Undersea Defence Technology Conference, Stockholm, 13-15 May 2019

Towards a Safety Argument for Autonomous Systems that use Machine Learning

James McCloskey, Chris Allsopp and Chris Carter

Presented by Chris Carter – Business Manager Underwater Technologies

SYSTEMS AND ENGINEERING TECHNOLOGY

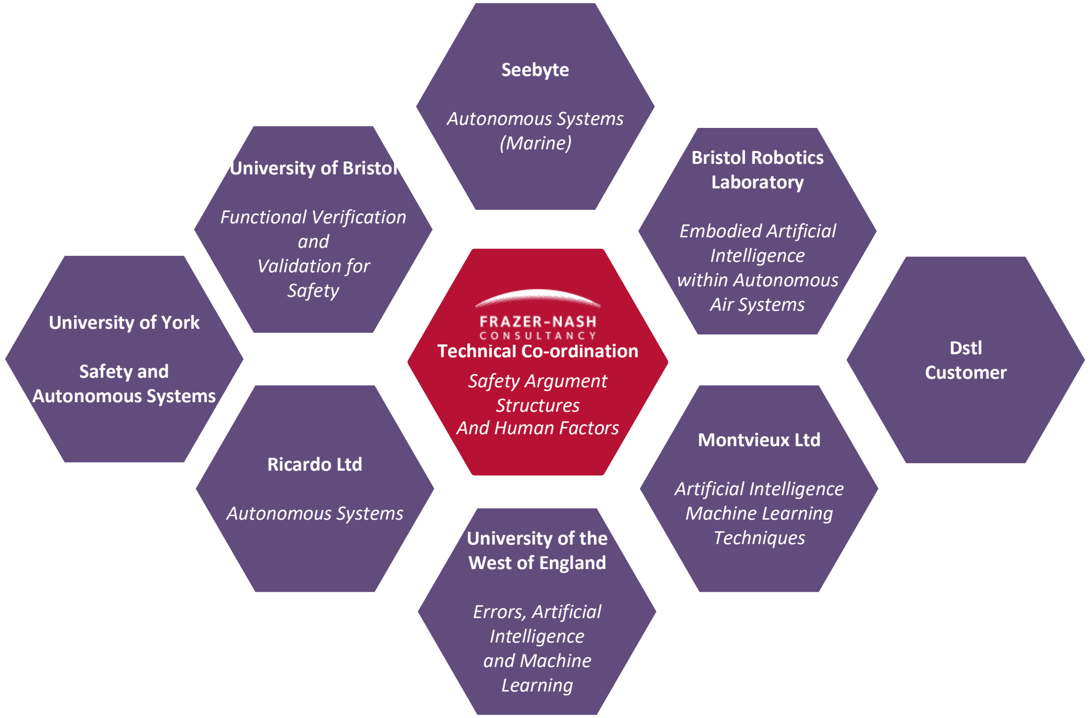
FRAZER-NASH CONSULTANCY

Overview

- Automation is set to bring benefits in efficiency, accuracy, and safety, but first we need to understand how to prove autonomous systems are safe to operate.

- The Defence Science and technology Laboratory (DSTL) commissioned this study, which Frazer-Nash led, with support from a team of academics and equipment manufacturers.

- The aim was to create a credible safety argument structure that can apply to autonomous systems of all types.



- This presentation will cover:

Concepts → Challenges → Frazer-Nash Approach → Way Ahead



Possible Underwater Autonomous System Applications

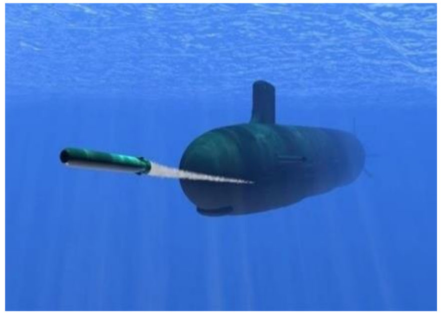
Manned Platforms

- Manoeuvring and navigation
- Platform management
- Protection systems (fire and flood)

Unmanned Platforms

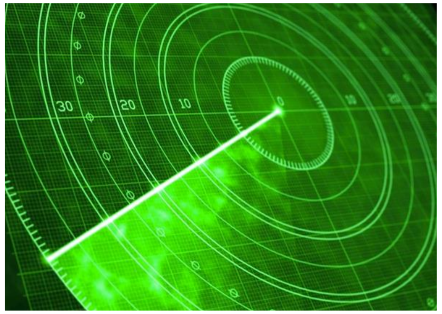
- Mine Sweeping
- Unmanned ASW
- Intelligence gathering
- Situation awareness

