UK Atomic Energy Authority

RACE

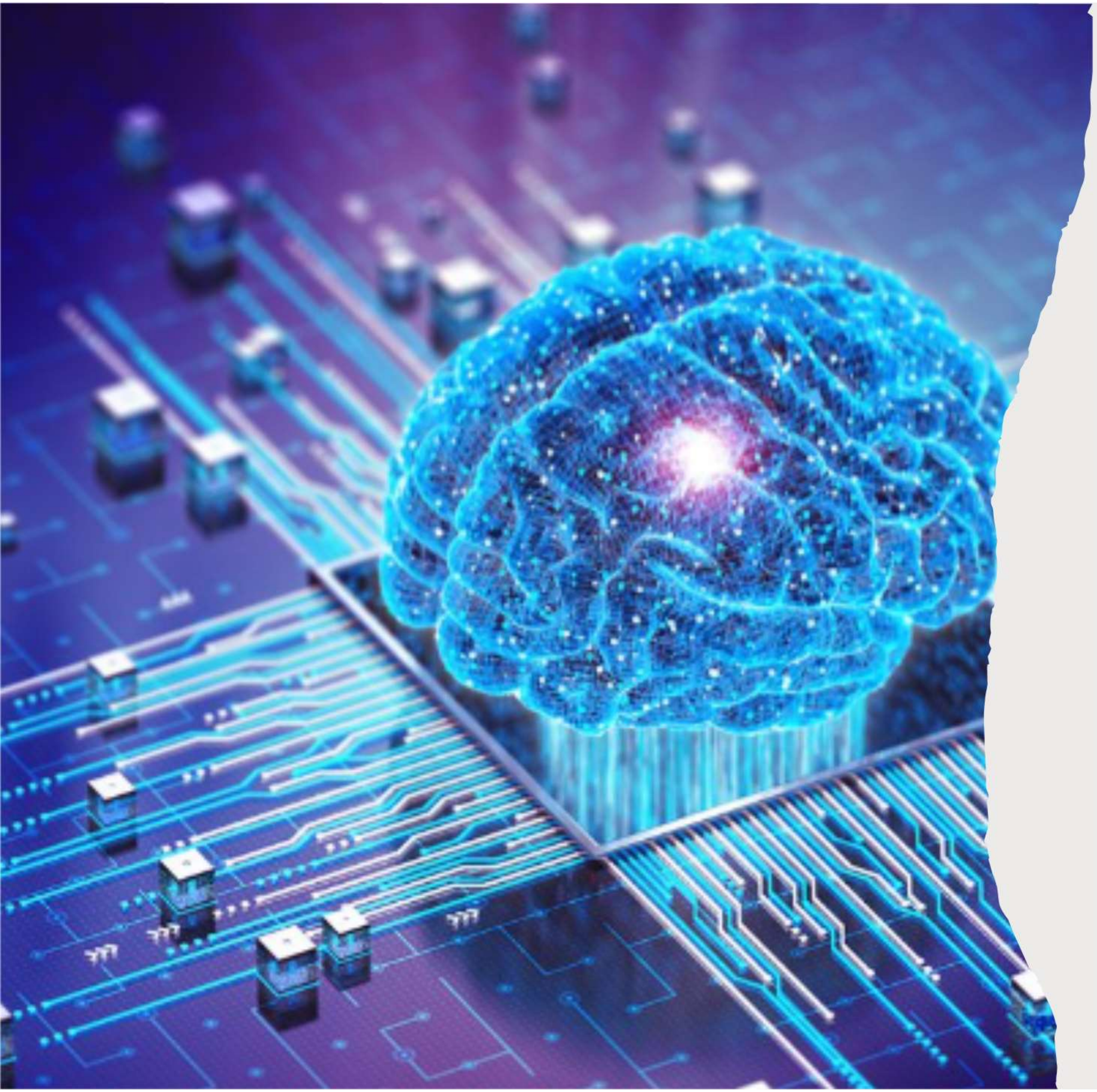
1981
IDEAL

1981 CONSULTANTS
jon.wiggins@1981consultants.com
07737 606836

ARTIFICIAL INTELLEGENCE AND FUNCTIONAL SAFETY

HAZARD ANALYSIS – JON WIGGINS (1981 CONSULTANTS)

SAFETY ASSURANCE – CHRISTOPHER HUME (UK ATOMIC ENERGY AUTHORITY)



WHAT IS AI?

