IFIP WG10.4 Research Report, Qeenstown Jan 2017
Shonan workshop number 90, 22-25 November 2016
"Implicit and explicit semantics integration in proof based developments of discrete systems," based on
Marktoberdorf NATO Summer School 2016, Lecture 2,
based on AAA15

The Indefeasibility Criterion for Assurance

John Rushby

Computer Science Laboratory
SRI International
Menlo Park, California, USA

Introduction

- We have some claim about a (software) system
  - E.g., correctness (wrt. specification), safety, etc.
- Want to know that it is true
- But truth is known only to the omniscient
- How about knowledge?
- Plato said it is justified true belief
- Accepted for 2300 years

Justified True Belief

- Russell, 1912:

  Alice sees a clock that reads two o'clock, and believes that the time is two o'clock. It is in fact two o'clock. However, unknown to Alice, the clock she is looking at stopped twelve hours ago.

- Alice has a justified belief
  - But the justification is not very good
  - It happens to be true, but by accident
- In 1963 Gettier published additional examples of poorly justified beliefs that are accidentally true
- The most widely cited modern work in epistemology
  - Over 3,000 citations; he wrote nothing else
- Much work in response attempts to adjust the definition of knowledge by replacing or augmenting justified true belief

Indefesibility Criterion

- The best we can do in assurance is justified belief
- Want a good criterion for justified
  - One that excludes Alice's justification
  - She did not consider possibility of faulty clock
  - Should have sought evidence about this
- Recent work in epistemology proposes indefeasibility
  - For a belief to be justified indefeasibly, we must be so sure that all contingencies have been identified and considered that there is no (or, more realistically, we cannot imagine any) new evidence that would change our belief
- Paraphrasing Barker:
  - If you have an indefeasibly justified belief, then what you don't know can't hurt you!

Assurance Cases

- We use a structured argument to justify the claim, based on evidence about the system
- A structured argument is a tree (usually°) of argument steps, each of which justifies a local claim on the basis of lower level subclaims and/or evidence
  - Need not be a tree if some subclaims or items of evidence support more than one argument step
- There are widely-used graphical notations

  CAE: Claims-Argument-Evidence (Adelard/City U)

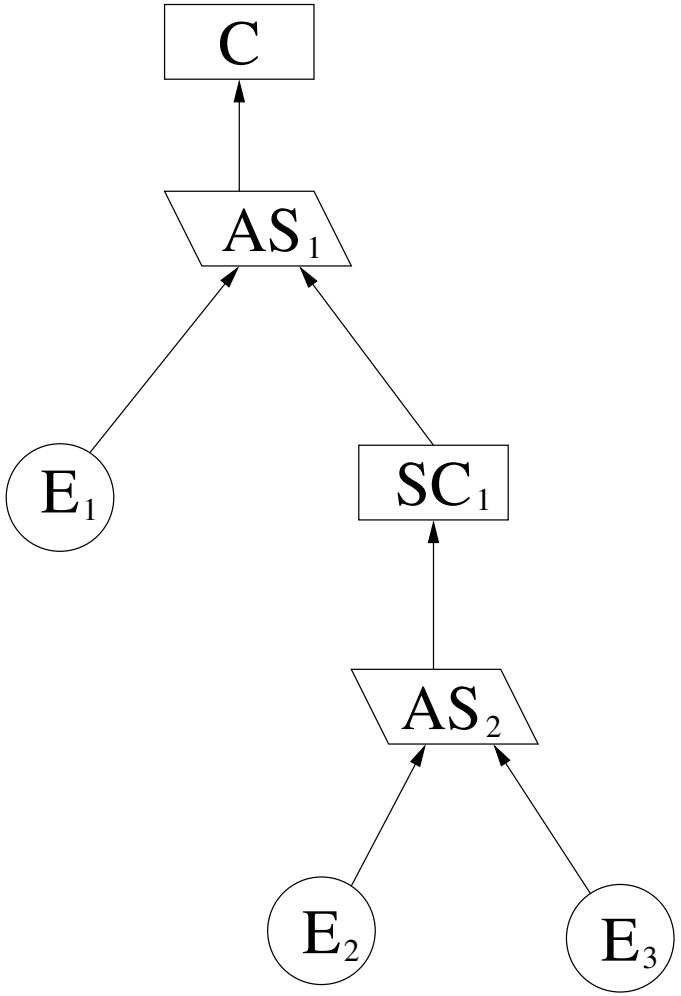
  GSN: Goal Structuring Notation (U York) [nb. Goal=Claim]