Weapons and CM



- Untethered torpedoes
- Smart countermeasures

Combat Systems

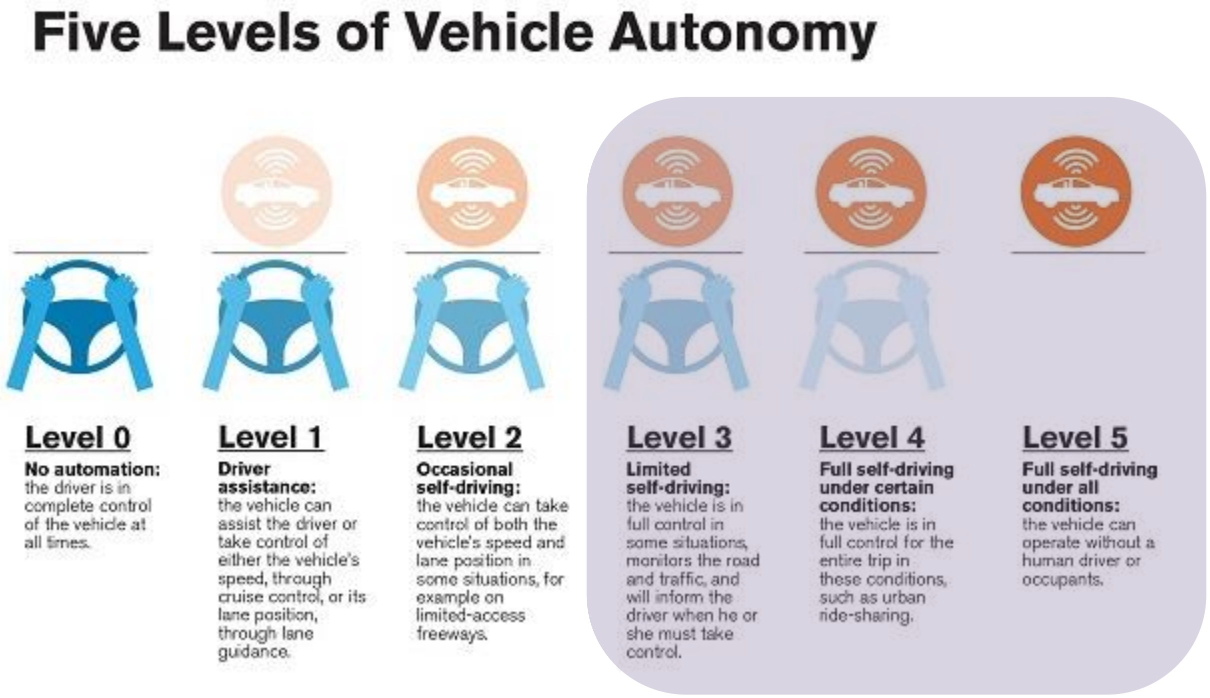


- Data triage, feature detection and analysis
- Contact tracking (radar, sonar, visual)
- Tactical support

Many of these applications are likely to need a robust safety argument

Concepts - What is Autonomy?

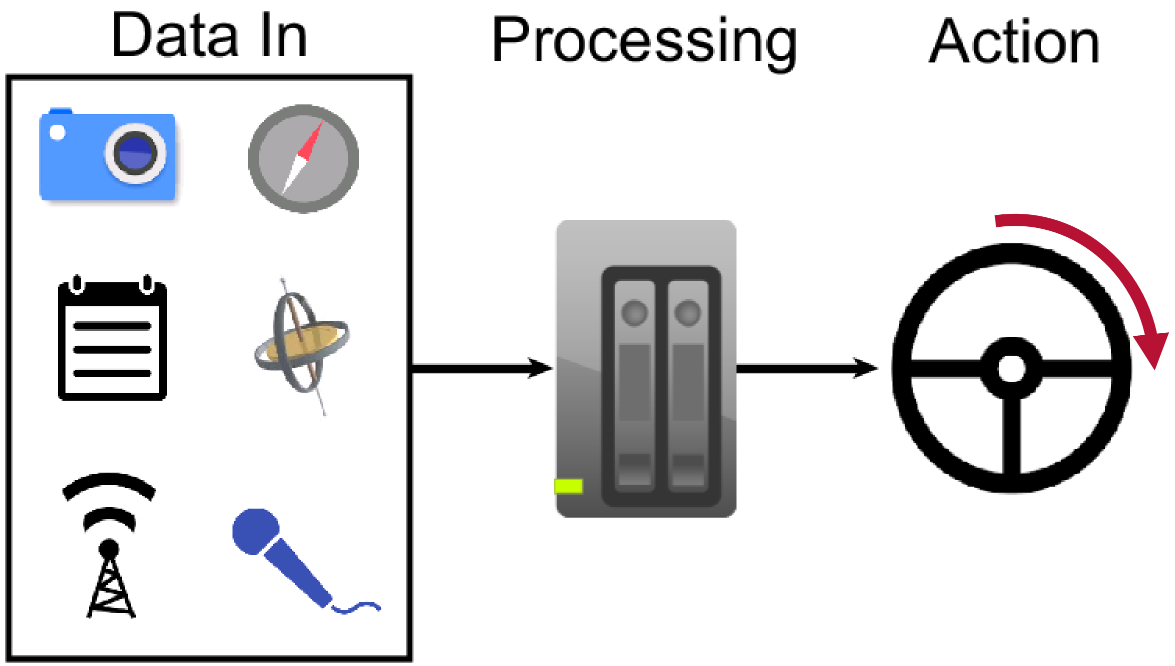
- Autonomy is when a machine performs a task without human assistance.
- The task can be simple (e.g. turning the brightness down when it's dark) or very complex (e.g. flying a UAV.)
- Simple autonomy can be achieved by a set of rules or behaviours
- Complex autonomy requires more complex approaches, e.g. machine learning.
- A platform can have a 'level' of autonomy
  - In this project we are concentrating on the more complex systems that cannot be covered by a simple rule set.



