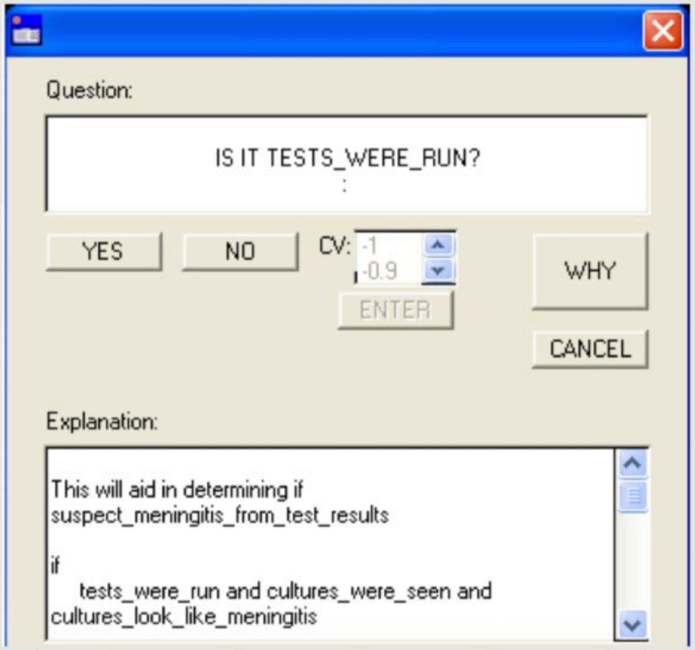
Presentation to the HFES Annual Meeting
Seattle, WA — October 2019

Explainable and Transparent AI

Mica R. Endsley
SA Technologies

Explainable AI circa 1990

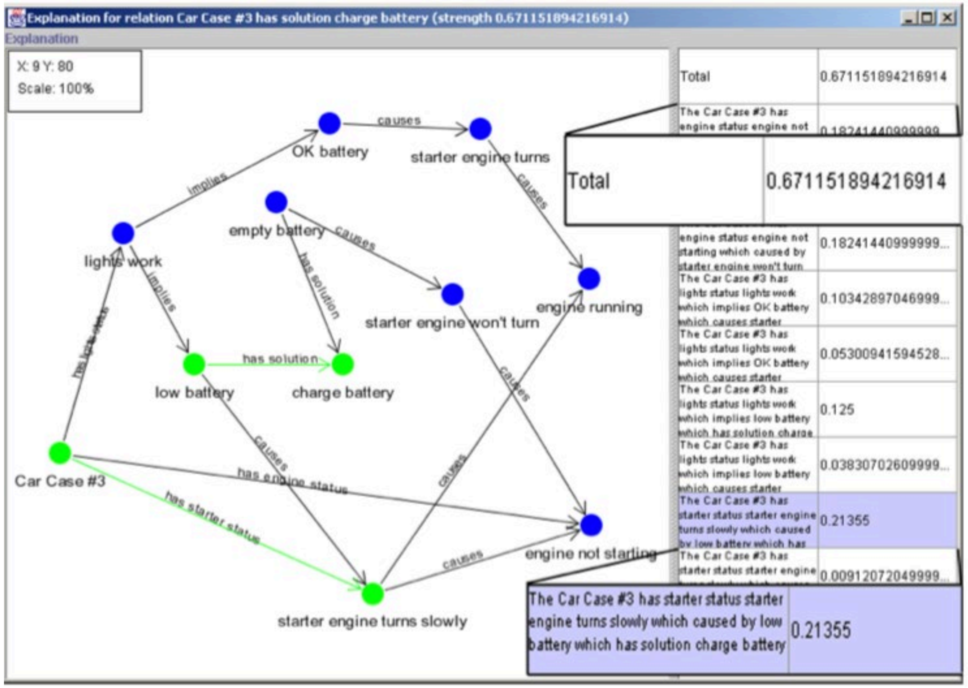
"WHY"



IS IT TESTS_WERE_RUN?
:

This will aid in determining if
suspect_meningitis_from_test_results

if
    tests_were_run and cultures_were_seen and
cultures_look_like_meningitis



Explanation for relation Car Case #3 has solution charge battery (strength 0.671151894216914)