  Ashtar is a popular tool in Japan

  Actually, industrial assurance cases are usually free-form

Structured Argument

In a generic notation (GSN shapes, CAE arrows)



C: Claim

AS: Argument Step

SC: Subclaim

E: Evidence

A hierarchical arrangement of argument steps, each of which justifies a claim or subclaim on the basis of further subclaims or evidence

Complications: Inductive vs. Deductive Arguments

- The world is an uncertain place (random faults and events)
- Our knowledge of the world is incomplete, may be flawed
- Same with our knowledge of the system
  (even though we designed it)
- Our methods and tools may be flawed, or rest on unexamined assumptions
- Our reasoning may be flawed also
- So an assurance case cannot expect to prove its claim
- Hence, the overall argument is inductive
  - Evidence & subclaims strongly suggest truth of top claim
  - Unfortunate overloading of the term inductive: many other meanings in science and logic
- Rather than deductive
  - Evidence & subclaims imply or entail the top claim

Complications: Confidence Items SKIP

- If the overall argument is inductive
- Does that mean all its steps may be inductive too?
- Traditionally, yes!
  - Considered unrealistic to be completely certain
  - cf. ceteris paribus hedges in science
- Can add ancillary confidence items to bolster confidence in inductive steps
  - Evidence or subclaims that do not directly contribute to the argument
  - i.e., their falsity would not invalidate the argument
  - But their truth increase our confidence in it
- Eh?

Complications: Graduated Assurance

- An Assurance Case should be "compelling, comprehensible and valid" [00-56]
- Assurance is expensive, so most standards and guidelines allow less assurance effort for elements that pose lesser risks
- E.g. DO-178C
  - 71 objectives for Level A, 33 with independence
  - 69 objectives for Level B, 21 with independence
  - 62 objectives for Level C, 8 with independence
  - 26 objectives for Level D, 5 with independence
- So if Level A is "compelling, comprehensible and valid"
- The lower levels must be less so, or not so
- We need some idea what is lost, and a measure of how much
- Suggests we try to quantify confidence in assurance cases

Quantifying Confidence in Assurance Cases

- Many proposals for quantifying confidence in assurance cases
  - Don't you need a semantics first? Yes, but. . .
  - Some based on Bayesian Belief Networks (BBNs)
  - Others on Dempster-Shafer (or other) Evidential Reasoning
- Graydon and Holloway (NASA) examined 12 such proposals
- By perturbing the original authors' own examples, they showed all the methods can deliver implausible results
- My interpretation:
  - The methods they examined all treat an assurance case as a collection of evidence (that's their implicit semantics)
  - They are blind to the logical content of the argument

Evaluating Confidence in Assurance Cases

- I propose we separate soundness of a case from its strength
  - i.e., start with a semantics for interpreting assurance cases