This is the area of focus for this project

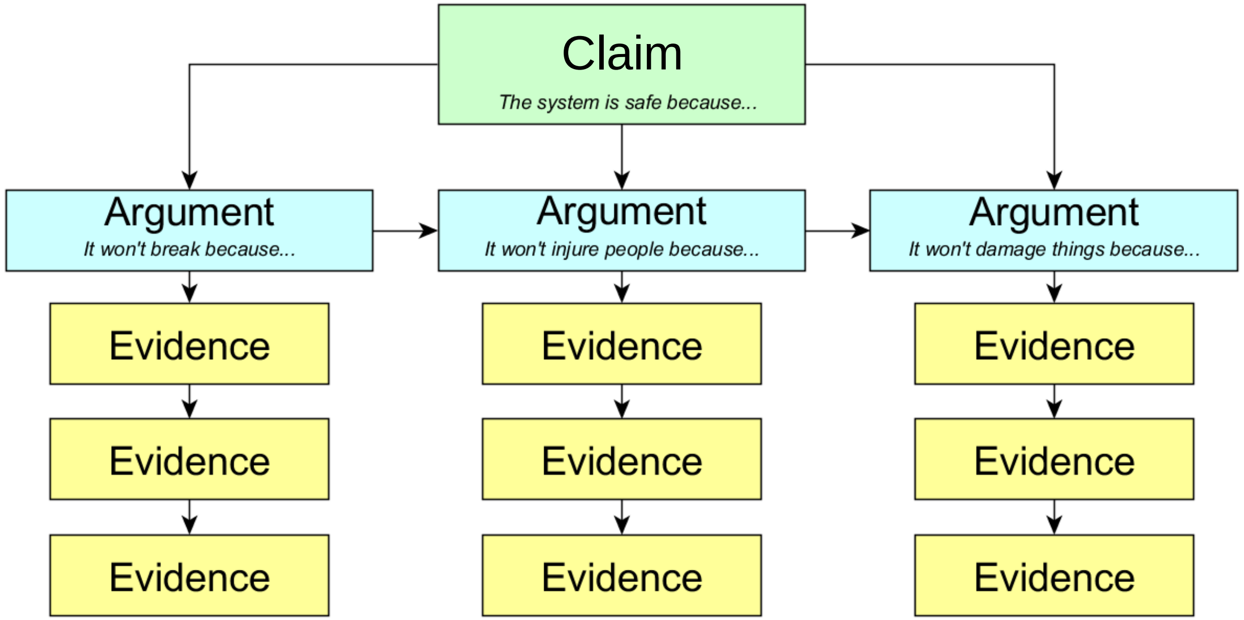
Concepts - What is AI?

- AI is **Complex automation**
- Artificial intelligence has many definitions:
  - Here we define it as a computer capable of making complex decisions and acting on them without input from a human.
- Systems can be *trained* or *learning*
  - *Trained* systems have fixed behaviour after leaving the factory.
  - *Learning* systems update their behaviour either during or between use.
- Field is rapidly developing.



Concepts - What is Safety?