- A series of algorithms which in combination allow a machine to learn
  - About its environment
  - About the expected output
  - About the anticipated input
- A statistical approach to pattern recognition and categorization
- Bound to the task it is taught to preform

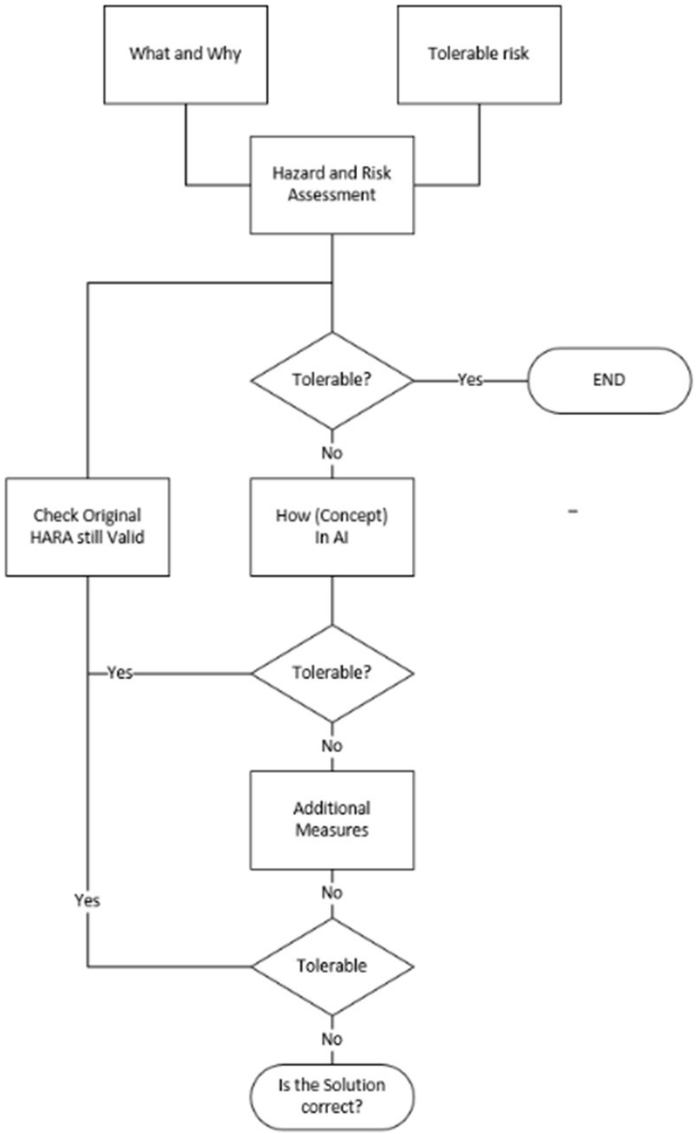
- AI may be used in two ways:
  - **Direct** – Where the AI is in the direct control of the EUC,
  - **Indirect** – Where the AI is in the preparation of a safety case or presentation of safety data (e.g. Data analysis).

Hazards of AI

- Variability
  - The Output will not be the same for the same input
- Drift
  - The output will change over time
- Interpretability
  - Understanding the output
- Accuracy
  - The output is “correct”

Approach to Understanding Risk

- Understand the use cases and Operational Domain,
- Understand the specific reasons for using AI in terms of risk reduction,
- Assess the consequential hazards and risks,
- Determine suitable measures to eliminate or mitigate these risks,
- Assess the consequential risks.