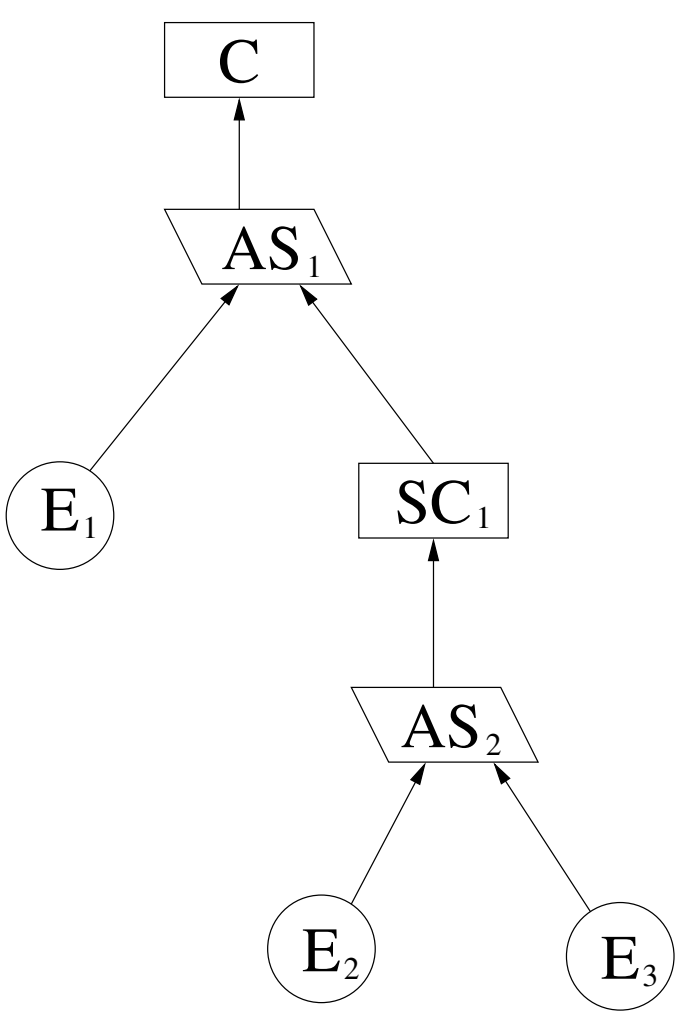
- It's easiest to understand the approach when there are just two kinds of argument steps
  - Reasoning steps: subclaim supported by further subclaims
  - Evidential steps: subclaim supported by evidence

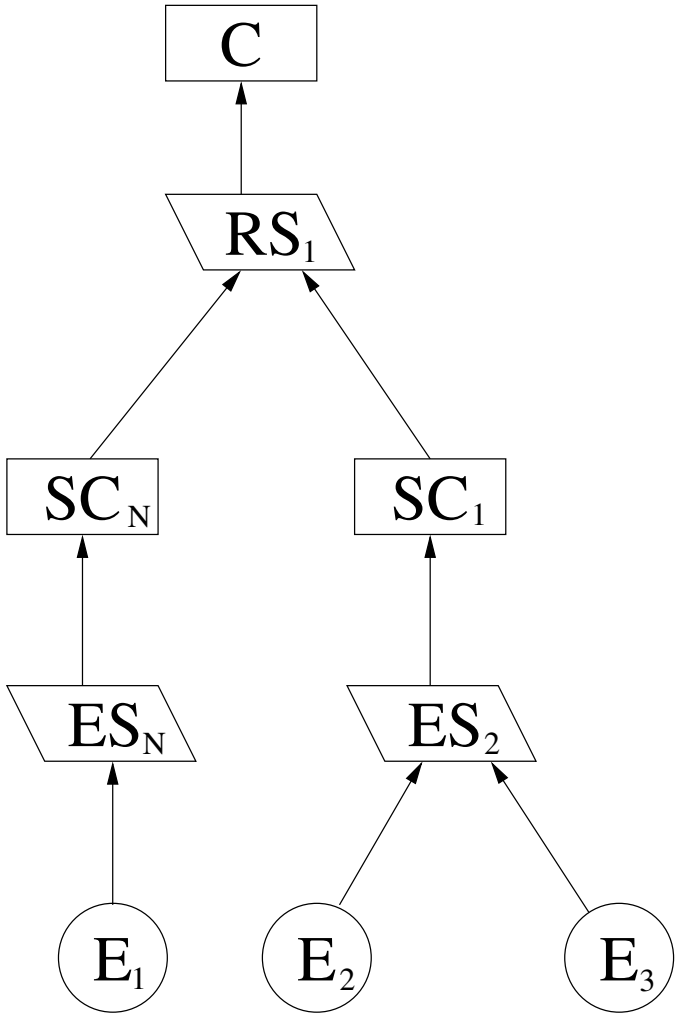
  No steps supported by combination of subclaims and evidence

- Call this a simple form argument
  - Can normalize to this form by adding subclaims

  (in AAA15 paper I outline treatment for general cases)

Normalizing an Argument to Simple Form





RS: reasoning step; **ES**: evidential step

Why Focus on Simple Form?

