Towards Efficacy Hypotheses for Safety Cases

Dr. Mallory S. Graydon

NASA Langley Research Center
Hampton, Virginia, USA

Do safety cases work?

- A *safety case* is an *assurance case* focused on safety
  - Typically defined as an argument supported by evidence etc.
- There's plenty of history of their use, particularly in the UK:
  - UK nuclear* (post-Windscale), offshore oil & gas (post-Piper Alpha), and defense (via Mil Std 00-56)
  - Automotive electrical & electronic systems (via ISO 26262)
- … but …
  - … it's hard to show that something makes rare disasters rarer
  - … there have been accidents (notably the loss of a Nimrod aircraft)
  - … there are no empirical studies of efficacy (even by proxy)

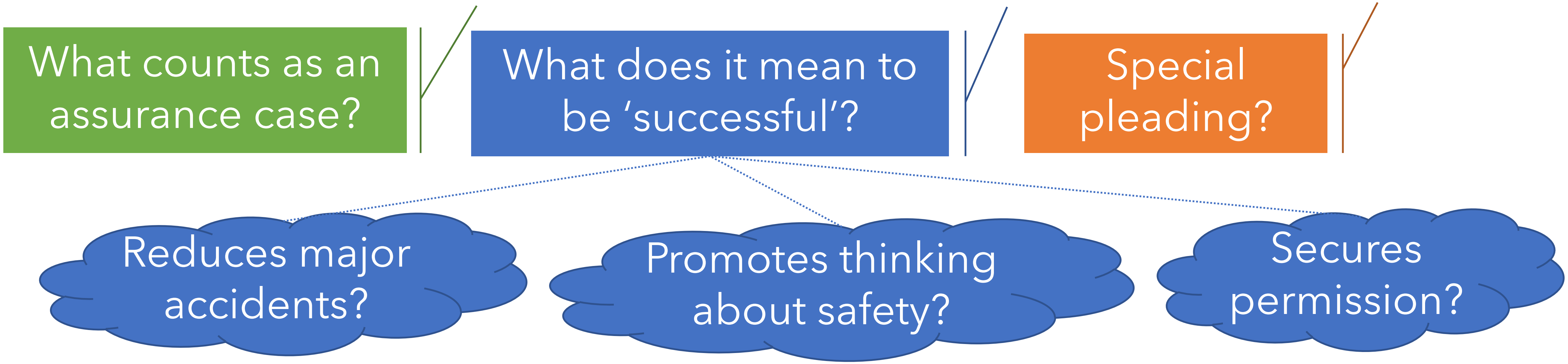
The academic debate

- Safety engineering & regulation is itself safety-critical
- There have been pro/con papers and mailing list discussions
  - But these have not cited solid empirical evidence of harm or benefit
- There are many papers proposing (unevaluated) patterns
- Several writers propose formalizing parts of arguments
  - But it isn't even clear there is a problem formalization could solve
- Others propose computing confidence in conclusions
  - But the evaluation is typically a smoke test of a toy example, and there are examples of proposals producing the wrong results

Towards a trustworthy answer

- Empirical studies require an efficacy question that can be operationalized in terms of observable parameters
- This is trickier than it looks:

Hypothesis: "*Assurance cases are successful where suitable*"



What counts as an assurance case?

What does it mean to be 'successful'?

Special pleading?

Reduces major accidents?

Promotes thinking about safety?

Secures permission?

First attempt to gather hypotheses

- NASA contractors surveyed the literature and practitioners, looking, in part, for efficacy hypotheses
  - They didn't find much that was concrete or testable (let alone empirical studies to assess them)
- I tried to carry this on with a broad literature survey
  - What I found instead were broad themes

Another attempt

- Maybe we can identify hypotheses like we do hazards
- Consider each combination of:
  - The kind of safety case
  - The value of interest
  - The actors who produce or consume arguments
  - Plausible alternatives
  - Relevant costs or side effects
- Pattern for hypotheses:

  <*form of argument*> produces more/better
  <*kind of value*> for <*actor*> than <*alternative*>

