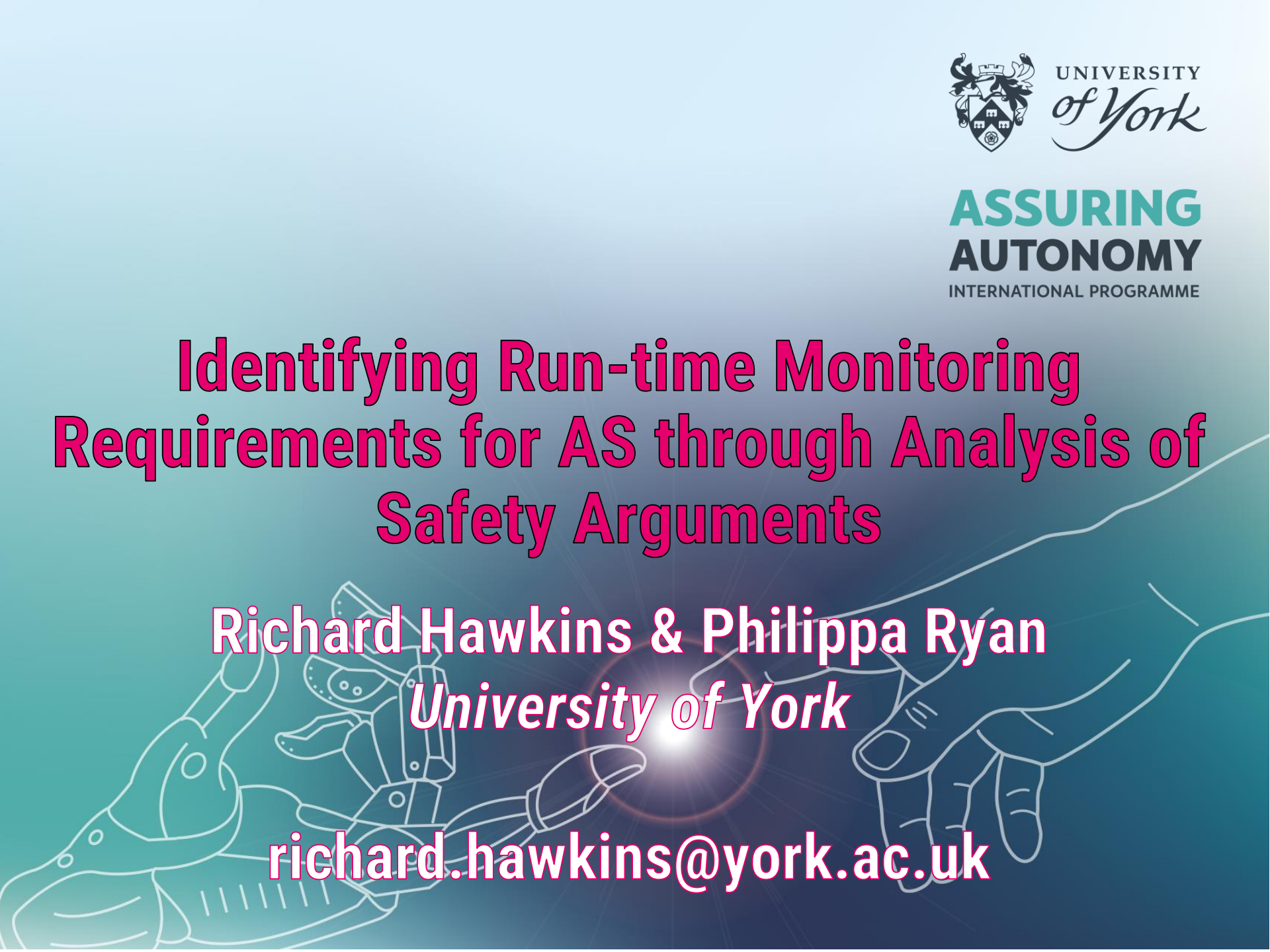


UNIVERSITY of York

ASSURING AUTONOMY INTERNATIONAL PROGRAMME

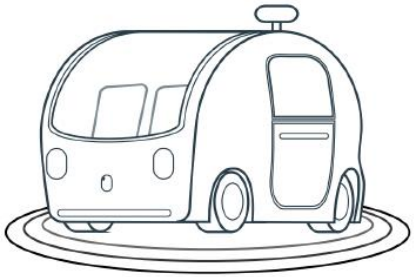
Identifying Run-time Monitoring Requirements for AS through Analysis of Safety Arguments

