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

Pre-Deployment

System Development

Assurance Activities

Decision to Deploy

Safety Case
Safety Monitoring for AS

Pre-Deployment

System Development

Assurance Activities

Safety Case

Decision to Deploy

Post-Deployment

System Operation

Operational Safety Management
Safety Monitoring for AS

- Environment change
- System change
- Operation change

Pre-Deployment

System Development

Assurance Activities

Decision to Deploy

Safety Case

Post-Deployment

System Operation

Operational Safety Management

- Environment change
- System change
- Operation change
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 identified

S1
Argument over the structured brainstorm approach

G2
Brainstorm controlled by keywords

Sn1
Keyword definition & interpretation doc

G3
All threats are recorded in the threat register

Sn2
Threat register
Dialectics

CSn1
A threat that is not in the threat register

G1
All threats have been identified

S1
Argument over the structured brainstorm approach

G2
Brainstorm controlled by keywords
Sn1
Keyword definition & interpretation doc

G3
All threats are recorded in the threat register
Sn2
Threat register
Dialectics

CSn1
A threat that is not in the threat register

CG3
That threat is not realistic due to the operating environment of the system

Sn4
Description of operating environment

G1
All threats have been identified

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

CG3
That threat is not realistic due to the operating environment of the system

CSn1
A threat that is not in the threat register

G1
All threats have been identified

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

Sn4
Description of operating environment
Dialectics

CSn1
A threat that is not in the threat register

CG3
That threat is not realistic due to the operating environment of the system

Sn4
Description of operating environment

G1
All threats have been identified

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

CG1
A structured brainstorm is not a good approach
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

G2.1
SR1 is addressed by the implementation of the object detection component

C8.2
Operating context definition

G2.1.1
SR1 is demonstrably satisfied by testing evidence

G2.1.2
The design and development of the object detection component ensures SR1 can be met

G2.1.3
Test platform is sufficiently representative of the target AV platform

G2.1.2.1
ML object detection model satisfies its allocated system safety requirements in the defined environment

G2.1.2.2
Design decisions taken for object detection are appropriate to ensure SR1 can be met

G2.1.2.3
The development process followed is sufficient to ensure hazardous errors are not introduced into the object detection component

G2.1.2.4
Object detection component design does not contain errors that could contribute to hazards

Sn2.1.1.1
Object detection test results

Sn2.1.1.2
Test report

Sn2.1.3
Test report

C2.1.1.1
Object detection test cases

Sn2.1.1.3
[Object detection test results]
Example Operational Dialectics

G2.1
SR1 is addressed by the implementation of the object detection component

G2.1.1
SR1 is demonstrably satisfied by testing evidence

G2.1.2
The design and development of the object detection component ensures SR1 can be met

G2.1.2.4
Object detection component design does not contain errors that could contribute to hazards

G2.1.3
Test platform is sufficiently representative of the target AV platform

Sn2.1.2.4
[Design Review Report]

Sn2.1.1.2
The test cases are sufficient to demonstrate the satisfaction of SR1 within the defined operating context of the AV

Sn2.1.1.3
[Test report]

Sn
Example Operational Dialectics

CC1
The observed performance in operation does not satisfy SR1

OpEv1
[operational object detection performance measures]

G2.1
SR1 is addressed by the implementation of the object detection component

G2.1.1
SR1 is demonstrably satisfied by testing evidence

G2.1.2
The design and development of the object detection component ensures SR1 can be met

G2.1.2.4
Object detection component design does not contain errors that could contribute to hazards

Sn

Sn2.1.1.2 [Test report]

Sn2.1.1.3 [Test report]

G2.1.1.2
The test cases are sufficient to demonstrate the satisfaction of SR1 within the defined operating context of the AV

G2.1.1.3
Test platform is sufficiently representative of the target AV platform

Sn2.1.2.4
[Design Review Report]
Example Operational Dialectics

CC1
The observed performance in operation does not satisfy SR1

OpEv1
[operational object detection performance measures]

Goal

Sn

CC2
The observed operating context during operation deviates from that predicted during development

OpEv2
[observations of the context of operation]

G2.1
SR1 is addressed by the implementation of the object detection component

G2.1.1
SR1 is demonstrably satisfied by testing evidence

G2.1.2
The design and development of the object detection component ensures SR1 can be met

G2.1.3
Test platform is sufficiently representative of the target AV platform

G2.1.4
Object detection component design does not contain errors that could contribute to hazards

Sn2.1.2
[Test report]

Sn2.1.3
[Test report]

Sn2.1.4
[Design Review Report]
Example Operational Dialectics

CC1
The observed performance in operation does not satisfy SR1

OpEv1
[operational object detection performance measures]

Goal

Sn

G2.1
SR1 is addressed by the implementation of the object detection component

G2.1.1
SR1 is demonstrably satisfied by testing evidence

Goal

G2.1.2
The test cases are sufficient to demonstrate the satisfaction of SR1 within the defined operating context of the AV

Sn2.1.2
[Test report]

Sn2.1.3
[Test report]

G2.1.3
Test platform is sufficiently representative of the target AV platform

Goal

G2.1.4
Object detection component design does not contain errors that could contribute to hazards

Sn2.1.4
[Design Review Report]

CC2
The observed operating context during operation deviates from that predicted during development

OpEv2
[observations of the context of operation]

G2.1.2
The design and development of the object detection component ensures SR1 can be met

CC3
Changes have been made to the AV platform after deployment of the AV

OpEv3
[vehicle change reports]
Example Operational Dialectics

CC1
The observed performance in operation does not satisfy SR1

G2.1
SR1 is addressed by the implementation of the object detection component

G2.1.1
SR1 is demonstrably satisfied by testing evidence

G2.1.2
The design and development of the object detection component ensures SR1 can be met

Sn2.1.2.4
Object detection component design does not contain errors that could contribute to hazards

Sn2.1.2.4
[Design Review Report]

Sn2.1.2.2
[Test report]

Sn2.1.3
[Test report]

Sn

CC2
The observed operating context during operation deviates from that predicted during development

OpEv2
[observations of the context of operation]

OpEv3
[vehicle change reports]

OpEv4
[Software Bug reports]

CC3
Changes have been made to the AV platform after deployment of the AV

CC4
Problems with the object detection software have been found during operation

OpEv1
[operational object detection performance measures]
Example Operational Dialectics

CC1
The observed performance in operation does not satisfy SR1

OpEv1
[operational object detection performance measures]

G2.1
SR1 is addressed by the implementation of the object detection component

G2.1.1
SR1 is demonstrably satisfied by testing evidence

Goal

Sn

G2.1.2
The test cases are sufficient to demonstrate the satisfaction of SR1 within the defined operating context of the AV

G2.1.3
Test platform is sufficiently representative of the target AV platform

G2.1.2.4
Object detection component design does not contain errors that could contribute to hazards

Sn2.1.2.4
[Design Review Report]

CC8
Modifications have been made to the object detection component since deployment

OpEv8
[Object detection software modification]

OpEv3
[vehicle change reports]

OpEv4
[Software Bug reports]

CC2
The observed operating context during operation deviates from that predicted during development

OpEv2
[observations of the context of operation]

CC3
Changes have been made to the AV platform after deployment of the AV

Sn2.1.1.2
[Test report]

Sn2.1.1.3
[Test report]

CC4
Problems with the object detection software have been found during operation
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

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

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

### 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

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

### Process

**Issues:**
- How can we be sure this happens?
- Who is responsible for checking?
## Example Monitoring Requirements

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

### 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

• It's 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