- Safety cases demonstrate that a system is safe to operate in a certain way because of a number of provable factors.
- Depending on the impact of failure, proof can be demanding and required failure rates to be extremely low.
- It is difficult to take human performance uncertainty into account.
  - A human is a natural analogue to a complex AI system.
  - We have used experience of civil aviation and road vehicle safety cases to consider these issues and the interaction between AI and the human operator



Construction of a traditional safety case

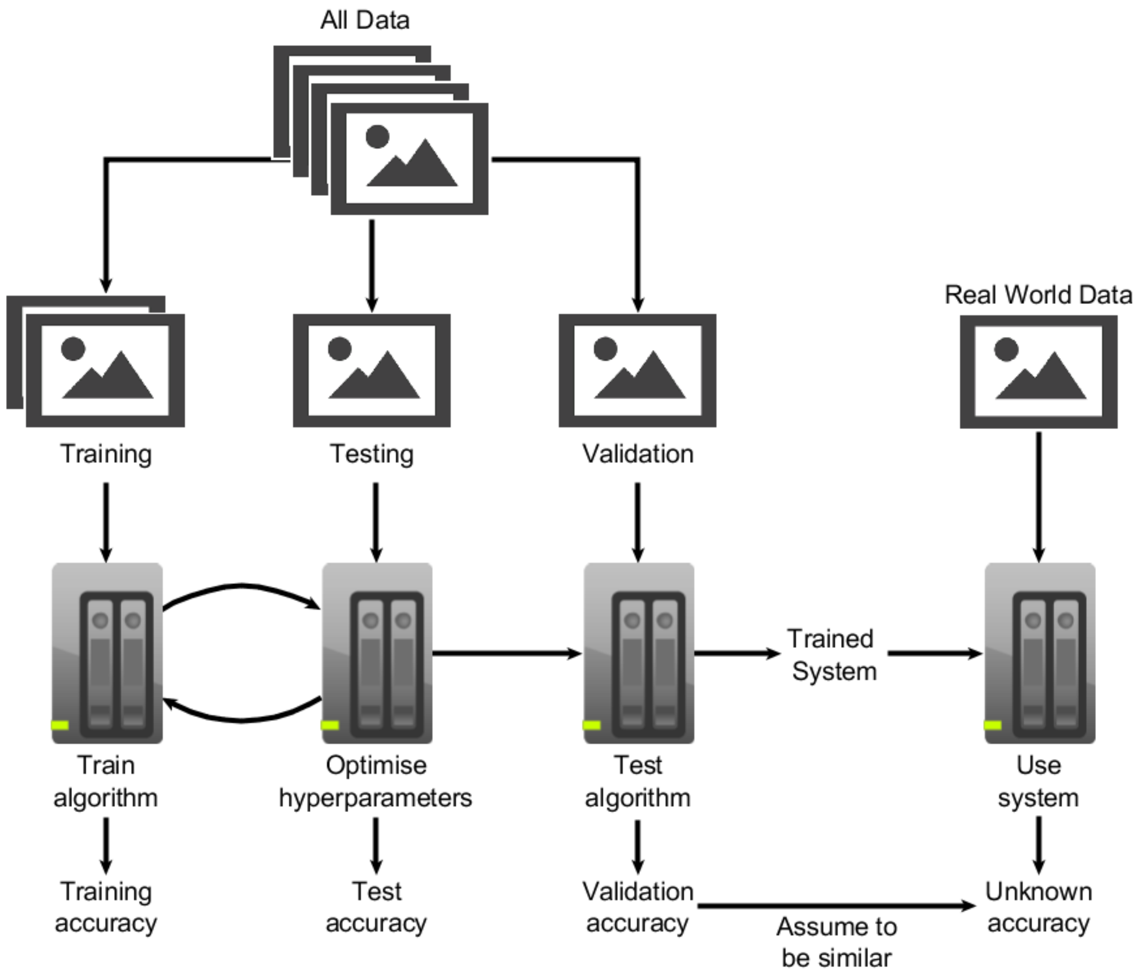
Problems - What are the main challenges with AI and Autonomy?

- Safety has no obvious way of handling AI:
  - Too opaque to consider as software.
  - Too unpredictable to consider as a component.
  - Too unaware to consider as a human.
- **Safety**
  - Rigorous and provable
  - Very detailed requirements
- **AI**
  - Lack of context
  - High failure rate (in safety context)
  - Very ‘black box’, even for developers
  - Tested rather than proven.
  - Designers tend to not think of safety first.

Defaced images recognised as ‘Speed Limit 45’.

