ADELARD

CITY UNIVERSITY LONDON

ASSURANCE 2.0: A MANIFESTO

THE DEVELOPMENT AND APPLICATION OF ASSURANCE 2.0

Prof Robin E Bloomfield FREng
Adelard LLP and City, University of London
reb@adelard.com

Joint paper with John Rushby, SRI

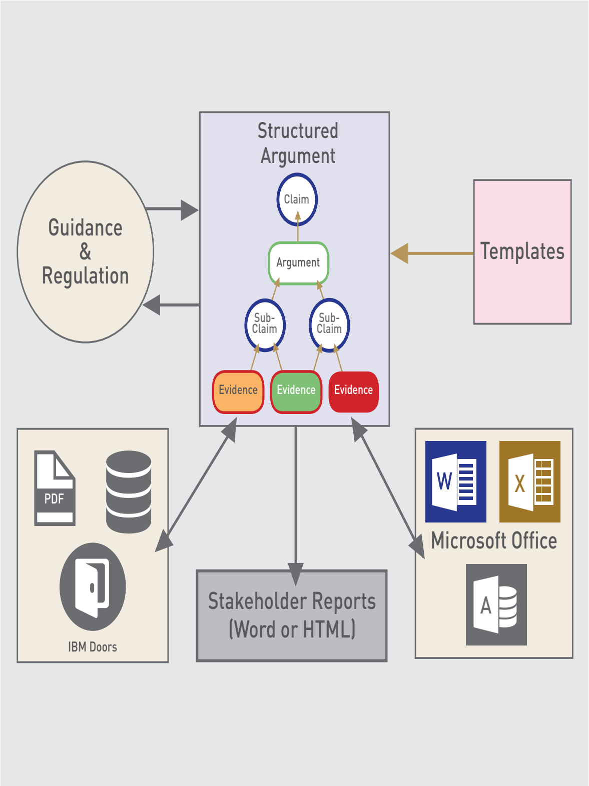
Presentation to SSS'21.
Feb 10th 2021

PT/908/180001/9

ADELARD

- Adelard is a specialized, influential product and services company working on safety, security and resilience
- Wide-ranging experience of assessing computer-based systems and components
- Work across different industrial sectors, including nuclear, transport, defence, financial, medical
  - Policy, methodology, technology
  - Product for managing safety and assurance cases (ASCE)
- Consultants PhD level, international team

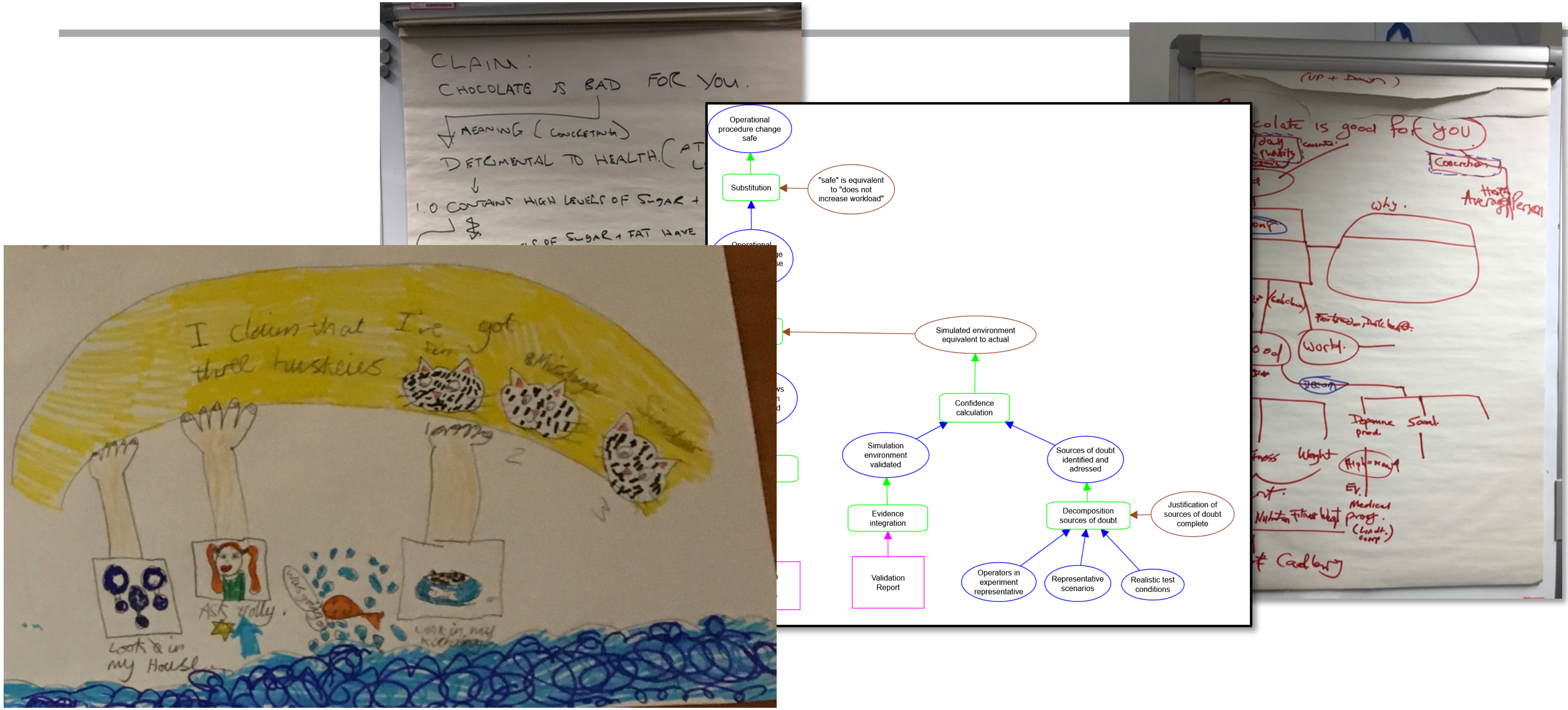
ASCE - in the wider environment



OUTLINE

- Motivation
  - Briefly, why is Assurance 2.0 needed
- Summary of Assurance 2.0
  - Joint work with John Rushby, SRI
- Some application experience
  - Templates and guidance for Autonomous systems
  - Tool support
  - Industry courses
- Conclusions – from manifesto to methodology

WHAT DOES GOOD LOOK LIKE?



Operational procedure change safe

Substitution

"safe" is equivalent to "does not increase workload"

Simulated environment equivalent to actual

Confidence calculation

Simulation environment validated

Sources of doubt identified and adressed

Evidence integration

Decomposition sources of doubt

Justification of sources of doubt complete

Validation Report

Operators in experiment representative

Representative scenarios

Realistic test conditions

DRIVERS FOR CHANGE

- Trustworthy systems expensive and often slow to produce
  - And still have failures
- Assurance is essential – gaining confidence in the system
  - Essential for legal, reputational, market, ethical, commercial reasons
  - Can be slow to produce, slow to change
- Innovation challenges
  - New lifecycles, new technology
  - Higher tempo, varied supply chains. increased threats
- Address existing and emerging requirements for safety and assurance arguments
  - ISO26262, PAS11281, UL4600, EU Pegasus project, Safety First For Automated Driving, UK Regulation for the Fourth Industrial Revolution White Paper

DRIVERS FOR NEW APPROACH

- Challenge from broadening approach to security and engineering justifications
  - The “non safety case” world using the approach
  - Long term study CAE adoption and CAE role in supporting innovation
- Commoditisation of risk assessment, loss of mindset
  - UK NCSC withdrawal of risk assessment guidance IS1 and IS2
  - https://www.ncsc.gov.uk/guidance/critical-appraisal-risk-methods-and-frameworks
- Challenge of
  - autonomous systems and those using AI/ML
  - automated certification
- Evolution of research on argumentation and assurance
- Overall need for
  - understanding, explanation, challenge, and learning

ASSURANCE 2.0

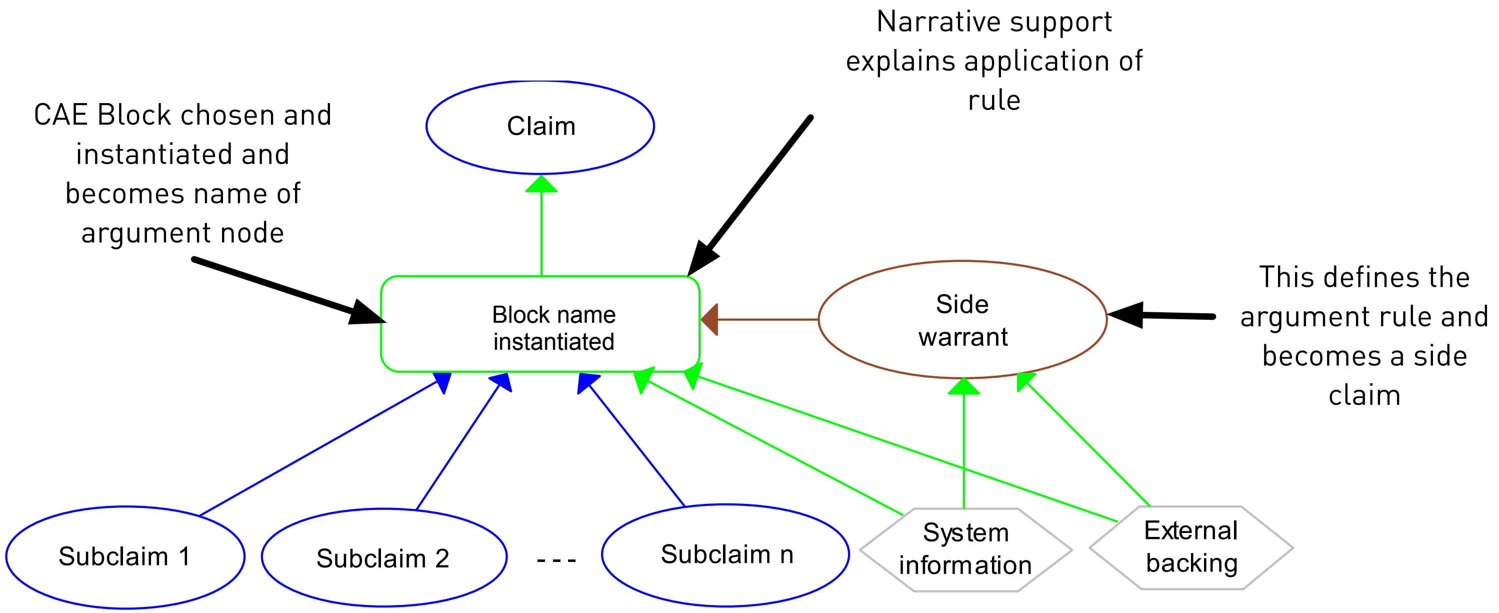
- Our idea is to make assurance an enabler for innovation, not a brake
- Paradoxically, we think we can achieve this by making it more rigorous
  - Keep structure of traditional assurance cases
  - Strengthen focus on evidence and reasoning
  - Bring assurance thinking forward within life-cycle
    - makes it clear what must be done and makes you do it earlier
  - Support assurance with known best practices
    - reduce the bewildering choice of free form cases with "pre-validated" blocks or templates

ASSURANCE 2.0 - MANIFESTO

- Making explicit inference rules and the separation of inductive and deductive reasoning.
  - empirically based CAE Blocks provides a mechanism for separating inductive and deductive aspects of the reasoning. *Natural language deductivism*. (NLD)
- Explicit use of doubts and defeaters
  - both undercutting and rebuttal, that confidence an integral part of the justification
  - indefeasibility criterion
- Focus on evidence integration, addressing both the relevance and provenance of evidence.
  - evidential threshold, in which a claim can be reasoned about deductively might be used when considering the role of automated reasoning
- Confirmation theory to evaluate the strength of evidence and arguments.
- Explicit approach to reduce bias by the use of counter-cases and confirmation theory.
- Recognition of importance of both mindset and methodology