- The two kinds of argument step are interpreted differently
- Evidential steps
  - These are about epistemology: knowledge of the world
  - Bridge from the real world to the world of our concepts
  - Have to be considered inductive
  - Multiple items of evidence are "weighed" not conjoined
- Reasoning Steps
  - These are about logic/reasoning
  - Conjunction of subclaims leads us to conclude the claim
    - ⋆ Deductively: subclaims imply claim (my preference)
    - ⋆ Inductively: subclaims suggest claim
- Combine these to yield complete arguments
  - Those evidential steps whose weight crosses some threshold of credibility are treated as premises in a classical deductive interpretation of the reasoning steps

How To Group Evidence?

- Have a choice whether related items of evidence are combined in a single evidential step, or used to support separate subclaims that are conjoined in a higher-level reasoning step

- My take:
  - Items of evidence that are not independent are combined in evidential steps
  - Independent (sets of) items support their own subclaims and these are conjoined in reasoning steps

Weighing Evidential Steps SKIP

- We measure and observe what we can
  - e.g., test results
- To infer a subclaim that is not directly observable
  - e.g., correctness
- Different observations provide different views
  - Some more significant than others
  - And not all independent
- "Confidence" items can be observations that vouch for others
  - Or provide independent backup
- Need to "weigh" all these in some way
- Probabilities provide a convenient metric
- And Bayesian methods and BBNs provide tools
  - Example in a few slides time

The Weight of Evidence SKIP

- What measure should we use for the weight of evidence?
- Plausible to suppose that we should accept claim C given collection of evidence E when $P(C \mid E)$ exceeds some threshold
- These are subjective probabilities expressing human judgement
- Experts find $P(C \mid E)$ hard to assess (so do juries)
- And it is influenced by prior $P(C)$, which may reflect ignorance. . . or prejudice
- Instead, factor problem into alternative quantities that are easier to assess and of separate significance
- So look instead at $P(E \mid C)$
  - Related to $P(C \mid E)$ by Bayes' Rule
  - But easier to assess likelihood of observations given a claim about the world than vice versa

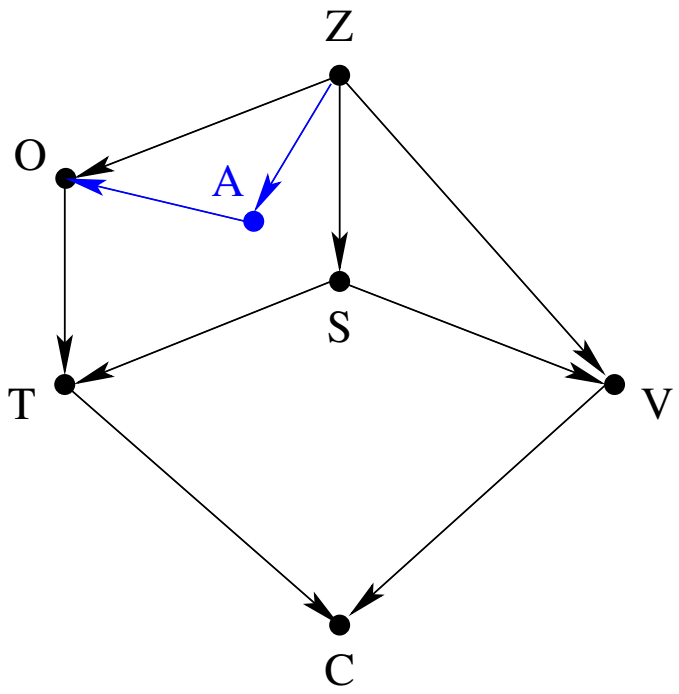
Confirmation Measures SKIP

- We really are interested in the extent to which E supports C rather than its negation $\neg C$
  - Also want $P(E \mid C)$ is not vacuous (e.g., E is a tautology)
- So focus on the ratio or difference of $P(E \mid C)$ and $P(E \mid \neg C)$, . . . or logarithms of these
- These are called confirmation measures
- They weigh C and $\neg C$ "in the balance" provided by E
- Good's measure: $\log \dfrac{P(E \mid C)}{P(E \mid \neg C)}$
- Kemeny and Oppenheim's measure: $\dfrac{P(E \mid C) - P(E \mid \neg C)}{P(E \mid C) + P(E \mid \neg C)}$
- Much discussion on merits of these and other measures
- Suggested that these are what criminal juries should be instructed to assess (Gardner-Medwin)

