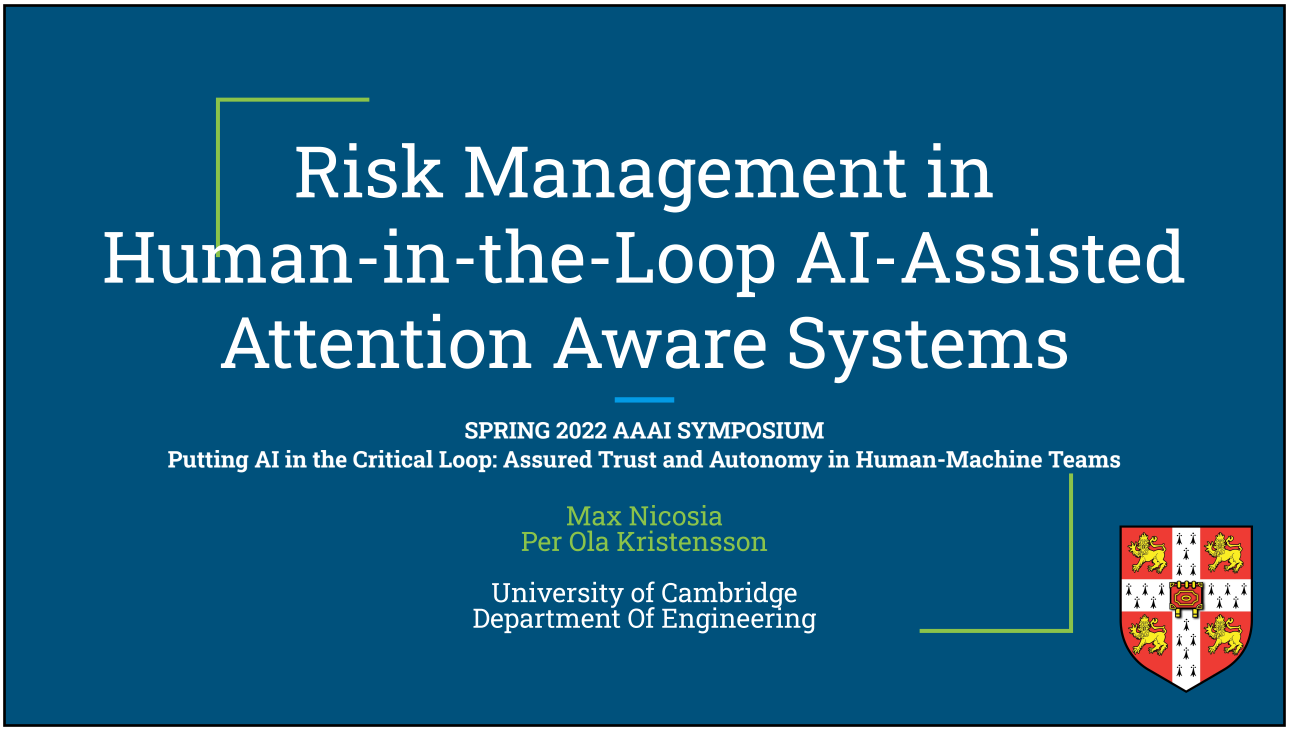
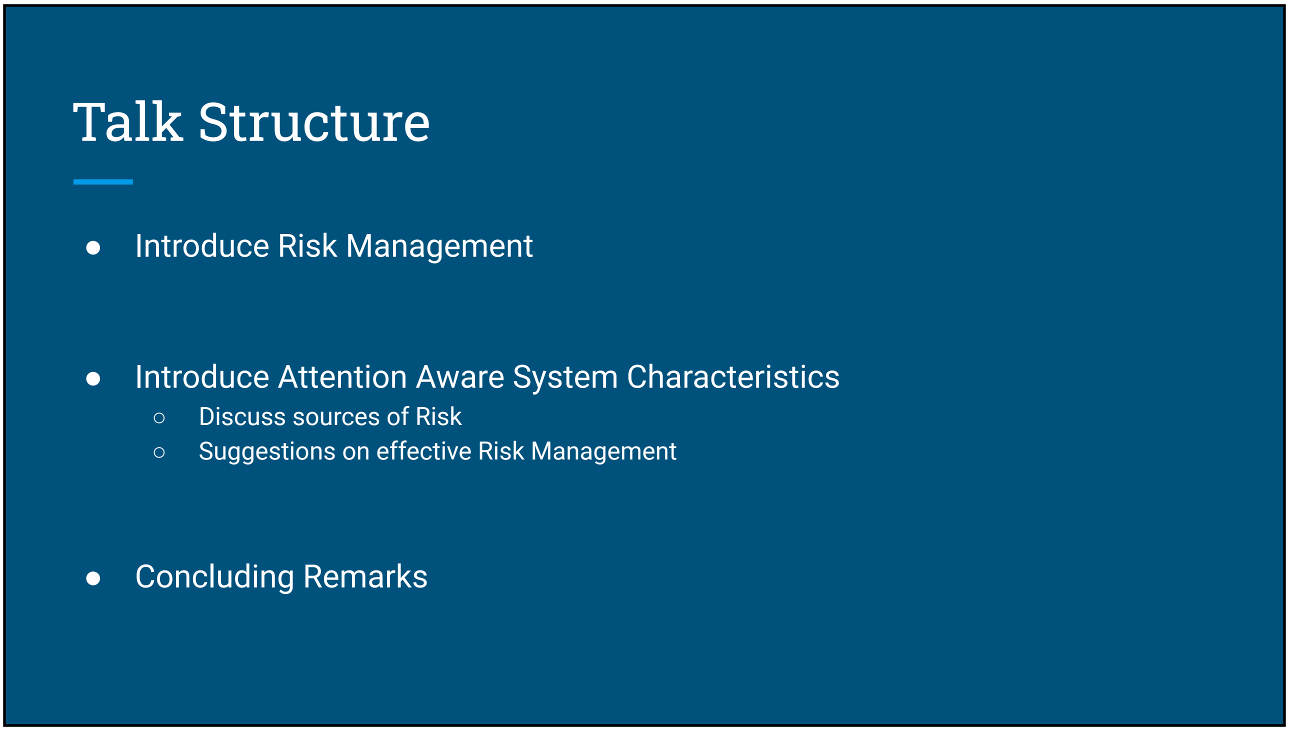


