that paper:

- It's focused on operational safety rather than safety engineering

[1] Indeed, we now have so many such grand theories that people should stop creating them, and instead put their effort into serious specification, evaluation, and comparison of the existing ones, *or* (and that's the point of this category) creating narrower, more specific theories. (cf the "manifesto" that I co-wrote with Rae, Provan, and Aboelssaad [4])

[2] See elsewhere in this document for details

Category — Literature review for theories about how some given kind of work happens

This category is for literature review of papers from the category “Theory specification…”.

This category is important because, as noted earlier, we need to understand how the world works before we try to change it. And to do that efficiently, without wasting the effort of countless prior researchers, we bring together what is already known and do our new work in the light of that.

Current favourite exemplar

D. J. Provan, S. W. A. Dekker, and A. J. Rae, ‘Bureaucracy, influence and beliefs: A literature review of the factors shaping the role of a safety professional’ [6]

Why is that paper good:

- It is in the safety space
- It provides an extensive literature review about how the work of safety professionals is shaped
- It gives some detail about its method (sec 1 para “The present review…”)
- It is uses explicit research methods (thematic analysis and cognitive mapping) to go over the reviewed papers and derive a useful organising structure, and then uses this to organise itself. (e.g. secs 2–4 correspond to the three “categories” in the map) (sec 1 para “The present review…”). These methods probably made it easier to write, by giving them methods to follow rather than just leaving them with a big stack of papers and only their intuition to guide them.

  NB I would call this organising structure a “theory” in the broad sense, in that it provides a way to split up phenomena into categories (which is one of the important roles of theory). Ralph would probably call this “a theory for understanding” [7].
- Sometimes (though not as much as I might like) it spells out how the reviewed empirical studies were done, thus helping the reader to gauge their validity (e.g. search for “in the UK”, “undertook a review”, “conducted the largest study”, “conducted a questionnaire”). This is a small thing, but important.
- It identifies differences between the beliefs of practitioners and those of researchers (sec 4.1 especially paras “Despite the changing…” and the one after it). It thus reveals the impact that researchers are (not) having.
- Throughout, it embodies a deep understanding of the actors and entities and phenomena in the safety domain, and of how they tend to interact. This could of course be the *product* of doing such a review, if done well, as well as something that the researchers bring to it.
- Overall, it discovers and explains (a) how safety professional’s work seems to happen and (b) where there are gaps or controversy in our understanding of that.

Caveats:

- It's not about safety *engineering* — its focus is on safety operations
- They could have described their literature review method more completely and in more detail.[3]
- It's not completely clear how sub and subsub -section headings relate to the cognitive map in figure 1.
- It is not clear whether empirical studies (or any kind of paper) were systematically screened and rated for quality, and thus it's not clear how likely those are to be valid and trustworthy. It's likely that these experienced researchers did use a critical eye, but you can't tell that from the text. Contrast Vilela et al and Martins et al (both are exemplars elsewhere in this document).
- A few paras read like a summary of claims rather than a coherent narrative (e.g. sec 3.2.2 first para)

[3] Although it's important we don't get too hung up on "all reviews must be rigorously and conventionally systematic (e.g. Cochrane-ready)" because that makes several kinds of thinking and argument impossible.

Category — Empirical evaluation of a single theory

Current favourite exemplar

The previously discussed *'An ethnography of the safety professional's dilemma: Safety work or the safety of work?'* does some of this. I could have split discussion of this perspective into its entry here, but I chose to keep to "one paper, one entry".

Category — Comparative empirical study of rival theories

See the paper *Safety Cases: An Impending Crisis?* [8] for an explanation of what this means (particularly sec 4.2). It's potentially a very important category, but it's not one for which I have truly good examples.