Application of Confirmation Measures SKIP

- I do not think the specific measures are important
- Nor is quantification necessary for individual arguments
  - Informal evaluation and narrative description can be OK
- Rather, use BBNs and confirmation measures for what-if investigations to develop insight and sharpen judgement
  - Can help guide selection of evidence for evidential steps
  - e.g., refine what objectives DO-178C should require
  - Example (next slides) explores use of "artifact quality" objectives as confidence items in DO-178C
    - ⋆ e.g., "Ensure that each High Level Requirement (HLR) is accurate, unambiguous, and sufficiently detailed, and the requirements do not conflict with each other" [§ 6.3.1.b]

Weighing Evidential Steps With BBNs SKIP



Z: System Specification

O: Test Oracle

S: System's true quality

T: Test results

V: Verification outcome

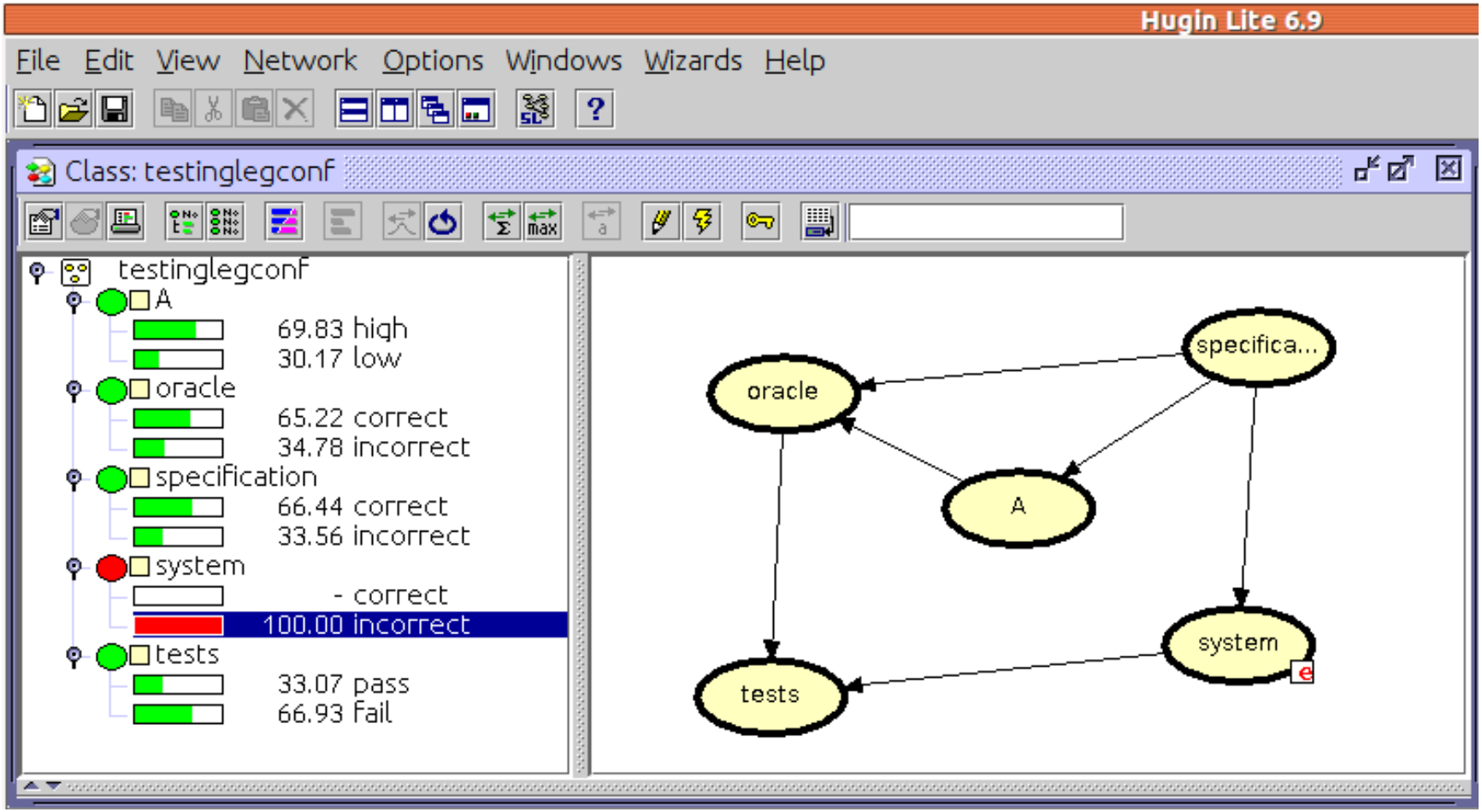
A: Specification "quality"