CAE BUILDING BLOCKS - NLD

- Well defined argument fragments, empirically based, but rigorously defined, supporting reasoning both deductive and inductive
- Fragment that support a combined graphical and narrative approach



Narrative support explains application of rule
CAE Block chosen and instantiated and becomes name of argument node
Claim
Block name instantiated
Side warrant
This defines the argument rule and becomes a side claim
Subclaim 1
Subclaim 2
- - -
Subclaim n
System information
External backing

DEDUCTIVE AND INDUCTIVE ARGUMENTS

- For valid deductive arguments the premises *logically entail* the conclusion, where the entailment means that the truth of the premises provides a *guarantee* of the truth of the conclusion
- An inductive logic is a system of evidential support that extends deductive logic to less-than-certain inferences
- In a good inductive argument the premises should provide some *degree of support* for the conclusion, where such support means that the truth of the premises indicates with some *degree of strength* that the conclusion is true.
  - acceptability, relevance and sufficiency

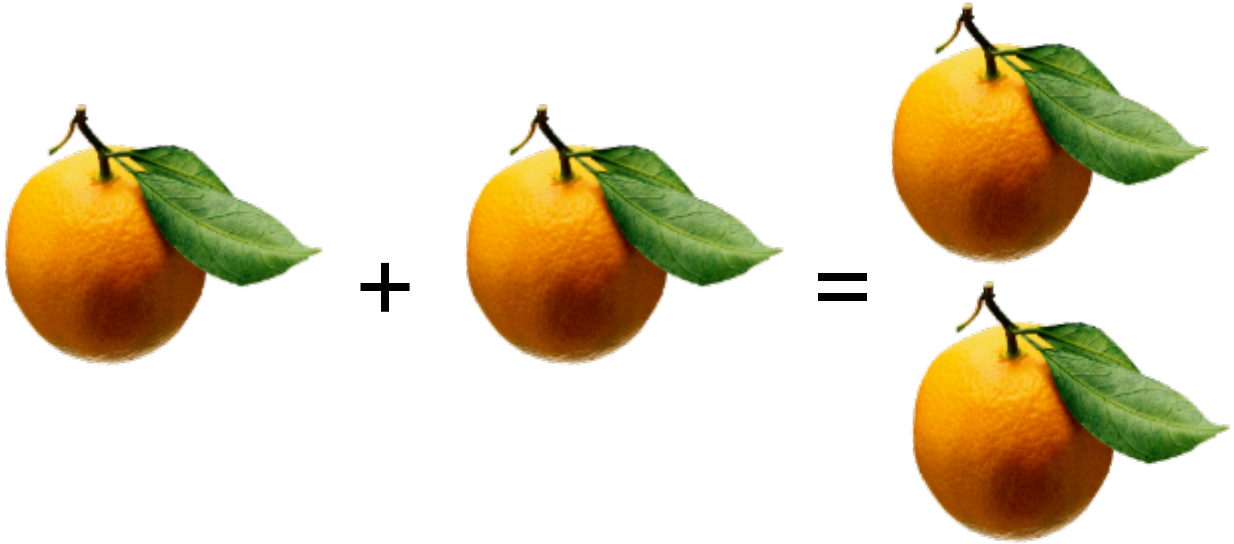
Adapted from *https://plato.stanford.edu/index.html*

DEDUCTIVE AND INDUCTIVE ARGUMENTS

- For valid deductive arguments the premises *logically entail* the conclusion, where the entailment means that the truth of the premises provides a *guarantee* of the truth of the conclusion
- An inductive logic is a system of evidential support that extends deductive logic to less-than-certain inferences
- In a good inductive argument the premises should provide some *degree of support* for the conclusion, where such support means that the truth of the premises indicates with some *degree of strength* that the conclusion is true.
  - acceptability, relevance and sufficiency

Adapted from *https://plato.stanford.edu/index.html*

EXAMPLE



DEDUCTIVE AND INDUCTIVE ARGUMENTS –WHY SEPARATE OUT?

Science of security – importance of deductive/inductive split

“We now detail security research failures to adopt accepted lessons from the history and philosophy of science.

A. Failure to observe inductive-deductive split

Despite broad consensus in the scientific community, in Security there is repeated failure to respect the separation of inductive and deductive statements “

SoK: Science, Security, and the Elusive Goal of Security as a Scientific Pursuit

Cormac Herley
Microsoft Research, Redmond, WA, USA
cormac@microsoft.com

P.C. van Oorschot
Carleton University, Ottawa, ON, Canada
paulv@scs.carleton.ca

DOI: 10.1109/SP.2017.38
Conference: 2017 IEEE Symposium on Security and Privacy (SP)

DEDUCTIVE AND INDUCTIVE ARGUMENTS – WHY SEPARATE OUT?

- Side claim provides a mechanism for factoring
  - Inductive argument-A = Deductive argument + Inductive argument-B
  - Where deductive gives some leverage e.g. analysis, tool support
  - Inductive argument-B is easier to show than Inductive argument-A (then we have made progress!

- Examples
  - Application of deductive models
    - Infer properties
      - Testing evidence -> reliability
      - Abstract interpretation -> run time errors
    - Architecture
      - Property distributes over components (e.g. confidentiality)
    - System properties
      - Fire, flood, earthquakes
  - Each time need to address validity of model and proper application via a side claim

FIVE CAE BUILDING BLOCKS

- Well defined argument fragments
  - Empirically based, but rigorously defined
  - Supporting both deductive and inductive reasoning
- Fragments support a combined graphical and narrative approach

Decomposition

Partition some aspect of the claim
Divide and conquer

Substitution

Refine a claim about an object into claim about an equivalent object

Evidence incorporation

Evidence supports the claim
Emphasis on direct support

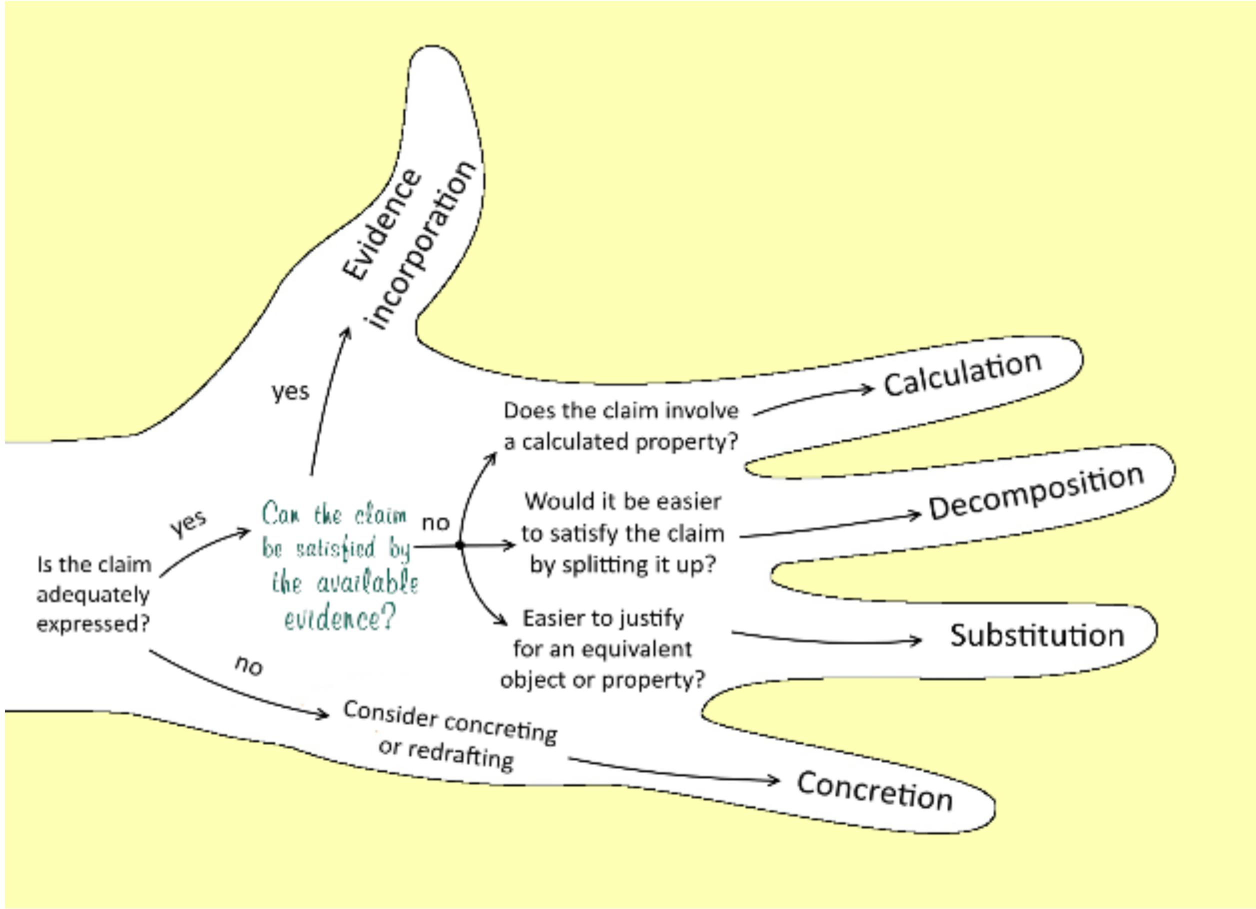
Concretion

Some aspect of the claim is given a more precise definition

Calculation or proof

Some value of the claim can be computed or proved

‘HELPING HAND’ - GUIDANCE ON SELECTING BLOCKS



Is the claim adequately expressed?
yes
Can the claim be satisfied by the available evidence?
yes
Evidence incorporation
no
Does the claim involve a calculated property?
Calculation
Would it be easier to satisfy the claim by splitting it up?
Decomposition
Easier to justify for an equivalent object or property?
Substitution
no
Consider concreting or redrafting
Concretion

DEFEATERS – EXPLICITLY DEALING WITH SOURCES OF DOUBT

- One concept used to address stopping rules and over-confidence is “defeaters”. The concept of defeaters is used to articulate reasons why a claim might **not** be supported.
- Two kinds of defeaters:
  - Rebutting defeaters, which are reasons for believing the negation of the conclusion, and
  - Undercutting defeaters, which provide a reason for doubting that claim.
- Identification and mitigation of defeaters are foundational to assurance
  - Think of as hazard analysis applied to arguments
- In CAE
  - Rebutting defeaters can be addressed with negated subclaims
  - Undercutting defeaters can be addressed by explicitly showing them in the CAE structure

CONFIDENCE

- The purpose of an assurance case is to assist in making, justifying, and communicating the *decision* to deploy a system or service in a given context
- Top level requirement is that the justification should be indefeasible.
  - Meaning it is so well supported and all credible doubts & objections have been so thoroughly considered & countered
  - That no credible doubts remain that could change the decision
- Confidence is strength of our belief that case is indefeasible
- We do not think is can be reduced to some single assessment of the case
- Instead, we identify three perspectives, and assessments and measures within those
  - Assessment of confidence based on all three perspectives

