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

Assurance Cases and their ~~Arguments~~ Evidence

John Rushby

Computer Science Laboratory
SRI International
Menlo Park, California, USA

Introduction

- Assurance must ensure that serious failures are very rare
- Typically this is done by ensuring the absence of faults
- There is a relationship between confidence in absence of faults (expressed as a subjective probability P_{nf}) and probability of failure. . . see Littlewood & Rushby TSE 2012
- Combined with modest observation of failure-free operation, this can deliver credible assurance for critical systems
- But how do we go about estimating and justifying confidence in absence of faults?
- Formal demonstrations like verification are subject to caveats that themselves need to be investigated and justified
- Overall, we need evidence that everything has been considered and examined
- And a rationale that ties it all together
- These are provided by an assurance case

Assurance Cases

- The key idea in an assurance case is that the rationale that ties things together takes the form of a structured argument
- More specifically, the argument "makes the case" that some claim is satisfied, 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)

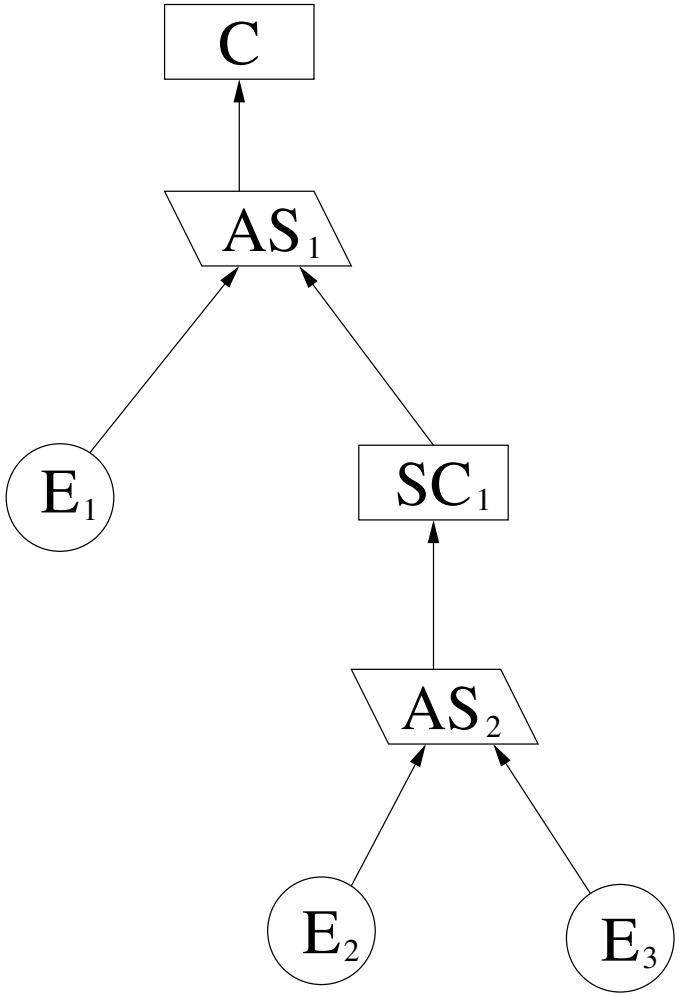
  GSN: Goal Structuring Notation (U York) [nb. Goal=Claim]

  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

Claims for Systems SKIP

- For a system-level assurance case, top claim usually concerns some critical requirement such as safety, security, reliability, etc.
  - Assurance cases generalize safety cases
- Basically, think of everything that could go wrong
  - Those are the hazards

  Design them out, find ways to mitigate them
  - i.e., reduce consequences, frequency

