The Indefeasibility Criterion For Effective Assurance Cases

John Rushby

Computer Science Laboratory
SRI International
Menlo Park, California, USA

Introduction

- "A safety case is a structured argument, supported by a body of evidence that provides a compelling, comprehensible and valid case that a system is safe for a given application in a given operating environment" [00-56]
- What does valid (or as I prefer sound) mean here?
- We know a case is a structured argument, so we could fix the notion of argument (e.g., as deduction or Toulmin-style) and import its notion of validity/soundness
- Or look for a larger context in which a suitable form of soundness can be defined that is independent of the style of argument employed
- I will try the latter

The Purpose of an Assurance Case

- Patrick Graydon tells us there are many purposes
- I will fix on one: purpose of a case is to give us justified belief in the properties of interest (not only safety)
- In the limit, we want to know that our system is good
- Epistemology links these concepts (since Plato)
  - Knowledge is justified true belief
- But recently doubts have arisen. . . Gettier (1963)
  - Over 3,000 citations, 3 pages, he wrote nothing else
  - Gives 2 examples of justified true belief that do not correspond to to intuitive sense of knowledge
  - The 3,000 papers give variant examples
  - All have same form: "bad luck" followed by "good luck"
  - Anticipated by Russell (1912)

The Case of the Stopped Clock

- 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 exactly twelve hours ago
- Alice has a justified belief
  - But the justification is not very good
  - And some of her beliefs are false (bad luck)
  - But critical one is true, by accident (good luck)
- Diagnosis: need a criterion for good justification
- Lots of attempts: e.g., "usually reliable process" (Ramsey)
- Indefeasibility:
  - Must be so confident in justification that there is no new information that would make us revise our opinion
  - More realistically: cannot imagine any such information
  - Such information is called a defeater

The Indefeasibility Criterion

- There are various nits and quibbles
  - e.g., new information that is false

  But basic idea is good

- Part company with philosophers: truth requires omniscience
  - So this is a criterion for justification, not knowledge