Problems - Training & Performance Measurement

- A safety case will state that system failures occur at levels such as $10^{-5} – 10^{-9}$ (per hour) for high integrity systems.
- This covers all operational environments.
- A good AI program will demonstrate >95% accuracy on an industry standard challenge:
  - Probably worse in real operation
  - Failure rate is at least 1000 times higher than a typical high integrity system at 99% accuracy.
- AI performance measurement is done on a particular set of data.
  - Performance outside of set is assumed
  - Does the set cover all expected scenarios?

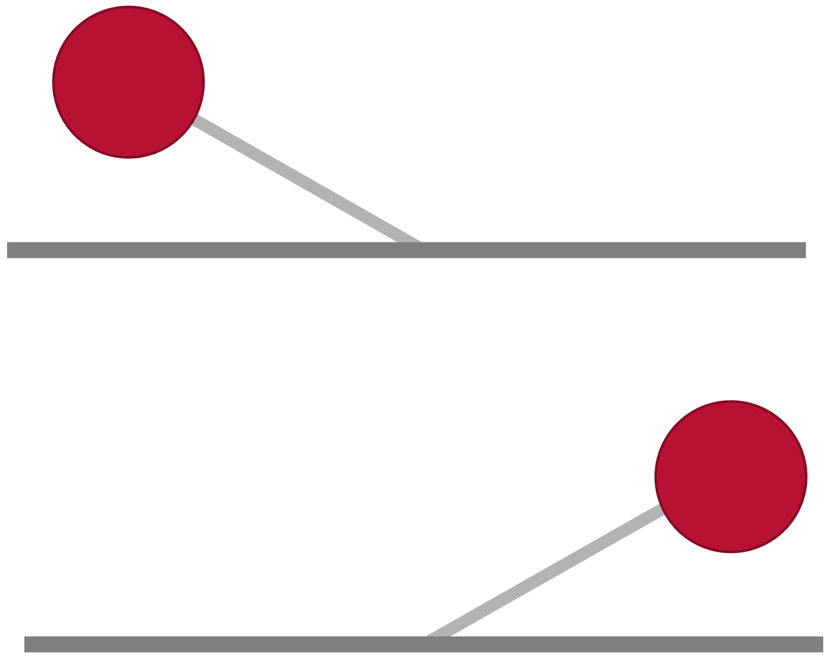


Problems - Example: Access Control

- An access control system has two functions:
  1. Allow access to a few specific people;
  2. Deny everyone else access.
- If I have 1000 people, 10 of whom have access, the system can achieve 99.9% accuracy by denying everyone access
  - Not good at function 1 though!
- For AI systems, performance is always a trade-off – no system is perfect!
- You either incorrectly:
  - Deny some people access (false negative)
  - Grant some people access (false positive)
- You choose which (and to what extent) based on the outcome of each error.

Problems - State Space

- Safety cases often demonstrate the outcome of all possible system states.
  - E.g. two levers with a set number of positions
- The total number of states can grow quickly if the number of dimensions (e.g. levers) and allowable states (e.g. lever position) increases.
  - E.g. 4 levers with 3 positions = 81 states
- The safety case can define what happens in each of these states and prove that it is safe.
- How does this apply in AI systems?



Problems - Data Coverage

- A typical image is made of pixels which have a value between 0 and 255 (3 values for RGB).
- For 4 pixels, the state space is 256*256*256*256 = 4.3 billion.
- Input space in AI can often be effectively infinite:
  - E.g. 512 x 512 pixel RGB image
  - Each pixel has 256*256*256 = 16.7M possible values
  - $(16.7M)^{(512*512)}$ is a <MATH ERROR>, or “very big number”
- How can we demonstrate adequate training / testing coverage in a space that large?
  - …but a lot of the input space is incoherent noise
- How can we say that our system has enough experience?