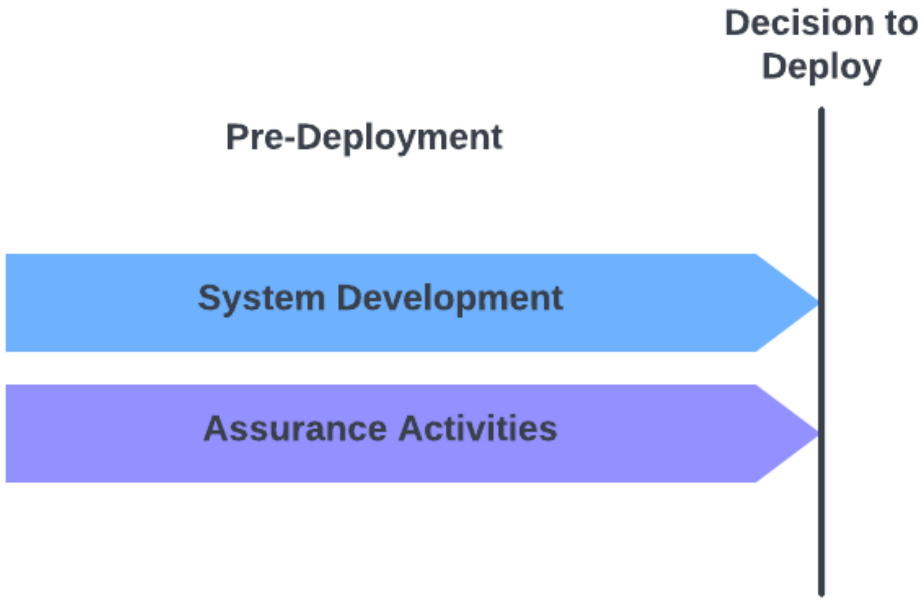
Richard Hawkins & Philippa Ryan

University of York

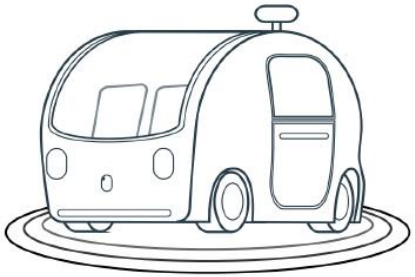
richard.hawkins@york.ac.uk

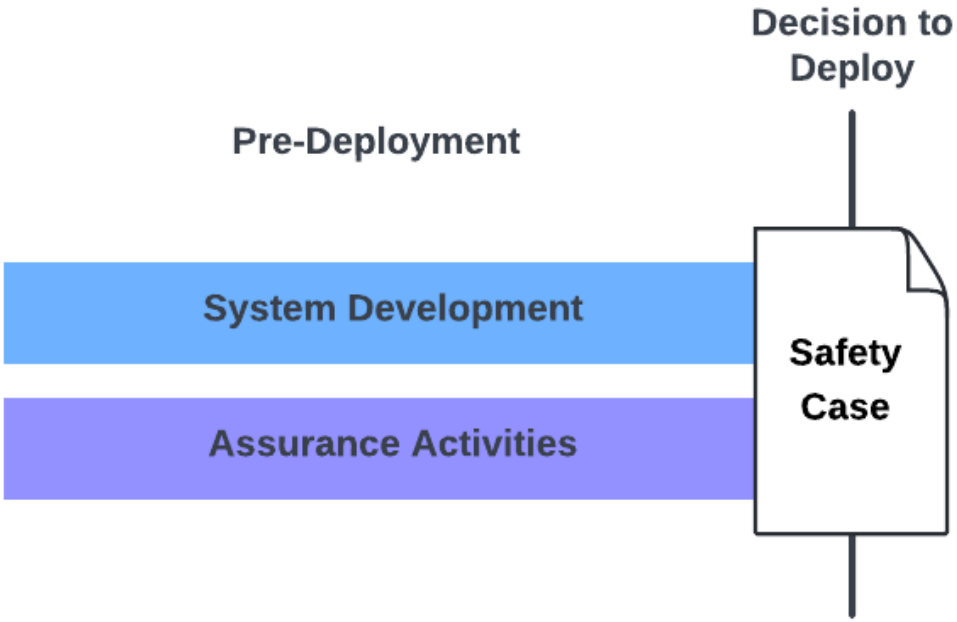
Safety Monitoring for AS



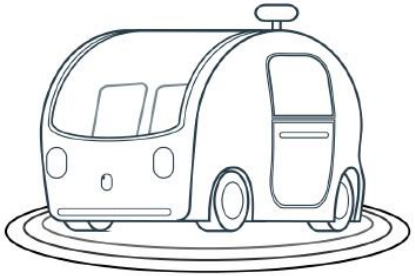


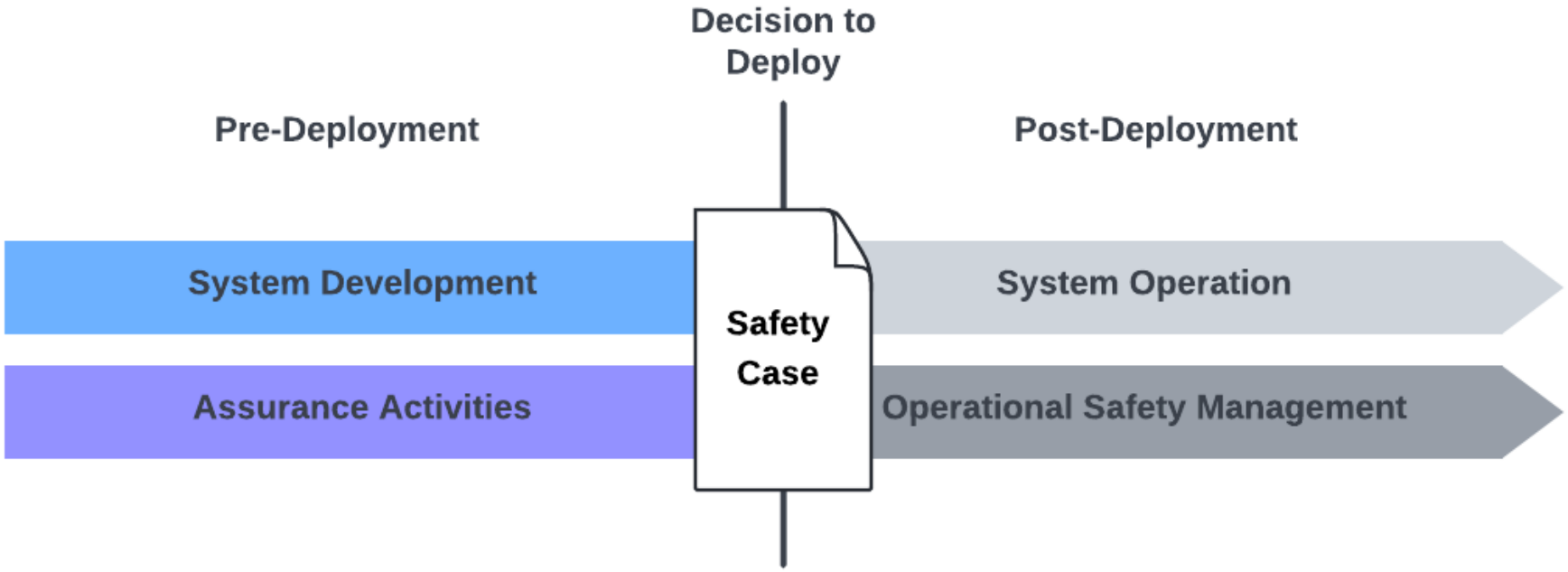
Safety Monitoring for AS



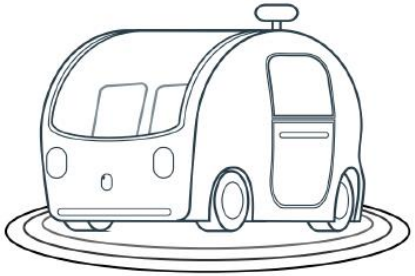


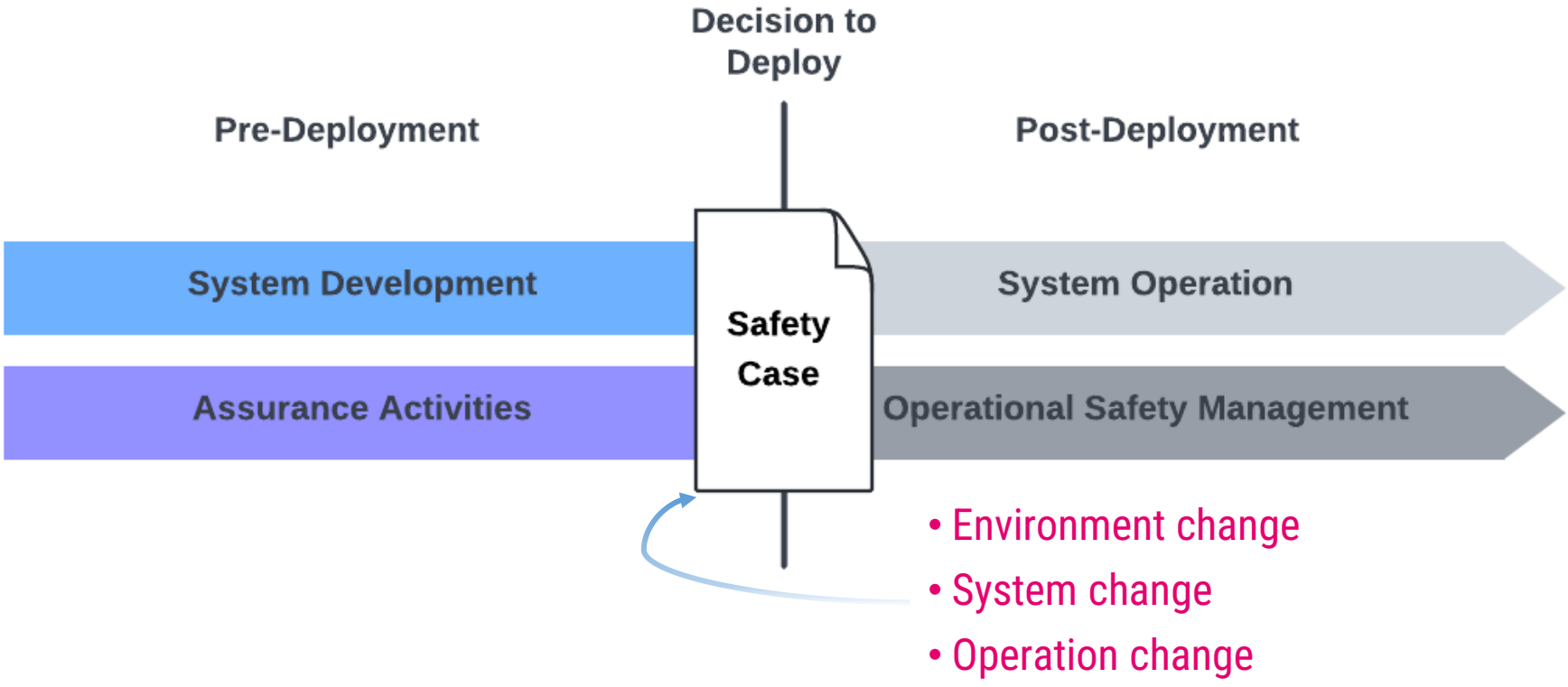
Safety Monitoring for AS





Safety Monitoring for AS

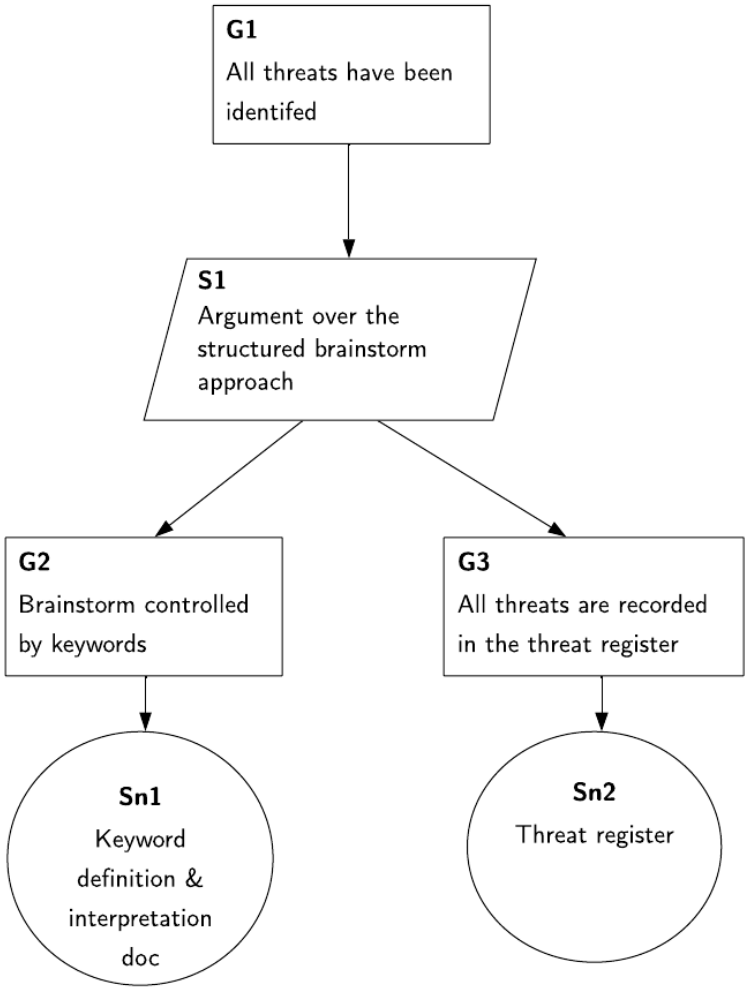




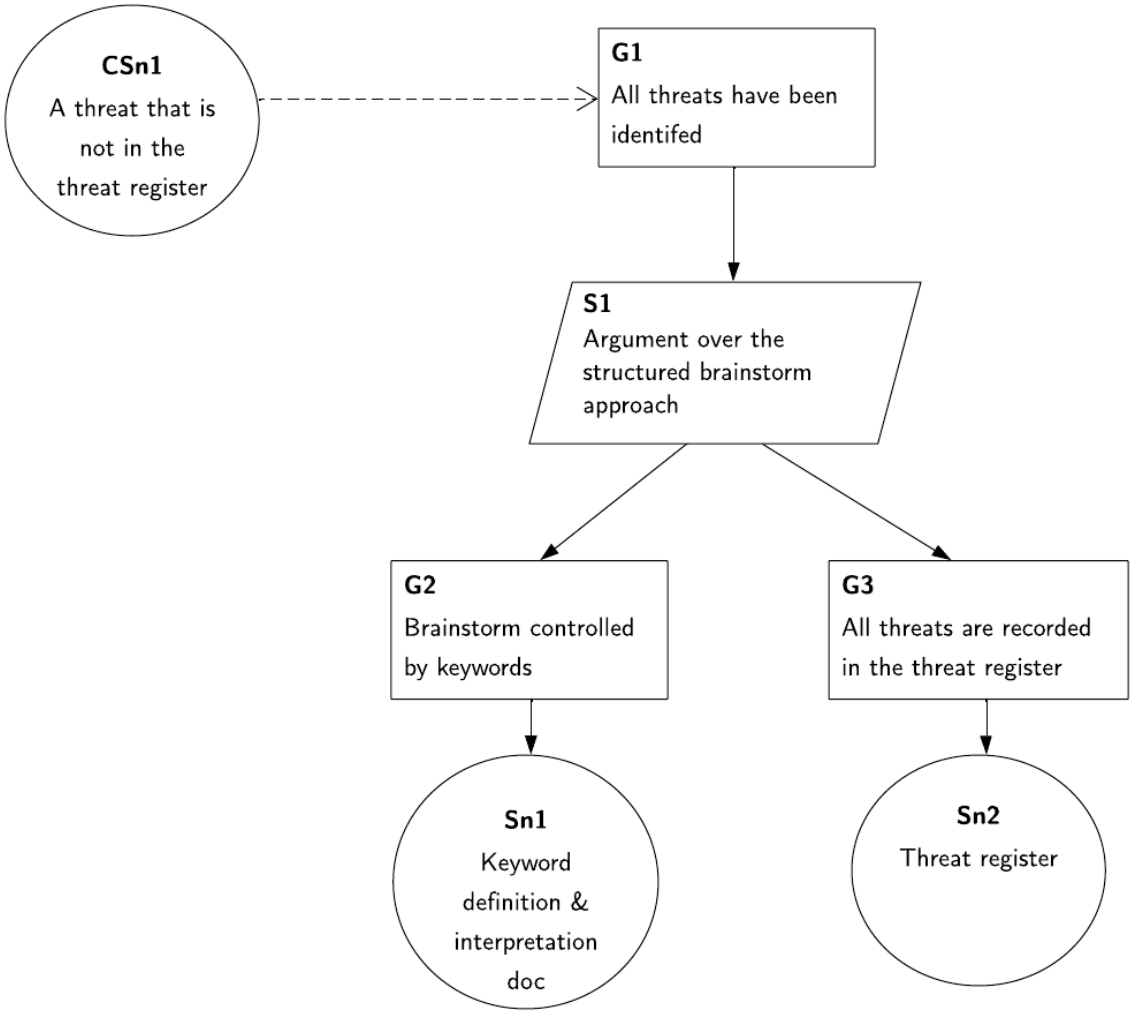
Our approach

- We currently rely heavily on engineering judgement to define monitoring requirements for AS
  - Difficult to justify the sufficiency of the monitoring