C: Conclusion

Example joint probability table: successful test outcome

Correct System		Incorrect System	
Correct Oracle	Bad Oracle	Correct Oracle	Bad Oracle
100%	50%	5%	30%

Example Represented in Hugin BBN Tool SKIP



www.hugin.com

Indefeasibility of Evidential Steps

- Two questions about evidential steps
  1. How strongly does the evidence support the subclaim?
     - e.g., if evidence is testing: did we do enough tests?
  2. Is there anything that could defeat the evidence?
     - e.g., did we test the real software?
     - Or how can we trust the oracle?

- The first is weighed informally, or using BBNs as described
  - Accept when above some threshold

- The second requires additional evidence
  - Either as part of the same evidential step,
    or as separate subclaims, conjoined in reasoning steps

Interpretation of Reasoning Steps

- When all evidential steps cross our threshold for credibility, we use them as premises in a classical interpretation of the reasoning steps
  - Deductive: p_1 AND p_2 AND $\cdots$ AND p_n IMPLIES c
  - Inductive: p_1 AND p_2 AND $\cdots$ AND p_n SUGGESTS c
- I advocate the deductive interpretation
  - Because it corresponds to the indefeasibility criterion
  - But there are also "local" reasons

Local Reasons for Preferring Deductive Reasoning Steps

1. There is no agreed interpretation for inductive reasoning
   - Many proposals: Dempster-Shafer, fuzzy logic, probability logic, etc.
   - But none universally accepted
   - And they flatten the argument (recall earlier slide)

2. Inductive reasoning is not modular
   - Must believe either the gap is insignificant (so deductive)
   - Or taken care of elsewhere (so not modular)

3. There is no way to evaluate the size of the gap in inductive steps (next slide)

The Inductive Gap

- Must surely believe inductive step is nearly deductive and would become so if some missing subclaim or assumption a were added (otherwise surely fallacious)
  - p_1 AND p_2 AND $\cdots$ AND p_n SUGGESTS c
  - a AND p'_1 AND p'_2 AND $\cdots$ AND p'_n IMPLIES c
- If we knew anything at all about a it would be irresponsible not to add it to the argument
- Since we did not do so, we must be ignorant of a
- It follows that we cannot estimate the doubt in inductive argument steps
- Hence should strive for deductive reasoning steps

But Aren't Deductive Reasoning Steps Unrealistic?

- Standard inductive example is a step concerning hazards

> Hazard_1 eliminated AND . . . AND Hazard_n eliminated
>
> SUGGESTS system safe

- How can we be sure there are no other hazards?
- Add this as an assumption (logically, another subclaim)
  - $A \supset (B \supset C) \equiv (A \wedge B) \supset C$

> $\text{Hazard}_1, \ldots, \text{Hazard}_n$ are the only hazards
>
> AND Hazard_1 eliminated AND . . . AND Hazard_n eliminated
>
> IMPLIES system safe