Current favourite exemplar

I don't actually know of many of these. The best I've got right now, which will serve to illustrate the point, is:

J. D. Shaw, N. Gupta, and J. E. Delery, 'Alternative Conceptualizations of the Relationship Between Voluntary Turnover and Organizational Performance' [9] [4]

Why is that paper good:

- It's a clear example of pitting rival theories against each other using a quantitative approach
- They start the paper by listing four rival theories and citing statements of them and giving advocacy for them (first para of introduction).
- They describe each theory, marshall the extant evidence in favour of each theory, and derive a testable hypothesis for each theory using common terminology and variables (see the four subsecs within section "Voluntary Turnover and Workforce Performance: …").
- Going beyond the four theories, they define a possible mechanism by which workforce performance (response variable in the first four theories) contributes to overall financial performance of firms (see the section "Voluntary Turnover and Financial Performance: …"). They then produce two more hypotheses which are tested in Study 2 (only).
- They carefully scope their population of study so as to match the population that they think was assumed by the creators of those theories (final para in intro).

Caveats

- It's not in safety engineering
- It's not *primarily* in safety at all
    - They do try to measure both workforce performance and workforce safety, but their safety measure was solely the number of lost time accidents per production employee per year, which is not a great measure.

[4] I found this via Leavitt et al [10], who gave it as a positive example of "strong inference" and theory pruning). (Their paper is about how a field that creates theories can avoid keeping too many of them.)

Exemplars for Axis 2 — Methods, Processes and Tools

Category — Description and illustration of an unevaluated and barely-validated new method, process, or tool that is appropriately humble in its claims

As Wieringa et al put it, a paper in this category *"proposes a solution technique and argues for its relevance, without a full-blown validation. The technique must be novel, or at least a significant improvement of an existing technique."*, adding later that *"A proof-of-concept may be offered by means of a small example, a sound argument, or by some other means."* [11] (Such a proof-of-concept is what we mean by "illustration" in the category title.)

What does "appropriately humble" look like?

- Having numerous explicit statements of limitations of the work and residual doubts about it — perhaps collected together in a prominent section along the lines of the "Threats to validity" that's common now in empirical papers. (This is similar to how I have noted "Caveats" about the papers I recommend in this document.)
- Being honest and clear about the nature of any example - e.g. by calling it a "worked example" and saying *"this is obvious a greatly simplified example which does not have a clear correspondence to a particular real-world case"*, and doing both very clearly *in the abstract and conclusions* where skimming readers will see it.

Current favourite exemplar

I don't have a good candidate in mind for this category. New method, process, and tool papers are a dime a dozen, but appropriate humility is rare. I welcome reader submissions for this category, in the hope of filling this gap in the next revision of this document.

Category — Validation of an existing method, process, or tool

The category is about studies that ask "how well is this likely to work in real-life practice?". There are a few ways to do validation studies, but most commonly they involve applying the method, process, or tool to a realistic-but-not-real example in a research (rather than practice) setting.

(Contrast with the similar category "Evaluation of…" which asks how well a thing *actually* works in real-life practice. I take the validation-evaluation distinction from Wieringa et al [11].)

Why do we need this category, given that we have "Solution proposal with illustrative example" and "Evaluation of…" categories? Firstly, having a clear third point on an axis can make the axis easier to understand — the extra middle point helps to clarify what is changing, and how, as we move from one extreme point to the other. The progression *illustration–validation–evaluation* is thus more informative than just thinking about *illustration–evaluation* alone. Secondly, if we want to get enough access to practice to do evaluation, we likely need good validation first to convince people it's worth it. Jumping from illustration to evaluation is likely to be harder.

Current favourite exemplar

P. Delgado-Pérez, I. Habli, S. Gregory, R. D. Alexander, J. A. Clark, and I. Medina-Bulo, 'Evaluation of Mutation Testing in a Nuclear Industry Case Study' [12]

This paper is here, rather than in "Evaluation…", because although it was on real software it was carried out by researchers based in a university. So e.g. it misses out the social dynamics of real-world testing as part of a safety-related software project.

Why is this paper good?

- It uses real software provided by a company that is working in this area
- It provides insight into the practice-relevant question "could mutation testing improve the fault-finding power of existing industrial test suites"? And, crucially, it also addresses the follow on question "could it do so affordably"?[5]

Caveats:

- It's not in safety per se (although it's in an area of software engineering that's strongly associated with safety).
- It's in an area of software engineering where a large part of the thing being studied (the actual code that is to be tested and its associated test sets) can be packaged up and copied cheaply. This is rarely true. Similarly, it was easier to do validation on this than (most) safety methods because the technique is largely done by the computer on computer data (program code) and so it's not that hard to take that out of its real environment and into the lab.