- Our approach uses an explicit analysis of the pre-deployment safety case to systematically identify run-time monitoring requirements
- Advantages of this approach
  - A) systematic
  - B) provides a way to justify the sufficiency of those monitoring requirements
  - C) Helps to distinguish real safety measures from performance measures
    - Correlation between metric and system-level safety of AS
- Based around the use of ***dialectic arguments***

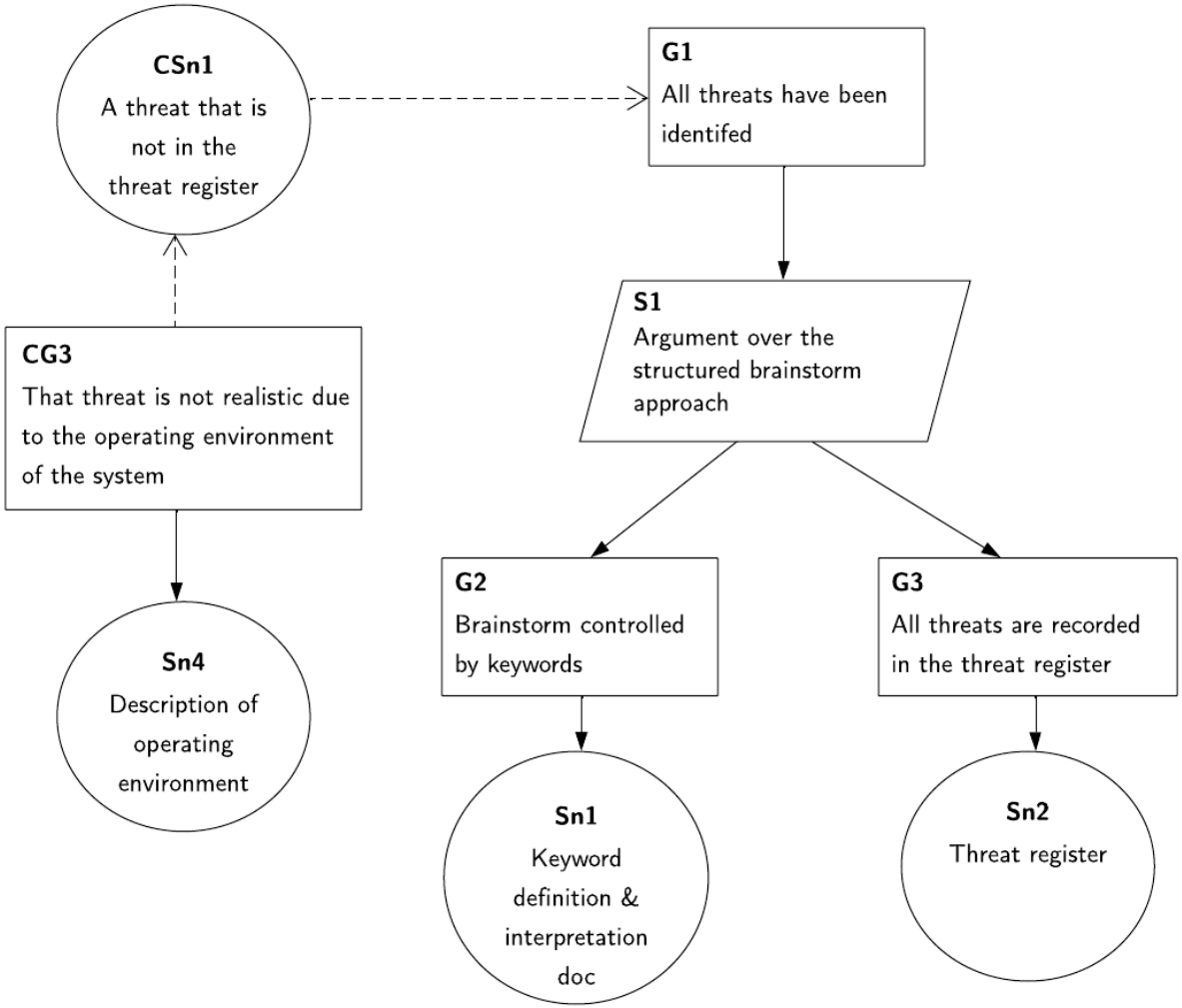
Dialectics


G1
All threats have been identifed
S1
Argument over the structured brainstorm approach
G2
Brainstorm controlled by keywords
G3
All threats are recorded in the threat register
Sn1
Keyword definition & interpretation doc
Sn2
Threat register