Structure

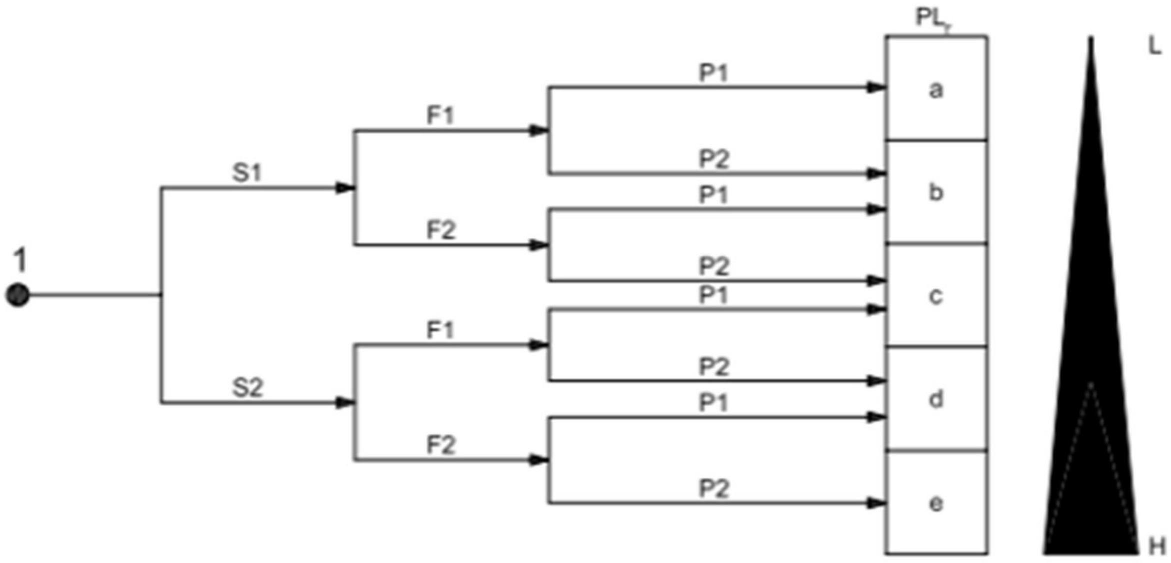
- Definition
  - Define What and Why the system is being designed.
  - Define the tolerable risk
- Initial Hazard Identification
  - What does the now look like? Is now tolerable?
  - Identify the hazards being mitigated and tolerable levels of risk
- Concept
  - Add complexity in layers and assess each layer
  - If additional safety Measures are needed add these but assess the overall system
- Revisit
  - Check that the Risk assessments and safety case still holds against the original criteria.
  - Check the original criteria is still valid
- Specify
  - Performance metrics (e.g. DIN SPEC 92005)
  - Integrity levels (PL, SIL, ASIL etc)
  - System Constraints (e.g. AI classes)

Foreseeable Misbehaviour

- AI not quite working as expected
- Misbehavior or errors within the constraints of the system
  - Hardware
  - Software
- Reasonably foreseeable for the given task