THREE PERSPECTIVES ON CONFIDENCE

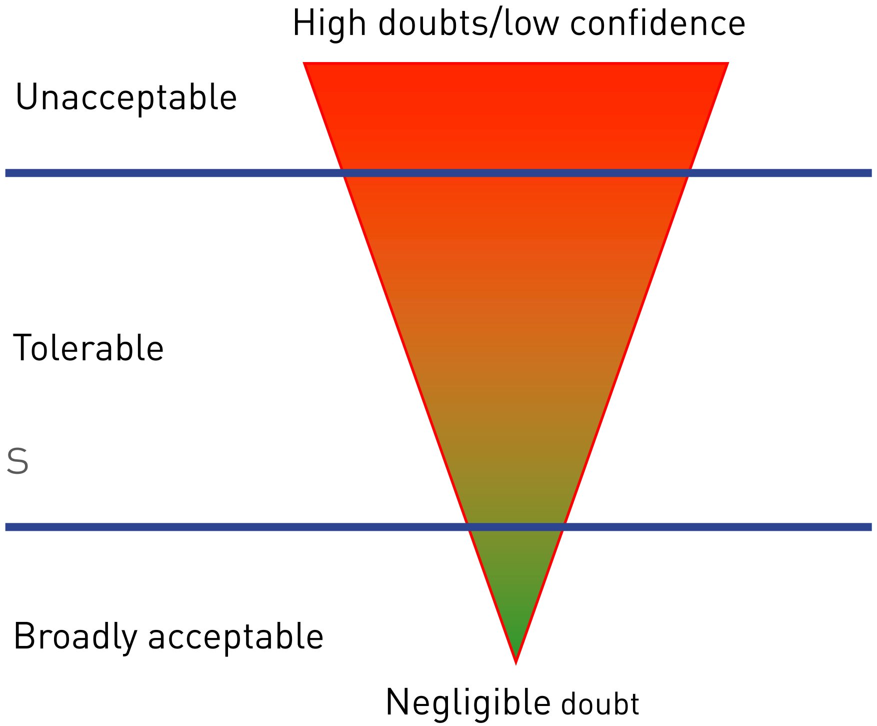
- Positive: extent to which case makes positive case to justify belief in its claims
  - Soundness: logical criterion using Natural Language Deductivism (NLD)
    - Based on weight of evidence, deductive reasoning
  - Probabilistic valuation: probabilistic criterion using Bayesian framework (CBI, BBN)
    - This is what many others mean by confidence: usually flawed (Graydon &Holloway)
    - We require case to be sound, only 5 argument blocks: avoids flaws

- Negative: extent to which doubts have been investigated and addressed
  - Doubts are vague, become defeaters when sharpened, recorded in the case
    - Together with justification for their own defeat (eliminative argumentation)
    - Use systematic methods to find credible defeaters (cf. hazard analysis)
    - May also be possible to invert positive perspective on counterclaims

- Residual Risks: cannot eliminate all doubt (world is uncertain)
  - So must assess risk (likelihood and cost) posed by residual doubts. Tiny ones that do not aggregate, small ones that do, Significant ones that must be quantified

ACARP - ANALOGY WITH ALARP

- Residual Risks: cannot eliminate all doubt (world is uncertain)
  - So must assess risk (likelihood and cost) of residual doubts. Tiny ones that do not aggregate, small ones that do, significant ones that must be quantified
- Confidence implicit in most discussions
  - ACARP concept
    - to promote discussion of what level of confidence is needed
    - consider whether regions of confidence might be useful and introduce the idea of proportionality
- Consider range of claim and confidence claimed
  - Weak claim - high confidence to strong claim - weak confidence



WEIGHT OF EVIDENCE – STRENGTH OF CLAIM

- It's not enough for evidence to support a claim
- It must distinguish a claim from its negation
- Confirmation measures do this: e.g., Kemeny-Oppenheim
  - Goes back to work of Good and Turing in WW2 codebreaking
- These force you to look at counterclaims
  - These are potential defeaters
- Can do this informally/qualitatively, don't need numerical probabilities

$$confirmation\_ratio(Evidence, Claim) = \frac{\Pr(Evidence \mid Claim\_true) - \Pr(Evidence \mid Claim\_false)}{\Pr(Evidence \mid Claim\_true) + \Pr(Evidence \mid Claim\_false)}$$

Probability that you see the evidence if the claim is true

Probability that you see the evidence if the claim is false

CONFIRMATION – ROLE OF DIFFERENT EVIDENCE

Probability see evidence if claim true

Probability see evidence if claim false

		very unlikely	perhaps	quite probable	very likely
		0.05	0.1	0.6	0.95
very unlikely	0.05	0.00	0.33	0.85	0.90
perhaps	0.1	-0.33	0.00	0.71	0.81
quite probable	0.6	-0.85	-0.71	0.00	0.23
very likely	0.95	-0.90	-0.81	-0.23	0.00

$$confirmation\_ratio\left(Evidence, Claim\right)$$

$$= \frac{\Pr\left(Evidence \mid Claim\_true\right) - \Pr\left(Evidence \mid Claim\_false\right)}{\Pr\left(Evidence \mid Claim\_true\right) + \Pr\left(Evidence \mid Claim\_false\right)}$$

CREATING COUNTER CASES

Group #1

- Chocolate is good for you

Group #2

- Chocolate is bad for you

SUMMARY – ASSURANCE 2.0 MANIFESTO

- Assurance 2.0 – key components
- Basic Concepts CAE
- CAE Blocks
  - Empirically based
  - Potential for deductive/inductive split
- Defeaters and confidence
  - Indefeasibility and residual rikss
- Evidence
  - Relevance and provenance
  - Confirmation theory and strength of arguments and evidence
- Explicit approach to bias
  - Counter-cases and confirmation theory

DEVELOPMENT AND APPLICATION – WILL IT WORK?

- Security applications
- Impact on regulation of systems incorporating AI/machine learning
- Developed autonomous system “templates and guidance”
- Tool support
  - building on Adelard ASCE tool within a program on automated certification
- Teaching concepts to professional engineers
  - many disciplines

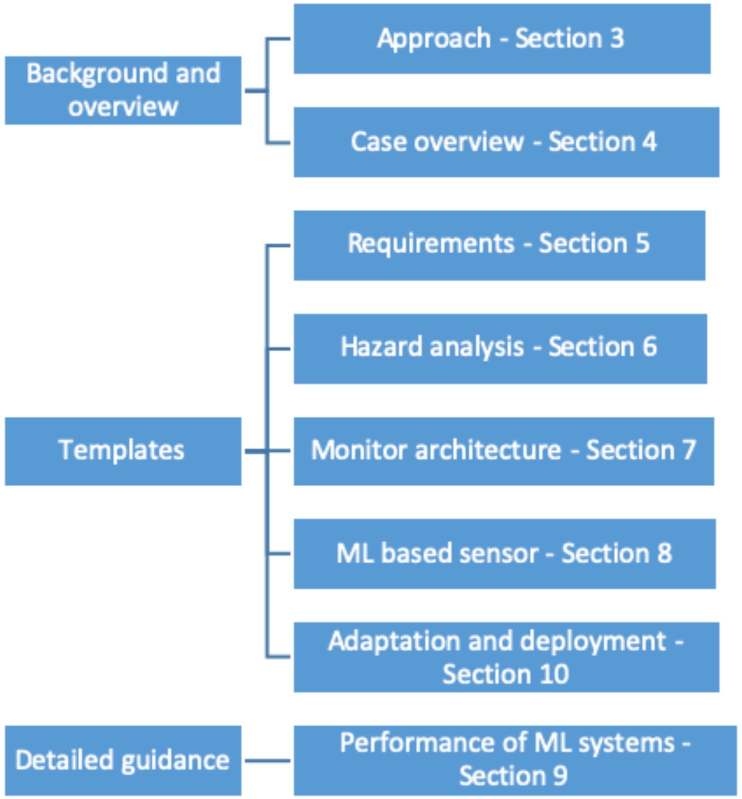
Theory into practice

DSTL sponsored research

SAFETY CASE TEMPLATES FOR AUTONOMOUS SYSTEMS

http://arxiv.org/abs/2102.02625

DEVELOPMENT OF TEMPATES FOR AV



Background and overview

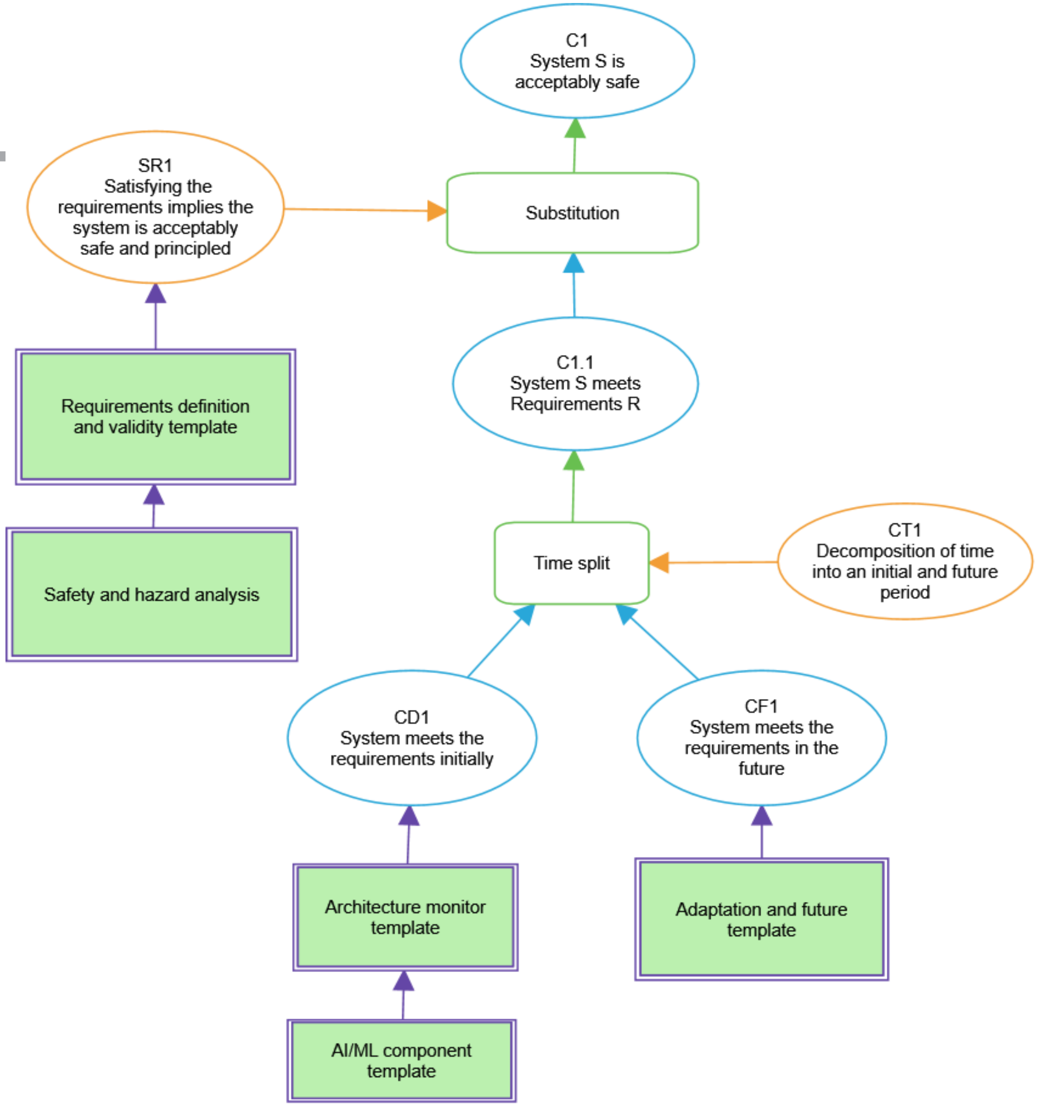
- Approach - Section 3
- Case overview - Section 4