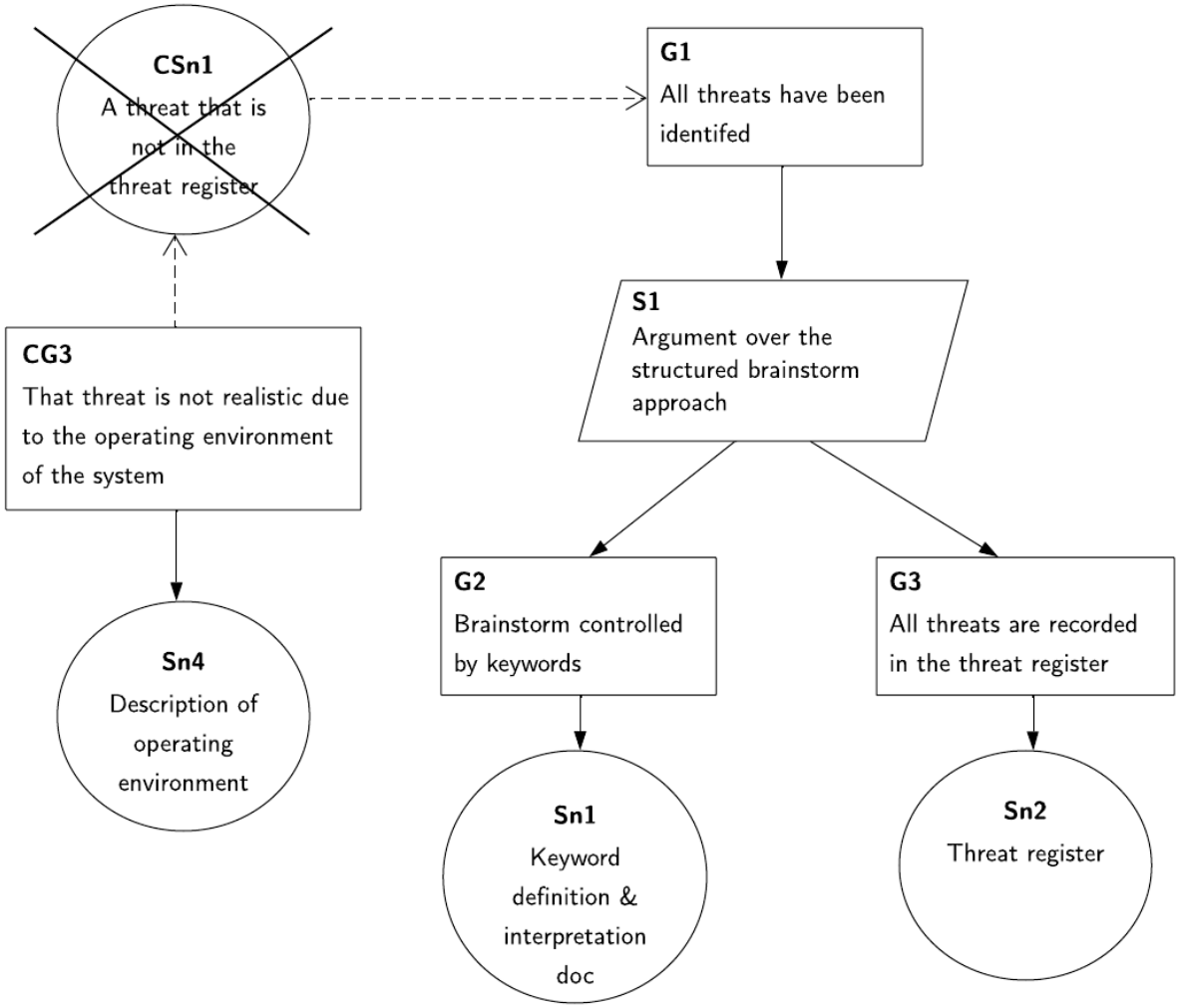
Dialectics


CSn1
A threat that is not in the threat register
G1
All threats have been identifed
S1
Argument over the structured brainstorm approach
G2
Brainstorm controlled by keywords
G3
All threats are recorded in the threat register
Sn1
Keyword definition & interpretation doc
Sn2
Threat register

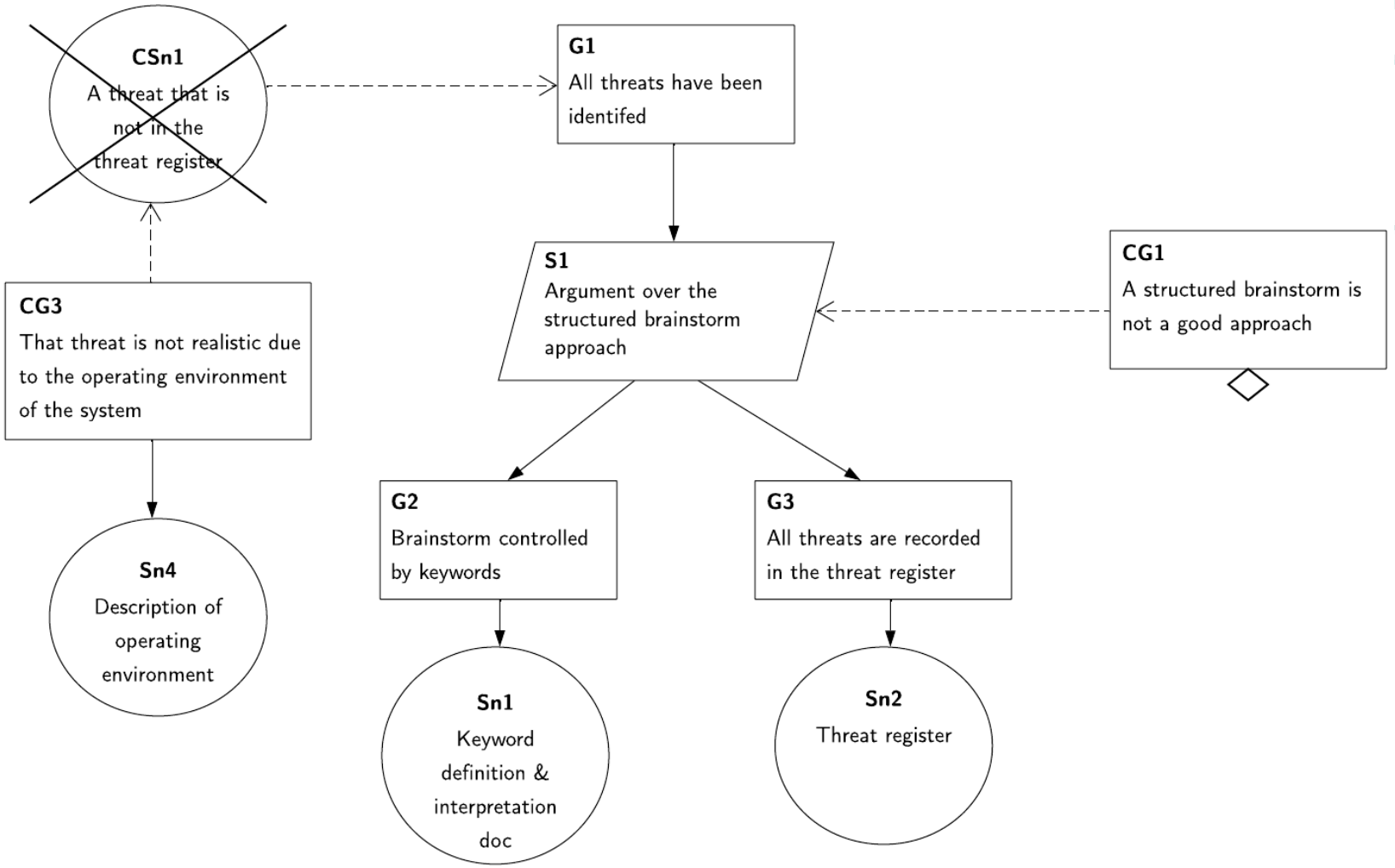
Dialectics


CSn1
A threat that is not in the threat register
G1
All threats have been identifed
S1
Argument over the structured brainstorm approach
CG3
That threat is not realistic due to the operating environment of the system
Sn4
Description of operating environment
G2
Brainstorm controlled by keywords
G3
All threats are recorded in the threat register
Sn1
Keyword definition & interpretation doc
Sn2
Threat register

