Computer Science meets Philosophy: Ethics and Epistemology in Assurance and Certification of Autonomous Systems

John Rushby

Computer Science Laboratory
SRI International
Menlo Park, CA

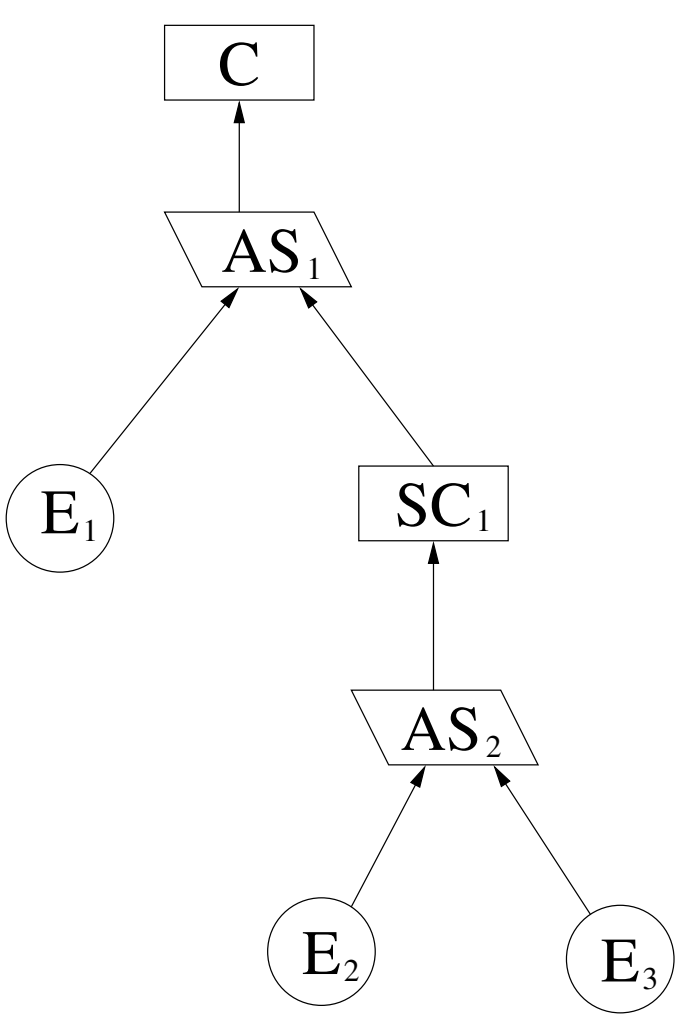
Starting Point: Assurance Cases

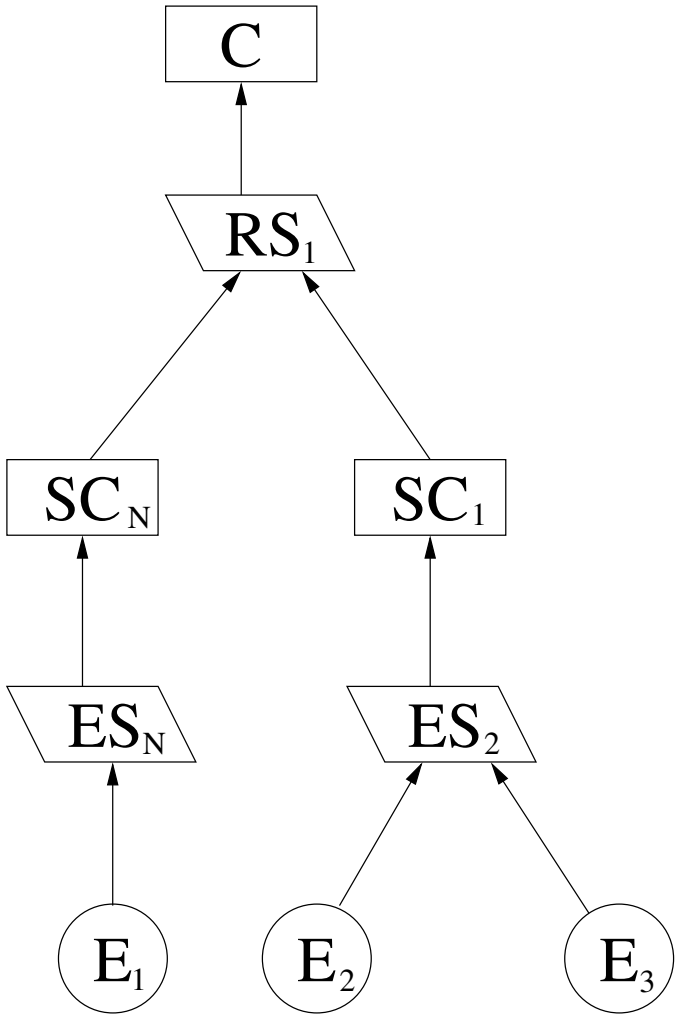
- The state of the art in "classical" (i.e., pre-autonomy) assurance is a safety or (more generally) an assurance case
- Assurance case: a structured argument, based on evidence, that certain claims hold
  - CAE: claims, argument, evidence
- Structured argument: hierarchical arrangement of argument steps
- Argument step: local claim supported by a collection of subclaims or evidence
- Simple form arguments: either subclaims or evidence, not both
  - Reasoning step: claim supported by subclaims
  - Evidential step: claim supported by evidence