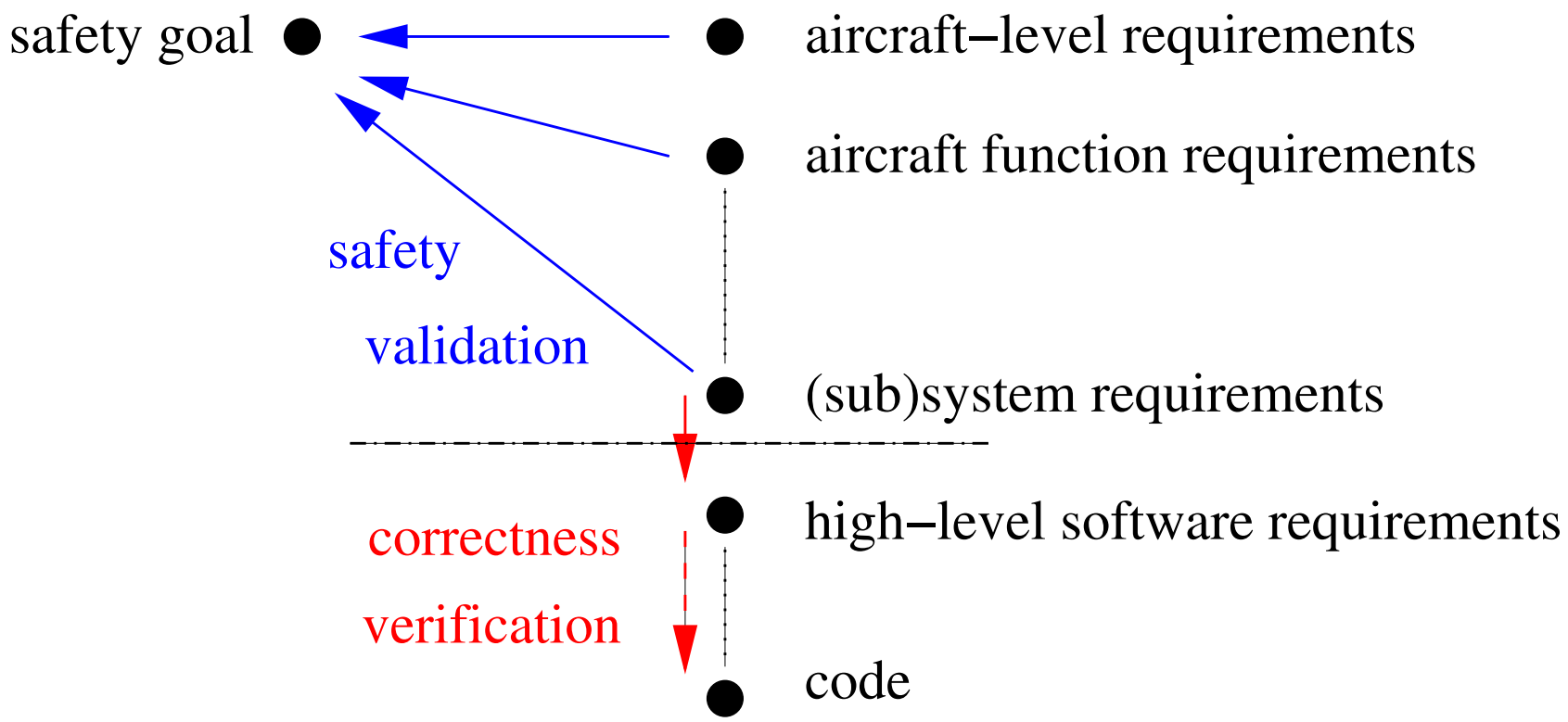
  This may add complexity (a source of hazards)
  - So Iterate
- And then recurse down through subsystems
- Until you get to widgets (small things, no internal structure)
  - Build those correctly
- Provide subarguments and evidence have done all this successfully

Claims for Software SKIP

- In some fields (e.g., aircraft), software is a widget
- So we don't analyze it for safety, we build it correctly
- In more detail. . .
  - Systems development yields functional and safety requirements on a subsystem that will be implemented in software; call these (sub)system requirements
    - ⋆ Often expressed as constraints or goals
  - From these, develop high level software requirements (HLR)
    - ⋆ How to achieve those goals
    - ⋆ Nonstandard terminology: these are really specifications
  - Elaborate through more detailed levels of specifications
  - Until you get to code (or something that generates code)
- Provide subarguments and evidence have done all this successfully
- Top claim is correctness wrt. (sub)system requirements

Aside: Software is a Mighty Big Widget SKIP

The example of aircraft



- As more of the system design goes into software
- Maybe the widget boundary should move
- Safety vs. correctness analysis would move with it