Risk assessment of Foreseeable Misbehavior

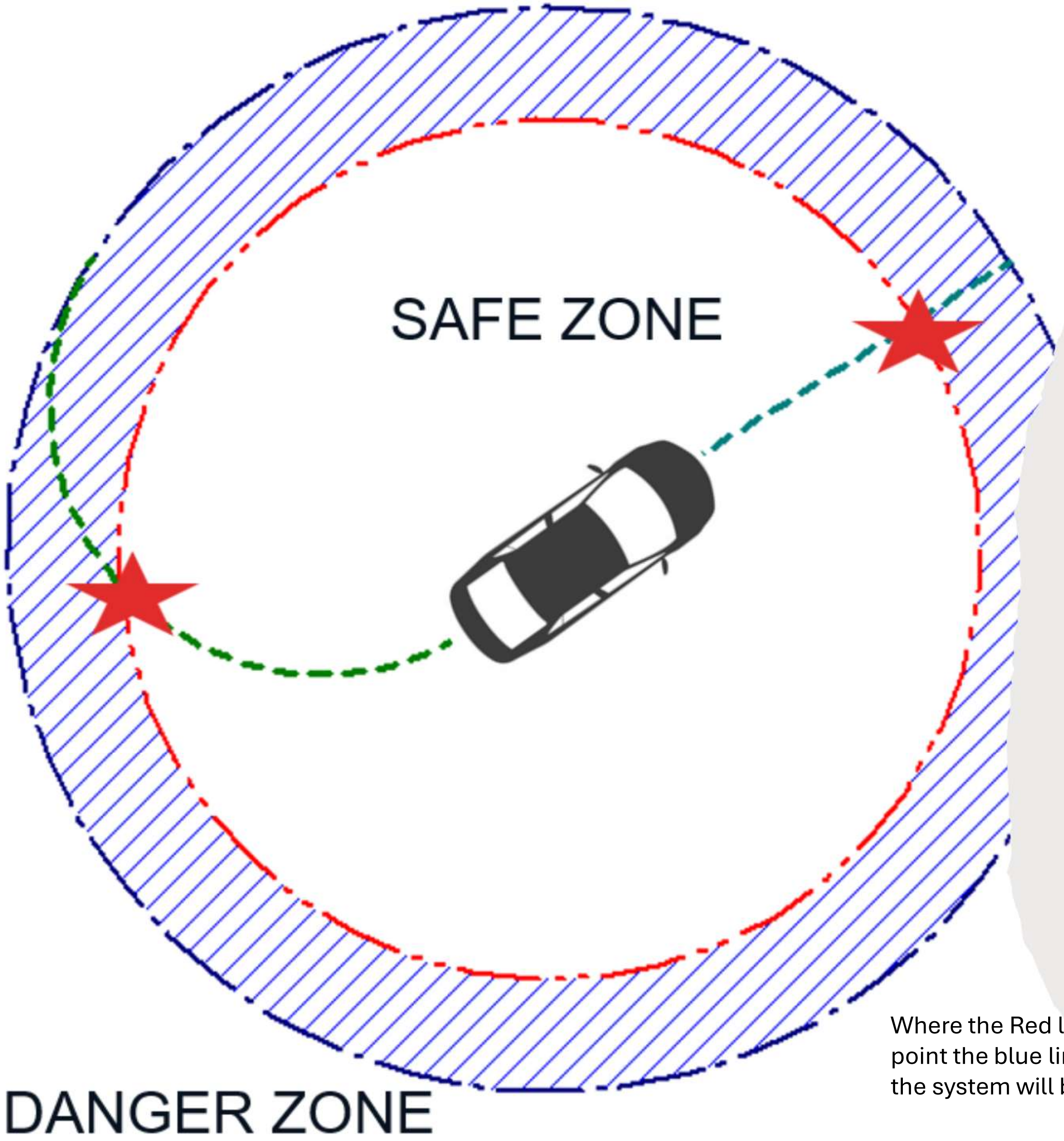
- Consider
  - Severity –
    - What are the consequences
  - Frequency –
    - How often is the AI system used
    - How often to the trigger conditions occur
  - Exposure
    - How often will this be an issue causing a hazard
  - Avoidance
    - Can the hazard be detected and limited



Foreseeable Misbehavior - Example

- ➢ What if…
- ➢ It spends too long on one section of carpet?
- ➢ Repeatedly misses a section of carpet?