[5] It happens to answer both in the affirmative, but that's not necessary for this to be a good paper.

Category — Evaluation of an existing method, process, or tool

Here, "evaluation" means asking "how well does this *actually* work in real-life practice?". It's for papers reporting primary studies of safety work happening under real-world conditions.

(Contrast with the category "Validation of an…", which is about "how well would this *likely* work in real-life practice?".)

Current favourite exemplar

J. Havinga, M. I. Shire, and A. Rae, 'Should We Cut the Cards? Assessing the Influence of "Take 5" Pre-Task Risk Assessments on Safety' [13]

Why is that paper good:

- It is in (operational) safety
- It studies a technique that's very widely used in practice but has been little studied (perhaps because it sounds almost trivial but obviously useful)
- It actually does a (somewhat) controlled experiment
- Because the method under evaluation ("take 5") isn't well-defined in the literature, they define it (single-line para "The essential feature…" in sec 1, the following two bullet lists, and the para after that). Note the careful structure there – it defines
    - "What properties all 'take 5' must have" in order to count as it, in their terms
    - "What properties some 'take 5' has" i.e. what extra or different properties it can have and still count
    - "What 'take 5' is not" e.g. it's not "use a checklist".
- Because "take 5" isn't well-studied in the literature, they identity several possible mechanisms by which "take 5" could plausibly be having benefit (sec 2.2). This is "theory specification", in the small. They then go on to test those mechanisms (secs 5.2–5.6).
- It addresses not just the obvious (a) "does this work?" but also (b) "if it doesn't work, why is it so widely used?" (sec 6.1). We can note that (b) may be a valuable question for lots of techniques.
- They make an explicit argument for why they think their results will generalise well (sec 6.2 para "The relevance defence…" and the two following para)

Caveats

- It's not in safety *engineering*

Other notes

- Sec 2.1 gives a history of the technique that they're evaluating. The value of this is that it at least establishes that the technique *has* been widely used. Combined with knowing that it *is still* widely used, this suggests us that it's a response to longstanding needs.

Runner-up/honourable mention

M. Sujan, 'An organisation without a memory: A qualitative study of hospital staff perceptions on reporting and organisational learning for patient safety' [14]

Why is that paper good:

- It's in (operational) safety
- It evaluates a well-defined existing thing — the "NRLS" incident reporting system https://report.nrls.nhs.uk/nrlsreporting/)
- It is a good example of *qualitative* evaluation i.e. discovering the qualitative features of a deployed system. Much evaluation work is *quantitative*, or tries to be. Good qualitative evaluation is also important, it is sometimes all that you can do, and there aren't so many papers out there that show how to do it.
- It has a good example of an introduction for such a paper. It opens with an extensive rationale for why that system was created, what it's supposed to achieve, and why you might reasonably expect it to achieve that. (sec 1 paras 1–5). It follows that with a brief account of the reports that raise suspicion over whether the system is working (sec 1 para 6 "While an early…"). It leads explicitly from the above combination of rationale and concerns to the specific study reported in the paper (sec 1 para 7).
- It explains the diversity and thus validity contribution of using multiple sites (here, hospitals and departments within them) and thus human participants (sec 2.1).
- It clearly describes the strategies they adopted to address validity concerns with their method (sec 2.2 final para).
- It gives example interview prompts for this kind of work (table 2).
- It extracts a small set of qualitative themes from the interview results, and presents them next to concrete examples that illustrate them (quotes from the interviews) (table 3)
- It complements its explicit evaluative aim with *descriptive* insights about parallel, informal, "locally owned" processes that complement the system being evaluated. (sec 3.2). I.e. in Havinga et al's terms it is partly a "soft normative" study [15]. One aspect of this is that it shows that the study team understands how real organisations work in practice — specifically that they never run on formal processes alone, instead relying heavily on informal, locally-created processes. This should help to maintain their credibility in the eyes of an organisationally-astute audience (a category which of course includes most experienced practitioners as they will have lived this knowledge).
- Notes consistency with existing third-party theories, albeit only briefly (sec 4 para "The literature…").

