What Does V&V Actually Achieve?

John Rushby
Based on joint work with Bev Littlewood (City University UK)

Computer Science Laboratory
SRI International
Menlo Park CA USA

System Claims vs. Software V&V

- Critical systems are those where failures can have unacceptable consequences
- Cannot eliminate failures with certainty, so required top-level claims are stated quantitatively
  - E.g., no catastrophic failure in the lifetime of all airplanes of one type
- And these lead to probabilistic requirements for software-intensive subsystems
  - E.g., probability of failure in flight less control than 10^{-9} per hour
- But V&V is all about showing the absence of faults
- For stronger claims, we do more V&V
- So how does amount of V&V relate to probability of failure?

Software Reliability

- Software contributes to system failures through faults in its requirements, design, implementation—bugs
- A bug that leads to failure is certain to do so whenever it is encountered in similar circumstances
  - There's nothing probabilistic about it
- Aaah, but the circumstances of the system are a stochastic process
- So there is a probability of encountering the circumstances that activate the bug
- Hence, probabilistic statements about software reliability or failure are perfectly reasonable
- Typically speak of probability of failure on demand (pfd), or failure rate (per hour, say)

Measuring/Predicting Software Reliability

- For pfds down to about 10^{-4}, it is feasible to measure software reliability by statistically valid random testing
- But 10^{-9} would need 114,000 years on test
- So how do we establish that a piece of software is adequately reliable for a system that requires, say, 10^{-6}?
- Most standards for system safety (e.g., IEC 61508, DO178B) require you to show that you did a lot of V&V
  - e.g., 57 V&V "objectives" at DO178B Level C (10^{-5})
- And you have to do more for higher levels
  - 65 objectives at DO178B Level B (10^{-7})
  - 66 objectives at DO178B Level A (10^{-9})
- What's the connection between amount of V&V and degree of software reliability?

Aleatory and Epistemic Uncertainty

- Aleatory or irreducible uncertainty
  - is "uncertainty in the world"
  - e.g., if I have a coin with $P(heads) = p_h$, I cannot predict exactly how many heads will occur in 100 trials because of randomness in the world

  Frequentist interpretation of probability needed here

- Epistemic or reducible uncertainty
  - is "uncertainty about the world"
  - e.g., if I give you the coin, you will not know p_h; you can estimate it, and can try to improve your estimate by doing experiments, learning something about its manufacture, the historical record of similar coins etc.

  Frequentist and subjective interpretations OK here

Aleatory and Epistemic Uncertainty in Models

- In much scientific modeling, the aleatory uncertainty is captured conditionally in a model with parameters
- And the epistemic uncertainty centers upon the values of these parameters
- As in the coin tossing example: p_h is the parameter

Aleatory and Epistemic Uncertainty for Software

- We have some probabilistic property of the software's dynamic behavior
  - There is aleatoric uncertainty due to variability in the circumstances of the software's operation
- We examine the static attributes of the software to form an epistemic estimate of the property
  - More examination refines the estimate
- For what kinds of properties does this work?

Perfect Software

- Property cannot be about individual executions of the software
  - Because the epistemic examination is static (i.e., global)
  - This is the difficulty with reliability
- Must be a global property, like correctness
- But correctness is relative to specifications, which themselves may be flawed
- We want correctness relative to the critical claims
- Call that perfection
- Software that will never experience a failure in operation, no matter how much operational exposure it has

Possibly Perfect Software

- You might not believe a given piece of software is perfect
- But you might concede it has a possibility of being perfect
- And the more V&V it has had, the greater that possibility
- So we can speak of a probability of perfection
- Think of all the software that might have been developed by comparable engineering processes to solve the same design problem as the software at hand
  - And that has had the same degree of V&V
- The probability of perfection is then the probability that any software randomly selected from this class is perfect

Probabilities of Perfection and Failure

- Probability of perfection relates to correctness-based V&V
- And it also relates to reliability:

  By the formula for total probability

$$
\begin{aligned}
& P(\text{s/w fails [on a randomly selected demand]}) \qquad (1)\\
= \; & P(\text{s/w fails} \mid \text{s/w perfect}) \times P(\text{s/w perfect})\\
& + P(\text{s/w fails} \mid \text{s/w imperfect}) \times P(\text{s/w imperfect}).
\end{aligned}
$$

- The first term in this sum is zero, because the software does not fail if it is perfect (other properties won't do)
- Hence, define
  - p_{np} probability the software is imperfect
  - p_{fnp} probability that it fails, if it is imperfect
- Then $P(\text{software fails}) < p_{fnp} \times p_{np}$
- This analysis is aleatoric, with parameters p_{fnp} and p_{np}

