DSIWG

SCSC Data Safety Initiative – WG Meeting 78

26th July 2023, Zoom

Minutes

Attendees

Mike Parsons (MP) – AAIP, Dale Callicott (DC) – BAE Systems, Dave Banham (DB) – Blackberry, Paul Hampton (PH) – CGI, Nick Hales (NH) – Consultant, Mike Standish (MS) – Dstl, Michael Green (MG) – Ecomergy, Paolo Giuliani (PG) – EDF, Arch McKinlay (AM) – NGA.

Apologies

Tim Rowe (TR) – Consultant, Jennifer Kracht (JK) – TomTom, Brent Kimberley (BK) – Durham, Roland Rosier (RR) – TomTom, Martin Atkins (MA) – MCA, Oscar Slotosch (OS) – Validas

Agenda

1. Welcome
2. Fires in Europe
3. Email from MA
4. Update on Moon Landings from OS?
5. Data Affecting Data?
6. Email from MS
7. ML Data Hazard Analysis
8. Oldest Data Safety Error
9. Calculus of DSALs
10. DSITN (Data Safety in the News)
11. Update on Tooling
12. Actions
13. Next meeting
14. AOB

NOTE: All comments or opinions in these notes are attributed only to individual attendees of the meeting, not to their respective organisations.

*[Note that actions are presented in the form **N.Mx** where **N** is the meeting number, **M** a reference number for the action raised in that meeting and **x** is an optional letter that differentiates related actions arising from the same discussion point].*

The meeting slides are available at: https://scsc.uk/file/gd/78th_DSIWG_Slides_v1-1599.pptx

1. Welcome

MP opened the meeting and welcomed those attending.

2. Fires in Europe

MP noted the ongoing wildfire situation in Europe, with many fires burning out of control. He showed there was a significant data contribution. He showed the monitoring site: https://effis.jrc.ec.europa.eu/apps/effis_current_situation/ which is very data-intensive.

PH mentioned that CGI had done related work. He said that the information provided (including predictions on fire spread which may cause evacuations) was given with strong legal disclaimers. He noted this was different to the UK radiation monitoring network, RIMNET (which has now been replaced), https://www.gov.uk/government/collections/radioactive-incident-monitoring

3. Email from MA

MP reported on an email received from MA asking if the guidance covers the appropriate choice of representation [for data] for the operations that are going to be done on it?

MP said that this an issue and we may need to add Format, Resolution and Accuracy too as operation-dependent aspects. It was noted that Phil Koopman had highlighted a problem with rounding errors in currency transactions, where a different representation should have been used: https://www.linkedin.com/posts/philip-koopman-0631a4116_the-extra-0000000000000003-on-electric-activity-7081975605478453248-WJw9?utm_source=share&utm_medium=member_android

ACTION 78.1 (MP/MA) – Investigate representation / format issues further. Consider adding another issue to the next version of the guidance on representation.

4. Update on Moon Landings from OS?

OS was not present so this was not discussed.
[Note subsequent email from OS added the following: *I checked the Lunar lander story: basically, they changed the landing position to a place they did not simulate. Then it turned out that the surface there was too steep for their software and it wrongly concluded the height sensor must be defective…If they had simulated this place too, they probably would have detected and avoided it.*]

5. Data Affecting Data

MP said that the issue of data-data interactions is not really addressed in the guidance, i.e. how one data item might affect another, possibly after a failure.

This is clear in calculations, but there are more complex situations, e.g.

- A bad config parameter data item causes other data items to be unused/unselected
- Tuning/filter/selection data settings are incorrect (e.g. bad date range) causing the wrong data to be selected or used
- Tags or metadata are incorrect or lost causing the original data to be unused/wrongly used
- Noisy or bad data causes correct data to be rejected (Dazzle Data: Guidance section L.3.2, 3.3, 3.4, etc.)
- Sentinel values are incorrect causing wrong data to be selected for onward use.