Templates

- Requirements - Section 5
- Hazard analysis - Section 6
- Monitor architecture - Section 7
- ML based sensor - Section 8
- Adaptation and deployment - Section 10

Detailed guidance

- Performance of ML systems - Section 9



C1
System S is acceptably safe

Substitution

SR1
Satisfying the requirements implies the system is acceptably safe and principled

Requirements definition and validity template

Safety and hazard analysis

C1.1
System S meets Requirements R

Time split

CT1
Decomposition of time into an initial and future period

CD1
System meets the requirements initially

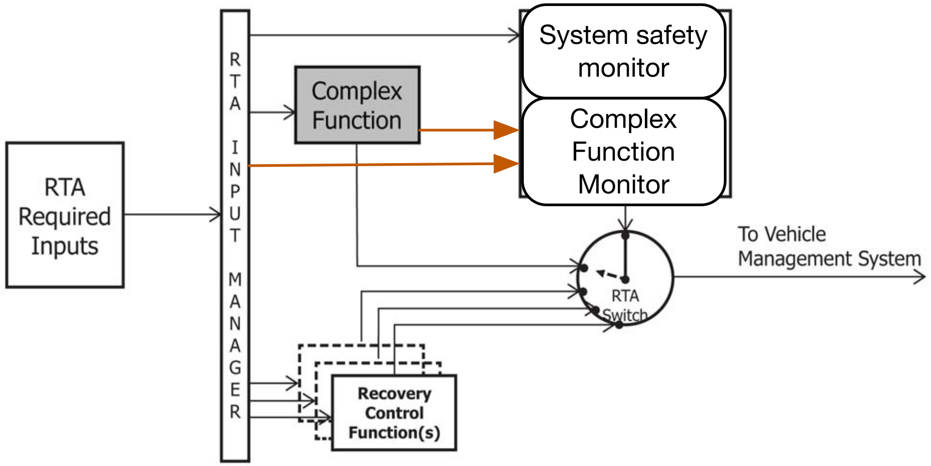
CF1
System meets the requirements in the future

Architecture monitor template

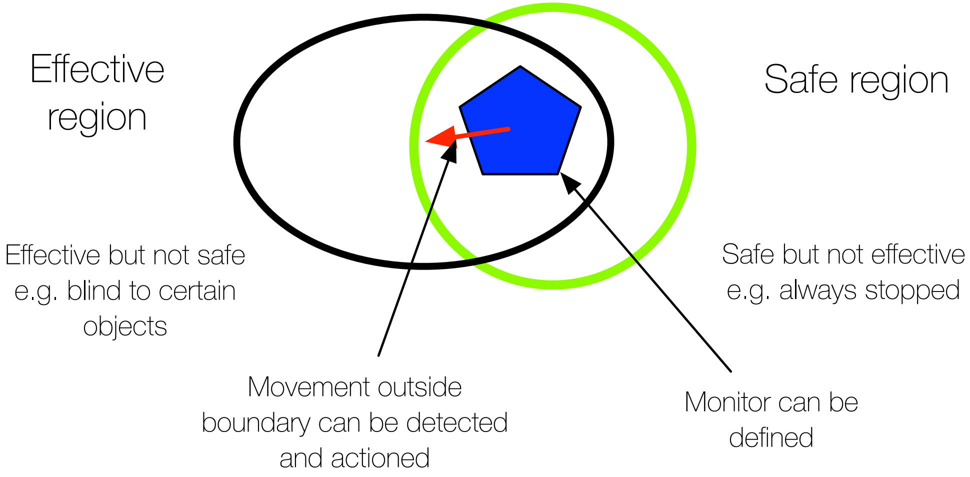
Adaptation and future template

AI/ML component template

GENERIC MONITOR GUARD ARCHITECTURE



F3269-17 Standard Practice for Methods to Safely Bound Flight Behavior of Unmanned Aircraft Systems Containing Complex Functions, ASTM International

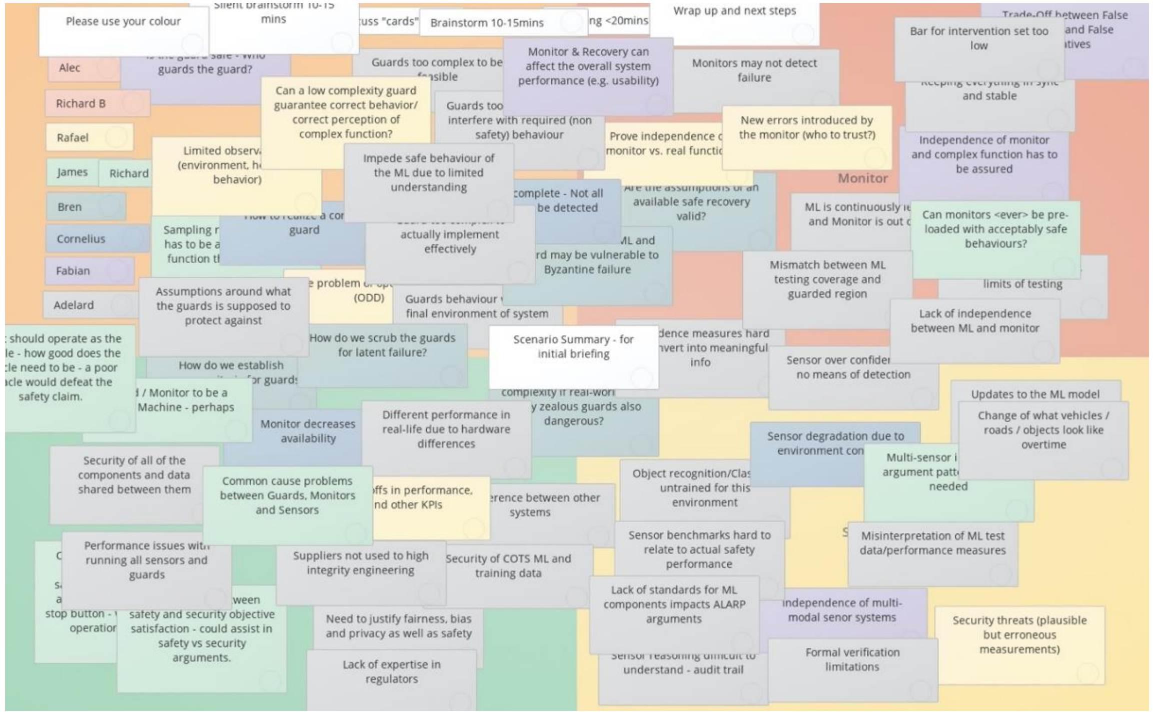


Effective region
Safe region
Effective but not safe e.g. blind to certain objects
Safe but not effective e.g. always stopped
Movement outside boundary can be detected and actioned
Monitor can be defined

Monitor feasibility

DEFEATER WORKSHOP – MONITOR/GUARD ARCHITECTURE

- Colour coded issues and organisations
- Identified issues on-line with international team
  - Briefing
  - Silent brainstorm
  - Collaborative
  - Grouping, sentencing
- Work in progress
  - Still exploring how to capture and present defeaters



Please use your colour

Silent brainstorm 10-15 mins

uss "cards"

Brainstorm 10-15mins

ng <20mins

Wrap up and next steps

Alec

Richard B

Rafael

James

Richard

Bren

Cornelius

Fabian

Adelard

guards the guard?

Guards too complex to be feasible

Monitor & Recovery can affect the overall system performance (e.g. usability)

Monitors may not detect failure

Trade-Off between False and False tives

Bar for intervention set too low

and stable

Can a low complexity guard guarantee correct behavior/ correct perception of complex function?

Guards too interfere with required (non safety) behaviour

Limited observ (environment, h behavior)

Impede safe behaviour of the ML due to limited understanding

Prove independence c monitor vs. real functic

New errors introduced by the monitor (who to trust?)

Independence of monitor and complex function has to be assured

Monitor

complete - Not all be detected

available safe recovery valid?

ML is continuously and Monitor is out c

Can monitors <ever> be pre-loaded with acceptably safe behaviours?

Sampling r has to be a function tl

guard

actually implement effectively

ML and rd may be vulnerable to Byzantine failure

Mismatch between ML testing coverage and guarded region

limits of testing

e problem (ODD)

Assumptions around what the guards is supposed to protect against

Guards behaviour final environment of system

Lack of independence between ML and monitor

should operate as the le - how good does the cle need to be - a poor acle would defeat the safety claim.

How do we scrub the guards for latent failure?

Scenario Summary - for initial briefing

dence measures hard nvert into meaningful info

Sensor over confider no means of detection

How do we establish or guards

d / Monitor to be a Machine - perhaps

y zealous guards also dangerous?

Updates to the ML model

Change of what vehicles / roads / objects look like overtime

Monitor decreases availability

Different performance in real-life due to hardware differences

Sensor degradation due to environment con

Multi-sensor i argument patt needed

Security of all of the components and data shared between them

Common cause problems between Guards, Monitors and Sensors

offs in performance, nd other KPIs

erence between other systems

Object recognition/Clas untrained for this environment

Performance issues with running all sensors and guards

Suppliers not used to high integrity engineering

Security of COTS ML and training data

Sensor benchmarks hard to relate to actual safety performance

Misinterpretation of ML test data/performance measures

stop button - operatior

ween safety and security objective satisfaction - could assist in safety vs security arguments.

Need to justify fairness, bias and privacy as well as safety

Lack of standards for ML components impacts ALARP arguments

independence of multi-modal senor systems

Security threats (plausible but erroneous measurements)