Kinds of safety case

- **Kind of safety case**
- Value of interest
- Actor
- Alternative

- *Implicit*: Data/evidence given w/o elaboration
- *Loosely structured*: Case w/ narrative prose
- *Structured, story focused*: Argument (prose or graphical) that accessibly explains the safety concept, key hazards & mitigations, etc.
- *Structured, detailed*: Argument tracing *each* hazard to requirements & mitigations
- *Computable*: Argument in machine-readable notation for automated analysis

Values of interest (1/2)

- Kind of safety case
- **Value of interest**
- Actor
- Alternative

- *Risk/design/evidence insight*: Insight into:
  - Risks (new or residual)
  - How the system gives rise to or mitigates risk
  - The purpose, meaning, or limitation of evidence
- *Telling the story*: Communicating what safety means, how it is achieved, and how we know
- *Confidence*: Argument might establish or justify confidence in safety claims
- *Flexibility*: The ability to use mitigations more appropriate to a given circumstance

Values of interest (2/2)

- Kind of safety case
- **Value of interest**
- Actor
- Alternative

- *Applicability*: Applicability to more systems
- *Stimulation of criticism*: Suggesting critical questions about mitigations, assumptions, designs, implementations, and V&V
- *Revelation of understanding*: Revealing, *e.g.*, what an author thinks matters most
- *Predictability of certification*
- *Pedagogy*: Teaching readers what matters in a given circumstance or what pitfalls to avoid

…

Actors

- Kind of safety case
- Value of interest
- **Actor**
- Alternative

- *Analysts* perform analyses & write arguments
- *Engineers* design, implement, verify, & validate systems & services
- *Managers* have authority to spend
- *Operators* use the system or service
- *Regulators* make certification, licensing, and involuntary recall decisions
- *Impacted third parties* might be harmed despite no direct involvement

Alternatives (1/2)

- Kind of safety case
- Value of interest
- Actor
- **Alternative**

- Traceability matrices
- Plan documents such as DO-178C PSACs
- Other forms of assurance argument
- Hazard or risk analyses such as FHA or FTA
- Standards that prescribe hazard management such as SAE ARP4754A
- Standards that prescribe risk mitigations or engineering techniques

Alternatives (2/2)

- Kind of safety case
- Value of interest
- Actor
- **Alternative**

- Accident analysis methods like FRAM
- Independent conformance or safety assessments
- Safety case report content standards
- A bespoke briefing or document (e.g., engineers walking an audience through key hazards, mitigations, etc.)

Example: Facing the blank page

Value: Risk insight

Form: Implicit

Actors: Analysts, engineers

Alternatives: Plan documents

Hypothesis: The process of planning what evidence to gather, gathering it, and reporting it forces analysts and engineers to more effectively identify the meaning of safety, hazards, risks, and residual risk than generating plan documents

Relevant costs: The costs of writing an implicit safety case versus those of writing plan documents

Example: Framing the story

Value:	Risk insight
Form:	Structured, story focused
Actors:	Analysts
Alternatives:	Loosely structured or implicit argument
Hypothesis:	Thinking through the overarching safety narrative–and identifying what should be highlighted for which readers–causes analysists to more effectively identify the meaning of safety hazards, risks, and residual risk than if the safety argument were loosely structured or implicit
Relevant costs:	The cost of thinking through the narrative and identifying what needs special attention

Example: Automated analysis

Value:	Risk insight
Form:	Computable
Actors:	Analysts
Alternatives:	Story-focused safety argument
Hypothesis:	Writing and analyzing a computable safety argument reveals more insight into risks, their mitigations, and residual risk than writing a story-focused safety argument
Relevant costs:	The costs of writing and analyzing both kinds of arguments

The next steps are studies

Consider this hypothesis:

"Writing and analyzing a formalized safety argument reveals more insight into risks, their mitigations, and residual risk than writing a story-focused safety argument"

How might we test this?

- Recruit representative volunteers and randomly assign them to teams in control and treatment groups
- Control/treatment groups write a story-focused/formalized argument for a specimen system
- Ask all teams to self-report insights derived through the writing process and score these

Conclusion

- Safety analysis, documentation, and certification are crucial; out practice should be evidence-based
- Studies requires testable hypotheses
- We can generate hypotheses by selecting:
  - A kind of safety case
  - A value of interest
  - An actor
  - An alternative

… and thinking through what the combination means