Evidence

- Includes reviews, tests, analyses of all development artifacts (specifications, code, test plans, you name it) and supporting documentation (e.g., how hazard analysis was done)
  - Formal verification is evidence (not part of the argument)
- Prior to assurance cases, assurance was performed by following standards and guidelines
  - These specify just the evidence to be produced
  - With no (explicitly documented) rationale
- Aviation software is still done this way
  - DO-178C enumerates 71 "objectives" that must be satisfied for the most critical software
  - 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]
- Seems to work: no aircraft incidents due to s/w implementation
- But several due to faults in s/w requirements (ARP 4754A)

Guidelines vs. Assurance Cases

- Guidelines are very slow moving
  - Took a decade to evolve DO-178B into DO-178C
- But the environment is changing fast
  - NextGen integrates once separate air and ground systems
  - Unmanned vehicles in same airspace
  - More autonomous systems
  - New methods of software development and assurance
- We don't really know why DO-178B worked
  - So difficult to predict impact of changed environment
- Consider Assurance Cases as a possible way forward
  - Trains, nuclear, infusion pumps, others already done this way
  - Prototype: retrospective reformulation of DO-178C as an assurance case (Michael Holloway)
- But then need a scientific basis for assurance cases

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

- 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

Flattened Arguments

- There's a reason we don't do this
  - An assurance case is not just a pile of evidence
    - ⋆ That's DO-178C, for example
  - It is an argument
  - With a structure based on our reasoning about the system

- So although probabilities make sense for evidence

- The reasoning should be interpreted in logic

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

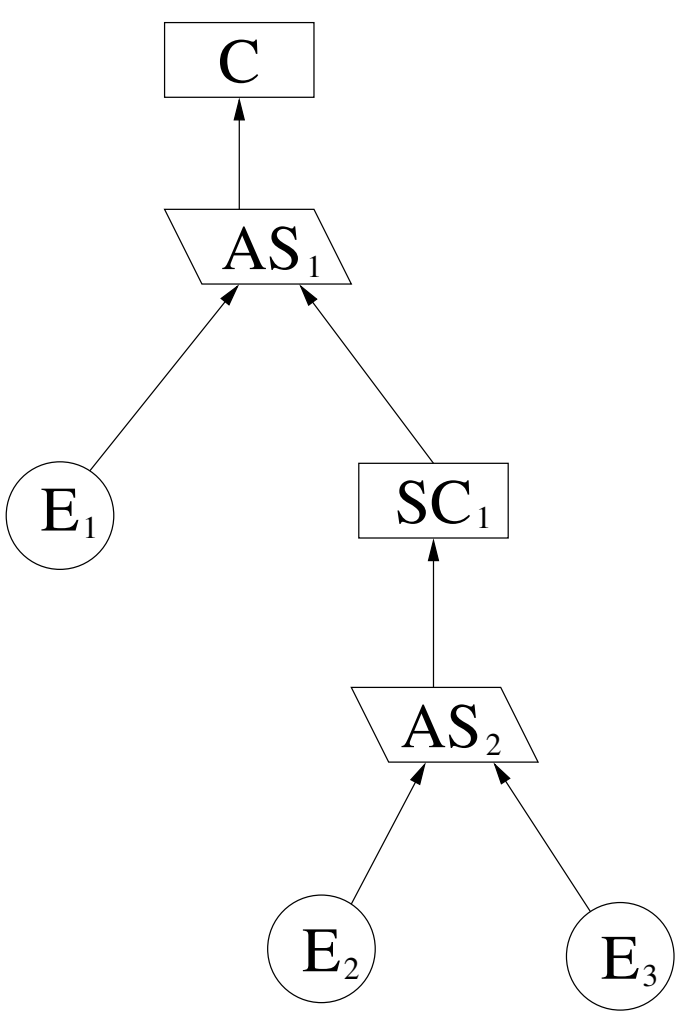
- It's easiest to understand the approach when there are just two kinds of argument steps
  - Reasoning steps: subclaim supported by further subclaims
  - Evidential steps: subclaim supported by evidence