  The two kinds of step are interpreted 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 judgment
- Or could add discipline of quantification: subjective probability
  - Strength of belief represented by numbers that obey the axioms of probability
- Elementary threshold of credibility: $P(C \mid E) > \theta$
- Difficult to estimate, better is $P(E \mid C) > \nu$ (use Bayes' rule)
- 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

Philosophy of Confirmation

- Confirmation measures for weight of evidence developed by Turing and Good in WWII codebreaking
- Now part of Bayesian Epistemology
- There are many measures, not ordinally equivalent
- But there is one that is ordinally equivalent to all "good ones"
  - Shogenji's measure: $1 - \frac{\log P(C \mid E)}{\log P(C)}$
  - Homework: find one expressed in terms of $P(E \mid C)$
- Suggestion for use in practice
  - Hone judgment by doing numerical what-if exercises
  - Then use informal judgment in practice

Aside: Conjunction Fallacy

- Evidence: Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations
- Claim 1: Linda is a bank teller
- Claim 2: Linda is a bank teller and active in feminist movement
- Which claim is more likely?
- People overwhelmingly favor Claim 2
- But it must be less probable than Claim 1
- So people are irrational, cannot do simple probabilistic reasoning
- No! They are using confirmation
- People evolved to weigh evidence

Requirement for "Sound" Assurance Cases

- Purpose of a case is to give us justified belief in its top claim
- In the limit, we want to know that the claim is true
- 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 true belief, but is it knowledge?
  - 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 criterion for knowledge:
  - 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

- Assurance case argument must have no undefeated defeaters
- Part company with philosophers: truth requires omniscience
  - So this is a criterion for justification, not knowledge
- But it is also consonant with Peirce's limit theory of truth
  - "truth is that concordance of a . . . statement with the ideal limit towards which endless investigation would tend to bring . . . belief"
- Suggestion for use in practice
  - Validate argument by seeking defeaters
  - And defeating them
  - It's a strong criterion: reasoning steps must imply their claim, not merely suggest it

Claims

- We've looked at evidence and argument, now claims
- Assurance case is typically about absence of serious faults
  - So top claim is "no catastrophic faults"
- But for most classical systems, social need is a bound on the rate/probability of serious failure
  - E.g., for airplanes: "no catastrophic failure condition in entire operational life of all airplanes of one type"
- What's the connection between assurance for absence of faults and low probability of serious failure?
- Failures are caused by faults
- Assurance case gives us confidence in absence of faults
- So high confidence in assurance yields low probability of failure
- Really?