Risk Management in Human-in-the-Loop AI-Assisted Attention Aware Systems

SPRING 2022 AAAI SYMPOSIUM
Putting AI in the Critical Loop: Assured Trust and Autonomy in Human-Machine Teams

Max Nicosia
Per Ola Kristensson

University of Cambridge
Department Of Engineering



Talk Structure

- Introduce Risk Management
- Introduce Attention Aware System Characteristics
  - Discuss sources of Risk
  - Suggestions on effective Risk Management
- Concluding Remarks

Definitions

- Risk = probability x impact
  - Likelihood of undesired behaviour multiplied by impact of undersidered behaviour

- Risk analysis
  - Systematic identification of potential sources of harm (hazards) and their risks.

- Risk Management:
  - “The identification, analysis, and prioritization of risks followed by coordinated and economical application of resources to reduce, monitor, and control the probability and/or impact of unfortunate events” – Hubbard [5].

Bit of History

1. Failure Mode and Effect Analysis (FMEA)
   a. Originally used by the U.S. Armed Forces in 1949 [7]
   b. Variants used in the 60s by NASA and in the 70s for petroleum exploration, wastewater treatment plants and food industry [6]
2. Fault Tree Analysis (FTA)
   a. Started by Bell Labs in 1962 to evaluate ballistic launch control systems [3]
   b. In 1970 was incorporated by the Federal Aviation Administration (FAA) into CFR §25.1309 airworthiness regulations for aircraft
   c. Some years later the FAA extended their use to other areas within the U.S. National Airspace Systems [1]
3. Event Tree Analysis (ETA)
   a. Introduced in The Reactor Safety Study (WASH-1400) (~1975). To simplify analysis [2]
   b. Alternative to FTA: Assumes system units either are working or failing.

Why should we do Risk Management?

1. Avert unnecessary costs
   a. Early detection of potential human-machine system failures
2. Improve safety and reliability:
   a. Create appropriate mitigation and monitoring strategies
   b. Increasing redundancy
   c. Incorporating checks to prevents miscommunications or errors
3. Improve understanding about operational limits
   a. Potentially improve the design of a system
4. Prevent litigation and reputational damage
   a. Meet obligations / regulations / ethical standards

The High Cost of Ignoring Risk Management

- Challenger disaster (1986):

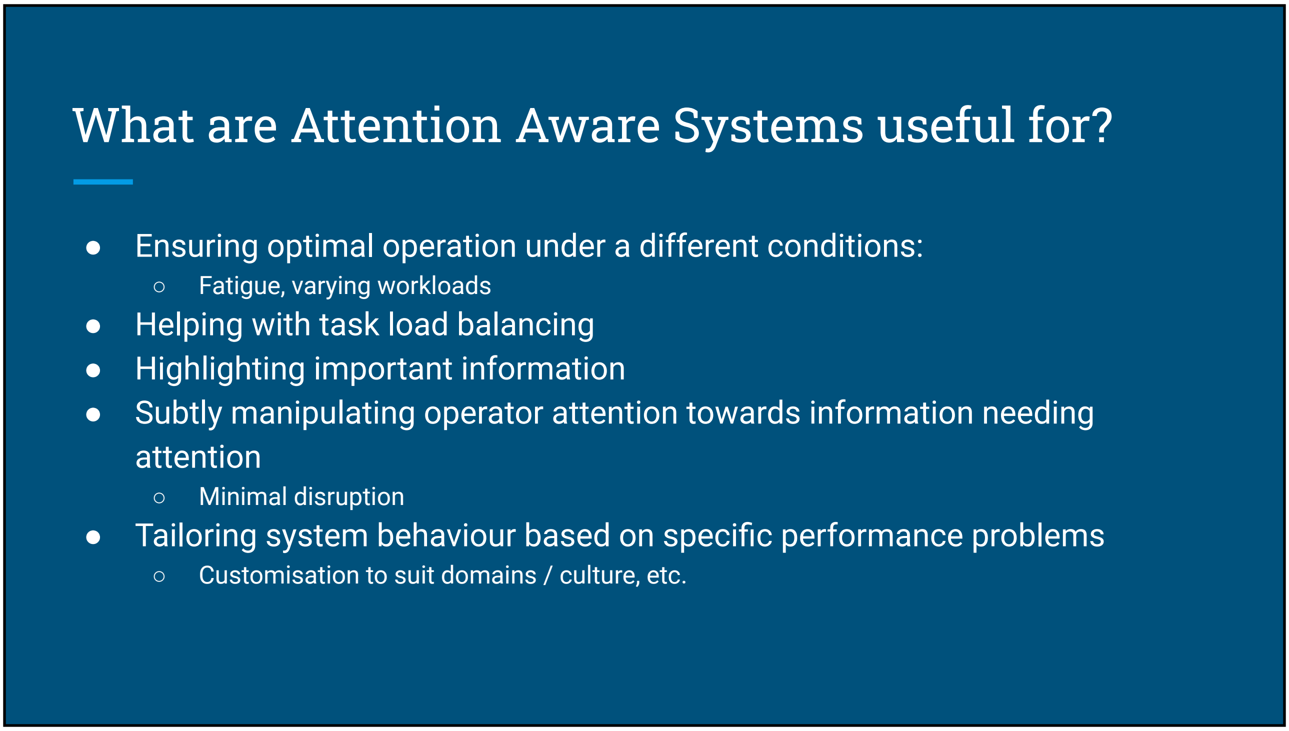
NASA considered using FTAs the program but decided against it due to calculations giving unacceptably low-reliability values. This proved to be a major oversight after the accident and NASA resumed their use afterward [8]