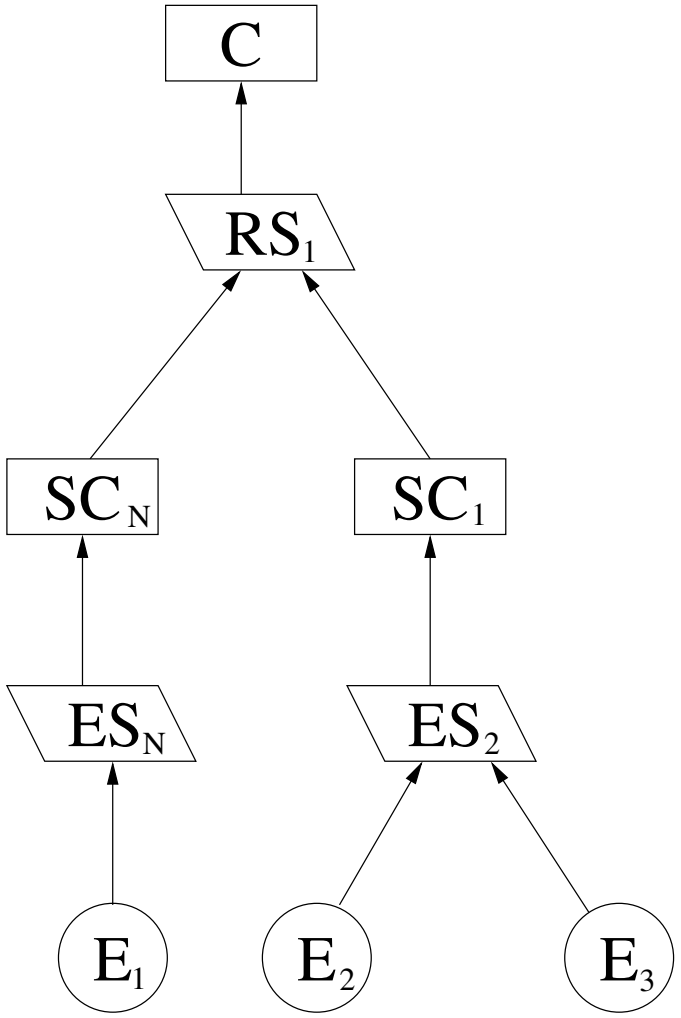
  No steps supported by combination of subclaims and evidence

- 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

Weighing Evidential Steps

- 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

- 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

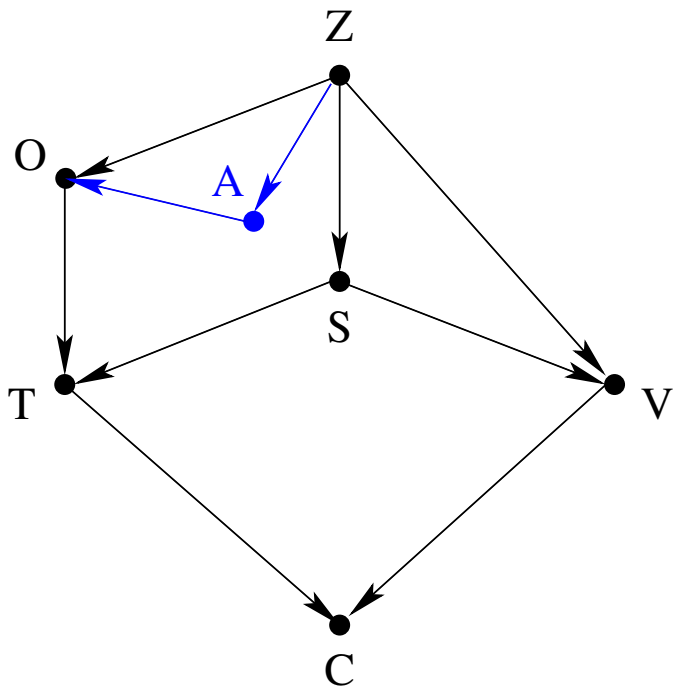
Confirmation Measures

- We really are interested in the extent to which E supports C rather than its negation $\neg C$
  - Also want $P(E \,|\, C)$ is not vacuous (e.g., E is a tautology)
- So focus on the ratio or difference of $P(E \,|\, C)$ and $P(E \,|\, \neg C)$, . . . or logarithms of these
- These are called confirmation measures
- They weigh C and $\neg\, C$ "in the balance" provided by E
- Good's measure: $\log \dfrac{P(E \,|\, C)}{P(E \,|\, \neg\, C)}$
- Kemeny and Oppenheim's measure: $\dfrac{P(E \,|\, C) - P(E \,|\, \neg\, C)}{P(E \,|\, C) + P(E \,|\, \neg\, C)}$
- 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