Epistemic Estimation

- To apply this result, we need to assess values for p_{fnp} and p_{np}
- These are most likely subjective probabilities
  - i.e., degrees of belief
- Beliefs about p_{fnp} and p_{np} may not be independent
- So will be represented by some joint distribution $F(p_{fnp}, p_{np})$
- Probability of system failure will be given by the Riemann-Stieltjes integral

$$\int_{\substack{0 \le p_{fnp} \le 1 \\ 0 \le p_{np} \le 1}} p_{fnp} \times p_{np} \, dF(p_{fnp}, p_{np}). \tag{2}$$

- If beliefs can be separated F factorizes as $F(p_{fnp}) \times F(p_{np})$
- And (2) becomes $P_{fnp} \times P_{np}$

  Where these are the means of the posterior distributions representing the assessor's beliefs about the two parameters

Crude Epistemic Estimation

- If beliefs cannot be separated, we can make conservative approximations
- Assume software always fails if it is imperfect (i.e., $p_{fnp} = 1$)
- Then, very crudely, and very conservatively,

$$P(\text{software fails}) < P(\text{software imperfect})$$

  Dually, probability of perfection is a lower bound on reliability

- Alternatively, can assume software is imperfect (i.e., $p_{np} = 1$)
  - This is the conventional assumption
  - Estimate of p_{fnp} is then taken as system failure rate
  - Any value $p_{np} < 1$ would improve this

Less Crude Epistemic Estimation

- Littlewood and Povyakalo show that if we have
  - $p_{np} < a$ with doubt A (i.e., confidence $1 - A$)
  - $p_{fnp} < b$ with doubt B

  Then system failure rate is less than $a \times b$ with doubt $A + B$
- e.g., p_{np}, p_{fnp} both 10^{-3} at 95% confidence,
  gives 10^{-6} for system at 90% confidence
- They also show (under independence assumption) that large number of failure-free runs shifts assessment from imperfect but reliable toward perfect
- Also some evidence for perfection can come from other comparable software

Two Channel Systems

- Many safety-critical systems have two (or more) diverse "channels"
  - E.g., nuclear shutdown, flight control
- One operational channel does the business
- A simpler channel provides a backup or monitor
- Cannot simply multiply the pfds of the two channels to get pfd for the system
  - Failures are unlikely to be independent
  - E.g., failure of one channel suggests this is a difficult case, so failure of the other is more likely
  - Infeasible to measure amount of dependence

So, traditionally, difficult to assess the reliabilty delivered

Two Channel Systems and Possible Perfection

- But if the second channel is possibly perfect
  - Its imperfection is conditionally independent of failures in the first channel
- Hence, system pfd is conservatively bounded by product of pfd of first channel and probability of imperfection of the second
- i.e., $P(\text{system fails on randomly selected demand} \leq pfd_A \times pnp_B$
- Epistemically, assuming beliefs can be separated

$$P(\text{system fails on randomly selected demand} \leq P_A \times P_B$$

- Joint work with Bev Littlewood:
  http://www.csl.sri.com/~rushby/abstracts/csl-09-02
  - Who originated the idea of possible perfection

Type 1 and Type 2 Failures

- So far, only considered failures of omission
  - Type 1 failure: both channels fail to respond to a demand
- Must also consider failures of commission
  - Type 2 failure: either channel responds to a nondemand
- Demands are events at a point in time; nondemands are absence of demands over an interval of time
- Unify by considering rates in formalism of Poisson process

Overall Failure Rate

- 1oo2 system Type 1 failure rate $\leq d \times (P_A \times P_B)$ where d is demand rate
- 1oo2 system Type 2 failure rate due to $A \leq F_{A2}$ where F_{A2} is the failure rate of Channel A wrt. Type 2 failures
- 1oo2 system Type 2 failure rate due to $B \leq F_{B2|np} \times P_{B2}$ where P_{B2} is epistemic probability of B being imperfect with respect to Type 2 failures, and $F_{B2|np}$ is its epistemic failure rate wrt. Type 2 failures, assuming it's imperfect
- Risk is the sum of these 3 cases, each multiplied by their cost

Total Risk

- Usually, cost of Type 1 failure is high, so a lot of system focus is in reducing demand rate d and failure rate $P_A \times P_B$
- Costs of Type 2 failures are not reduced by a demand rate, so either costs or failure rates must be small
- Cost of Type 2 failures by A is inherent in any safety system, so presumably F_{A2} is acceptable
- Cost of Type 2 failures by B is a new factor, due to choice of 1oo2 architecture: either its cost must be small or $F_{B2|np} \times P_{B2}$ must be small