Risk Management Steps

1. Identify system boundaries
2. Map system components / entities
3. Identify Hazards
4. Estimate Risks
5. Devise strategy to manage/mitigate risks

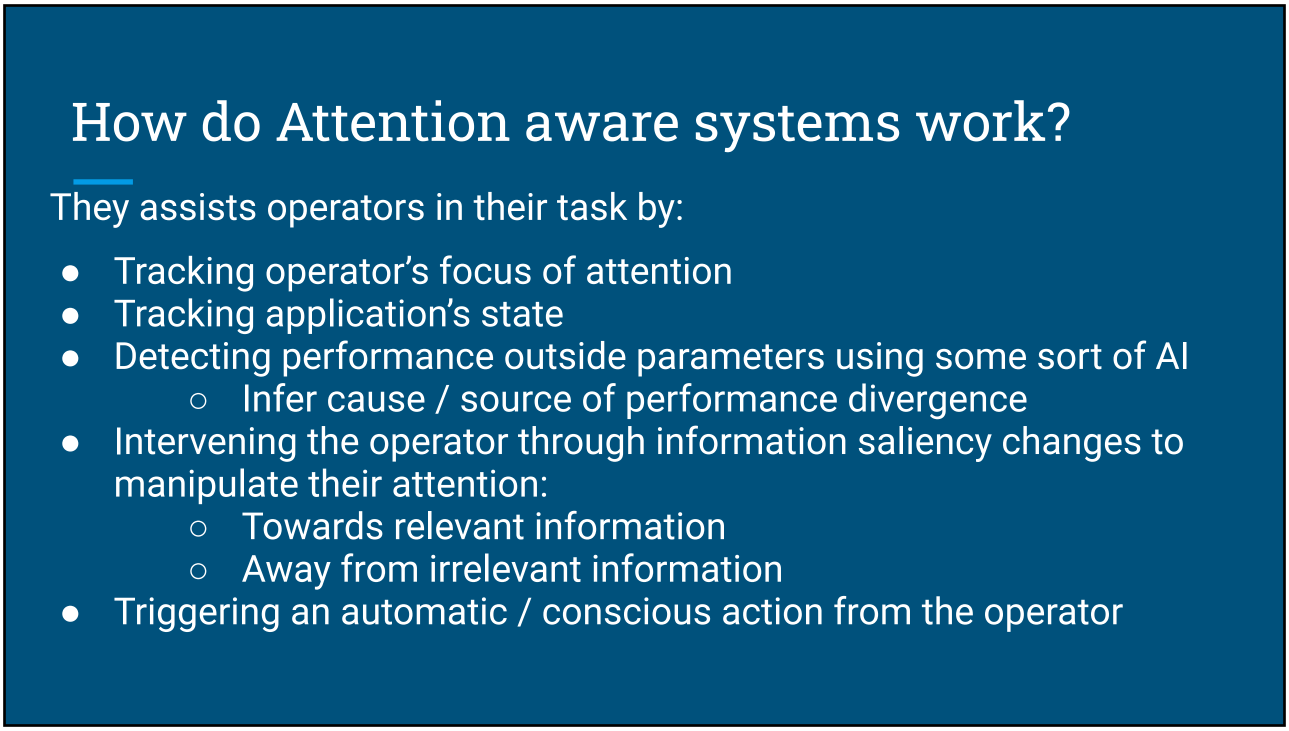
Risk Management Methods

- Qualitative or quantitative
  - Both needed, in particular systematic tend to help control for human judgement that can underestimate or overlook risks
  - However, quantitative can also failed if values assigned to risks depend on misunderstandings. Then the risks can still be underestimated



What are Attention Aware Systems useful for?