- 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



Z: System Specification

O: Test Oracle

S: System's true quality

T: Test results

V: Verification outcome

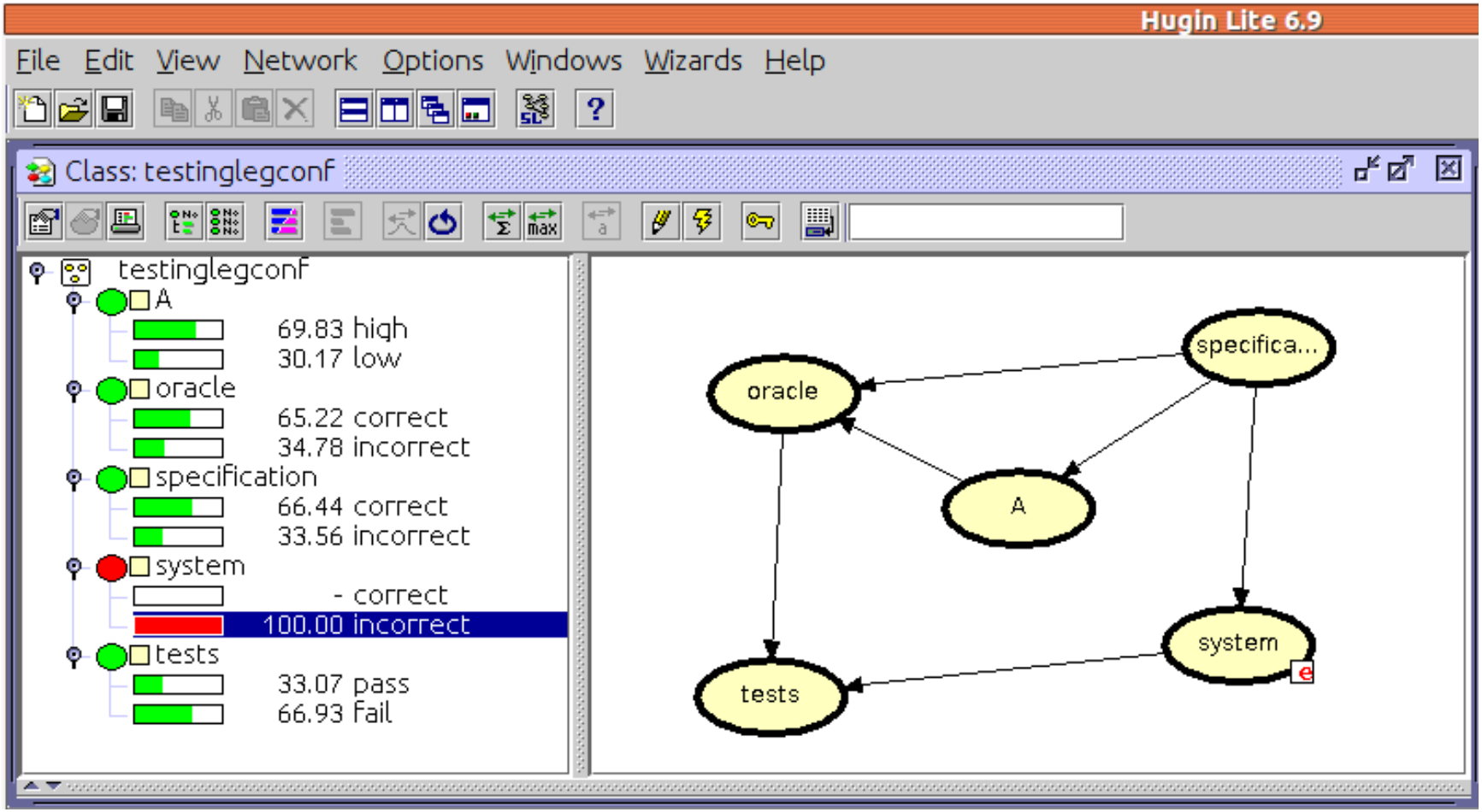
A: Specification "quality"