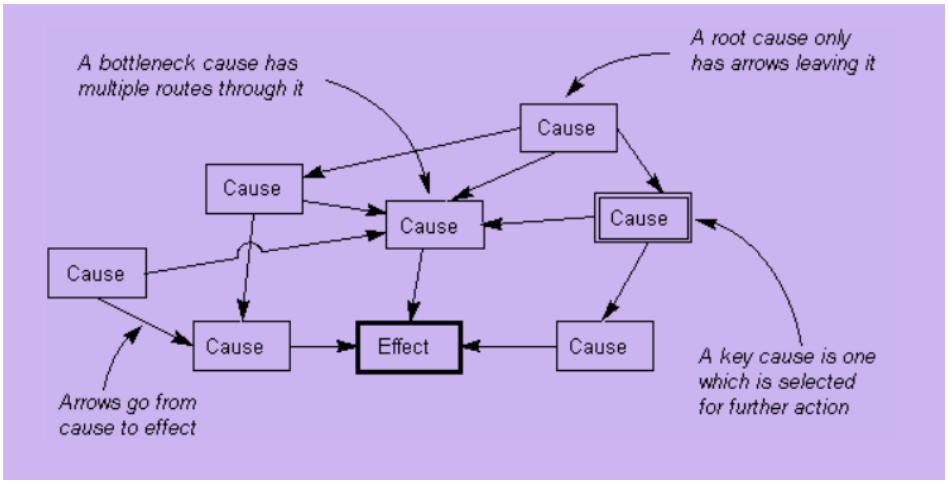
There was general agreement that this could be an issue and should be investigated further.

PH said that the 'sphere of influence' of the data should be considered, i.e. what other data items one faulty item could affect and how to limit any damage. He also said there were often protection measures in place, e.g. parity bits or error correcting codes which could limit any effects.

DC said some data may be hard to encapsulate to protect it, and that more mitigations or tests may be needed to detect the problem.

PG said that there may chains of data with subsequent knock-on effects. It is necessary to think of the origin of the data.

DB said it would be useful to have analysis methods for this problem. He though data-data dependency analysis might be a good place to start. He thought the diagram [that MP found on the web] was interesting to study:



NH asked if there was a notation for a data-data dependency model. DC said that fault tree analysis might be used. MP suggested that a data flow diagram is useful.

ACTION 78.2 (MP) – Investigate data-data interactions further.

6. Email from MS

MS reported an example of a data safety issue with the email typo between the US military '.mil' domain and that of Mali's '.ml' causing thousands of emails over many years to go to the wrong destination: https://www.bbc.co.uk/news/world-us-canada-66226873

7. ML Data Hazard Analysis

MP reported that an academic paper (via John McDermid) reported an ML-data FMEA technique: https://jsystemsafety.com/index.php/jss/article/view/253/235
However he said the description in the paper was very brief.

Several of the data issues from the guidance were thought relevant to large data sets used in ML, e.g.

- Reuse
- Ageing

- Ownership
- Falsification
- Masking

In general the source of the training data is critical - whether acquired in the real world or done by simulation, etc. For the FMEA it is suggested to take the properties defined in the guidance and negate them to get a more general set of hazards, so, using Table 5 and Appendix F, e.g:

- The property Traceability could have failure modes: loss of tracing information, incorrect tracing information, etc.
- The property Timeliness could have failure modes: too soon, too late, missing, etc.

Note that the data FMEA for ML data is going to be different - it is going to be very hard to see what effects particular defects in the data have on the resulting ML model, and therefore what safety impact they will have.

He thought it would be good to have a data FMEA example in the next guidance issue. Existing action 43.4 covers this.

MG mentioned an example of data issues in healthcare where conditions were coded differently across different systems, and inconsistent labelling was used. The way this is mitigated currently is for clinicians to manually review data before incorporation.

8. Oldest Data Safety Error

MP said that the abstract (below) had been accepted for SSS'24 main programme and now volunteers were needed to help write the paper.
Any help is appreciated! Contact mike.parsons@scsc.uk

> *"Data-intensive systems are now around us everywhere and so the opportunities for data to cause accidents has increased. However such data safety errors are not new and research by members of the Data Safety Initiative Working Group has identified some important cases throughout history where data-related errors have had a major impact: from the recent Covid-19 spreadsheet silent loss of rows causing an estimated 1,500 deaths, the Gemini space mission where the landing site was miscalculated due to an error in the value used for the Earth's rotation rate, to the loss of sailing ships caused by longitude errors due to the lack of accurate time fixes, to early civilisations when pyramid dimensions were likely incorrect. This paper offers a timeline of data safety problems in systems and shows that they have been with humankind for millennia"*

As per the reviewer comments it was agreed that the paper would also include relevance to today's data problems and include mitigations where possible.

9. Calculus of DSALs

There was some further discussion on this. There was no obvious decomposition technique that works for a general case.