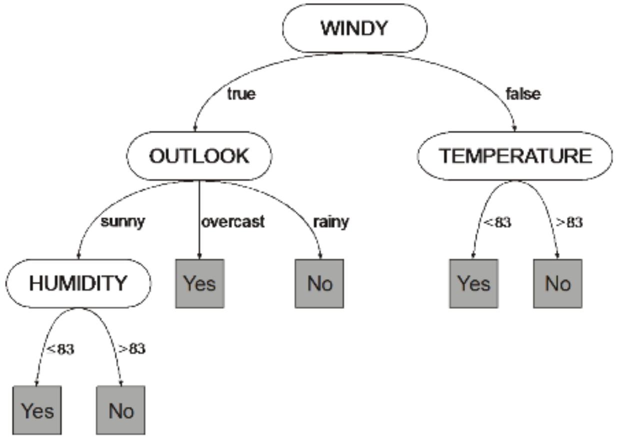
FRAZER-NASH CONSULTANCY

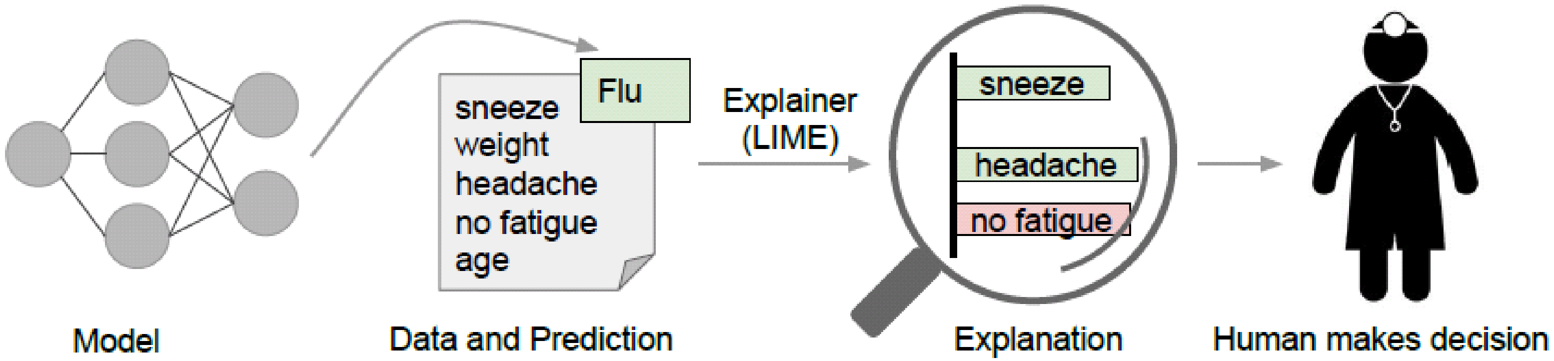
Problems - Data Coverage

- Instead of pure coverage of possible states, can we instead think of concepts and challenges?
- What can my system experience?
  - **Objects** (scale, position, number, orientation, occlusion)
  - **Lighting** (brightness, contrast, colour, saturation, reflections)
  - **Noise** (sensor, dirt)
  - **Motion** (blurring, shearing, jitter)
  - **Weather** (rain, sun, fog)
  - **Background**
- This space is much smaller and more understandable
- Still difficult to be exhaustive in a category, but can demonstrate resilience.
- Could industries or regulators assemble standard training / testing / validation sets?

Problems - Understanding AI

- In Safety, all parameters / Line-of-code can be traced back to a high level requirement.
- In a deep learning model, can we say with any confidence what a single parameter does?
- Situation is improving – ongoing research into explaining and visualising *why* the AI has made a decision:

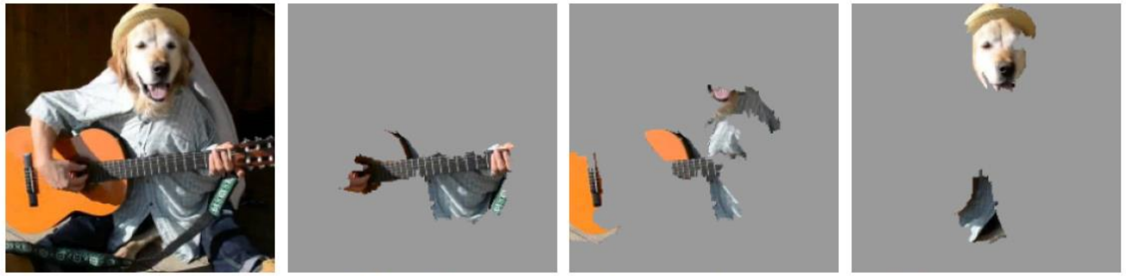




Problems - Understanding AI

- These ‘understanding’ techniques aren’t universal, and are focussed on imagery / classifiers at the moment
- They often require specific model types and need to be specified at the requirement stage.
- Different ways of explaining:
  - By reason: I think the image is a dog because of the nose and ears
  - By analogy: I think the image is a dog because it looks like this other image of a dog

- Understanding builds trust in the system and allows us to improve safety integrity



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

A woman is throwing a frisbee in a park. A dog is standing on a hardwood floor. A stop sign is on a road with a mountain in the background.

A little girl sitting on a bed with a teddy bear. A group of people sitting on a boat in the water. A giraffe standing in a forest with trees in the background.