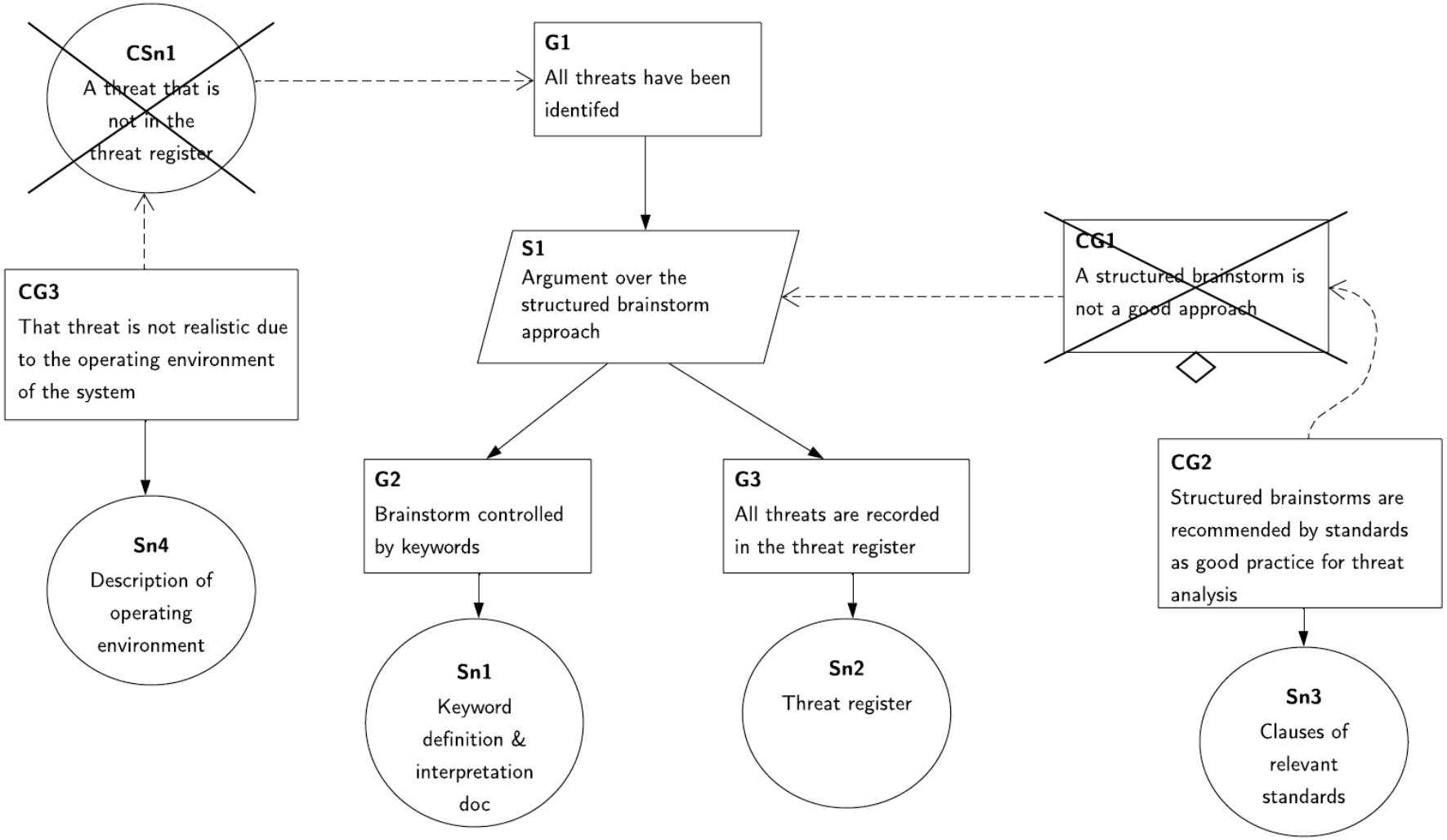
Dialectics


CSn1
A threat that is not in the threat register
G1
All threats have been identifed
CG3
That threat is not realistic due to the operating environment of the system
S1
Argument over the structured brainstorm approach
Sn4
Description of operating environment
G2
Brainstorm controlled by keywords
G3
All threats are recorded in the threat register
Sn1
Keyword definition & interpretation doc
Sn2
Threat register

Dialectics


CSn1
A threat that is not in the threat register
G1
All threats have been identifed
S1
Argument over the structured brainstorm approach
CG1
A structured brainstorm is not a good approach
CG3
That threat is not realistic due to the operating environment of the system
Sn4
Description of operating environment
G2
Brainstorm controlled by keywords
G3
All threats are recorded in the threat register
Sn1
Keyword definition & interpretation doc
Sn2
Threat register

Dialectics

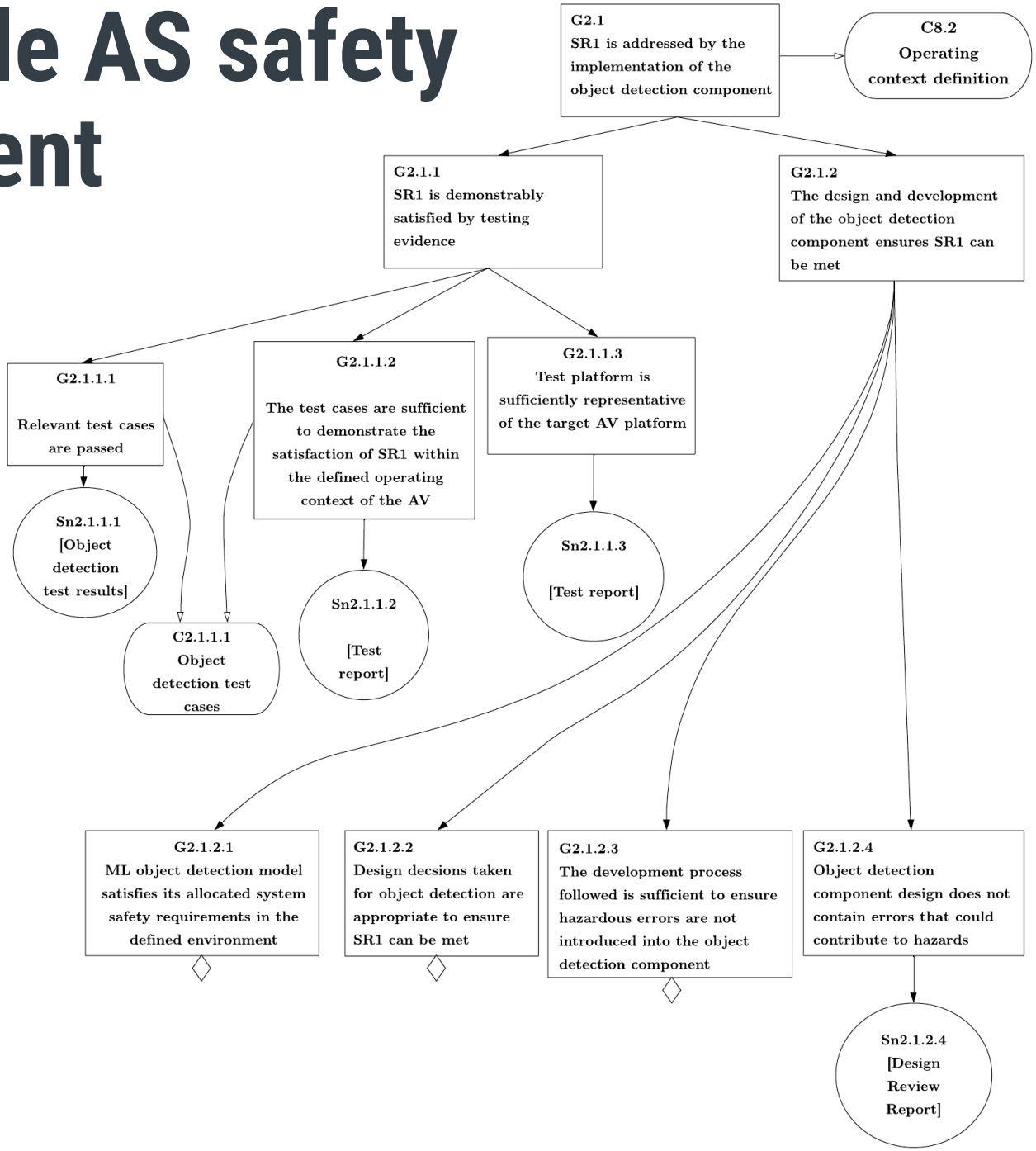

CSn1
A threat that is not in the threat register
G1
All threats have been identifed
CG3
That threat is not realistic due to the operating environment of the system
S1
Argument over the structured brainstorm approach
CG1
A structured brainstorm is not a good approach
Sn4
Description of operating environment
G2
Brainstorm controlled by keywords
G3
All threats are recorded in the threat register
CG2
Structured brainstorms are recommended by standards as good practice for threat analysis
Sn1
Keyword definition & interpretation doc
Sn2
Threat register
Sn3
Clauses of relevant standards

Operational Dialectic Argument

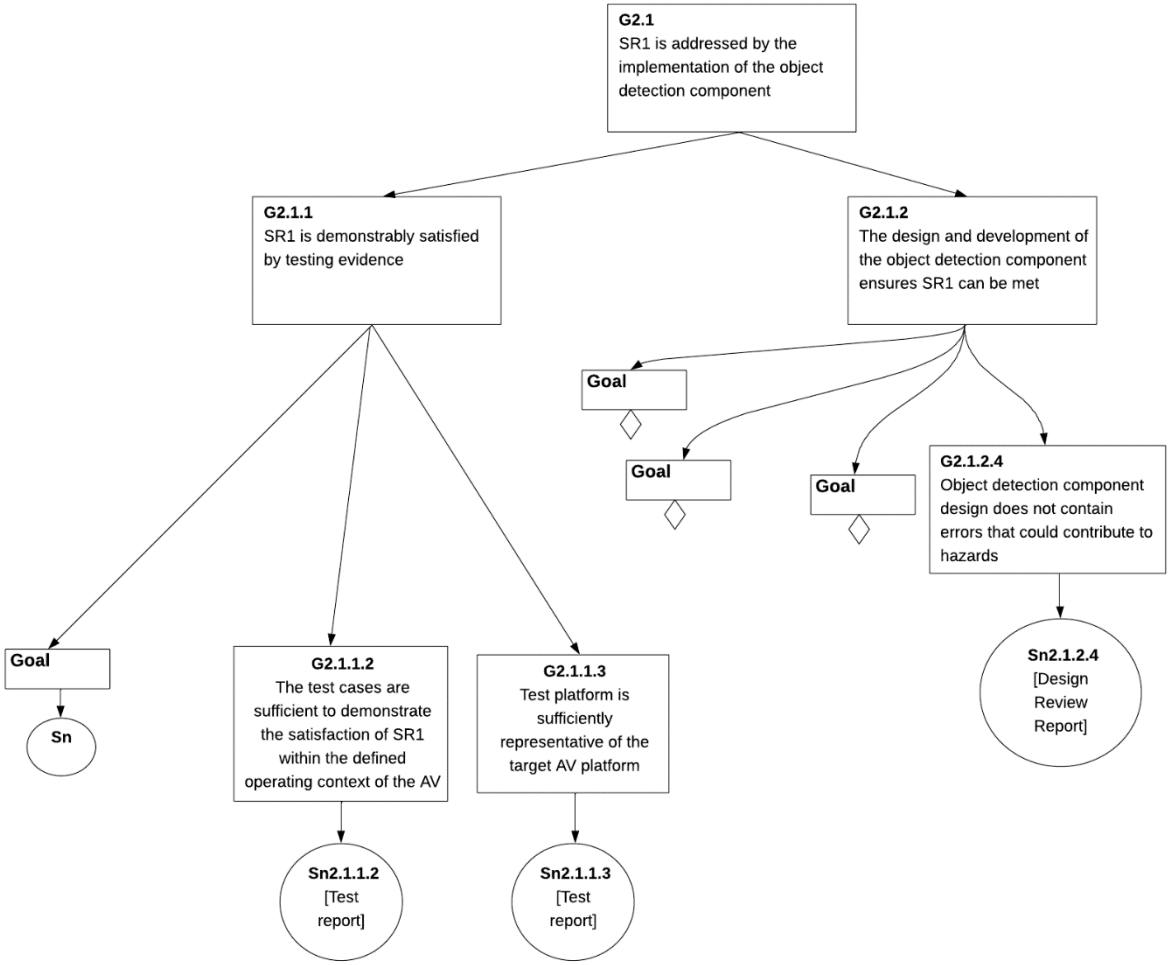
A systematic identification of ***potential*** run-time challenges to elements of the safety case.