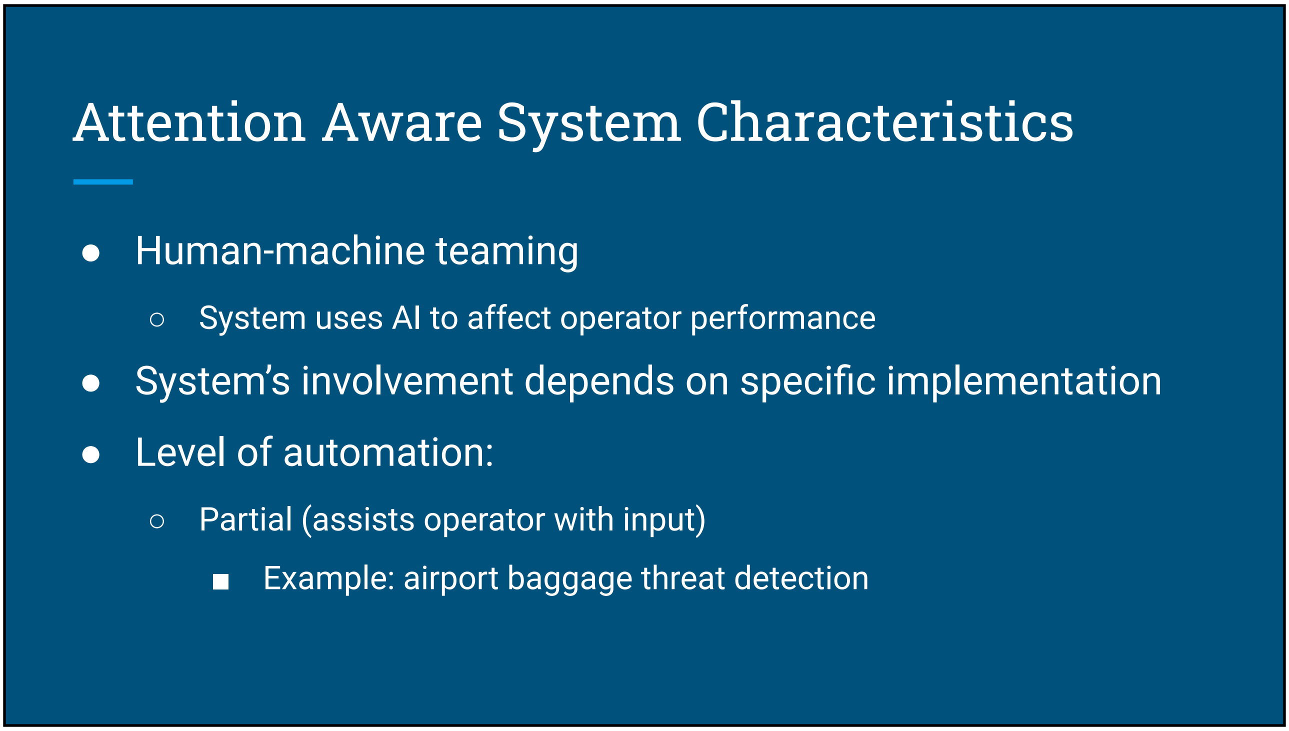
- Ensuring optimal operation under a different conditions:
  - Fatigue, varying workloads
- Helping with task load balancing
- Highlighting important information
- Subtly manipulating operator attention towards information needing attention
  - Minimal disruption
- Tailoring system behaviour based on specific performance problems
  - Customisation to suit domains / culture, etc.



How do Attention aware systems work?

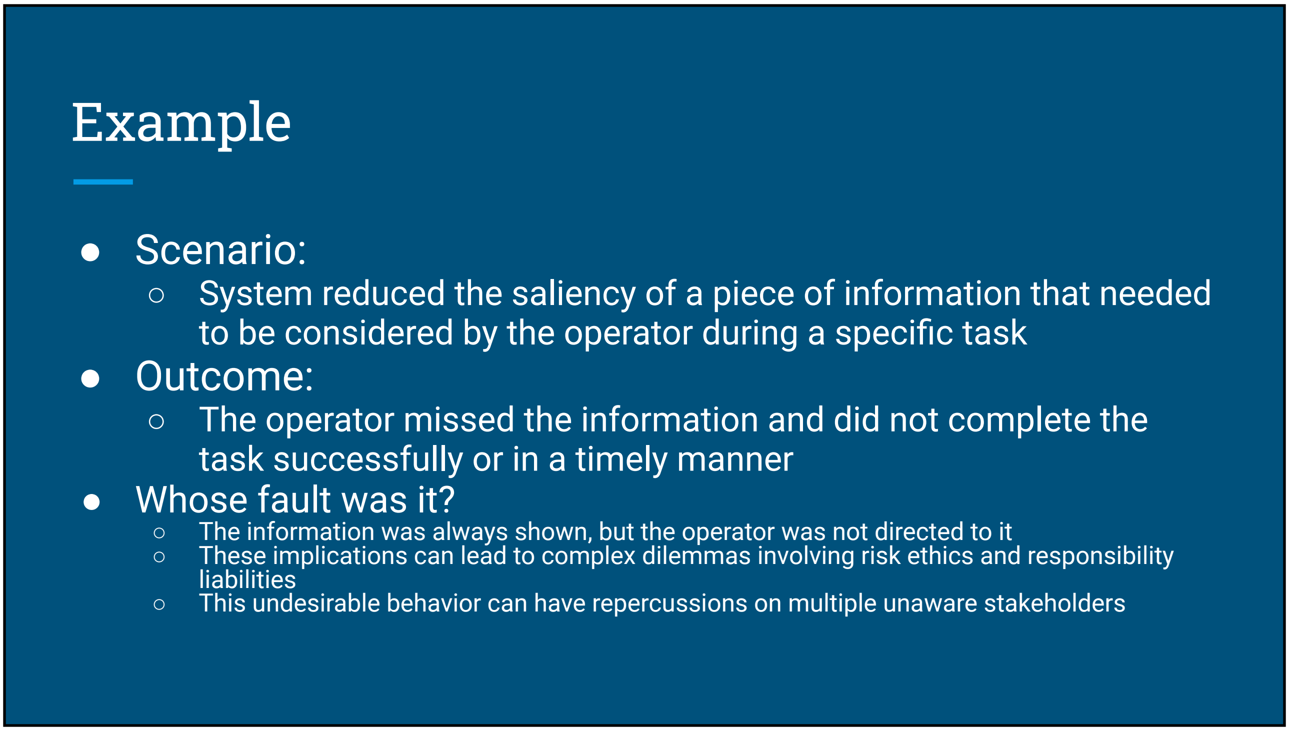
They assists operators in their task by:

- Tracking operator’s focus of attention
- Tracking application’s state
- Detecting performance outside parameters using some sort of AI
  - Infer cause / source of performance divergence
- Intervening the operator through information saliency changes to manipulate their attention:
  - Towards relevant information
  - Away from irrelevant information
- Triggering an automatic / conscious action from the operator



Attention Aware System Characteristics

- Human-machine teaming
  - System uses AI to affect operator performance
- System’s involvement depends on specific implementation
- Level of automation:
  - Partial (assists operator with input)
    - Example: airport baggage threat detection



Example

- Scenario:
  - System reduced the saliency of a piece of information that needed to be considered by the operator during a specific task
- Outcome:
  - The operator missed the information and did not complete the task successfully or in a timely manner
- Whose fault was it?
  - The information was always shown, but the operator was not directed to it
  - These implications can lead to complex dilemmas involving risk ethics and responsibility liabilities
  - This undesirable behavior can have repercussions on multiple unaware stakeholders

Our Objective

1. Raise awareness about the complexity of these systems
    a. The very fact that they are not well understood yet
2. Point out sources of risks
3. Advocate for the use of standard risk management methods as a starting point
4. Prevent history from repeating itself due to complacency

Considerations

- Complex AI-assisted joint human-machine systems with multiple stakeholders
- Depend on multiple external systems
- Require extensive parameterization
- Highly specific: each deployment will have their own associated specific risks
- Factors relating to the individual operator, or team of operators, such as skill level, experience and culture
- System can generate actions that are unthinkable or otherwise confusing to human operators (hiding information because it disagrees with operator)

This leads to VERY high risks!

Attention Aware Systems - Sources of Risks

1. System failures
2. Incorrect parameterisation
3. Operator characteristics
4. Domain context characteristics

System Failures

System not operating as intended despite being correctly parameterised

- Software failures:
  - Program bugs, etc.
- Hardware failures:
  - Loose cables, etc.

System Failures (2)

Example of system signals:

- External performance data (baseline evaluation)
- Application output (state)
- External classifier (prioritisation of targets / information)
- Operator sensor data

Incorrect Parameterisation

- Parameterisation errors can lead to unexpected behaviour
- Not monitoring right metrics
- Not having the right thresholds
- Intervention parameters need to match the domain requirements (right colors, symbology, etc)

Operator Characteristics

Operator becomes out-of-phase with the system leading to undesirable behavior

- Operator misunderstanding the system's interventions
  - Unaware of the specifics of an intervention, despite correct parameterisation
- System misunderstanding the operator's actions
  - Caused by parameterisation errors, system failures, or unexpected or unknown operator behavior

Domain Context Characteristics

System reacting unexpectedly to unaccounted inputs and as such, the system behaves in an undefined way

- System mislabeling / misidentifying data because it was not configured for a specific case
- Unexpected task complexities arising (rapid explosion of targets on the display)
  - Breaks assumptions of visualisation parameters, mechanisms, etc.

Additional Sources of Risks

- Interactions between stakeholders:
  - Operators
  - The people responsible for parameterizing the system
  - Entity that is directly or indirectly controlled by the operator through the system
  - The organization providing the system
  - The organization managing and employing the operator

- Miscommunication:
  - Not keeping proper record of what is needed / operation / parameterisation / etc.