Our Approach – The Safety Argument Scope

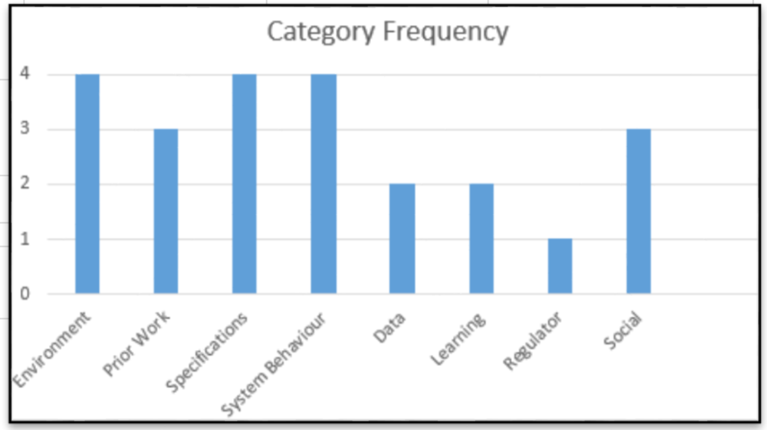
The approach aimed to:

- Facilitate discussion on existing AI safety problems
- Cover a range of scenarios
- Be realistic, solvable, and applicable to wider systems

Starting assumptions around the system:

- A fully autonomous system
- A single contained embedded system
- A single unit/agent/platform
- Humans in proximity of the operation
- A trained system, not a learning one
- An environment which is sufficiently complex to require AI

Topic	Problem Spectrum			Category
Social issues	No social issues	Known social issues	Unknown social issues	Social
Where is the AI located	Embedded, single platform	Remote brain	Hive mind	Specifications
Learning	Not learning	Adaptable behaviour	Fully changeable behaviour	Learning
Isolated?	No one around	Aligned / involved people	Non-involved / antagonistic people	Environment
Physical interaction	None (digital system)	Limited / stationary	Fully mobile	System Behaviour
Where was it trained / is training data representative	In operational env	In simulated op. env.	In similar env.	Data
Is the platform suitable for the task	Not Applicable / Yes	Mostly	Barely	Specifications
How visible is this system (can you interpret its parameters)	Black box	Some parameters understood, some hidden	fully understanadble	Learning
Expected performance	< Human			
How good would a human need to be to successfully complete mission using this system (platform appropriateness)	Novice			
Training data coverage	Full / near full			
Deterministic	Deterministic			
Cultural differences (e.g. japan walking to the left)	None			



Our Approach – Conclusions

5 Safety Principles

Principle 1

- Software safety requirements shall be defined to address the software contribution to system hazards

Principle 2

- The software detailed design shall embody the intent of the software safety requirements