- Prior to deployment these challenges are ***hypothetical***
- However, if the counter-evidence becomes present during operation that challenge becomes valid
- So we must have sufficient monitoring for that counter-evidence
  - This must be put in place prior to deployment of the AS
  - Otherwise the system may be unsafe without system operator realising it
- The starting point is the AS safety case itself…

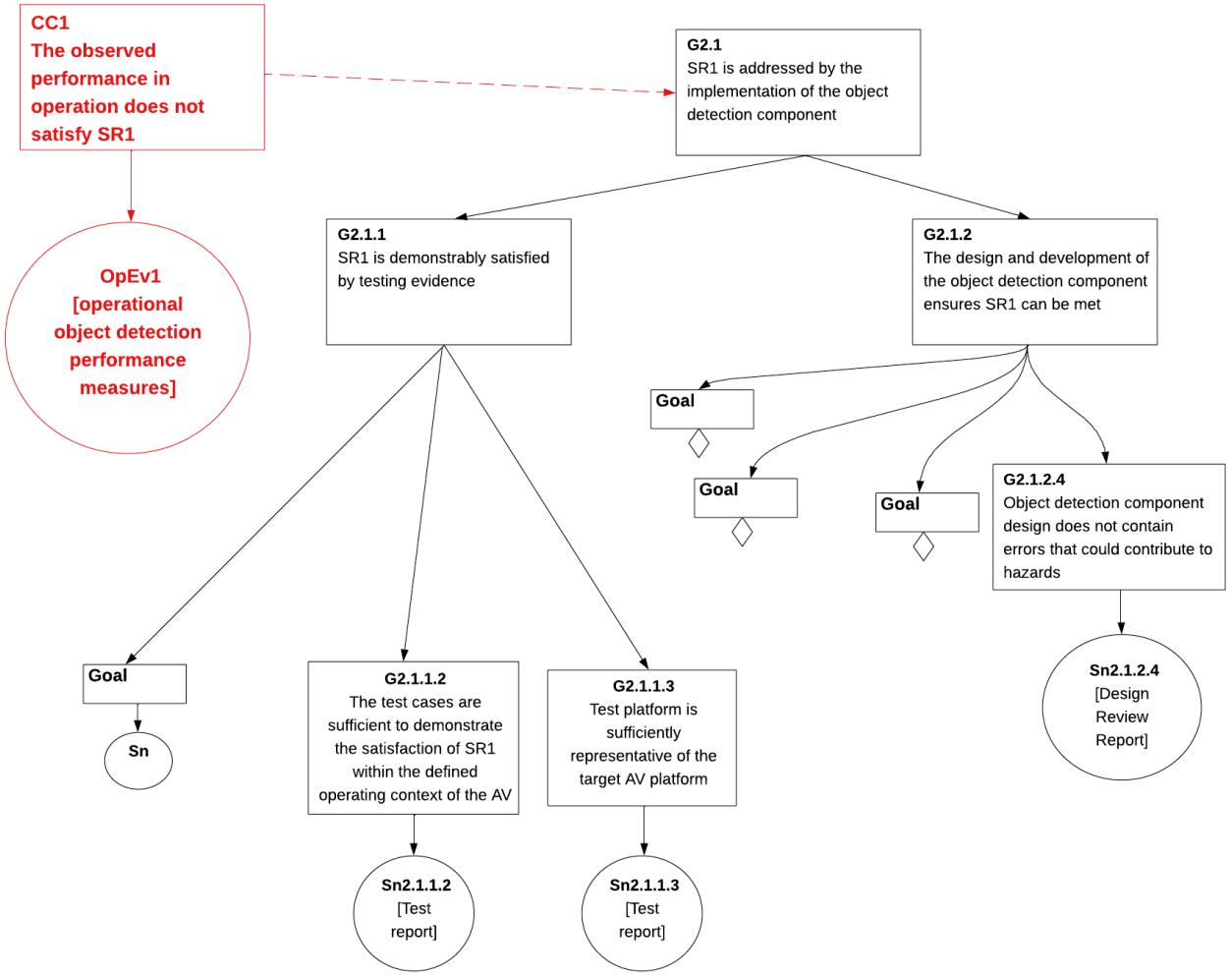
Example AS safety argument



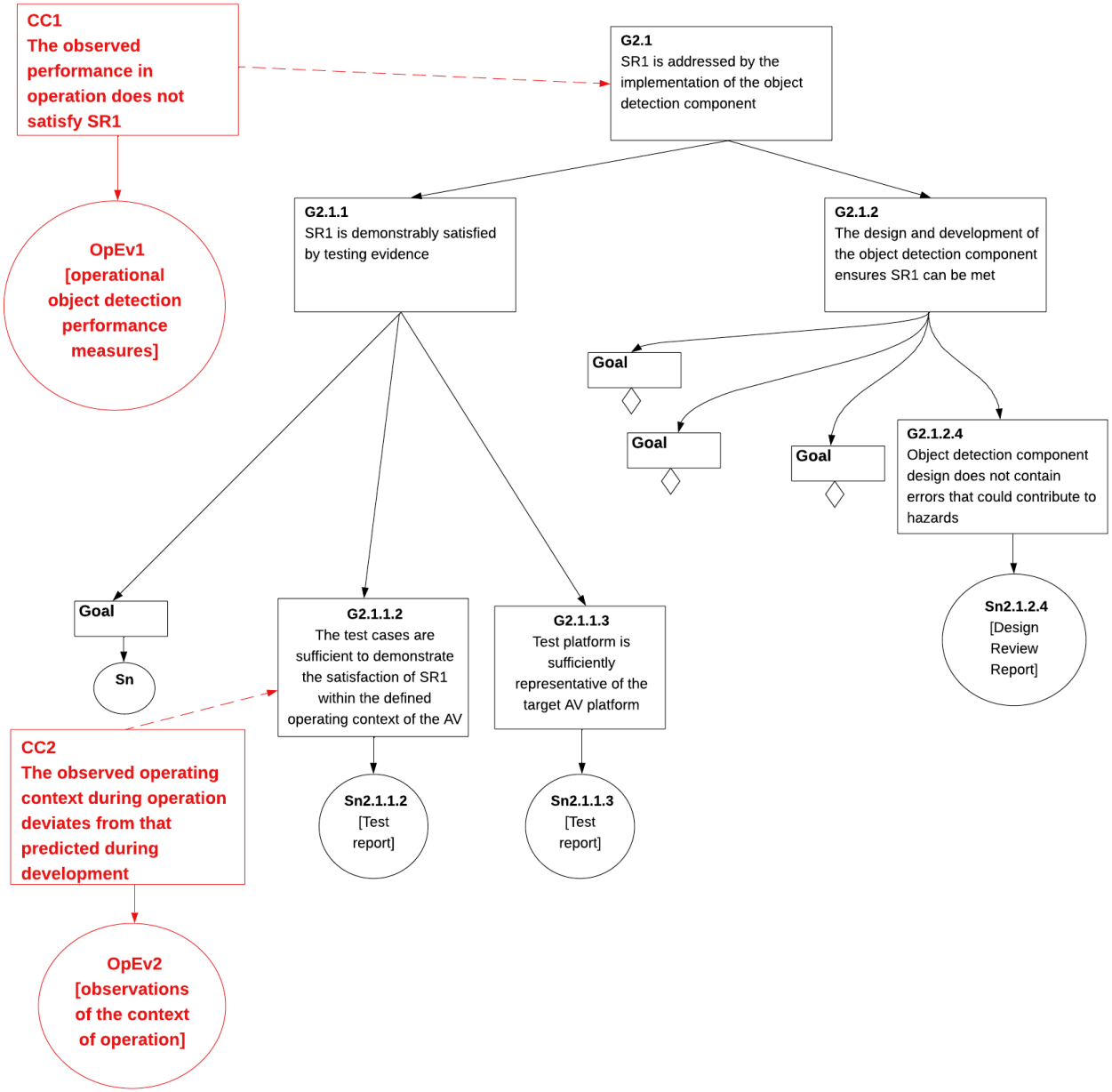
Example Operational Dialectics



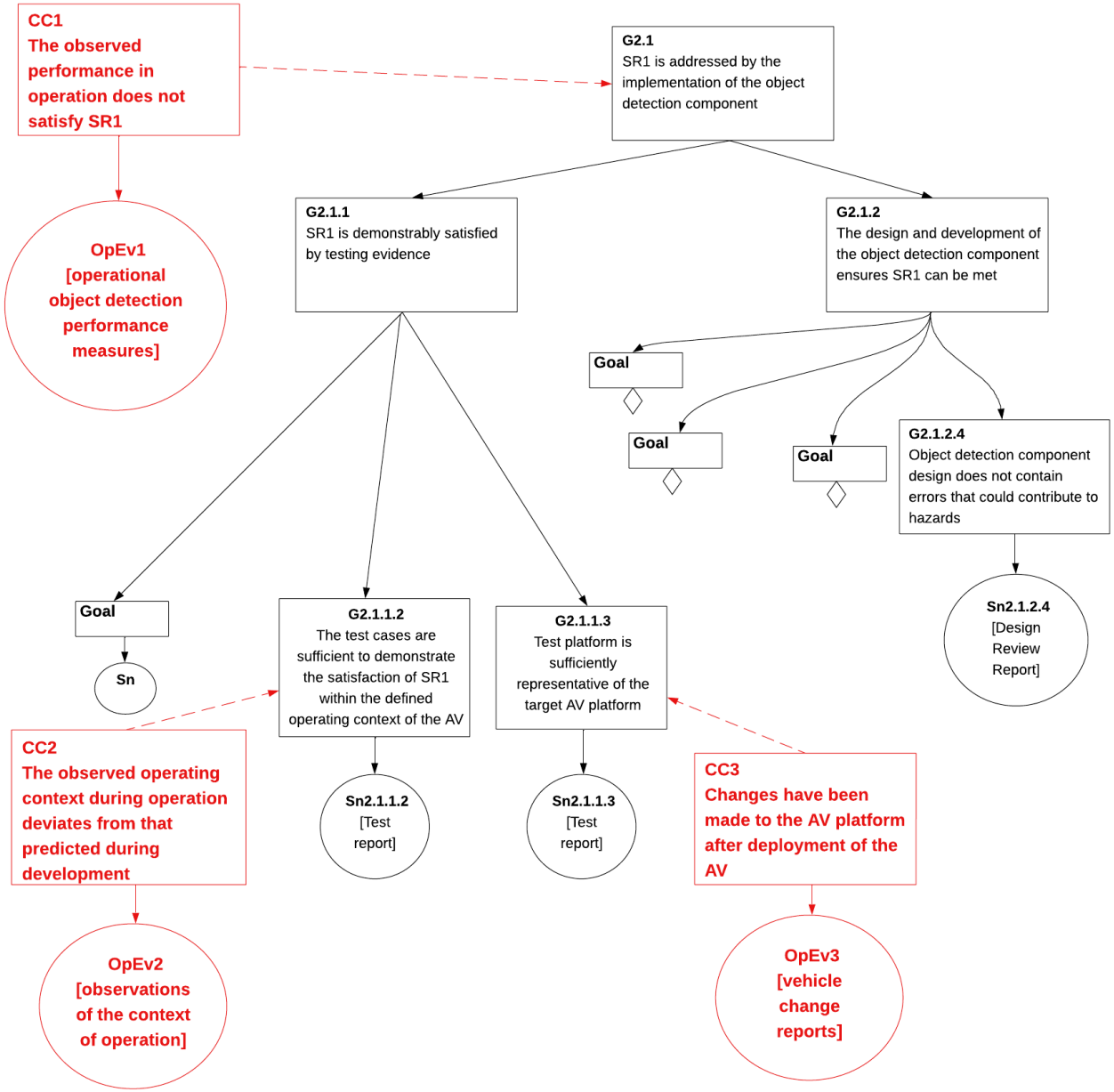
Example Operational Dialectics



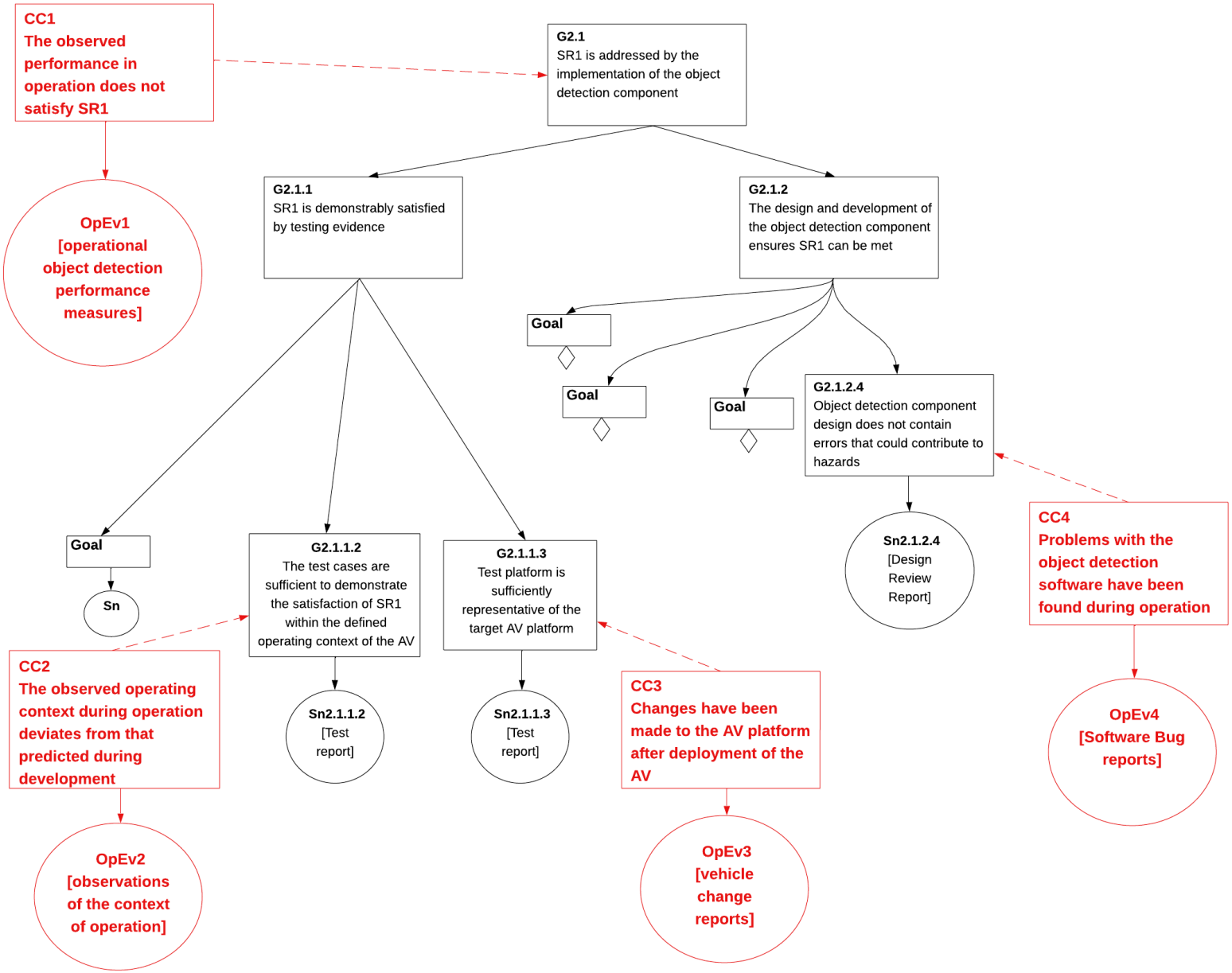
Example Operational Dialectics



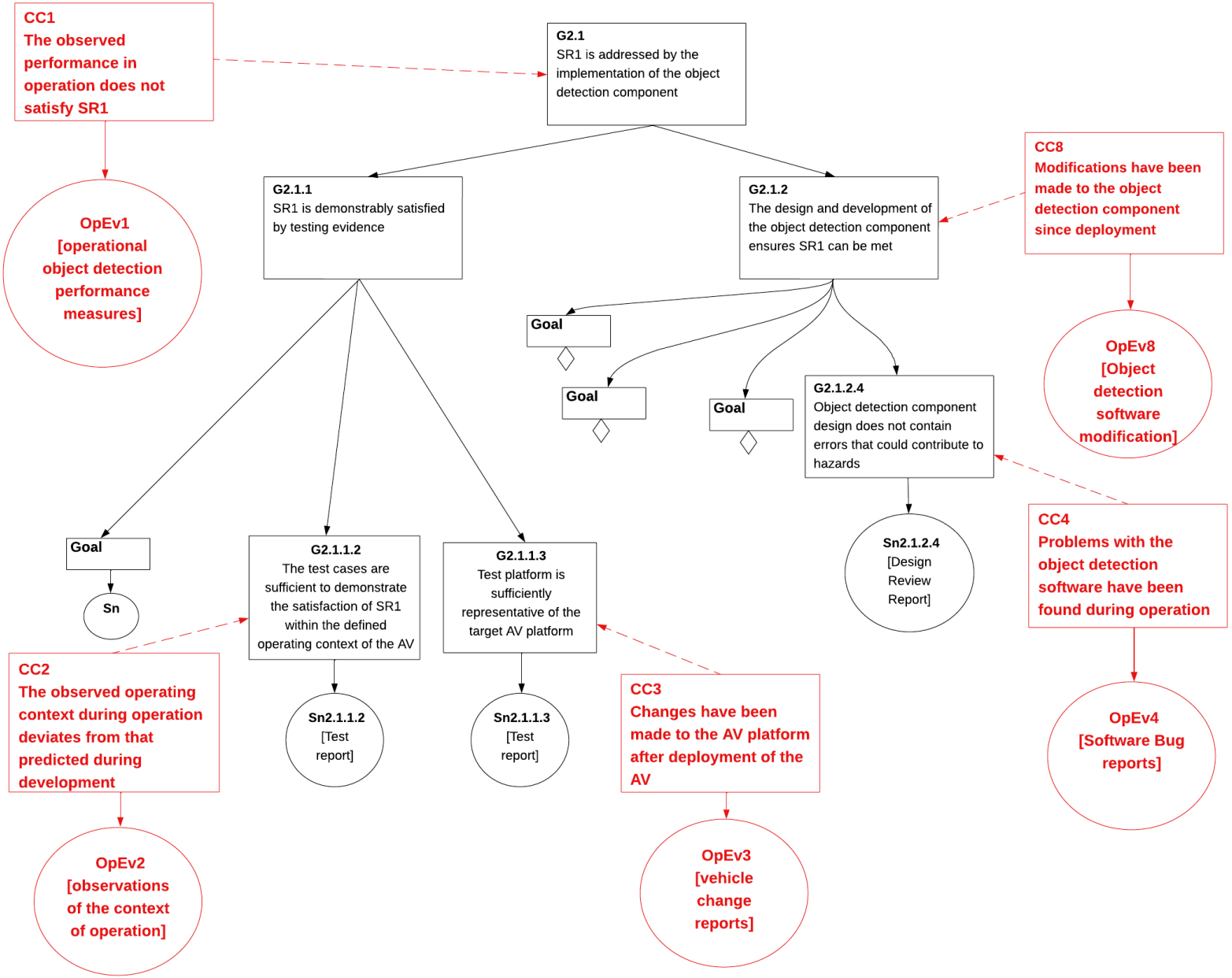
Example Operational Dialectics



Example Operational Dialectics



Example Operational Dialectics



Identifying Run-time Monitoring Requirements

- Based on the Operational Dialectic Argument we can define:
  - what needs to be monitored
    - System
    - Component
    - Process
    - Operation
  - How it can be measured
    - May require fleet-level aggregation
  - What is the trigger (threshold)

Example Monitoring Requirements

Op. Evidence	Monitor	Criteria	Trigger
OpEv1 - [operational object detection performance measures]	Number of missed pedestrian detections across the vehicle fleet	Missed detections observed per 1000 miles of operation	#misses/1000 miles exceeds rate reporting in test results by 10%
OpEv2 - [observations of the context of operation]	Input images arising from the camera for operation within defined ODD	Measurement of key parameters within images (e.g., light levels, surfaces, colours etc.)	Operational images outside of test distribution
OpEv3 - [vehicle change reports]	Physical changes to vehicle platform (such as updates to sensors, processors etc.)	Changes that may impact software performance	Notification of AV platform modification
OpEv4 - [Software bug report]	Software errors discovered during operation	Errors identified in object detection during operation	Notification of error found in object detection
OpEv5 - [AV incident reports]	Reports raised by operators of the vehicle	Incidents that relate to object detection	Notification of object detection incidents that may be hazardous
OpEv6 - [Camera maintenance records]	Calibration of camera	Time since last calibration	Greater than 6 months since last calibration
OpEv7 - [Camera drift measurements]	Drift measurement of camera images	Rate of drift in operation	Rate of drifting exceeds design assumption
OpEv8 - [Object detection software update]	Software version	Change to object detection software	Non-approved version of software running
OpEv9 - [Lidar error status]	Lidar health monitoring	Lidar availability	Lidar fails to provide output to object detection component

Example Monitoring Requirements

Op. Evidence	Monitor	Criteria	Trigger
OpEv1 - [operational object detection performance measures]	Number of missed pedestrian detections across the vehicle fleet	Missed detections observed per 1000 miles of operation	#misses/1000 miles exceeds rate reporting in test results by 10%