Handover Hazards

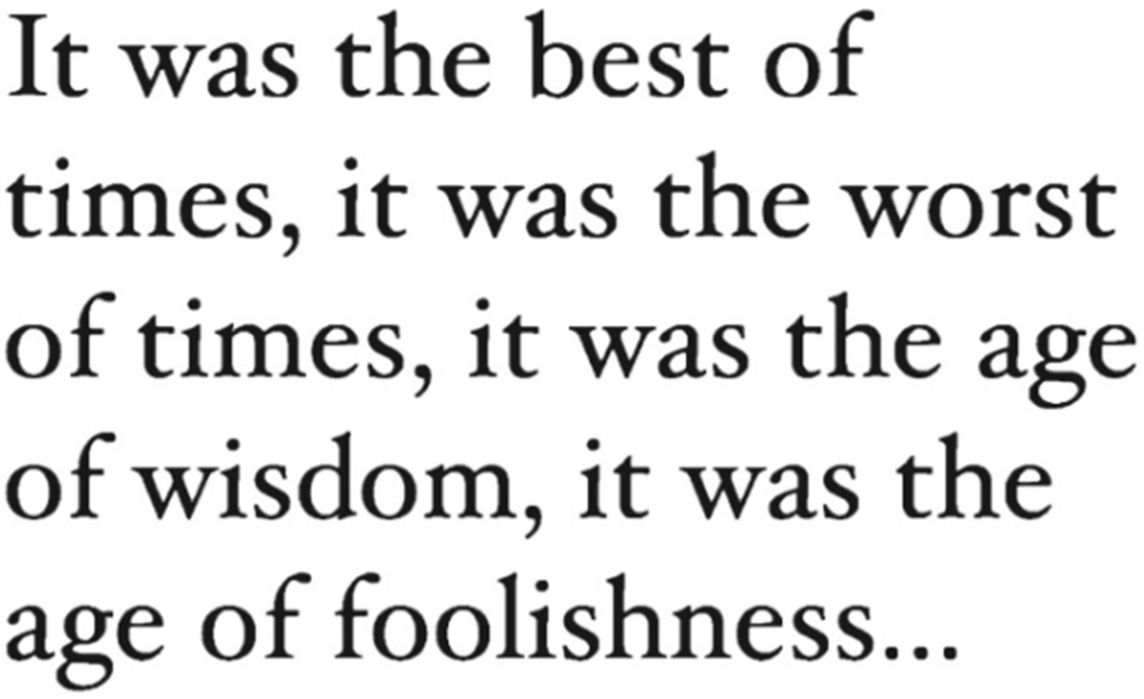


Where the Red line is the trigger point the blue line is the point the system will become safe.

- Consider the system to have a behavior vector. This must be kept within a safe area.
- To prevent the system entering the danger zone the E/E/PE safety system shut the system down
- Type of handover
  - System to System
  - System to Human
- Maximum speed of handover
  - Impact if this is too slow or too fast
- Maximum Frequency of handover
  - Impact if this is too high or too low
- Degree of uncertainty in the above
  - Could the system overwhelm the E/E/PE
  - Can the transition cause a failure