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



www.hugin.com

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, for three reasons
  - 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)
  - Inductive reasoning is not modular: must believe either the gap is insignificant (so deductive), or taken care of elsewhere (so not modular)
  - 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
- This is related to the indefeasibility criterion for knowledge in modern (post-Gettier) epistemology

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)
  - $A \supset (B \supset C) \equiv (A \wedge B) \supset C$

> Hazard$_1$, . . . , 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

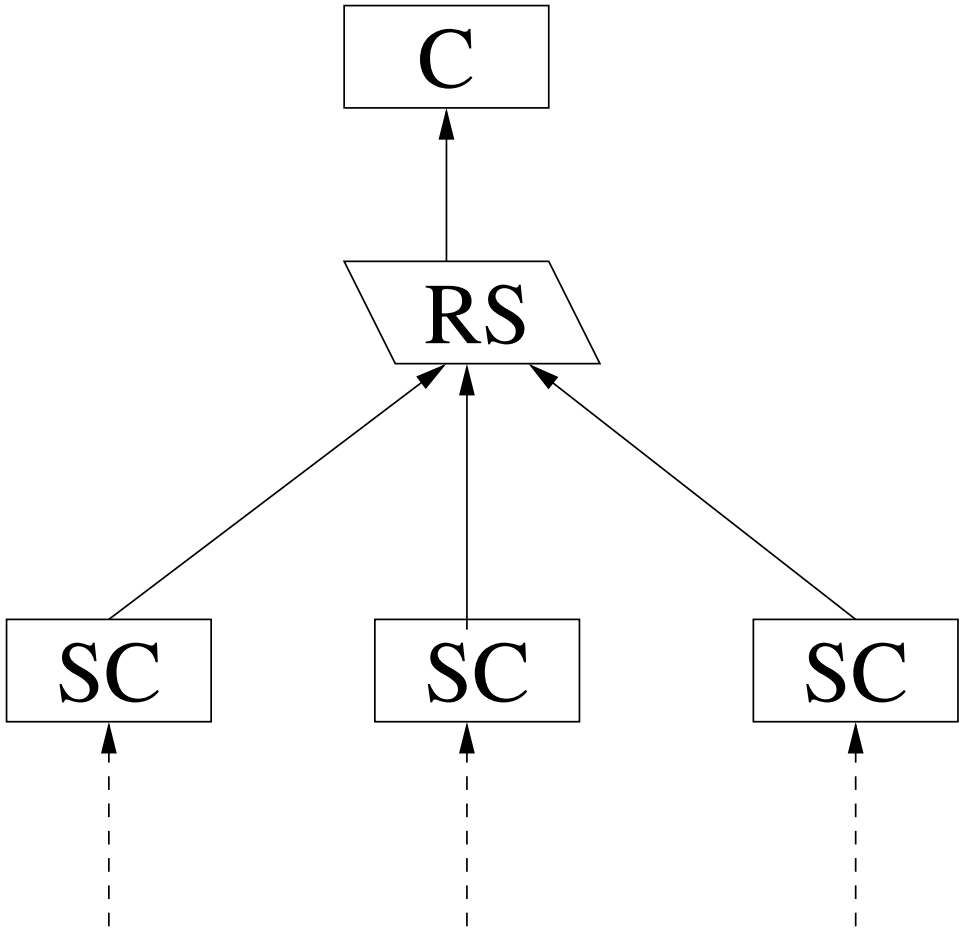
- Those evidential steps whose weight crosses some threshold of credibility are treated as premises in a classical deductive interpretation of the reasoning steps
- That tells what an assurance case argument means but how do we evaluate whether it is any good?
- Concern is confirmation bias (cf. Nimrod 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 a rebuttal for each
- Defeaters and rebuttals need to be recorded as part of the case
  - How?