Monitored Architectures

- One operational channel does the business
- Simpler monitor channel can shut it down if things look bad
- Analysis is a variant of 1oo2:
  no Type 2 failures for operational channel
- Monitored architecture risk per unit time
  $\leq c_1 \times (M_1 + F_A \times P_{B1}) + c_2 \times (M_2 + F_{B2|np} \times P_{B2})$
  where the Ms are due to mechanism shared between channels
- May provide justification for some of the architectures suggested in ARP 4754
  - e.g., 10^{-9} system made of Level C operational channel and Level A monitor

Aside: Monitors Do Fail

- Fuel emergency on Airbus A340-642, G-VATL, 8 February 2005
  - Type 1 failure
- EFIS Reboot during spin recovery on Airbus A300 (American Airlines Flight 903), 12 May 1997
  - Type 2 failure
- Current proposals are for formally synthesized/verified monitors for properties in the safety case

Formal Verification and the Probability of Perfection

- We want to assess P_{np}
- Context is likely a safety case in which claims about a system are justified by an argument based on evidence about the system and its development
- Suppose part of the evidence is formal verification
- ◦ i.e., what is the probability of perfection of formally verified software?
- Let's consider where formal verification can go wrong

  This is considered in the paper with Bev Littlewood, and in

  `http://www.csl.sri.com/~rushby/abstracts/sefm09`

The Basic Requirements For The Software Are Wrong

- This error is made before any formalization
- It seems to be the dominant source of errors in flight software
- But monitoring and backup software is built to requirements taken directly from the safety case
  - If these are wrong, we have big problems
- In any case, it's not specific to formal verification
- So we'll discount this concern

The Requirements etc. are Formalized Incorrectly

- Could also be the assumptions, or the design
- Formalization may be inconsistent
  - i.e., meaningless
  - Many Z specs are like this

  Can be eliminated using constructive specifications
  - In a tool-supported framework
  - That guarantees conservative extension

  But that's not always appropriate
  - Prefer to state assumptions as axioms
  - Consistency can then be guaranteed by exhibiting a constructive model (interpretation)
  - PVS can do this
- So we can eliminate this concern

The Requirements etc. are Formalized Incorrectly (ctd.)

- Formalization may be consistent, but wrong
- In my experience, this is the dominant source of errors in formal verification
    - There are papers on errors in my specifications
- Formal specifications that have not been subjected to analysis are no more likely to be correct than programs that have never been run
    - In fact, less so: engineers have better intuitions about programs than specifications
- Should challenge formal specifications
    - Prove putative theorems
    - Get counterexamples for deliberately false conjectures
    - Directly execute them on test cases
- Social process operates on widely used theories

The Requirements etc. are Formalized Incorrectly (ctd. 2)

- Even if a theory or specification is formalized incorrectly, it does not necessarily invalidate all theorems that use it
- Only if the verification actually exploits the incorrectness will the validity of the theorem be in doubt
  - Even then, it could still be true, but unproven
- Some verification systems identify all the axioms and definitions on which a formally verified conclusion depends
  - PVS does this

  If these are correct, then logical validity of the verified conclusion follows by soundness of the verification system
  - So can apply special scrutiny to them

The Formal Specification and Verification is Discontinuous or Incomplete

- Discontinuities arise when several analysis tools are applied in the same specification
  - e.g., static analyzer, model checker, timing analyzer