Total	0.671151894216914
The Car Case #3 has engine status engine not	0.18241440999999
Total	0.671151894216914
engine status engine not starting which caused by starter engine won't turn	0.18241440999999...
The Car Case #3 has lights status lights work which implies OK battery which causes starter	0.10342897046999...
The Car Case #3 has lights status lights work which implies OK battery which causes starter	0.05300941594528...
The Car Case #3 has lights status lights work which implies low battery which has solution charge	0.125
The Car Case #3 has lights status lights work which implies low battery which causes starter	0.03830702609999...
The Car Case #3 has starter status starter engine turns slowly which caused by low battery which has	0.21355
The Car Case #3 has starter status starter engine	0.00912072049999...
The Car Case #3 has starter status starter engine turns slowly which caused by low battery which has solution charge battery	0.21355

hard-driving *always-leads-to* extreme-engine-load *may-lead-to* abnormally-high-carburettor-pressure *causes* broken-carburettor-membrane

Today's challenges

- Learning algorithms
  - No "rules" to provide
    - Can be extracted (Huang & Endsley, 1997)
  - Frequent changes in how the system works
    - May be updated on a daily, weekly or monthly basis
  - May be applied to real-time control systems where time to assess is important
    - automobiles, aviation, power systems,...
  - Need more support for creating understanding of system
    - Can't just train once and be done

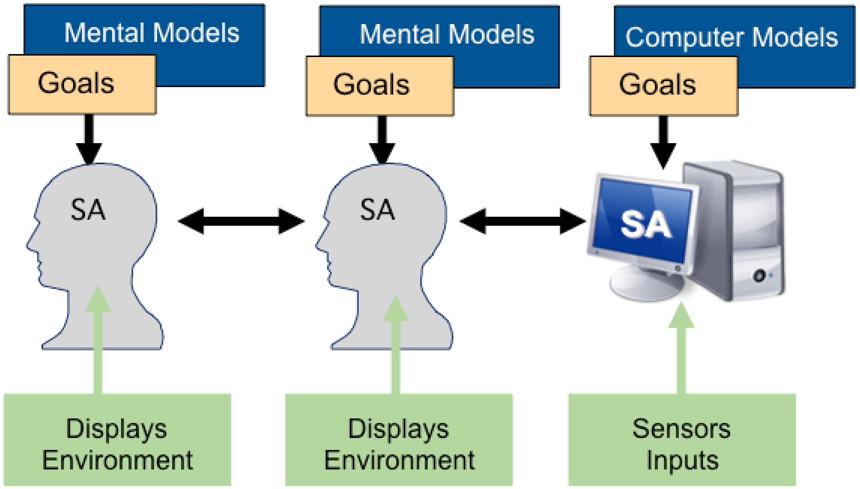
Explainability vs Transparency

- Explainability
  - Often backward looking
  - Focused on why
  - Separate from system actions/outputs
- Transparency
  - Real-time
  - Focused much more broadly on shared SA needs
  - Integrated into the operator interface

Human-Autonomy Teaming

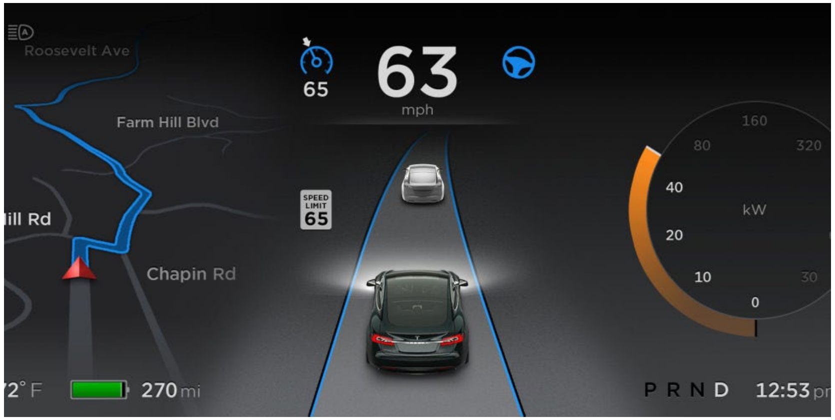
Shared SA between the system and the operators

- Understanding of its status
- How well is it functioning
- When interventions are needed and what kind
- How the system's status effects operator tasking and vice-versa



What do we need to understand?

- What does it know/not know about the situation?
- What is it doing?
  - In real time
- And why is it doing that?
  - Current goal and tasking
  - Logic that led to that behavior
- What will it (or can it) do in the near future?
- What are the limits of its performance?
  - Can it handle present and upcoming conditions or do I need to intervene?