Our Suggestions

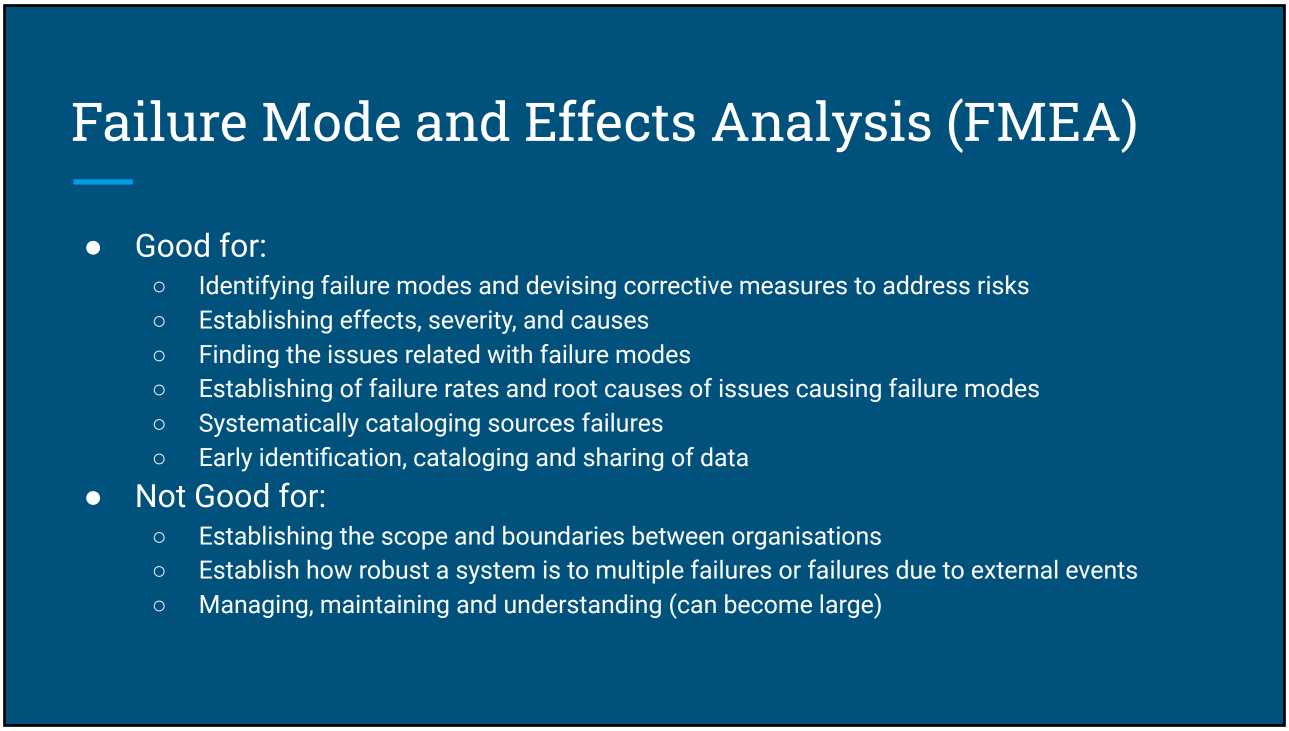
1. Use well established tools to understand the system, its components, entities and define the boundaries
2. Employ well tested methods for identifying and assessing risks
3. Develop a strategy to monitor and evaluate the efficacy of the risk management strategy
   a. Ensure resources not wasted
   b. Efforts transfer into added reliability
   c. Detect emergent behaviours caused by unexpected/misunderstood interactions

System Mapping Tools

- Organizational diagrams:
  - Determining people involved, their roles, and their relationships with other people
  - Understanding the scope and determining the boundaries of all involved entities
- Information diagrams:
  - Determining relationships between documents in the system
- Communication diagrams
  - Establishing information flow between stakeholders and other entities in the system
  - Helpful for parameterisation changes needed to be requested or acted upon
  - They also provide context for the system operator's role in the entire communication structure of the organization