Assurance and Probability of Failure

- Confidence in assurance can be expressed as a subjective probability that the system is fault-free or nonfaulty: p_{nf}
  - Frequentist interpretation possible
  - Research: how to get a credible estimate
- Define $p_{F|f}$ as the probability that it Fails, if faulty
- Then probability $p_{srv}(n)$ of surviving n independent demands (e.g., flight hours) without failure is given by

$$p_{srv}(n) = p_{nf} + (1 - p_{nf}) \times (1 - p_{F|f})^n \quad (1)$$

  A suitably large n (e.g., 10^9 hours) can represent "entire operational life of all airplanes of one type"
- First term gives lower bound for $p_{srv}(n)$, independent of n
- But we could be wrong (i.e., system has faults), so need contribution from second term, despite exponential decay

Assurance and Probability of Failure (ctd.)

- Useful 2nd term could come from prior failure-free operation
- Calculating overall $p_{srv}(n)$ is a problem in Bayesian inference
  - We have assessed a value for p_{nf}
  - Have observed some number r of failure-free demands
  - Want to predict prob. of $n - r$ future failure-free demands
- Need a prior distribution for $p_{F|f}$
  - Difficult to obtain, and difficult to justify for certification
  - However, there is a provably worst-case distribution
- So can make predictions that are guaranteed conservative, given only p_{nf}, r, and n
  - For values of p_{nf} above 0.9
  - The second term in (1) is well above zero
  - Provided $r > \frac{n}{10}$

Assurance and Probability of Failure (ctd. 2)

- So it looks like we need to fly 10^8 hours to certify 10^9
- Maybe not!
- Entering service, we have only a few planes, need confidence for only, say, first six months of operation, so a small n
- Flight tests are enough for this
- Next six months, have more planes, but can base prediction on first six months (or ground the fleet, fix things, like 787)
- And so on
- Theory due to Strigini, Povyakalo, Littlewood, Zhao at City U
- This is how/why airplane certification works
- Remember it, we'll return for autonomous cars

Now On To (Autonomous) Cars

- Let's take it as given that we have an assurance case for the general mechanical and software systems, largely ISO 26262
- And focus on assurance case for the autonomy elements
- Plausible top claim is "no worse than human drivers"
- Large variation worldwide. . . fatalities per billion vehicle kms
  - Sweden 3.5, UK 3.6, Germany 4.9, USA 7.1, Japan 8, SK 15.6
- 35,000 fatalities in USA, so 5 trillion kms per year
- How much testing to validate a claim of q per billion kms?
  - Recall (1); if $q = 5$, want $n = 200,000,000$ failure-free kms
- If we have no contribution from assurance case
  (i.e., all validation is by just "collecting miles")
  - $100,000$ kms gives us modest confidence in next $100,000$

  Whereas with assurance case, $20,000$ gives strong confidence

Assurance Case for Autonomy

- How to provide an assurance case for autonomous systems?
- Difficult! Typically we have deep learning, other ML elements
- Can use massive, representative training sets
  - But lots of corner cases in 10^{12} kms of exposure
  - Adversarial examples indicate ML doesn't work as we'd like
  - Hard to get beyond collecting miles
- I propose much of the case has to come from system architecture
- Philosophy (consciousness): architecture of the brain
  - Popular theory: Predictive Processing (aka. Predictive Coding)
  - It's Bayesian Estimators all the way down (and up)
  - Each level maintains a model, sends predictions down to lower levels (sense organs at bottom) as Bayesian priors, lower levels calculate posteriors, send up corrections to model
  - That's why more neural pathways go down than up

Predictive Processing for Autonomy

- Consider lane detector of an automated driving system
- Analyzes image frames, finds the lanes, sends them to upper levels
- But lanes don't change much from frame to frame
  - Will do better if we retain state
  - And use history to guide detection
  - Improved accuracy, ride out bad frames
- In Bayesian terms, express prior model as a *pdf*, and for each frame calculate posterior as update to model
- That's pretty standard, like a Kalman filter extended to images
- Now extend to a hierarchy, with explicit knowledge at top
  - Research: integration of NN models with logic

Partitioned Assurance Case for Autonomy

- The PP architecture maintains a (top) model of the world
- System actions driven off that model
- So assurance case can be partitioned
  1. Is the model accurate?
  2. Are the actions good/safe, given the model

