WG114/G34 Plenary meeting: Plenary session on Explainability

Prepared by:

Guillaume Soudain
EASA Software Senior Expert / AI Project Manager

François Triboulet
EASA ATM/ANS Expert

Marta Llobet
ECTL ATM Safety Expert

Ramesh V. Anapathur
Boeing Technical Fellow

Steve Beland
Boeing Technical Fellow

8th July 2021
Objectives of the breakout session

→ Debrief from previous session (30 - max 45 minutes)
  → Explainability – Definitions
  → Explainability – Scope and needs
  → Explainability – Split among SGs

→ Enable discussion on (20 minutes each):
  → Topic 1 – Further discussion on scope of the WG for operational explainaibility
    → Operational explainability
    → Human and procedures
  → Topic 2 - Proportionality in Explainability
  → Topic 3 - System versus output level of explainability
Explainability: definitions

→ Definition from the WG114/G34 SOC

EXPLAINABILITY: The extent to which humans can understand, interpret, and account for the causality of an AI system or algorithm: why a particular output is produced for a given set of inputs. (Note: In this document, we use explainability in a broad manner to cover both explainability, as in a system’s ability to show its work, and interpretability, as in a human’s ability to logically understand what is being explained.)

→ The definition proposed in EASA AI Roadmap is the following

**AI Explainability:** Capability to provide human with understandable and relevant information on how an AI/ML application is coming to its results.
Summary of the previous session

→ Slide ‘Definitions’: 
  → There are many definitions and research on the topic 
  → The definitions from the SOC and EASA AI Roadmap are compatible with each other 
  → Also with the 4 principles of Explainability (NISTR report) 
  → Discussion was triggered on 
    → link to “accountability” which is clearly a motivation for AI explainability, 
    → and on “causality” which may be a way to reach an explanation but not necessarily practicable with complex model (NNs) -> prefer “transparency”? 

Slide presented at April Plenary breakout
Explainability: Scope and needs

System level versus output level

-> Topic #3!
Summary of the previous session

→ Slide ‘Scope and needs’:
  → There is a consensus that definitions, as good as they can be, will not embed all necessary details to define what the WG should be working on
  → Scope split between development/post-ops and operations is necessary
  → Needs and motivation should be further clarified to enable work in the different SGs
Split among SGs

SG2,3,4,7 / MLDL
Support: SG5

SG5 / system + interface to HF

OPS / HF = out of scope? -> Topic #1!
Summary of the previous session

→ Slide ‘Split between SGs’:
  → Split is supported.
  → MLDL as is might cover a major portion of the “development/post-ops explainability” but may need to be complemented (interpretability, data recording…)
  → SG7 shall be added in the top box.
  → If out-scoping Operations/Human Factors aspects, who will tackle it?
Topic 1 – Human centric aspects of AI

Common aspects = MLDL / Learning Assurance

Domain specific aspects
On the ATM/ANS side:

- **Functional system** = human + procedures + equipment

'functional system' means a combination of procedures, human resources and equipment, including hardware and software, organised to perform a function within the context of ATM/ANS [...]

On the Airworthiness side:

- **System functional behaviour** = function + logic + interface to human

(1) The functional behaviour of an automated system results from the interaction between the crew members and the automated system, and is determined by:

- (i) the functions of the system and the logic that governs its operation; and
- (ii) the user interface, which consists of the controls that communicate the crew’s inputs to the system, and the information that provides feedback to the crew members on the behaviour of the system.
Topic 1 - Scope extension
Topic 1

→ Option 1: expand WG114/G34 scope/competence to cover human centric aspects of AI (create SG8?)
  → Refine need within WG114/G34
  → With competence and scope extension on Human Factors / procedures
  → Liaison with other groups

→ Option 2: task another existing group (e.g. G10/S7/S18/G35/G45)
  → Question for EUROCAE/SAE

→ Option 3: create a dedicated committee “human centric aspects of advanced automation”
  → Question for EUROCAE/SAE
The required « level of explainability » is anticipated to be modulated at least by two main criteria identified so far:

- Level of AI (see EASA AI Level 1 Concept Paper 4.2.2), in a nutshell:
  