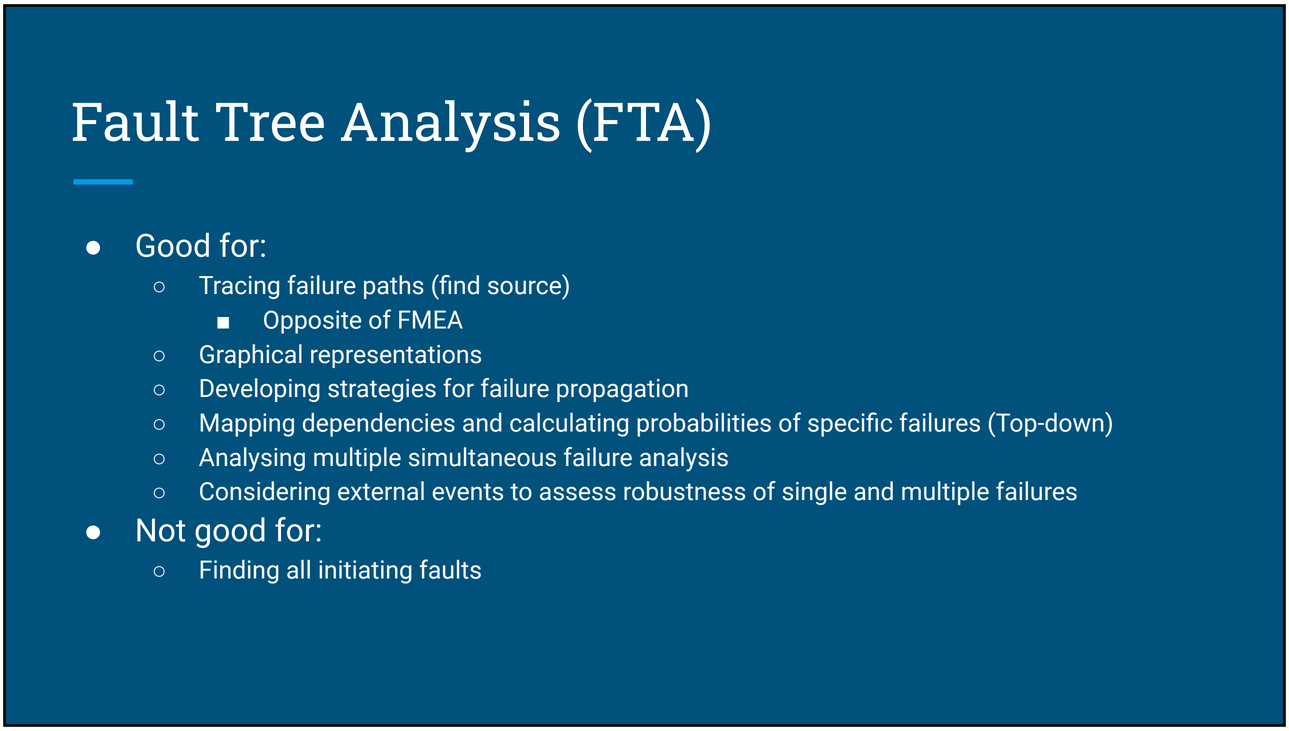
Identifying and Assessing Risks

1. Failure Mode and Effects Analysis (FMEA)
2. Fault Tree Analysis (FTA)
3. Event Tree Analysis (ETA)



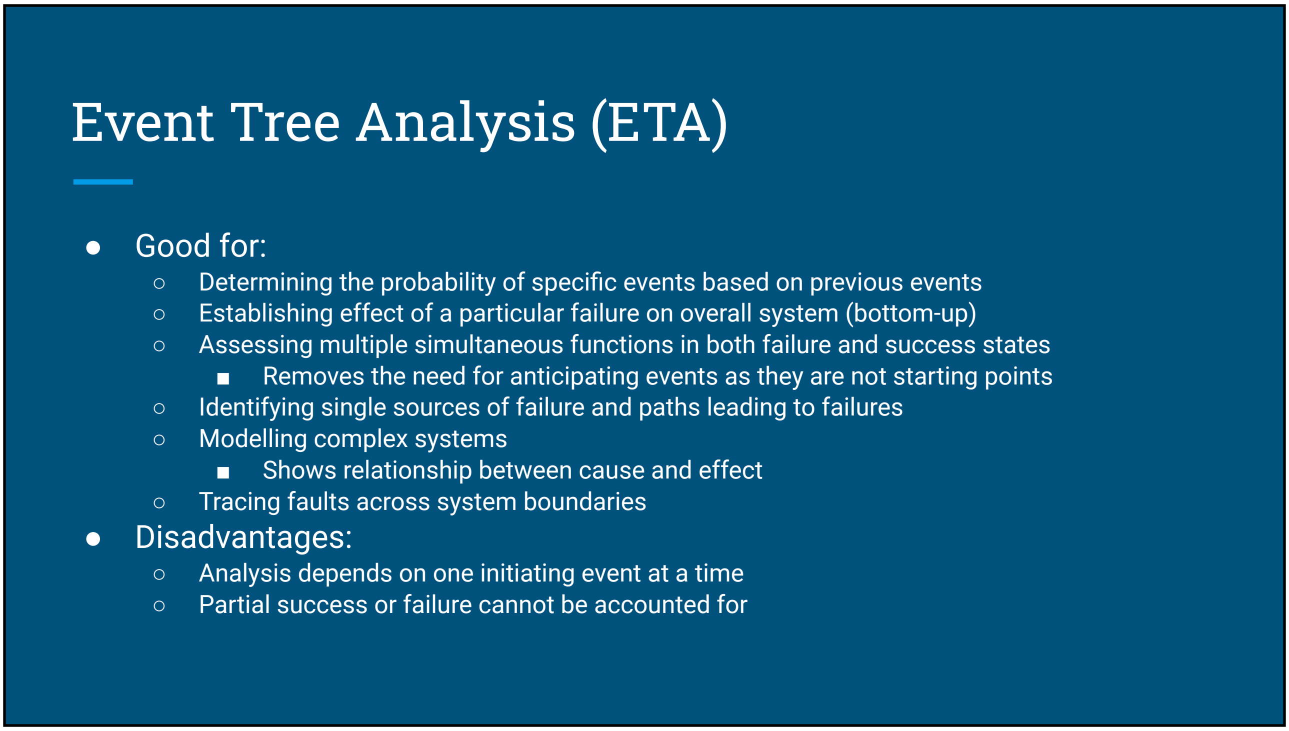
Failure Mode and Effects Analysis (FMEA)

- Good for:
  - Identifying failure modes and devising corrective measures to address risks
  - Establishing effects, severity, and causes
  - Finding the issues related with failure modes
  - Establishing of failure rates and root causes of issues causing failure modes
  - Systematically cataloging sources failures
  - Early identification, cataloging and sharing of data
- Not Good for:
  - Establishing the scope and boundaries between organisations
  - Establish how robust a system is to multiple failures or failures due to external events
  - Managing, maintaining and understanding (can become large)