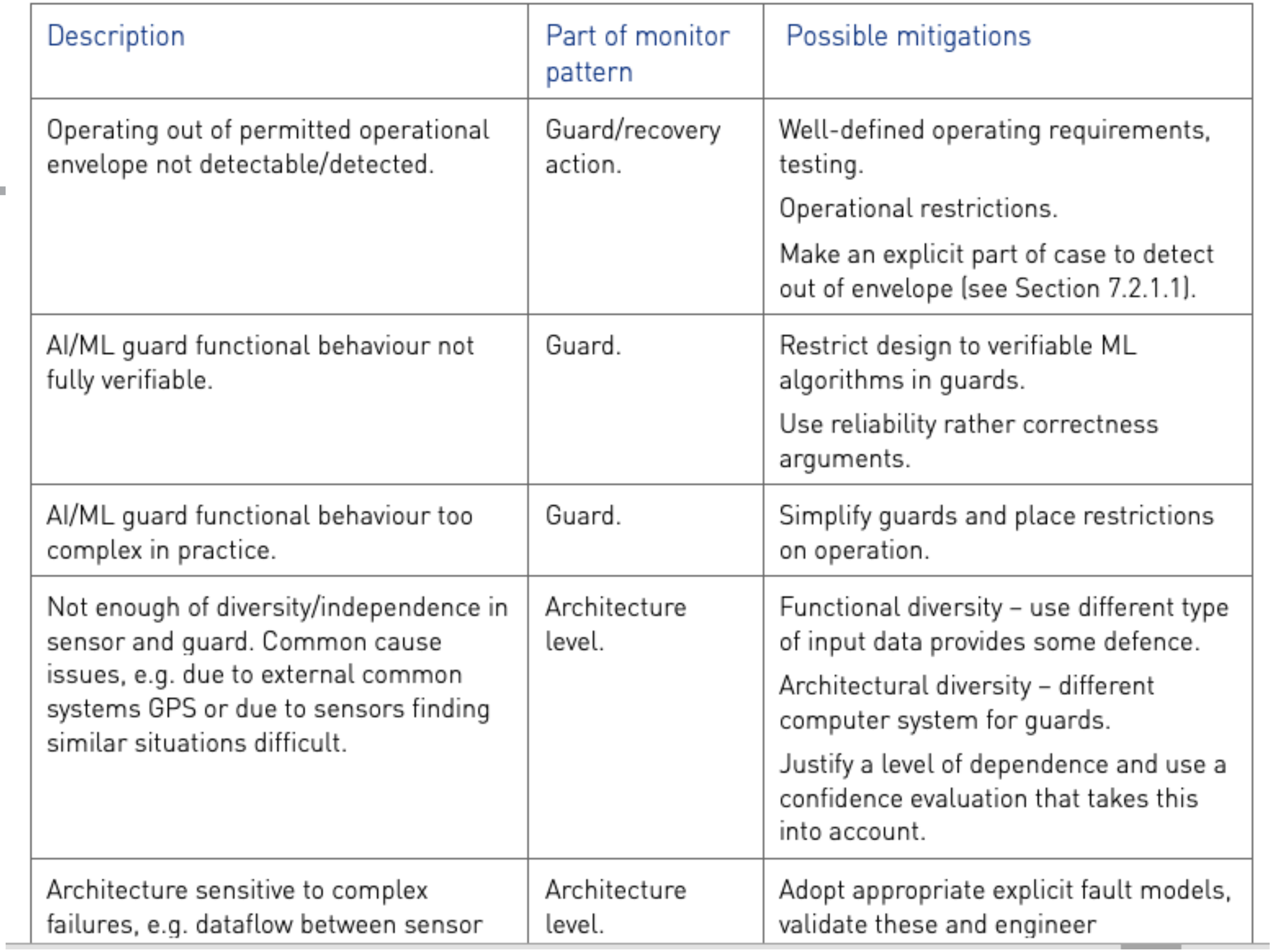
Lack of expertise in regulators

understand - audit trail

Formal verification limitations

DEFEATERS

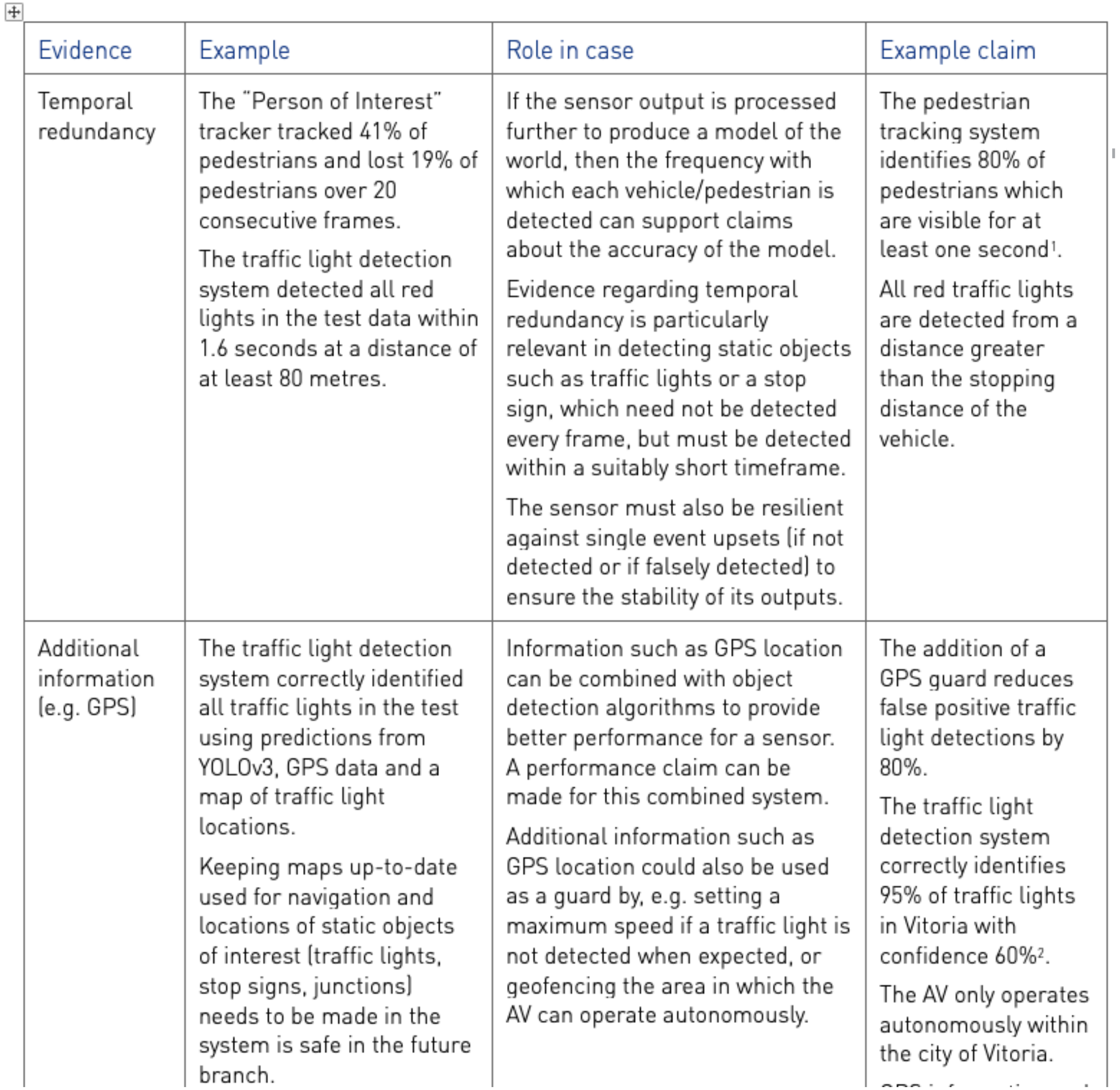
- Summary tables – with supporting narrative



Description	Part of monitor pattern	Possible mitigations
Operating out of permitted operational envelope not detectable/detected.	Guard/recovery action.	Well-defined operating requirements, testing. Operational restrictions. Make an explicit part of case to detect out of envelope (see Section 7.2.1.1).
AI/ML guard functional behaviour not fully verifiable.	Guard.	Restrict design to verifiable ML algorithms in guards. Use reliability rather correctness arguments.
AI/ML guard functional behaviour too complex in practice.	Guard.	Simplify guards and place restrictions on operation.
Not enough of diversity/independence in sensor and guard. Common cause issues, e.g. due to external common systems GPS or due to sensors finding similar situations difficult.	Architecture level.	Functional diversity – use different type of input data provides some defence. Architectural diversity – different computer system for guards. Justify a level of dependence and use a confidence evaluation that takes this into account.
Architecture sensitive to complex failures, e.g. dataflow between sensor	Architecture level.	Adopt appropriate explicit fault models, validate these and engineer

TECHNICAL GUIDANCE

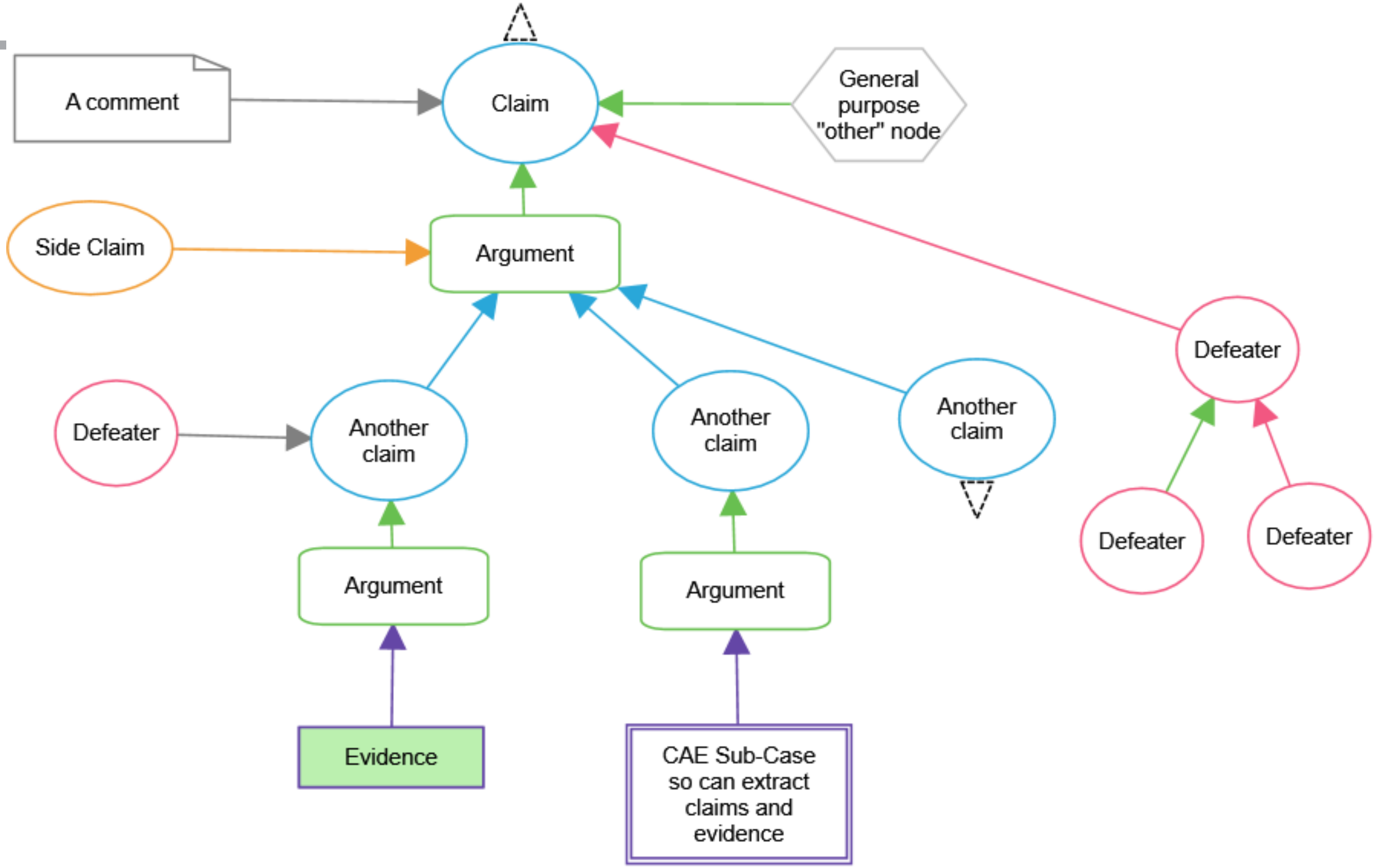
- Confidence measures for ML
  - Conformal Prediction
  - Inductive Conformal Prediction
  - Attribution-based confidence
  - Learning confidence
- Performance of ML based components
  - Performance metrics for binary classifiers
  - Object detection
  - Experimental performance



Evidence	Example	Role in case	Example claim
Temporal redundancy	The "Person of Interest" tracker tracked 41% of pedestrians and lost 19% of pedestrians over 20 consecutive frames. The traffic light detection system detected all red lights in the test data within 1.6 seconds at a distance of at least 80 metres.	If the sensor output is processed further to produce a model of the world, then the frequency with which each vehicle/pedestrian is detected can support claims about the accuracy of the model. Evidence regarding temporal redundancy is particularly relevant in detecting static objects such as traffic lights or a stop sign, which need not be detected every frame, but must be detected within a suitably short timeframe. The sensor must also be resilient against single event upsets (if not detected or if falsely detected) to ensure the stability of its outputs.	The pedestrian tracking system identifies 80% of pedestrians which are visible for at least one second[1]. All red traffic lights are detected from a distance greater than the stopping distance of the vehicle.
Additional information (e.g. GPS)	The traffic light detection system correctly identified all traffic lights in the test using predictions from YOLOv3, GPS data and a map of traffic light locations. Keeping maps up-to-date used for navigation and locations of static objects of interest (traffic lights, stop signs, junctions) needs to be made in the future system is safe in the future branch.	Information such as GPS location can be combined with object detection algorithms to provide better performance for a sensor. A performance claim can be made for this combined system. Additional information such as GPS location could also be used as a guard by, e.g. setting a maximum speed if a traffic light is not detected when expected, or geofencing the area in which the AV can operate autonomously.	The addition of a GPS guard reduces false positive traffic light detections by 80%. The traffic light detection system correctly identifies 95% of traffic lights in Vitoria with confidence 60%[2]. The AV only operates autonomously within the city of Vitoria.