Caveats:

- Not safety engineering

Category — Literature review of methods, processes, tools and the evidence for their quality

This category is for literature review of *validation* and *evaluation* work — evidence about whether some method, process, tool, or such (some *product*) achieves its stated aims. It's analogous to the use of meta-analysis in medicine, although in practice it can't be as straightforwardly quantitative because there's so much less clarity about how validation and evaluation in safety should be performed.

In other words, this category is for studies that review, summarise, analyse, and synthesize the studies from the previous three categories.

Current favourite exemplar

J. Vilela, J. Castro, L. E. G. Martins, and T. Gorschek, 'Integration between requirements engineering and safety analysis: A systematic literature review' [16]

Why is that paper good:

- It's about a safety engineering topic.
- It's a systematic literature review with a highly-explicit method, including some very useful diagrams for communicating said method (fig 1 and fig 2).
- It gives some useful numbers about the realism of the claims made in the work surveyed
    - E.g. only 23% of papers had a compelling empirical evaluation, and 16% had no evaluation at all (sec 4.2 para "Despite…" and next one)
    - E.g. 54% of papers were overall zero "relevance" — "they are examples of application of a proposal done by either students or researchers in academia in toy examples." (sec 4.3 para "31 studies…")
- It builds well on prior work e.g. rather than creating all its own scales for rating the qualities of papers surveyed (always a strong temptation) it reuses the rigour and relevance scales from Ivarsson and Gorschek [17] (sec 4.3 "we performed").

Runner-up/honourable mention 1

L. E. G. Martins and T. Gorschek, 'Requirements engineering for safety-critical systems: A systematic literature review' [16]

Why is that paper good:

- It's about a safety engineering topic
- It provides a template for evaluating research in engineering method (e.g. by giving specific questions to ask and scales/comparators to use, some of which Vilela reuses in the paper above)
- Gives some useful numbers about the realism of the claims made in the work surveyed
    - E.g. fully half of the studies talked about an approach that was never discussed in any of the other studies (suggesting that it was dead on arrival, or never practically credible in the first place) (sec 4.1.2 first three paras)
    - E.g. 95% of studies had only "weak" evidence about the actual usability of the techniques discussed (sec 4.3.1 para "On the other hand…")

Caveats:

- Their "rigor" scale, confusingly, doesn't measure any concept of rigour directly. Rather, it measures how clearly the paper communicates the level of rigour i.e. how adequately it communicates what exactly they did, methodologically.

Runner-up/honourable mention 2

R. Ashmore, R. Calinescu, and C. Paterson, 'Assuring the Machine Learning Lifecycle: Desiderata, Methods, and Challenges' [18]

Why is that paper good:

- It's talking about safety assurance of a technology
- It opens with evidence of machine learning uptake (sec 1 para 1) followed immediately by a summary of problems with it for safety-critical uses. Note how the latter paragraph shows they are very aware of their scope of concern versus that of most authors in the ML space
- They use an explicit structure for the field, for "what kinds of things do we need in order to have success in assurance for safety-critical ML"? (sec 1 para 3, all of sec 2). This provides a way for readers to interpret what they are reading, reducing the cognitive load of making sense of the paper. It also helps readers with completeness checks of what has been achieved, and thus with identifying gaps.
- They include (and include *early on*) a survey of related surveys (sec 3). This helps them argue that their survey is worthwhile (i.e. not just a duplicate of prior work). There are, after all, a lot of surveys out there.
- Each of their top-level review sections (4–7) has the same subsection structure (4.1, 4.2 etc…). This makes it clear that they are applying the same review approach to each top-level topic. Contrast the same thing written with heterogeneous subsection structure — the reader would be lead to think "they've approached each topic in a different way", even though they hadn't.