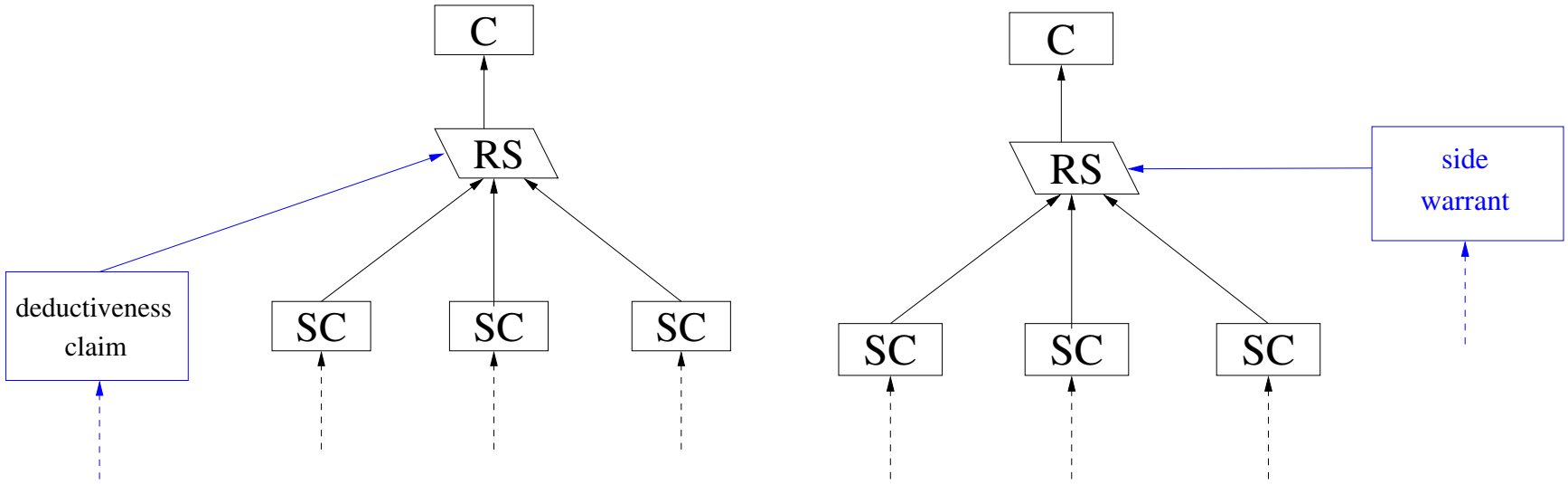
Documenting Evaluation of Reasoning Steps SKIP

- Each argument step has a narrative justification
  - Also called a side warrant
- Could put defeater rebuttals in there
  - But we surely want rebuttals organized as (sub)arguments
  - And these would be unconnected to the main argument
- Alternative is to add X-is-not-a-defeater as a subclaim
-
  - With the rebuttal for defeater X as its subargument
  - Then all subarguments are part of the main argument
- Of course, if X is a successful defeater
  - We will need to add `NOT` X as an assumption
  - Or make larger corrections to the argument
- Iterate until satisfied

Where to Attach the Claim of Deductiveness? SKIP



Two Reasonable Choices SKIP



Similarly for other refuted defeaters

Evaluation of Evidential Steps

- Either quantitatively (with confirmation measures and BBNs) or informally, assess credibility of the combination of evidence provided for each evidential step

- Encourage dialectical challenge with postulated defeaters
  - Consideration of proposed defeaters can be recorded in BBNs or informal narrative
  - Successful defeaters suggest new assumptions, or larger corrections

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
- 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)

References

[1] John Rushby. The interpretation and evaluation of assurance cases. Technical Report SRI-CSL-15-01, Computer Science Laboratory, SRI International, Menlo Park, CA, July 2015.

[2] John Rushby. On the interpretation of assurance case arguments. In *2nd International Workshop on Argument for Agreement and Assurance (AAA 2015)*, Kanagawa, Japan, November 2015. Postproceedings to be published by Springer LNCS.

[3] Bev Littlewood and John Rushby. Reasoning about the reliability of diverse two-channel systems in which one channel is "possibly perfect". *IEEE Transactions on Software Engineering*, 38(5):1178–1194, September/October 2012.

[4] John Rushby. The Ontological Argument in PVS. In Nikolay Shilov, editor, *Fun With Formal Methods*, St Petersburg, Russia, July 2013. Workshop in association with CAV'13