TOOL SUPPORT

NOTATION

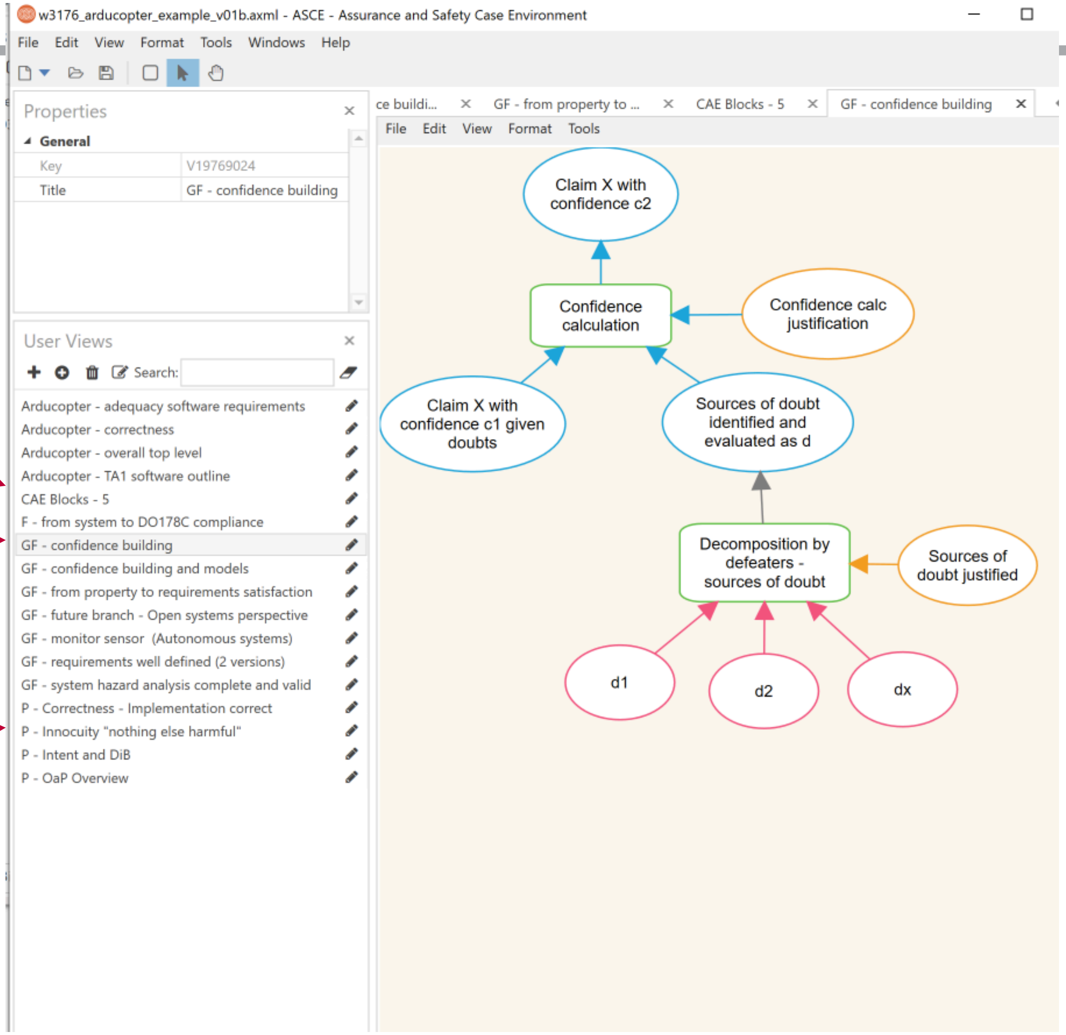

A comment
Claim
General purpose "other" node
Side Claim
Argument
Defeater
Defeater
Another claim
Another claim
Another claim
Defeater
Defeater
Argument
Argument
Evidence
CAE Sub-Case so can extract claims and evidence

BLOCKS AND PATTERNS

User View - CAE Blocks

User View - Confidence building

Pattern for innocuity



w3176_arducopter_example_v01b.axml - ASCE - Assurance and Safety Case Environment

Properties

General

Key	V19769024
Title	GF - confidence building

User Views

Search:

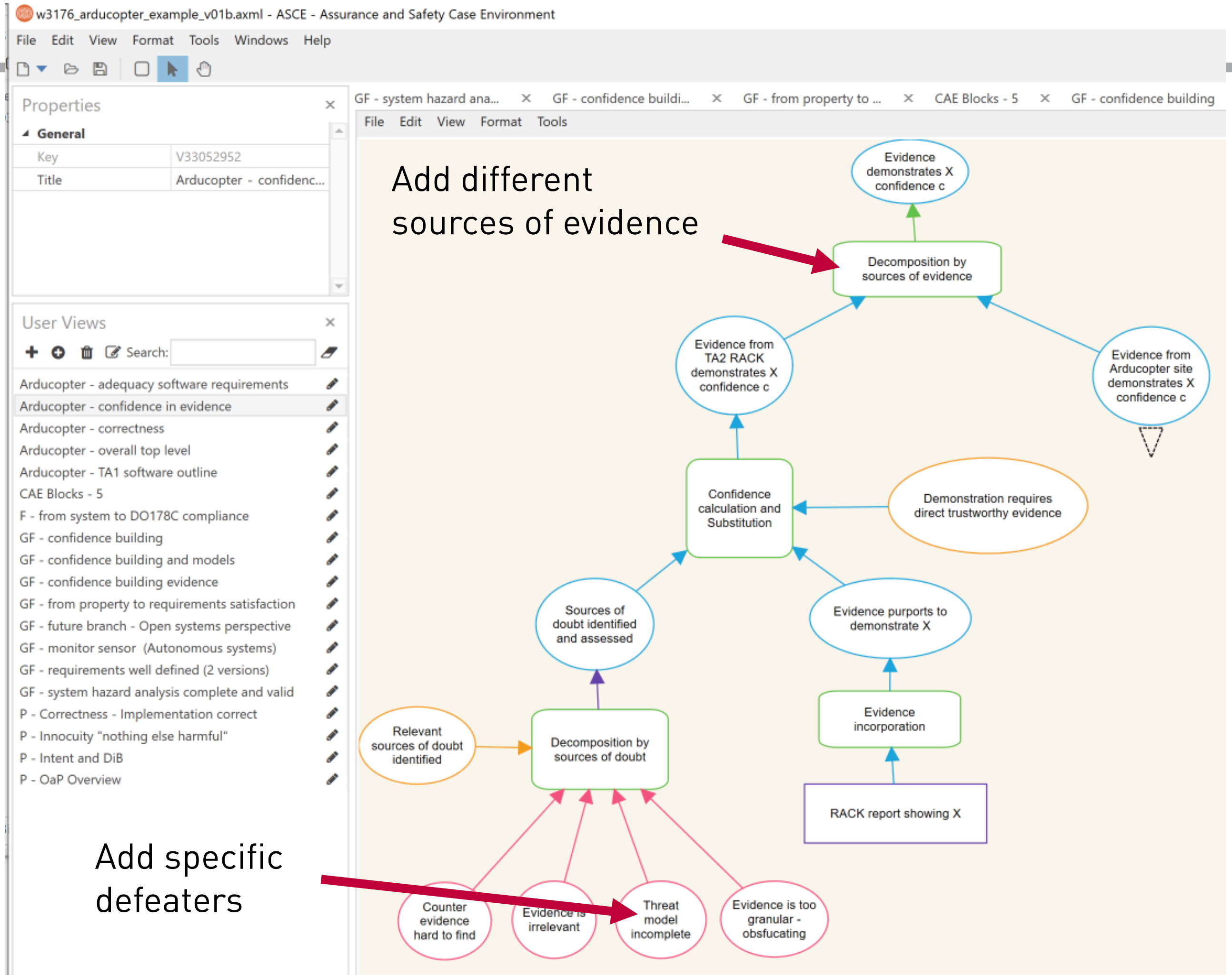
- Arducopter - adequacy software requirements
- Arducopter - correctness
- Arducopter - overall top level
- Arducopter - TA1 software outline
- CAE Blocks - 5
- F - from system to DO178C compliance
- GF - confidence building
- GF - confidence building and models
- GF - from property to requirements satisfaction
- GF - future branch - Open systems perspective
- GF - monitor sensor (Autonomous systems)
- GF - requirements well defined (2 versions)
- GF - system hazard analysis complete and valid
- P - Correctness - Implementation correct
- P - Innocuity "nothing else harmful"
- P - Intent and DiB
- P - OaP Overview

File Edit View Format Tools Windows Help
ce buildi...
GF - from property to ...
CAE Blocks - 5
GF - confidence building
File Edit View Format Tools
Claim X with confidence c2
Confidence calculation
Confidence calc justification
Claim X with confidence c1 given doubts
Sources of doubt identified and evaluated as d
Decomposition by defeaters - sources of doubt
Sources of doubt justified
d1
d2
dx

© 2021 ADELARD LLP

SYNTHESIS

- Evidence Integration + Confidence pattern
- Different sources of evidence
  - Added Decomposition
- Added specific defeaters



w3176_arducopter_example_v01b.axml - ASCE - Assurance and Safety Case Environment

Properties

General

Key	V33052952
Title	Arducopter - confidenc...