<table>
<thead>
<tr>
<th>AI level</th>
<th>HMI Impact</th>
<th>Expected Level of op. EXP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1A</td>
<td>None (tbc)</td>
<td>None (tbc)</td>
</tr>
<tr>
<td>1B</td>
<td>Medium</td>
<td>Important</td>
</tr>
<tr>
<td>2</td>
<td>Very High</td>
<td>Critical</td>
</tr>
<tr>
<td>3A</td>
<td>High</td>
<td>Critical</td>
</tr>
<tr>
<td>3B</td>
<td>No HMI</td>
<td>tbc</td>
</tr>
</tbody>
</table>

- Criticality of the AI application
  
  Modulation by DAL/SWAL will be introduced once objectives and means of compliance will have been refined.
Topic 3: System level versus output level

System level versus output level

Recipients
- Product/equipment Manufacturer
- Developer
- Tester
- Certification Authority
- Safety Investigator
- End user (Flight Crew, ATCO...)

AI/ML Model

- System level
- Output level
Debrief – Topic #1

→ Considerations on « procedures » do not appear clearly in the Explainability diagramme: should be one additional dimension in the operational view in addition to « human factors »
→ V&V is an essential topic for autonomous systems and could be developed more clearly in the diagrammes
→ Distinction between "black box" (a.k.a., adjunct explainabilitiy models) vs "white box" (a.k.a., interpretability) should be considered: anticipated in the guidance through considerations on model interpretability for high criticality/level of automation.
→ One of the main drivers of “Explainability’ is the level of Automation; we have to focus on the automation enabled by AI.
→ AI/ML Development/Post Ops Explainability is already addressed in the current scope of the WG114/G34. Operational Explainability is only partly addressed: monitoring part is covered but the End User aspects are not for the moment.
Debrief – Topic #1

→ Regarding the proposed options, the group is in favour of a need/gap analysis separately for both ATM & airborne domains related to human-centric aspects of AI, in order to better select the right option (it is proposed that the gap is to be performed in the frame of SG5).

→ Elements to be considered for the selection of the final Option:
  → The results from the Gap analysis will be an additional input to select the best option
  → Availability of Resources and expertise within WG114/G34 on those aspects (HF)
  → The timeline/planning for the current work done in WG114/G34

→ Current feedback on options:
  → Option 1 gathered support but also discussion on the limited resource in the WG114/G34 and need to account for impact on planning and scope of the current group.
  → Option 2 also was supported with limitations on the fact that no existing group appear to be covering automation in the wide range of domains that our group is targeting. The main challenge is on the “lack of explainability” of AI-based systems, that may not be addressed adequately by other WGs without adequate liaison
  → Option 3 seems to have less support overall
Debrief – Topic #2

→ Other criteria proposed to orient the guidance proportionality:
  → Complexity of the AI model
  → Dimensionality of the input space
  → Learning method / algorithm (e.g. decision tree versus neural network)
  → Determinism of the model (as applicable in a future version of the standard)

→ In creating a proportionality framework for explainability, the two types of explainability may need to be considered separately:
  → Operational Explainability proportionality appears more driven by the AI level
  → Development Explainability proportionality appears more driven by the criticality
  → Clear cut may however not be the final approach
Debrief – Topic #3

→ The position of roles on the scale « system / output level » triggered many discussions
→ End user interest (Flight Crew, ATCO...) is clearly focused at “output level”.
→ The other actors are more or less interested on the whole scale

→ Explainability is to be defined at « system level » even if the discussion was done at « model level »

→ Question whether this distinction is needed at high level in the explainability guidance
→ This characterisation is more meant as a criteria for an explanation as one for the classification
Thank you for your attention!
Topic 1 – Human centric aspects of AI

Common aspects = MLDL / Learning Assurance

Domain specific aspects

Operational Explainability
Four Principles for Explainable AI

(NISTIR 8312 Report, 2020)

• Explanation
  – The Explanation principle obligates AI systems to supply evidence, support, or reasoning for each output

• Meaningful or Interpretable
  – The system fulfills the Meaningful principle if the recipient understands the system’s explanation

• Explanation Accuracy
  – The system delivers an explanation that correctly reflects its process for generating the output

• Knowledge Limits
  – The system identifies cases it was not designed or approved to operate, or when its outputs are not reliable
XAI Techniques
(Final Report, AFE-87 – Machine Learning, AVSI, 2020)

Figure A-1: Classifying the different approaches to exposing the black box.\textsuperscript{44}
"This Looks Like That: Deep Learning for Interpretable Image Recognition", Chen et al., NuerIPS, 2019

Figure 3: The reasoning process of our network in deciding the species of a bird (top).