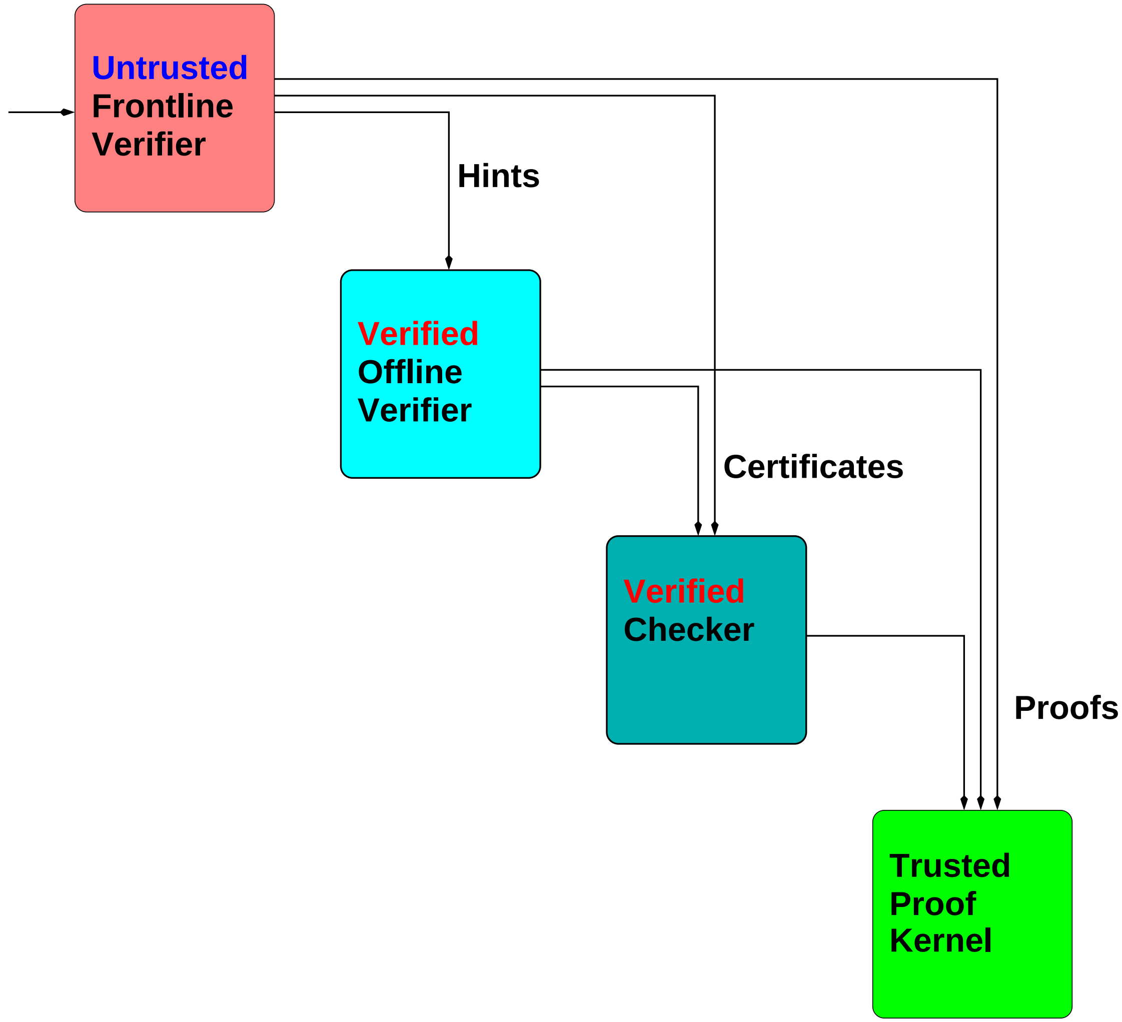
  Concern is that different tools ascribe different semantics
- Increasing issue as specialized tools outstrip monolithic ones
  - Need integrating frameworks such as a tool bus
- Most significant incompleteness is generally the gap between the most detailed model and the real thing
  - Algorithms vs. code, libraries, OS calls

  That's one reason why we still need testing
  - Driven from the formal specification
  - Cf. penetration tests for secure systems: probe the assumptions

Unsoundness In the Verification System

- All verification systems have had soundness bugs
- But none have been exploited to prove a false theorem
- Many efforts to guarantee soundness are costly
  - e.g., reduction to elementary steps, proof objects
- What does soundness matter if you cannot do the proof?
- A better approach is KOT: the Kernel Of Truth (Shankar)
  - A ladder of increasingly powerful verified checkers
  - Untrusted prover leaves a trail, blessed by verified checker
  - More powerful checkers guaranteed by one-time check of its verification by the one below
  - The more powerful the verified checker, the more economical the trail can be (little more than hints)

KOT: A Ladder of Verified Checkers



Shankar and Marc Vaucher have verified a modern SAT solver that is executable (modulo lacunae in the PVS evaluator)

Application

- Suppose we need P_{np} of 10^{-4}
- Bulk of this "budget" should be divided between incorrect formalization and incompleteness of the formal analysis, with small fraction allocated to unsoundness of verification system
- Through sufficiently careful and comprehensive formal challenges, it is plausible an assessor can assign a subjective posterior probability of imperfection on the order of 10^{-4} to the formal statements on which a formal verification depends
- Through testing and other scrutiny, a similar figure can be assigned to the probability of imperfection due to discontinuities and incompleteness in the formal analysis
- By use of a verification system with a trusted or verified kernel, or trusted, verified, or diverse checkers, assessor can assign probability of 10^{-5} or smaller that the theorem prover incorrectly verified the theorems that attest to perfection

Discussion

- These numbers are feasible and plausible
- Formal methods and their tools do not need to be held to (much) higher standards than the systems they assure
- But what are we to do about single channel systems that require 10^{-9}?
  - Type 2 failures of monitors (incorrect activation) may be in this class
  - Topic for investigation and discussion whether such assessments could be considered feasible and credible
  - The earlier single channel analysis holds promise for values approaching this

Conclusion

- Probability of perfection is a radical and valuable idea
- Provides the bridge between correctness-based verification activities and probabilistic claims needed at the system level
- Relieves formal verification, and its tools, of the burden of absolute perfection