OpEv2 - [observations of the context of operation]	Input images arising from the camera for operation within defined ODD	Measurement of key parameters within images (e.g., light levels, surfaces, colours etc.)	Operational images outside of test distribution
OpEv3 - [vehicle change reports]	Physical changes to vehicle platform (such as updates to sensors, processors etc.)	Changes that may impact software performance	Notification of AV platform modification
OpEv4 - [Software bug report]	Software errors discovered during operation	Errors identified in object detection during operation	Notification of error found in object detection
OpEv5 - [AV incident reports]	Reports raised by operators of the vehicle	Incidents that relate to object detection	Notification of object detection incidents that may be hazardous
OpEv6 - [Camera maintenance records]	Calibration of camera	Time since last calibration	Greater than 6 months since last calibration
OpEv7 - [Camera drift measurements]	Drift measurement of camera images	Rate of drift in operation	Rate of drifting exceeds design assumption
OpEv8 - [Object detection software update]	Software version	Change to object detection software	Non-approved version of software running
OpEv9 - [Lidar error status]	Lidar health monitoring	Lidar availability	Lidar fails to provide output to object detection component

Component; Multi-vehicle

Issues:

- How do we know there's been a missed detection?
- How does the data get shared and with whom?

Example Monitoring Requirements

Op. Evidence	Monitor	Criteria	Trigger
OpEv1 - [operational object detection performance measures]	Number of missed pedestrian detections across the vehicle fleet	Missed detections observed per 1000 miles of operation	#misses/1000 miles exceeds rate reporting in test results by 10%
OpEv2 - [observations of the context of operation]	Input images arising from the camera for operation within defined ODD	Measurement of key parameters within images (e.g., light levels, surfaces, colours etc.)	Operational images outside of test distribution
OpEv3 - [vehicle change reports]	Physical changes to vehicle platform (such as updates to sensors, processors etc.)	Changes that may impact software performance	Notification of AV platform modification
OpEv4 - [Software bug report]	Software errors discovered during operation	Errors identified in object detection during operation	Notification of error found in object detection
OpEv5 - [AV incident reports]	Reports raised by operators of the vehicle	Incidents that relate to object detection	Notification of object detection incidents that may be hazardous
OpEv6 - [Camera maintenance records]	Calibration of camera	Time since last calibration	Greater than 6 months since last calibration
OpEv7 - [Camera drift measurements]	Drift measurement of camera images	Rate of drift in operation	Rate of drifting exceeds design assumption
OpEv8 - [Object detection software update]	Software version	Change to object detection software	Non-approved version of software running
OpEv9 - [Lidar error status]	Lidar health monitoring	Lidar availability	Lidar fails to provide output to object detection component