- Documentation of the hazard analysis performed provides the evidential support for this subclaim
- In general, deductive doubts give rise to assumptions and we must seek evidence (or subarguments) to support them
  - Or find a better argument

From Interpretation to Evaluation

- Those evidential steps whose weight crosses some threshold of credibility are treated as premises in a classical deductive interpretation of the reasoning steps
- That tells what an assurance case argument means but how do we evaluate whether it is any good?
- Concern is confirmation bias (cf. Nimrod inquiry)
- Must be subjected to serious dialectical challenge
- Can be organized as a search for defeaters
  - Reasons the argument might be wrong
  - Cf. hazards to a system

  And construction of rebuttal for each (i.e., defeat the defeater)

Evaluating an Assurance Case

- Although the final case is indefeasible/deductive
- During development and challenge there will be gaps and doubts, indicated by potential defeaters
- Methods of defeasible reasoning and argumentation allow evaluation of these
- Takai and Kido build on these ideas to extend the Astah GSN assurance case toolset with support for dialectical reasoning

Argument Strength

- An assurance case is valid if its reasoning steps are judged to be deductively valid, and survive dialectical challenge
- A valid case is sound if in addition its evidential steps cross the threshold for credibility, and survive their own challenges
  - All inductive doubts located here
- Then want some measure of the strength of a sound argument
- Needed for overall estimates of fault freeness or failure rate
- Crudely, just accumulate confidence on evidential steps
- Could use an ordinal scale (low, medium, high, etc.)
- Or probabilities calculated by BBNs
  - Can sum them (Adams' Uncertainty Accumulation)
  - Or multiply (independence assumption)
- Note that it's a weakest link calculation
- Beware of gaming

  (e.g., combining subclaims to maximize strength measure)

Graduated Assurance

- Graduated assurance retains soundness, reduces strength
- One approach to weakening an argument for lower levels is to reduce the threshold on evidential steps
- But others actually change the argument
  - E.g., Level D of DO-1788C removes the Low Level Requirements (LLR) and all attendant steps
- Reason for LLR is not just more evidence, but the credibility of the overall argument strategy
  - More credible to go from HLR to EOC via LLR
  - Than in a single leap, unless machanized
- So there's more to it than just accumulated evidential strength
- Topic for future work
  - Likely related to ability to withstand defeaters
  - Would welcome input from philosophy
  - There's a whole field called argumentation

Summary

- Interpretation is a combination of probability and logic
- (Possibly informal) probabilities for evidential steps
- Logic for reasoning steps
- Case is sound **if** evidential steps cross some threshold **and** reasoning steps are deductively valid
  - All inductive doubt is located in the evidential steps
  - Inductive reasoning steps are too low a bar
- Graduated Assurance may weaken evidential support
  - Overall strength of a sound case is then determined by weakest evidential step
  - Can formalize this in probability logic, but I think the real appeal has to be to intuition and consensus. . .
- Deeper notion of strength needed for other forms of graduated assurance: defeaters and argumentation frameworks may be the way to go here

Caution

- My personal opinion is that bespoke assurance cases are likely to be unreliable
  - Insufficient dialectical challenge
- So best approach may be to reformulate future standards and guidelines as assurance cases
  - I think that will make them better
  - And provide a basis for customization
- Alternative: build assurance cases from accepted patterns (GSN) or blocks (CAE)

Challenges 1

More autonomous airplanes

- Currently, automation gives up when it is not safe
  - e.g., frozen pitot tubes, as in AF 447
- More autonomous plane should offer to fly sraight and level
  - By holding consant pitch and thrust
  - Which is what the pilot is instructed to do
- But cannot certify this as safe–it's not
- So what to do?
  - Relax indefeasibility?
    - ⋆ Touchstone is: no worse than human?
  - Or massively lower the bar on evidence?
    - ⋆ Use crude synthetic sensors

Challenges 2

Truly autonomous things. . . like self-driving cars

- Is it just a more extreme form of the previous case?
- So again relax indefeasibility?
  - And touchstone is: no worse than human?
- Or is it indefeasibly no worse than human?
- Or lower the bar on evidence?
  - Use somewhat unreliable sensors (like vision)?

I'm looking for ideas