User Views

- Arducopter - adequacy software requirements
- Arducopter - confidence in evidence
- Arducopter - correctness
- Arducopter - overall top level
- Arducopter - TA1 software outline
- CAE Blocks - 5
- F - from system to DO178C compliance
- GF - confidence building
- GF - confidence building and models
- GF - confidence building evidence
- GF - from property to requirements satisfaction
- GF - future branch - Open systems perspective
- GF - monitor sensor (Autonomous systems)
- GF - requirements well defined (2 versions)
- GF - system hazard analysis complete and valid
- P - Correctness - Implementation correct
- P - Innocuity "nothing else harmful"
- P - Intent and DiB
- P - OaP Overview

File Edit View Format Tools Windows Help
Search:
GF - system hazard ana...
GF - confidence buildi...
GF - from property to ...
CAE Blocks - 5
GF - confidence building
File Edit View Format Tools
Evidence demonstrates X confidence c
Decomposition by sources of evidence
Evidence from TA2 RACK demonstrates X confidence c
Evidence from Arducopter site demonstrates X confidence c
Confidence calculation and Substitution
Demonstration requires direct trustworthy evidence
Sources of doubt identified and assessed
Evidence purports to demonstrate X
Relevant sources of doubt identified
Decomposition by sources of doubt
Evidence incorporation
RACK report showing X
Counter evidence hard to find
Evidence is irrelevant
Threat model incomplete
Evidence is too granular - obsfucating

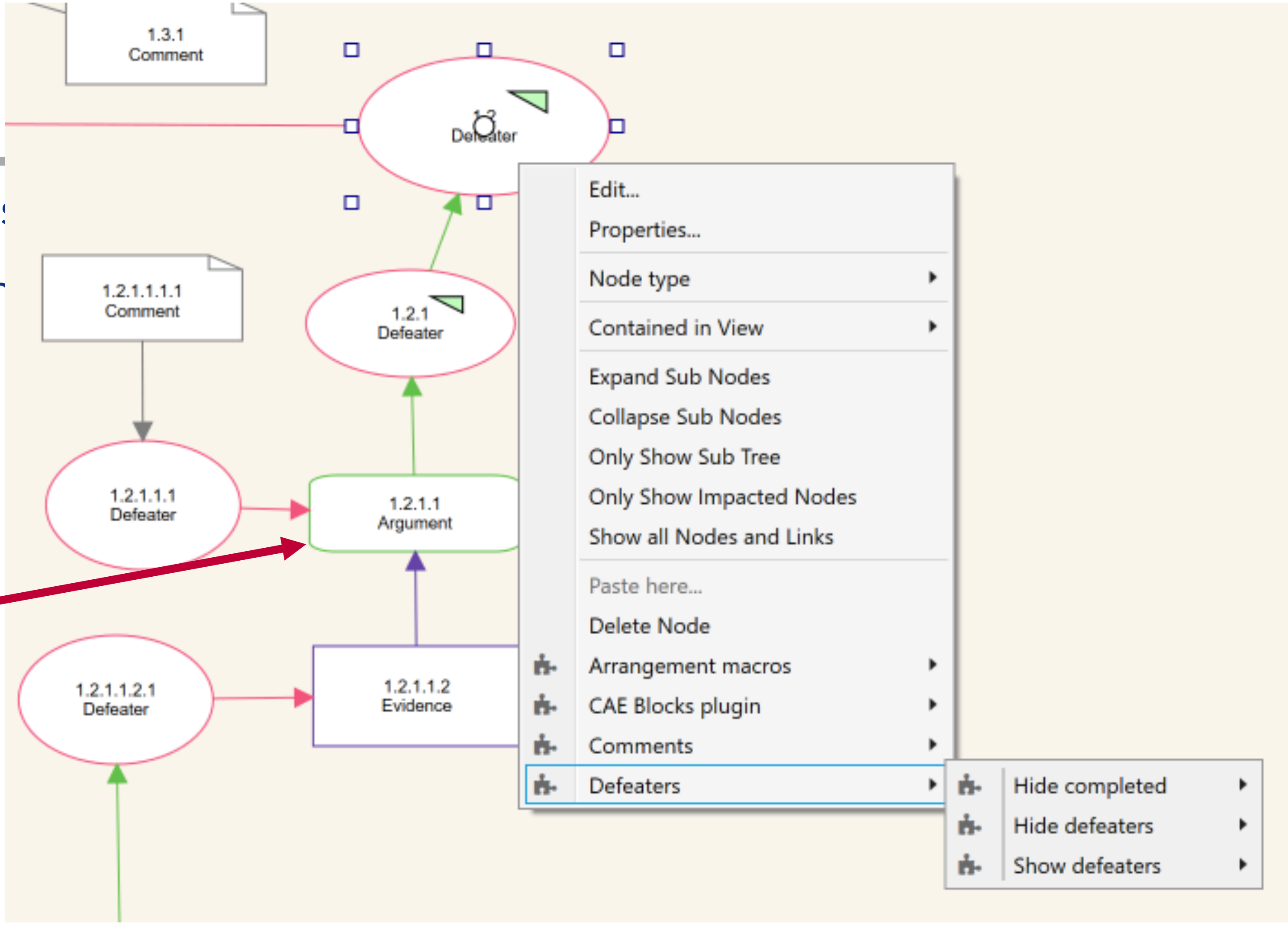
Add different sources of evidence

Add specific defeaters

© 2021 ADELARD LLP

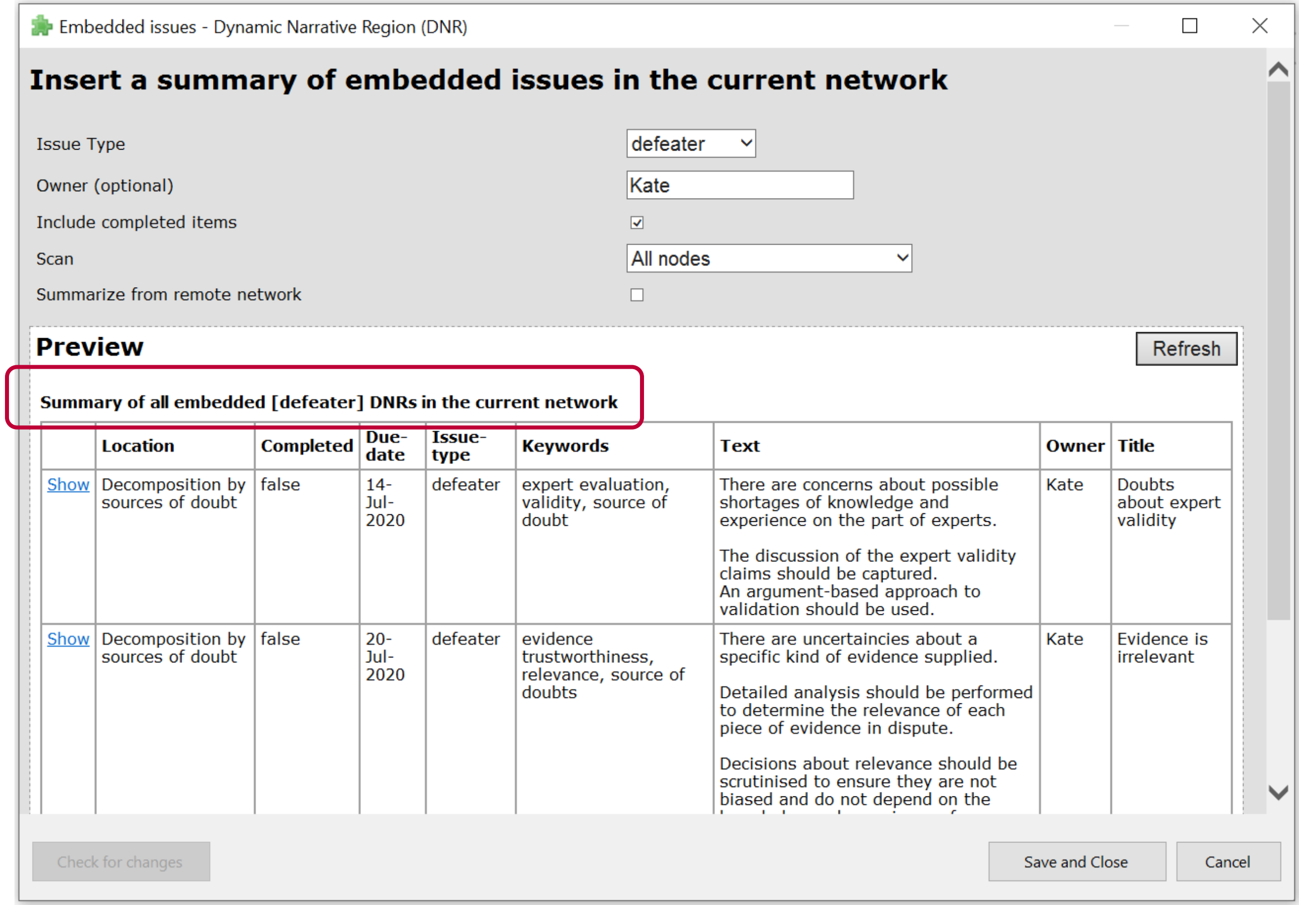
DEFEATER MANAGEMENT

- Use of issue management s
- Defeater node managemen



Defeaters
attacking CAE

EMBEDDED DEFEATERS



Embedded issues - Dynamic Narrative Region (DNR)

Insert a summary of embedded issues in the current network

Issue Type	defeater
Owner (optional)	Kate
Include completed items	☑
Scan	All nodes
Summarize from remote network	□

Preview

Summary of all embedded [defeater] DNRs in the current network

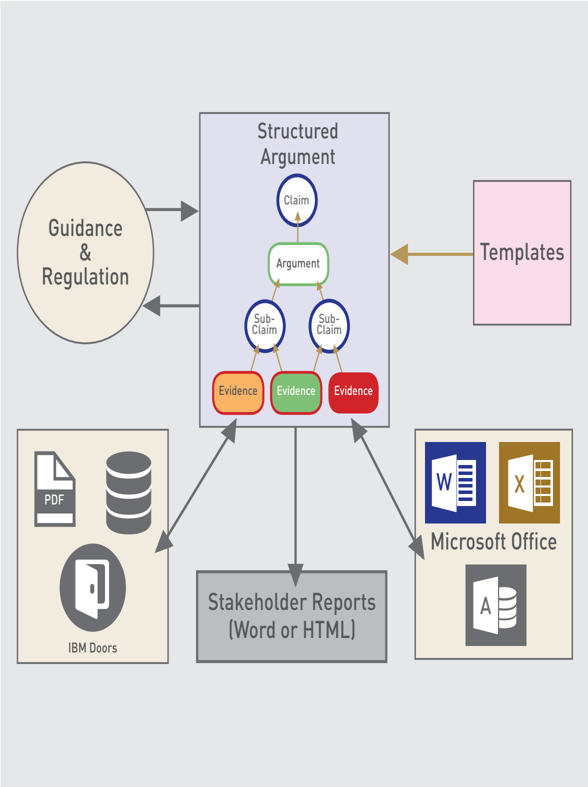
	Location	Completed	Due-date	Issue-type	Keywords	Text	Owner	Title
Show	Decomposition by sources of doubt	false	14-Jul-2020	defeater	expert evaluation, validity, source of doubt	There are concerns about possible shortages of knowledge and experience on the part of experts. The discussion of the expert validity claims should be captured. An argument-based approach to validation should be used.	Kate	Doubts about expert validity
Show	Decomposition by sources of doubt	false	20-Jul-2020	defeater	evidence trustworthiness, relevance, source of doubts	There are uncertaincies about a specific kind of evidence supplied. Detailed analysis should be performed to determine the relevance of each piece of evidence in dispute. Decisions about relevance should be scrutinised to ensure they are not biased and do not depend on the	Kate	Evidence is irrelevant

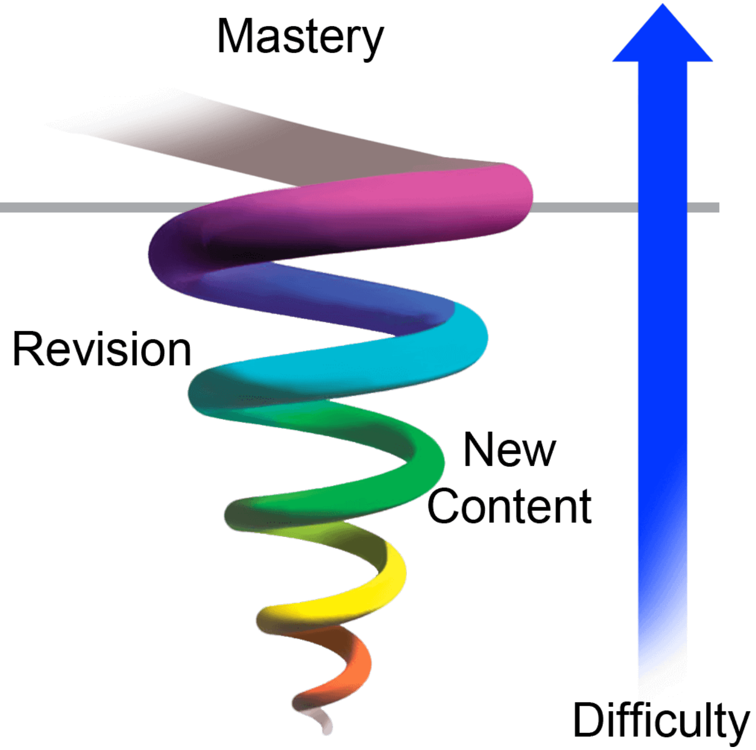
NEXT STEPS

- Assurance 2.0 support in Adelard ASCE tool
  - Available in new release, March 2021
  - If interested in beta versions please get in touch