Process

Issues:

- How can we be sure this happens?
- Who is responsible for checking?

Example Monitoring Requirements

Op. Evidence	Monitor	Criteria	Trigger
OpEv1 - [operational object detection performance measures]	Number of missed pedestrian detections across the vehicle fleet	Missed detections observed per 1000 miles of operation	#misses/1000 miles exceeds rate reporting in test results by 10%
OpEv2 - [observations of the context of operation]	Input images arising from the camera for operation within defined ODD	Measurement of key parameters within images (e.g., light levels, surfaces, colours etc.)	Operational images outside of test distribution
OpEv3 - [vehicle change reports]	Physical changes to vehicle platform (such as updates to sensors, processors etc.)	Changes that may impact software performance	Notification of AV platform modification
OpEv4 - [Software bug report]	Software errors discovered during operation	Errors identified in object detection during operation	Notification of error found in object detection
OpEv5 - [AV incident reports]	Reports raised by operators of the vehicle	Incidents that relate to object detection	Notification of object detection incidents that may be hazardous
OpEv6 - [Camera maintenance records]	Calibration of camera	Time since last calibration	Greater than 6 months since last calibration
OpEv7 - [Camera drift measurements]	Drift measurement of camera images	Rate of drift in operation	Rate of drifting exceeds design assumption
OpEv8 - [Object detection software update]	Software version	Change to object detection software	Non-approved version of software running
OpEv9 - [Lidar error status]	Lidar health monitoring	Lidar availability	Lidar fails to provide output to object detection component

Operation

Issues:

- How are the notifications generated
- Is it always obvious which incidents are relevant?

Post-deployment

What happens when a trigger occurs?

- This represents a “live challenge” in the safety case
  - E.g. OpEv1 - No. of missed pedestrian detections per 1000 miles is higher than was claimed in the safety case
- Are there any possible rebuttals to the challenge
- What should the response be?
- Must identify responsible organisations and create processes to track and review monitors and triggers
  - ***the effectiveness of these also needs to be justified in the safety case***

Conclusions

- Its imperative for safe operation of AS that we monitor for when things go wrong
  - Specifically we need to know that the safety case has not become invalid
- This requires that we can demonstrate that
  - we understand what will challenge validity of the safety case
  - We have sufficient monitoring in place for those things
- Monitors only have value for safety assurance if we can show that we are monitoring ***all of the right things***
- Our approach enables systematic identification of monitoring requirements from analysis of the safety argument
  - This allows us to argue about the sufficiency of the monitoring

ASSURING AUTONOMY

INTERNATIONAL PROGRAMME

Funded by

www.york.ac.uk/assuring-autonomy/