- They present explicit desiderata for each lifecycle stage —"what properties we need to assure at this stage" (secs 4.3, 5.3, etc). Note that these are *not* criteria to be applied to each paper — they've made an explicit choice to focus on what we need overall, not what we need from any particular paper. Contrast Vilela et al, elsewhere in this category
- At the end of each section, they:
    - Summarise the surveyed methods, processes and tools under clear and explicit but "generalised" names (sec 4 table 1, sec 5 table 3, …). They name things that don't give themselves a name, group broadly-equivalent things together, and strip away noise like author names or cute acronyms. This is an important knowledge-structuring activity — it makes the whole landscape of methods, processes, and tools more intelligible.
    - List the identified gaps ("open challenges") as pithy one-line titles (sec 4 table 2, sec 5 table 4, …).
- They review 184 papers — a very large number.

Caveats:

- The paper doesn't suggest that they used a systematic review approach in any sense. If they did, they don't document it here.
- Writing this kind of paper (with its strong structuring approach) is only viable for authors already deeply knowledgeable about an area
- The validity of the paper will depend strongly on the adequacy of the structure they used. In particular, work that somehow falls outside their structure would be easy for them to miss. Not that this is not really a solvable problem — structures provide insight, tractability, a way to assess completeness within their scope… you can't just do without them and thereby be entirely better off. E.g. a more discursive study with the same authors and budget might have found some sources that they missed, but in turn would likely have missed some that they found.

Conclusions

Perhaps the most important conclusion we can draw from the above is that there are many kinds of studies that researchers can useful carry out (and thus many kinds of papers that they can write). These diverse studies differ in their goals, differ in terms of the methods that are appropriate, and differ in terms of what criteria we should use to evaluate them. If system safety is to be a fully-functional research field, we need researchers to do all these kinds of studies.

It does not follow that every individual or team needs to work on every kind of study. But it does follow that researchers need to do the kinds of studies that need doing next for their topic. If there are no studies of how safety practice actually works with regard to problem X or phenomena Y, researchers should probably not be proposing method, processes, or tools that seek to solve problem X or manage phenomena Y. Conversely, if there are countless survey, observation, and theory-specification papers on problematic phenomenon Z, *and they seem to be good*, it is probably high time that someone did propose a method, process, or tool to help.[6]

Many of these types of study are difficult to do. Many observation and evaluation studies, in particular, can be expensive, require both machine-technical and social-science skills, and need privileged access to real safety practice which industry is often loathe to provide. But if we are to truly understand safety, and to be effective in advancing safety practice, then as a community we need to do them.

[6] The astute reader may observe that this latter situation is rare in system safety.

Acknowledgements

Thanks to John McDermid, Ibrahim Habli, Richard Hawkins, Matt Osbourne, Colin Paterson, and Drew Rae for their comments, criticisms, and suggestions on this document.

References

Exemplars are presented in the text with full reference information, so are not repeated here, except where cited directly.

[1] D. J. Provan, A. J. Rae, and S. W. A. Dekker, 'An ethnography of the safety professional's dilemma: Safety work or the safety of work?', *Saf. Sci.*, vol. 117, pp. 276–289, Aug. 2019, doi: 10.1016/j.ssci.2019.04.024.

[2] A. H. van de Ven, *Engaged Scholarship: A Guide for Organizational and Social Research*. Oxford University Press, 2007.

[3] J. Rooksby, M. Rouncefield, and I. Sommerville, 'Testing in the Wild: The Social and Organisational Dimensions of Real World Practice', *Comput. Support. Coop. Work CSCW*, vol. 18, no. 5–6, p. 559, Dec. 2009, doi: 10.1007/s10606-009-9098-7.

[4] A. Rae, D. Provan, H. Aboelssaad, and R. Alexander, 'A manifesto for Reality-based Safety Science', *Saf. Sci.*, vol. 126, p. 104654, Jun. 2020, doi: 10.1016/j.ssci.2020.104654.

[5] A. Rae and D. Provan, 'Safety work versus the safety of work', *Saf. Sci.*, vol. 111, pp. 119–127, Jan. 2019, doi: 10.1016/j.ssci.2018.07.001.