Principle 3

- Software safety requirements shall be satisfied

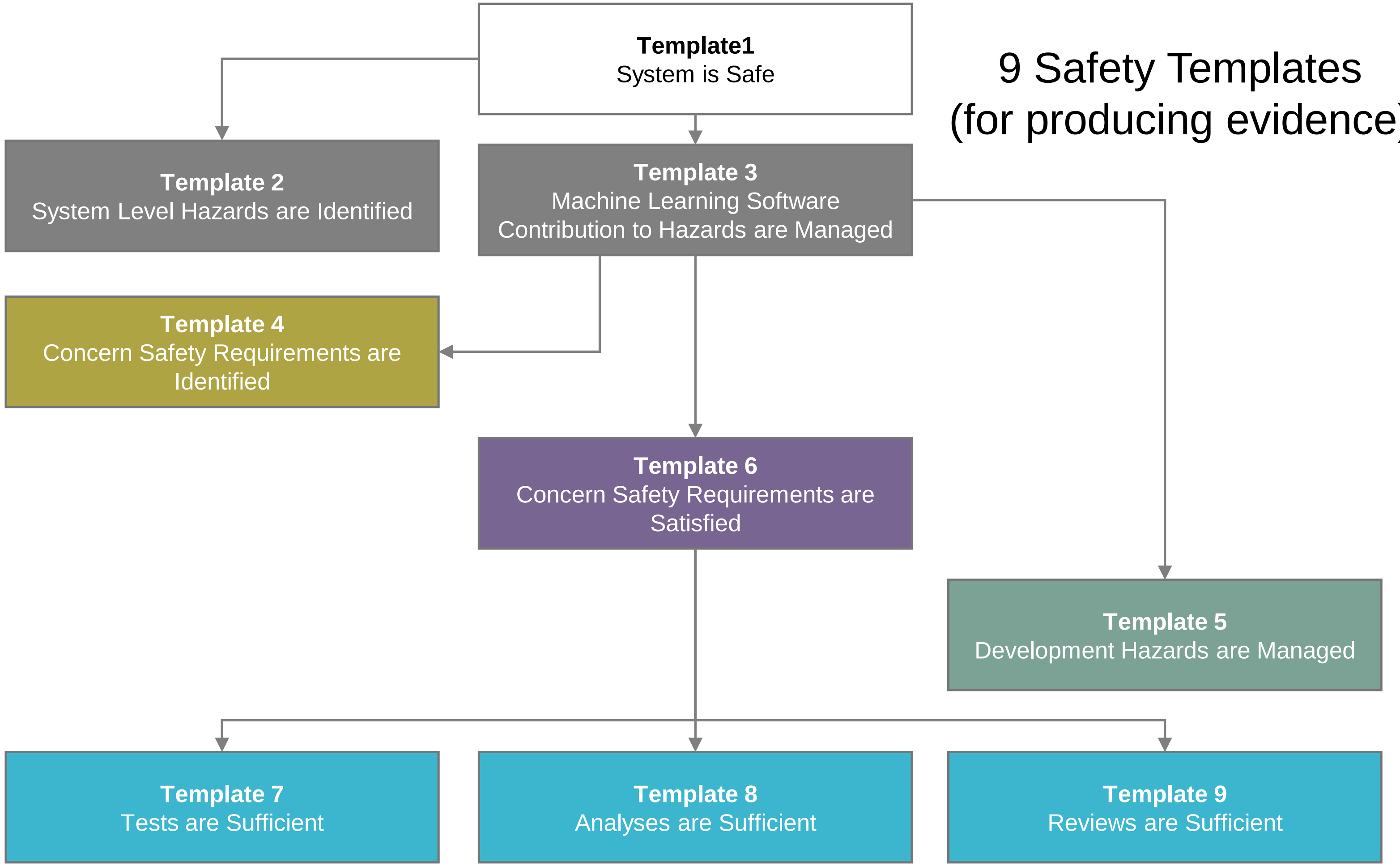
Principle 4

- Hazardous behaviour of the software shall be identified and mitigated

Principle 4+1

- The confidence established in addressing the software safety principles shall be commensurate to the contribution of the software to system risk

9 Safety Templates (for producing evidence)

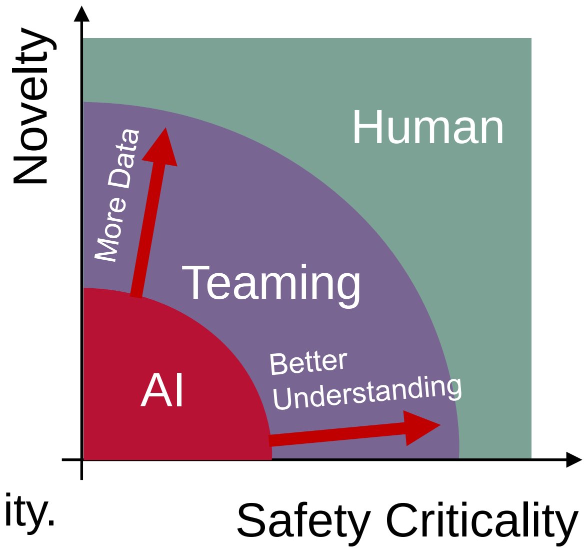


Concluding Thoughts

- Structure has been developed
  - Planning to publish output for use by government & industry
  - Looking for autonomy projects to demonstrate approach

- A number of interesting challenges for the future, not least:
  - Large volumes of data are required. Can we be smarter about generating this data?
  - Need to improve understanding of AI to enable higher integrity applications.
  - Focus on effective AI / Human Teaming for tasks with higher novelty or safety criticality.

- Enablers for AI in safety critical applications
  - Use of AI as an assistive technology, with fall back to traditional software to enforce the safety envelope (Control-monitor architecture).
  - Use of multiple and diverse ML software in a voting system – how to do quickly and consistently
  - Consideration of AI and ML as part of the operational safety case in place of the human operator



FRAZER-NASH CONSULTANCY

Naval Architecture
Marine Engineering
Requirements and MBSE
Operational Analysis
Safety and Environmental
Signatures, Shock and EMC
Technology Roadmapping
Structural Integrity
Materials and Corrosion
Submarine Propulsion
Hydrodynamics
Control & Instrumentation
ILS / AR&M
Electronics
Electrical Engineering
Platform Management Systems
Weapons Safety
Submarine Communications
Human Factors & Training
Safety Critical Software
Data Analysis
Geospatial Intelligence
Autonomy and Machine Learning
Information Security & Cyber

Thank you, any questions?

Chris Carter

Email: c.carter@fnc.co.uk

Tel: 01306 88 50 50

www.fnc.co.uk