- Safety Case Templates for Autonomous Systems
  - Example templates for autonomous systems will be available too based on work for DSTL. Report is
  - http://arxiv.org/abs/2102.02625

ASCE - in the wider environment





Teaching concepts to professional engineers (many disciplines)

- 4 pilot courses, 80 engineers and managers, 200 on waiting list

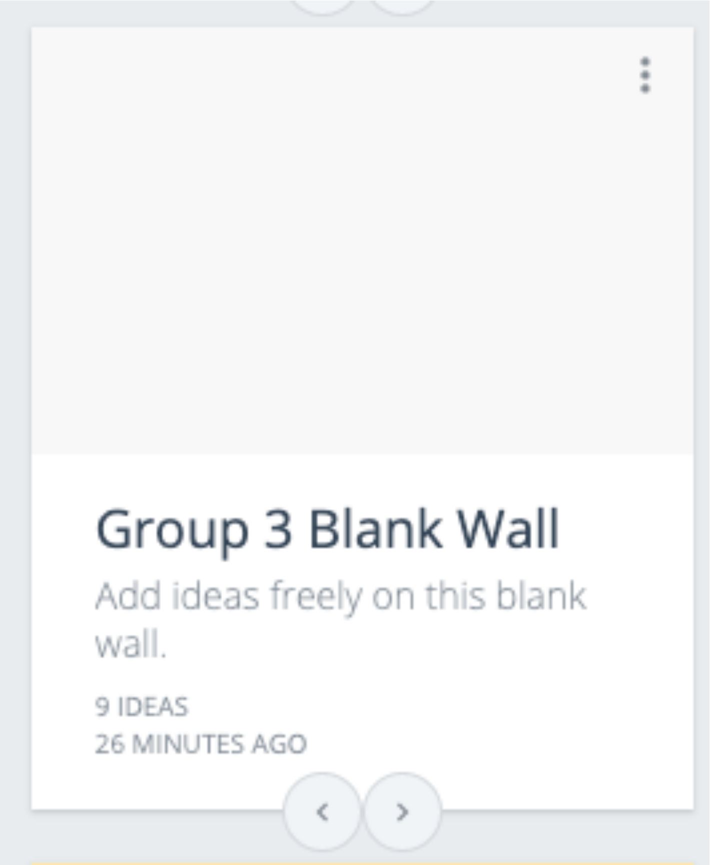
APPLICATION - MAJOR HAZARDS SITE

OUTLINE – ONLINE COURSE

- Session 1: CAE concepts
  - Claims, Arguments, Evidence (CAE): concepts and background
  - Inductive and deductive reasoning
  - Application of CAE concepts
  - Introduction to defeaters
  - Short exercise
- Session 2: Theory into practice
  - Short exercise
  - The CAE blocks and guidance
  - Discussion of Operations Room example
  - Workshop exercise and discussion
- Session 3: Learning by doing, workshop exercises and discussion
- Session 4: Challenge, review and deployment
  - Build confidence into the justification
  - Review and challenge
  - Summary
- Session 5: Wrap up and discussion
  - Putting it all together and next steps, work projects

EXERCISES

- Objective is to practice using the CAE Blocks
- Work in groups with a canvas per group
- Stages
  - Decomposition Block example
  - An example of putting the Blocks together
  - Examples of all 5 Blocks
- Add questions and comments to us as you go
- Review



Group 3 Blank Wall

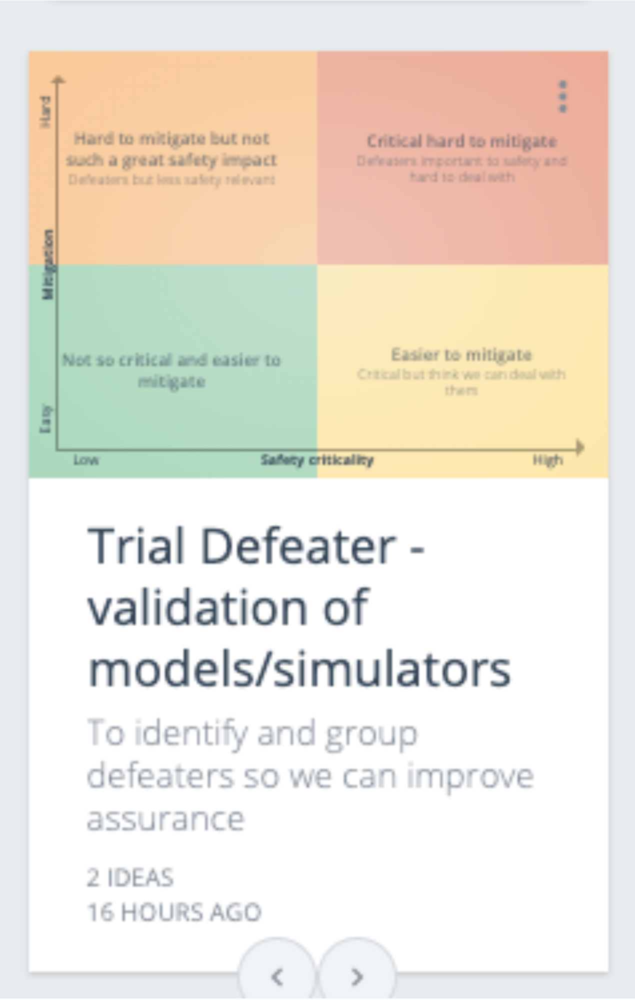
Add ideas freely on this blank wall.

9 IDEAS

26 MINUTES AGO

EXERCISE - DOUBTS AND SIMULATION VALIDATION

- Objective is to express defeaters
    - What might defeat the reasoning that the simulator is valid i.e. sufficiently realistic?
    - “Simulated environment equivalent to actual”
- Work individually
- Add questions and comments to us as you go


Hard
Mitigation
Easy
Hard to mitigate but not such a great safety impact
Defeaters but less safety relevant
Critical hard to mitigate
Defeaters important to safety and hard to deal with
Not so critical and easier to mitigate
Easier to mitigate
Critical but think we can deal with them
Low
Safety criticality
High

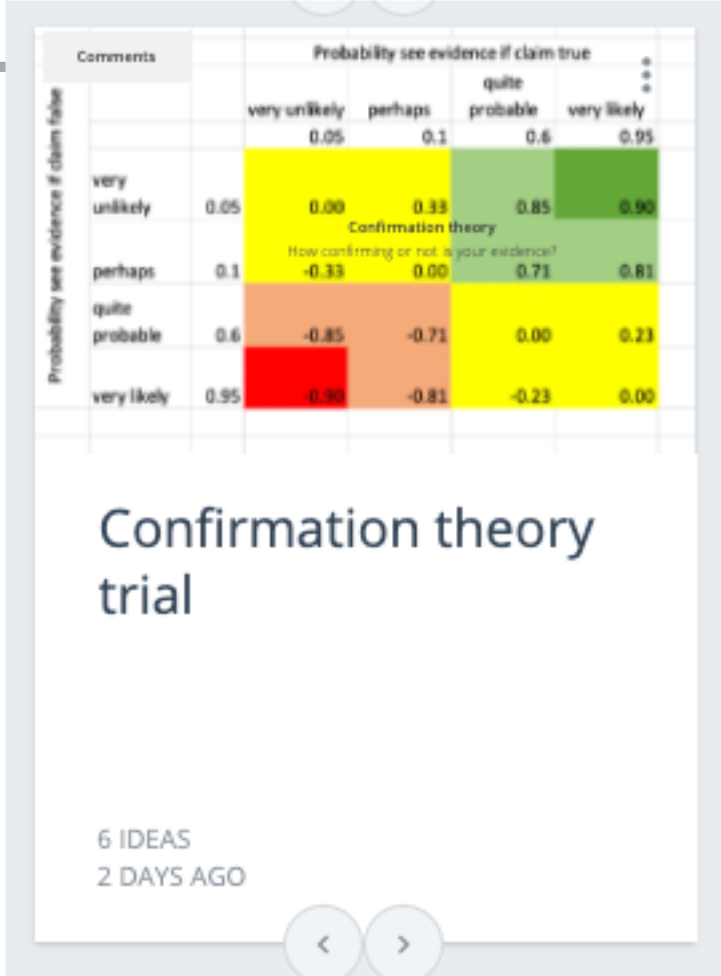
Trial Defeater - validation of models/simulators

To identify and group defeaters so we can improve assurance

2 IDEAS
16 HOURS AGO

EXERCISE

- In groups discuss examples of claims and evidence asking
  - How likely I am to see the evidence if the claim is true?
  - How likely I am to see the evidence if the claim is false?
- and put on the grid along with any comments



Comments		Probability see evidence if claim true			
		very unlikely	perhaps	quite probable	very likely
		0.05	0.1	0.6	0.95
Probability see evidence if claim false: very unlikely	0.05	0.00	0.33	0.85	0.90
perhaps	0.1	-0.33	0.00	0.71	0.81
quite probable	0.6	-0.85	-0.71	0.00	0.23
very likely	0.95	-0.90	-0.81	-0.23	0.00

Confirmation theory

How confirming or not is your evidence?

Confirmation theory trial

6 IDEAS
2 DAYS AGO

APPLICATION IN MAJOR HAZARDOUS SITE – CONCLUSIONS TO DATE

- Can get ideas across with a day course
  - Teaching concepts to professional engineers (many disciplines)
  - Often those *without* safety case background find it easier
  - Wide range of responses – struggle, OK, great
- Follow up application on real projects required
  - Over several months
  - Surgeries and support
- Experience and feedback
  - In progress
  - So far 4 pilot courses, 80 engineers and managers, 200 on waiting list
    - CAE Blocks , defeaters, counter cases ☺
  - Will review and publish experience after ~100 through course

FROM MANIFESTO TO MATURE METHODOLOGY

- Empirically based CAE Blocks separate inductive and deductive aspects
- Explicit use of doubts and defeaters
- Increased focus on evidence integration, addressing both relevance and provenance
- Confirmation theory to evaluate the strength of evidence and arguments.
- Explicit approach to bias by the use of counter-cases and confirmation theory.
- Recognition of both mindset and methodology

- Publish and apply
  - Different maturity
- Real applications
  - Engineering justifications, safety and security
- Teaching and learning - evaluation
  - >100 industry by April
- Further development of methodology
  - Defeater identification and management
  - Synthesis approaches
  - Confidence and defeaters
- Assurance 2.0 and templates + tools
  - Evaluation and further development

Prof Robin E Bloomfield FREng

Adelard LLP and City, University of London

reb@adelard.com

r.e.bloomfield@city.ac.uk

Joint paper with John Rushby, SRI

ADELARD

Presentation to SSS'21, Feb 10th 2021