1. I suspect model accuracy has to be validated by collecting miles
   - But PP architecture seems more robust than straight sense–interpret–actuate
   - Research: use ensembles (diversity),
     Could some be specialized for safety?
2. Safety of actions, given model: looks like classical verification
   - At least at lower levels
   - Higher levels may require ethical(?) judgment

Classical Assurance for HAD

- SOA seems to be simulations validated against scenarios
- Research on scenario languages, curated datasets etc.
- In one fairly small dataset
  - 26 scenarios cover half of all pedestrian accidents
  - 5,287 needed for the other half
- If goal is about 5 deaths per billion kms,
  you need massive scenario suite
- Huge number of "improbable" scenarios in the fat tail
- How do you find/create them?
- Could you do it by models rather than scenarios?
- Any scenario that is an instance of this model is OK

Computational Ethics

- Upper level decision making may concern degree of assertiveness (aggression?), accident scenarios (who dies?)
- Ethics are the basic rules by which societies maintain order and cohesion
- Wide variation, but "experimental ethics" finds that human moral sense is built on five basic principles that seem universal
  - Care, fairness, loyalty/ingroup, authority/respect, and purity
- These function like the five basic senses of taste
  - Different societies and individuals prefer some, and some combinations, to others
  - e.g., liberals stress fairness, conservatives favor authority

Computational Ethics: Example

- Our car is crossing a junction
- Another car jumps the lights on a cross street
- Choice:
  - Collide the cars (injure occupants)
    - ⋆ Care and ingroup principles say don't do this
  - Or swerve onto sidewalk (injure pedestrians)
    - ⋆ Fairness principle says don't do this
- Dial in your principles, then do utilitarian accounting
- Well, it may be more complicated than that
  - e.g., Doctrine of Double Effect (DDE)

Trolley Problems and DDE

- "Trolley problems" are thought experiments to probe human judgments on ethical dilemmas
- Classic case: runaway street car/trolley heading toward a group of five people
- You are standing by a switch/point, can throw this to redirect trolley to a track where it will hit just one person
  - OK? Most people say Yes
- Variant: you push another person onto track in front of the trolley, bringing it to a halt
  - OK? Most people say No
  - Same as first case by utilitarian accounting
- DDE: OK if harm is an unintended (even if predictable) side effect of another (preferably larger) good
  - Second case violates "unintended" condition
- How about throwing yourself in front of the trolley?

Computational Ethics (Neutrality)

- System will need model of human behavior and society
- Whether learned or programmed, it could be biased
- So repeat calculations with age, gender, race reassigned
  - Require decision is invariant under these changes
  - Computational version of Rawls' "Veil of Ignorance"

Computational Ethics (Variations)

- We assumed rule-based decisions, and utilitarian ethics
- Alternatives are game-theoretic formulations
- And "Virtue Ethics" (Aristotle)
- May also want to adjust behavior by rewards and punishment
  - Here's an example (click to activate)
  - Requires the fiction of free will (philosophy)
  - Does the car inhabit a guilt or a shame society?
    - ⋆ Need not be the same as human society
  - A reputation system might be viable

US and EU Perspectives

- What do I know? I drive a 15-year-old Japanese car
- I was strongly impressed by BBC piece "Why you have (probably) already bought your last car"
  - And underlying paper by RethinkX
- Transportation As A Service (TAAS) using autonomous electric cars reduces transportation costs 10 fold
  - Basically, Uber with self-driving Teslas
  - Each one replaces 10 personal cars, lasts 10 times as long
  - Car companies irrelevant, fleet operators dominate
- Will happen by 2030: utterly transformative
  - Liberates vast resources: land and consumer spending
  - Eliminates pollution, curbs global warming
- First mover advantage: be first big city to sanction TAAS
  - Contingent on real and perceived safety
- Where will it happen first?

Concluding Observations

- Advanced computer systems raise numerous questions that have traditionally been the province of philosophy
  - What does it mean to know something? Epistemology
  - Or to participate in a just and decent society? Ethics
  - Plus, of course, reasoning: logic
  - And, at bottom, metaphysical questions of machine consciousness and free will

- Can learn from 2,500 years of philosophical contemplation

- And contribute a less anthropomorphic perspective