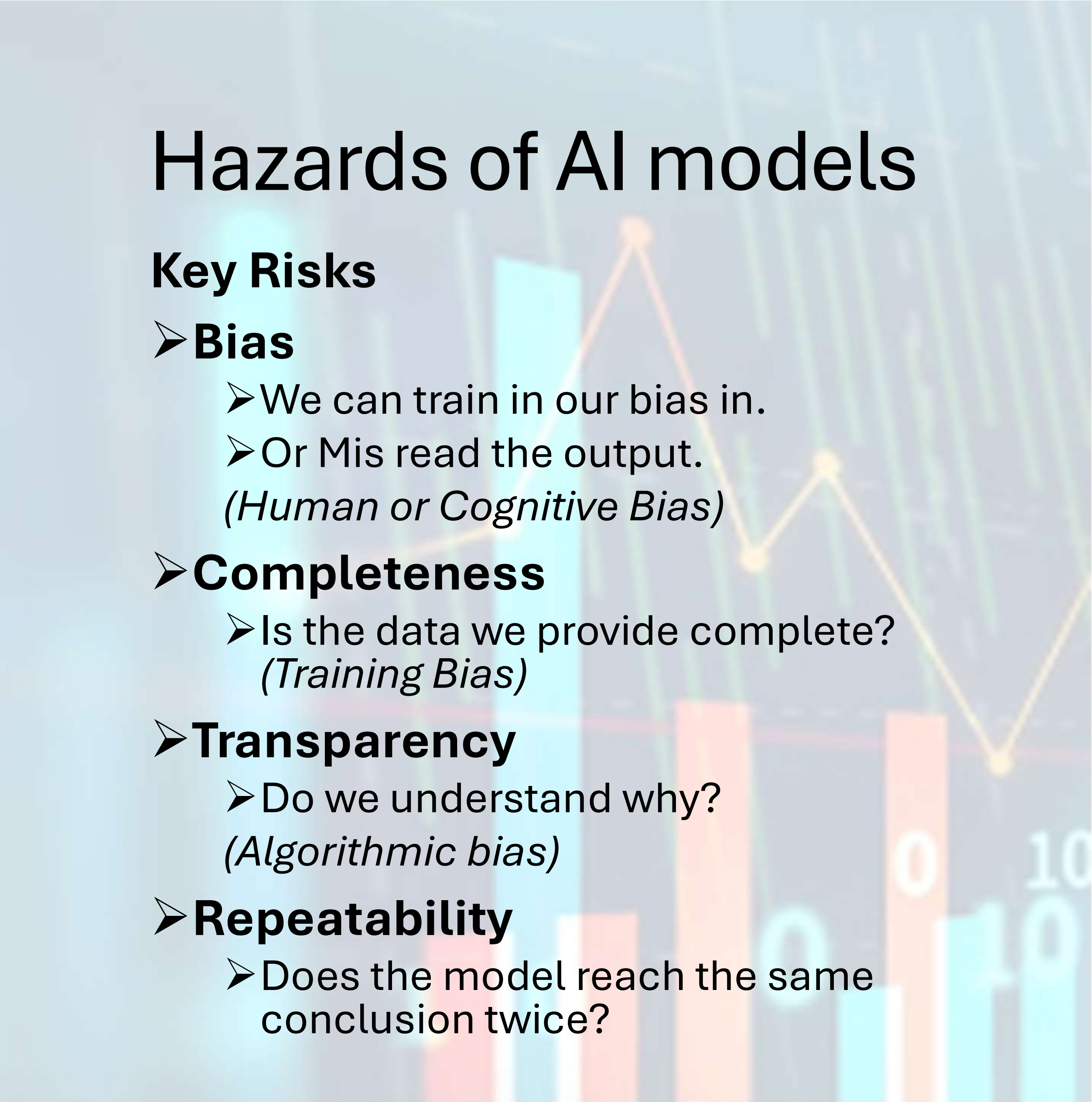
Indirect AI Use

- Using AI to analyze, interoperate and present data
- Preparation of Safety case
  - Language Models
- Analysis of system history
  - Data Models
- Presentation to Operator
  - Data / Presentation models
  - Alarms



It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness...

Charles Dickens – Google recommended this as a quote on AI...



Hazards of AI models

Key Risks

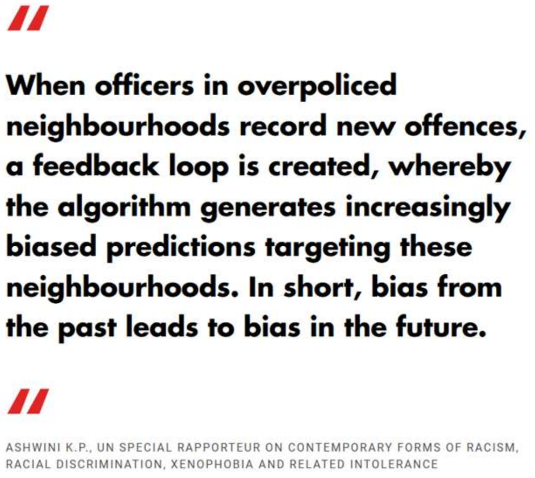
- ➢ **Bias**
  - ➢ We can train in our bias in.
  - ➢ Or Mis read the output.

  (Human or Cognitive Bias)
- ➢ **Completeness**
  - ➢ Is the data we provide complete? *(Training Bias)*
- ➢ **Transparency**
  - ➢ Do we understand why?

  (Algorithmic bias)
- ➢ **Repeatability**
  - ➢ Does the model reach the same conclusion twice?