DB mentioned that in ISO 26262 the higher-level ASIL is also quoted with the lower-level ASIL for sub-components, so it can be seen where it came from.

10. DSITN

No specific data safety in the news other than those items already covered in earlier sections.

11. Update on Tooling

PH was currently using the RADISH tool[1] from MCA and has raised some comments based on a real example. The assistance provided by the tool was considered very useful.

12. Actions

MP noted that some old actions were now current again (43.4, 46.1).
77.3 was closed.
See table at end.

13. AOB

NH mentioned that the 1830 fatal accident with the 'Rocket' locomotive https://rainhilltrials.co.uk/william-huskisson-and-the-first-railway-tragedy/ was a good example of 'knock-on' where lots of cause-effect chained together.

DB asked if there was any update on standardization. MP replied that nothing had happened recently although several bodies had shown an interest, including the IEE and BSI. He said that the real issues are: (i) lack of resources: who would do the editing, communication, liaison, etc. with the standards body? and (ii) possible lack of control as to what happens with the work going forward. DB said that the OMG had a good model for creating standards. He said ISO 26262 Issue 3 will have more data aspects.

AM said that the ISSS system safety conference was coming up, https://system-safety.org/page/2023-schedule

AM presented some Data Lifecycle Overview slides which were discussed by the WG.

14. Next Meeting

The next meeting will be held 20th September via Zoom, https://scsc.uk/gd

15. Thanks

Thanks to all who provided contributions.
Thanks to MP for chairing and taking minutes.

Summary of Open Actions

Actions greyed out are considered closed and will be removed from the list at next issue.

Ref	Owner	Description	Target Guidance Version
42.9	MP	Work out a matrix of data categories (previously 'types') and data properties (as per DB discussion)	N/A
43.4	MP	Write up a data focussed FMEA approach.	4.0

[1] Note RADISH is "Risk Assessor for Data Integrity and Safety Hazards"

Ref	Owner	Description	Target Guidance Version
44.2	MP	To discuss with AK on how to get the Wikipedia article published	N/A
46.1	MP	Review the application of DSALs to higher level forms of aggregation	N/A
53.1	MP	To talk to Kevin King about what we need to do in the guidance for digital twins.	4.0
61.2	AW	Research the relevance of digital currencies and report back to the group (with MA and MT)	N/A
63.1	CT	Look at both Dark Data and Dazzle Data for sensors (e.g. when a sensor is saturated, in noisy environment or when readings are below the detection level floor)	4.0
69.3	PMcK	Develop a scoping diagram that shows how the DSG fits into the overall lifecycle process and other standards	4.0
69.4	MA	Write a short note on the issues of aggregation	4.0
69.6	MA/DA	Update the data safety tool to use the latest version of the guidance document	-
71.3	PH/DA/RO	Develop security properties thinking further for next DSIWG	4.0
71.4	PH/DA/RO	Present security properties work to next SISWG meeting	-
71.5	AM	(i) Establish if any of this can be published within the DSIWG and (ii) Consider a structuring similar to that used in security standards or ISO26262	-
71.7	MP/CT	Consider impact of FAIR data on the guidance	4.0
73.1	JK, RR	Consider production of a short note which could be used as an appendix to the guidance on lessons learnt using the guidance at TomTom	4.0
73.2	MT	Consider how the guidance fits with different lifecycles considering 'V', Continuous Service, Agile and 'Data Pipeline'	4.0
73.4	DA	See whether any of the DSITN entries from previous minutes and slides could be used to enhance the list of accidents	-
75.3	MT	Fix the minor typos that have been reported with v3.5 and prepare updates for both the online version and also the KDP/Amazon hardcopy.	3.5 Update
76.1	MP, PH, MT	Consider whether Guidance Annexes can be integrated into the main body of the document.	4.0
76.2	TR	Look further at IEC 25012 to consider whether a new Annex to the Guidance would be beneficial.	4.0
76.3	MT	Provide slides from last October's SQEPtember presentation (an introduction to data safety) to MP.	-
77.1	OS	Investigate this moon landing accident further	
77.2	MP, PH, MT	Review proposed updates and add to list of changes to version 4.0 of the guidance	4.0
77.3	MP	Send MS the slides from the Data Safety Evolution seminar	
78.1	MP/MA	Investigate representation / format issues further. Consider adding another issue to the next version of the guidance on representation.	4.0
78.2	MP	Investigate data-data interactions further.	4.0