- Assurance case argument must have no undefeated defeaters

- Validate argument by seeking defeaters

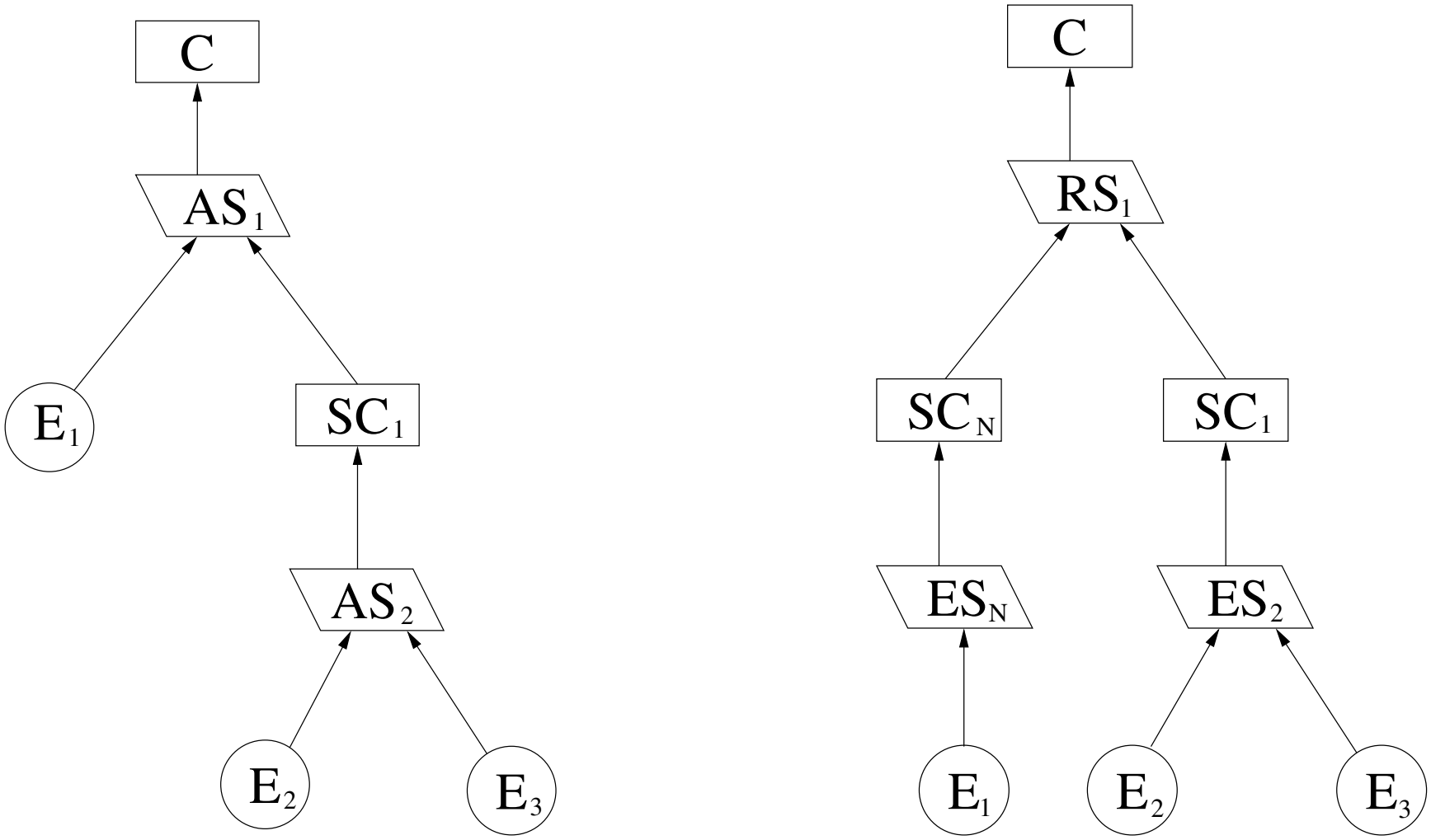
- And defeating them

Application of Indefeasibility To Assurance Case Arguments

- Although we speak of argument
- What we really need is evidence
  - About various aspects of the system
- Purpose of the argument is to be sure we cover all aspects
- Factor argument into
  - Interior reasoning steps
  - And evidential leaf steps
- Not necessary to do this (though can always do so), but it makes the presentation simpler
- The two kinds of step are evaluated differently

Normalizing an Argument to Simple Form

In a generic notation (GSN shapes, CAE arrows)



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

For Example

- The claim $\mathtt{C}$ could be system correctness
  - $\mathtt{E}_2$ could be test results
  - $\mathtt{E}_3$ could then be a description of how the tests were selected and the adequacy of their coverage

  So $\mathtt{SC}_1$ is a claim that the system is adequately tested

- And $\mathtt{E}_1$ might be version management data to confirm it is the deployed software that was tested

- Expect substantial narrative with each step to explain why the evidence or subclaims support the local claim

Evidential Steps

- Accept an evidentially supported claim when the "weight of evidence" crosses some threshold of credibility
- Could be informal judgement
- Or could add discipline of quantification
  - Strength of belief can be represented by numbers that obey the axioms of probability
- Elementary threshold of credibility: $P(C \mid E) > \theta$
- Difficult to estimate, better is $P(E \mid C) > \nu$
- But really want to distinguish between C and $\neg C$
- So use a confirmation measure: e.g., $\log \frac{P(E \mid C)}{P(E \mid \neg C)}$ (I. J. Good)
- Multiple items of evidence that are conditionally independent can each support their own claim (e.g., version management)
- Others support a single claim, dependencies managed by BBNs

Applying the Indefeasibility Criterion

- Need to be sure there is no reason our evidence could be invalidated
- Here, how were test results evaluated?
  - So need evidence for quality of test oracle
- In general, need to be sure there are no defeaters
- But notice evidence does not have to be perfect
  - e.g., expert opinion: but do need evidence expert is good
- And claim does not have to be of perfection
  - e.g., testing supports "adequately tested" not "fault free"

Reasoning Steps

- Evidential steps are the bridge between epistemology and logic
  - When multiple items of evidence support a single claim, they are "added up" (e.g., by BBN analysis)

- Reasoning steps are all logic
  - When multiple subclaims support a claim in a reasoning step, they are conjoined
  - And indefeasibility says the conjunction must imply the parent claim
  - Not "strongly suggest," but imply or entail
  - Because otherwise there is a defeater in the "gap"
  - Hence, deductive rather than inductive interpretation

Is Indefeasibility Realistic?

- Defeasible cases have gaps of unknown size
- Indefeasible cases have no gaps
- But can it be done?
- Many reasoning steps decompose over some structure
- Need to be sure decomposition is complete
- e.g., how do we know we have found all hazards?
- We do hazard analysis
  - Provides evidence we found them all
  - Evidence describes method of hazard analysis employed, diligence of its performance, historical effectiveness, standards applied, and so on
- This transforms potential gap into evidence there is no gap
  - And we can weigh that evidence
- No, it is not a trick

Another Perspective

- The indefeasibility criterion has led us to an approach called "natural language deductivism" (NLD)
- NLD is deductive logic, where premises are
  "reasonable or plausible" rather than certain (as in logic)
- Hence conclusions are reasonable or plausible
  rather than certain
- Our notion of "weight of evidence" formalizes what is meant by "reasonable or plausible"
- NLD is actually a pejorative term due to Trudy Govier (Canadian philosopher) in making the case that this is not what informal arguments are like
- But we are not interested in informal arguments, we seek a suitable notion of justification for assurance cases
- So NLD is just fine for us

Graduated Cases and Quantification

- Not all systems and properties need same assurance
- In NLD all uncertainty is located in weight of evidence
- So can graduate by lowering weight required for evidence
- May allow different evidence
  - e.g., manual review instead of static analysis
- Which may then remove subcases
  - e.g., soundness of static analyzer
- Ultimately need to support a real-world probabilistic claim
- e.g., not expected to occur in lifetime of all airplanes of one type
  - i.e., 10^{-9}
- That is a research challenge

Summary

- Indefeasibility is a natural requirement
- And leads directly to deductive interpretation of reasoning steps
- "Weight of evidence" is an accepted notion
- Together, these lead to a principled form of NLD
- As the criterion for soundness in an assurance case
- Indefeasibility also suggests how to explore and challenge an argument
  - Search for defeaters
- In early stages, and during challenge, argument may not be deductive
  - So merit in automation that tolerates defeasible cases and has ability to record challenges and refutations
  - cf. Astah GSN
- Altogether, indefeasibility ensures effective assurance cases