Mitigations

- ➢ **Ready Reckoning**
  - ➢ Using simple, if crude models to indicate gross accuracy
- ➢ **Muli-path validation**
  - ➢ Using Diverse sources of information.
- ➢ **Limitation of Scope**
  - ➢ Only using the information where specific complete information is available.

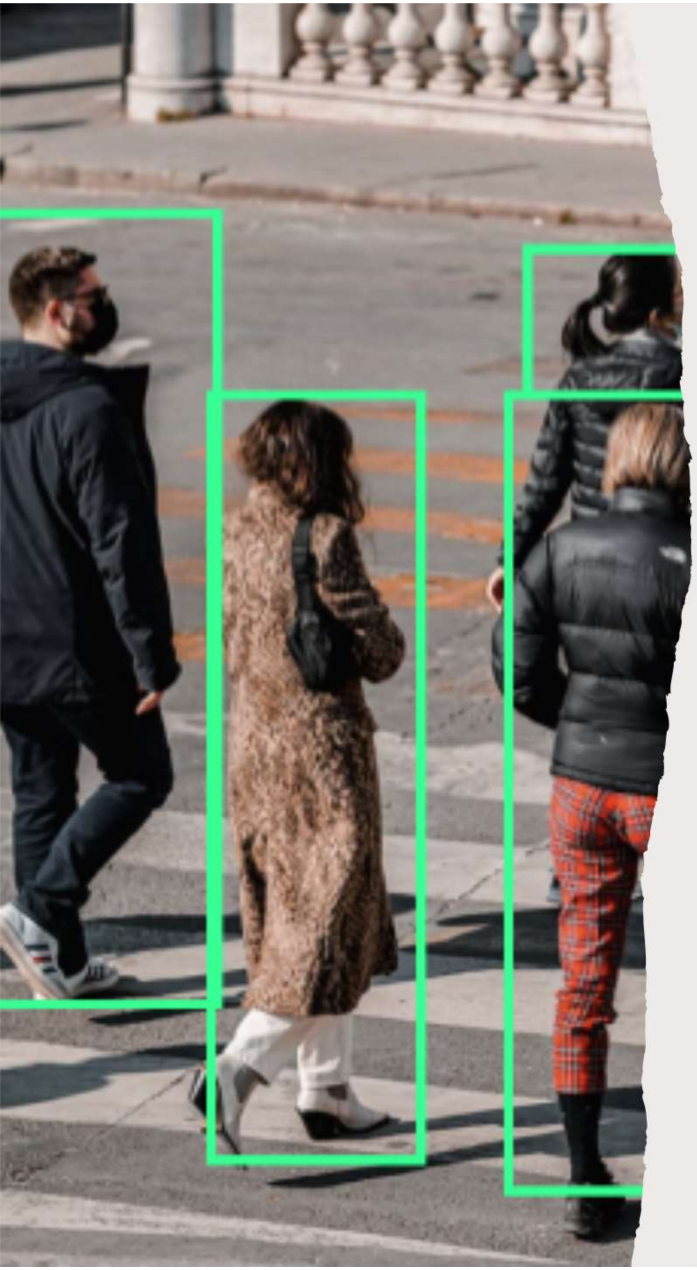


"

When officers in overpoliced neighbourhoods record new offences, a feedback loop is created, whereby the algorithm generates increasingly biased predictions targeting these neighbourhoods. In short, bias from the past leads to bias in the future.

"

ASHWINI K.P., UN SPECIAL RAPPORTEUR ON CONTEMPORARY FORMS OF RACISM, RACIAL DISCRIMINATION, XENOPHOBIA AND RELATED INTOLERANCE



Online Learning

- Online learning is seen as allowing the AI to refine it’s model “on the job”.
- This method presents AI technology with great potential to be able to refine to particular use conditions from a more generic model.
- The case may be a pipeline where the system learns how the flow, pressure, temperature sits within normal limits. This may change over time and online the AI will learn to adapt to these changes and optimise performance and uptime.
- This has a direct impact on a safety system in that the metrics of frequency of demand and rate of handover may change in time. This sets a time limit on the validity of assumptions made.