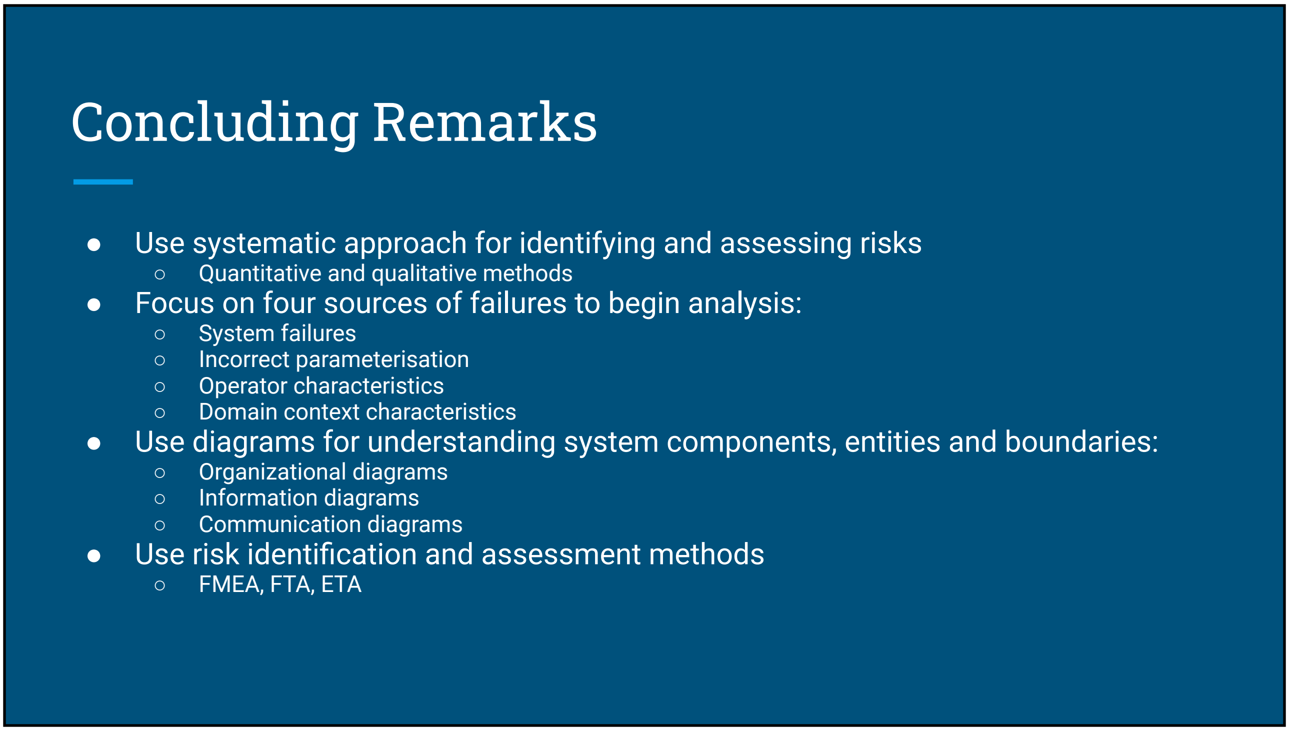
Fault Tree Analysis (FTA)

- Good for:
  - Tracing failure paths (find source)
    - Opposite of FMEA
  - Graphical representations
  - Developing strategies for failure propagation
  - Mapping dependencies and calculating probabilities of specific failures (Top-down)
  - Analysing multiple simultaneous failure analysis
  - Considering external events to assess robustness of single and multiple failures
- Not good for:
  - Finding all initiating faults



Event Tree Analysis (ETA)

- Good for:
  - Determining the probability of specific events based on previous events
  - Establishing effect of a particular failure on overall system (bottom-up)
  - Assessing multiple simultaneous functions in both failure and success states
    - Removes the need for anticipating events as they are not starting points
  - Identifying single sources of failure and paths leading to failures
  - Modelling complex systems
    - Shows relationship between cause and effect
  - Tracing faults across system boundaries
- Disadvantages:
  - Analysis depends on one initiating event at a time
  - Partial success or failure cannot be accounted for



Concluding Remarks

- Use systematic approach for identifying and assessing risks
  - Quantitative and qualitative methods
- Focus on four sources of failures to begin analysis:
  - System failures
  - Incorrect parameterisation
  - Operator characteristics
  - Domain context characteristics
- Use diagrams for understanding system components, entities and boundaries:
  - Organizational diagrams
  - Information diagrams
  - Communication diagrams
- Use risk identification and assessment methods
  - FMEA, FTA, ETA

Questions?

References:

[1] Federal Aviation Authority. System safety handbook, 2000.
[2] PL Clemens and Rodney J Simmons. System safety and risk management: Niosh instructional module. US Department of Health and Human Services, 1998.
[3] Clifton Ericson. Fault tree analysis–a history from the proceeding of the 17th international system safety conference, 1999.
[4] Gregory Falco, Ben Shneiderman, Julia Badger, Ryan Carrier, Anton Dahbura, David Danks, Martin Eling, Alwyn Goodloe, Jerry Gupta, Christopher Hart, et al. Governing ai safety through independent audits. Nature Machine Intelligence, 3(7):566–571, 2021.
[5] Douglas W Hubbard. Healthy Skepticism for Risk Management, chapter 1, pages 1–19. John Wiley & Sons, Ltd, 2020.
[6] RA Neal. Modes of failure analysis summary for the nerva b-2 reactor. westinghouse electric corporation astronuclear laboratory. Technical report, WANL–TNR-042, 1962.
[7] USDD. Mil-p-1629–procedures for performing a failure mode effect and critical analysis, 1949.
[8] W Vesely, J Dugan, and J Fragola. Minarick, and j. railsback. fault tree handbook with aerospace applications. handbook. National Aeronautics and Space Administration, Washington, DC, 38, 2002.