PI Meeting, Mesa AZ; 29, 30 June 2022

Confidence Measures for Assurance Cases in CLARISSA

John Rushby, on behalf of the CLARISSA Team
(Honeywell, Adelard PLC, SRI International, UT Dallas)

Computer Science Laboratory
SRI International
Menlo Park, CA

Confidence Measures

- We have an 80-page report on this topic: `https://arxiv.org/abs/2205.04522`
  - On arXiv search Computer Science abstracts for CLARISSA
- Want to know strength of confidence (i.e., justified belief) in top claim
- We use multiple measures, positive and negative, logical and numeric
  - Employ (approximations to) these throughout development
- Our primary positive measure is logical soundness
  - The weight of each evidence assembly crosses some threshold in support of its claim
  - The conjunction of subclaims to each argument node deductively entails the parent claim
- This is Natural Language Deductivism (NLD) — informal version of formal logic
- It corresponds to the epistemological notion of indefeasibility
  - So confident have considered all relevant facts that there is no new information that would change the decision
- More rigorous than other measures, but conceptually clear
- Authors and evaluators are less bewildered by choice

Weight of Evidence and Deductive Validity

- Weight of evidence is assessed by confirmation measures (from Bayesian Epistemology)
- Keynes: how much does the evidence increase my confidence in the claim

$$\log \frac{P(C \mid E)}{P(C)}$$

  These are subjective probabilities and can be qualitative; log is just there to normalize

- Good: how well does this evidence distinguish between claim and counterclaim (defeater)

$$\log \frac{P(E \mid C)}{P(E \mid \neg C)}$$

  Other measures & choice of posterior $P(C \mid E)$ vs. likelihood $P(E \mid C)$ are discussed in report

- CLARISSA has a widget to visualize these that is/was illustrated in another presentation
- Confirmation measures force careful appraisal of the contribution of evidence
  - And allow checks on consistency of evaluator's judgment about this
- Deductive validity (ensured by side-claims) forces careful appraisal of interior argument steps
  - And elimination or explicit recognition of defeaters due to nondeductive argument steps

Sum of Doubts

- Our secondary positive measure is probabilistic doubt (i.e., 1 - probabilistic confidence)
  - Accumulates bottom up, from evidence to top claim
- For each evidence assembly, assess (maybe qualitatively) subjective doubt (i.e., $1 - P(C \mid E)$)
- For each interior node, doubt for parent claim is sum of doubts for its subclaims
  - Makes sense to use numbers, should follow certain rules, hence subjective probabilities
  - Applied only to arguments already assessed to be sound
  - Then a valid combination of logic and probability under very weak assumptions
- Can adjust up or down for argument nodes supported by particularly strong or weak theories
  - Otherwise doubt at top claim is just sum of doubts over all evidence
- CLARISSA colors nodes as visualization of doubt as is/was illustrated in another presentation
- Probabilistic doubt is not used for overall decisions
  - But to help assessors keep track of weak and strong parts of an argument
  - And to compare arguments for graduated assurance (DALs, SILs, EALs etc.)
  - And to assess residual risks

Residual Risks

- Our primary negative measure is residual risk
  - May have some nondeductive argument steps
  - Questionable assumptions,
  - Unresolved defeaters

  These all pose risk: how likely, and how serious? Consider these throughout development
- Probabilistic doubt is one component in assessment of their likelihood/frequency
- Another is their multiplicity: e.g., weak static analysis might allow many instances of a flaw
- Categorize (i.e., measure) residual risks as

  Significant: must be eliminated or mitigated

  Minor: one such is below threshold for concern but many might exceed it

  Manageable: like minor but can limit number/collective severity

  Negligible: many such collectively remain below threshold of concern
- Only manageable and negligible risks may remain

Defeaters

- Our secondary negative measure is a qualitative assessment of the (number and significance of) defeaters considered and examined

- Again, these are considered and explored throughout development

- CLARISSA case retains record of defeaters considered
  - Anticipates evaluator questions and doubts
  - Avoids rework (rediscovery of previous defeaters)
  - Supports eliminative argumentation
  - Is/was demoed in another presentation

From Confidence to Safety

- Top claim is typically "system is safe" (or secure, or some other property)
- And we have some holistic confidence in that as a result of assurance case assessment
- Recorded in a "sentencing report"
- How do we get from confidence in a property
  (e.g., 95% confident system is safe),
  to a prediction of reliability wrt. that property
  (e.g., no hull loss in entire lifetime of all planes of the type)
- We use Conservative Bayesian Inference (CBI) and Bootstrapping from test and current operational experience to derive a sound conservative estimate of reliability wrt. safety (or other property) from a holistic estimate of confidence in the top claim
- We also support internal probabilistic assessments, where the claims include probabilistic assertions, and are grounded on evidence such as statistically valid random testing
- Reminder, all this is discussed in exquisite detail in our Confidence Report on arXiv