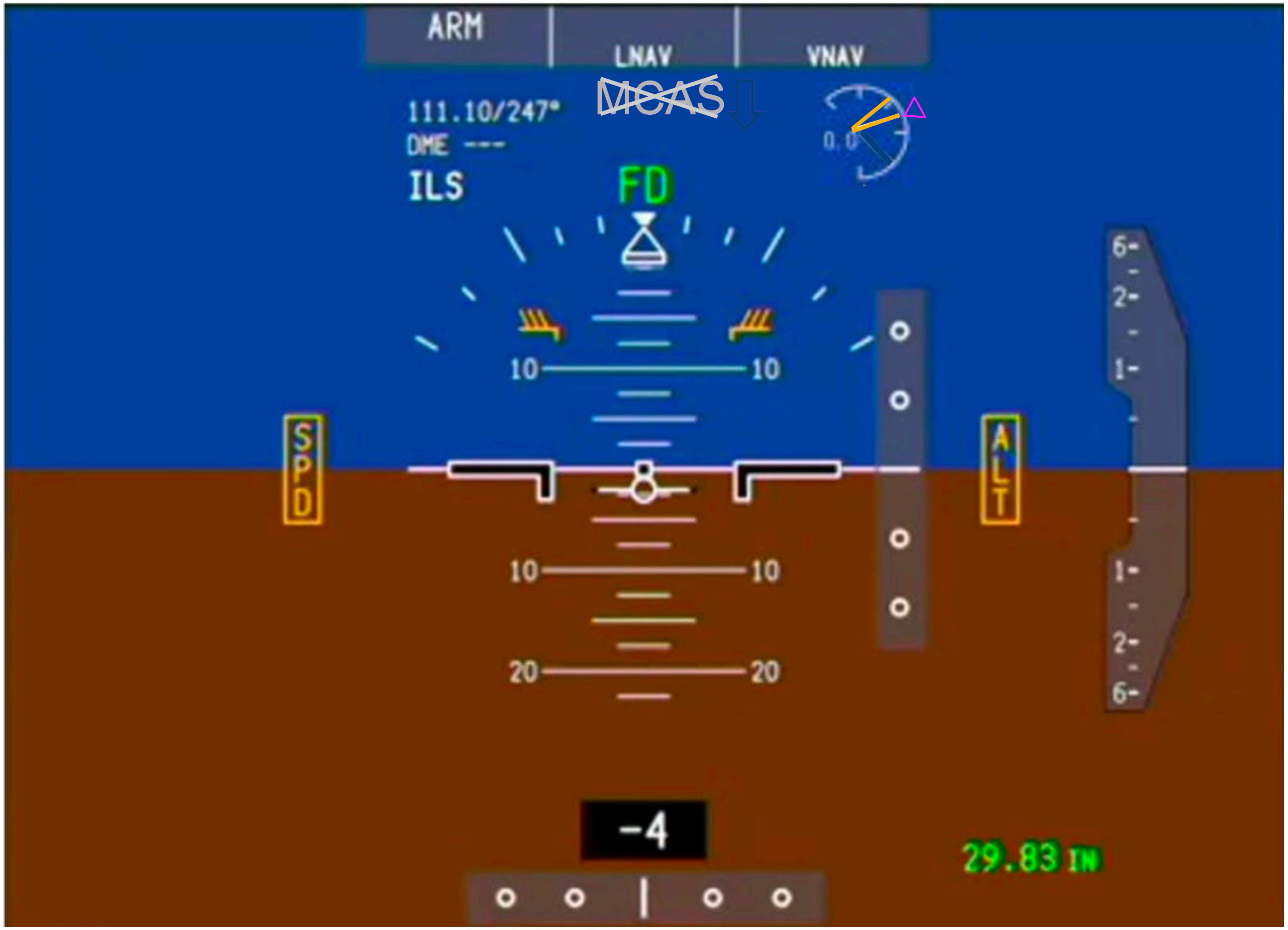


Goals for Transparent Interfaces

- Understandability
- Predictability
- Understanding of system states and mode transitions
- Understanding of system reliability
  - How well it is functioning
  - Level of confidence in fused data
  - Level of confidence in system assessments
- Robustness
  - Ability to handle current and upcoming situations
- Persistence
  - Ongoing reinforcement and presentation of information

Example of Transparency 737-Max8

AOA Disagree


ARM
LNAV
VNAV
~~MCAS~~
111.10/247°
DME ---
ILS
FD
0.0
6
2
1
1
2
6
SPD
ALT
10
10
10
10
20
20
-4
29.83 IN