[6] D. J. Provan, S. W. A. Dekker, and A. J. Rae, 'Bureaucracy, influence and beliefs: A literature review of the factors shaping the role of a safety professional', *Saf. Sci.*, vol. 98, pp. 98–112, Oct. 2017, doi: 10.1016/j.ssci.2017.06.006.

[7] P. Ralph, 'Toward Methodological Guidelines for Process Theories and Taxonomies in Software Engineering', *IEEE Trans. Softw. Eng.*, vol. 45, no. 7, pp. 712–735, Jul. 2019, doi: 10.1109/TSE.2018.2796554.

[8] I. Habli, R. Alexander, and R. D. Hawkins, 'Safety Cases: An Impending Crisis?', *Safety-Critical Systems Symposium (SSS'21)*, Feb. 10, 2021. https://eprints.whiterose.ac.uk/169183/ (accessed Dec. 15, 2021).

[9] J. D. Shaw, N. Gupta, and J. E. Delery, 'Alternative Conceptualizations of the Relationship Between Voluntary Turnover and Organizational Performance', *Acad. Manage. J.*, vol. 48, no. 1, pp. 50–68, Feb. 2005, doi: 10.5465/amj.2005.15993112.

[10] K. Leavitt, T. R. Mitchell, and J. Peterson, 'Theory Pruning: Strategies to Reduce Our Dense Theoretical Landscape', *Organ. Res. Methods*, vol. 13, no. 4, pp. 644–667, Oct. 2010, doi: 10.1177/1094428109345156.

[11] R. Wieringa, N. Maiden, N. Mead, and C. Rolland, 'Requirements engineering paper classification and evaluation criteria: a proposal and a discussion', *Requir. Eng.*, vol. 11, no. 1, pp. 102–107, Mar. 2006, doi: 10.1007/s00766-005-0021-6.

[12] P. Delgado-Pérez, I. Habli, S. Gregory, R. D. Alexander, J. A. Clark, and I. Medina-Bulo, 'Evaluation of Mutation Testing in a Nuclear Industry Case Study', *IEEE Trans. Reliab.*, vol. 67, no. 4, pp. 1406–1419, Dec. 2018, doi: 10.1109/TR.2018.2864678.

[13] J. Havinga, M. I. Shire, and A. Rae, 'Should We Cut the Cards? Assessing the Influence of "Take 5" Pre-Task Risk Assessments on Safety', *Safety*, vol. 8, no. 2, Art. no. 2, Jun. 2022, doi: 10.3390/safety8020027.

[14] M. Sujan, 'An organisation without a memory: A qualitative study of hospital staff perceptions on reporting and organisational learning for patient safety', *Reliab. Eng. Syst. Saf.*, vol. 144, pp. 45–52, Dec. 2015, doi: 10.1016/j.ress.2015.07.011.

[15] J. Havinga, S. Dekker, and A. Rae, 'Everyday work investigations for safety', *Theor. Issues Ergon. Sci.*, vol. 19, no. 2, pp. 213–228, Mar. 2018, doi: 10.1080/1463922X.2017.1356394.

[16] J. Vilela, J. Castro, L. E. G. Martins, and T. Gorschek, 'Integration between requirements engineering and safety analysis: A systematic literature review', *J. Syst. Softw.*, vol. 125, pp. 68–92, Mar. 2017, doi: 10.1016/j.jss.2016.11.031.

[17] M. Ivarsson and T. Gorschek, ‘A method for evaluating rigor and industrial relevance of technology evaluations’, *Empir. Softw. Eng.*, vol. 16, no. 3, pp. 365–395, Jun. 2011, doi: 10.1007/s10664-010-9146-4.
[18] R. Ashmore, R. Calinescu, and C. Paterson, ‘Assuring the Machine Learning Lifecycle: Desiderata, Methods, and Challenges’, *ArXiv190504223 Cs Stat*, May 2019, Accessed: May 15, 2019. [Online]. Available: http://arxiv.org/abs/1905.04223