UNIVERSITY of York

ASSURING AUTONOMY
INTERNATIONAL PROGRAMME

Assurance of ML: Challenges & Approaches

John McDermid

26th May 2022

Agenda

Key Topics

- Challenges of AI and ML
  - Fundamentals
- Approaches
  - Life cycle models
  - Assurance of AI and ML
- Wider Issues
  - Analysis of perception and decision-making
  - Ethical Issues
- Conclusions

Fundamental Challenges

AI/ML vs Human Decision-Making

- Autonomous systems
  - Transfer decision-making from human to machine (AI/ML)
  - ML learns future behaviour generalising from training data
- Humans have a semantic model, e.g., know what a bicycle is and its likely behaviour
  - Machines do not have these models
- Humans have contextual models, e.g., know what a roundabout is and the effects on driver behaviour …
  - Machines do not have these models

Fundamental Challenges

Trompe l'oeil

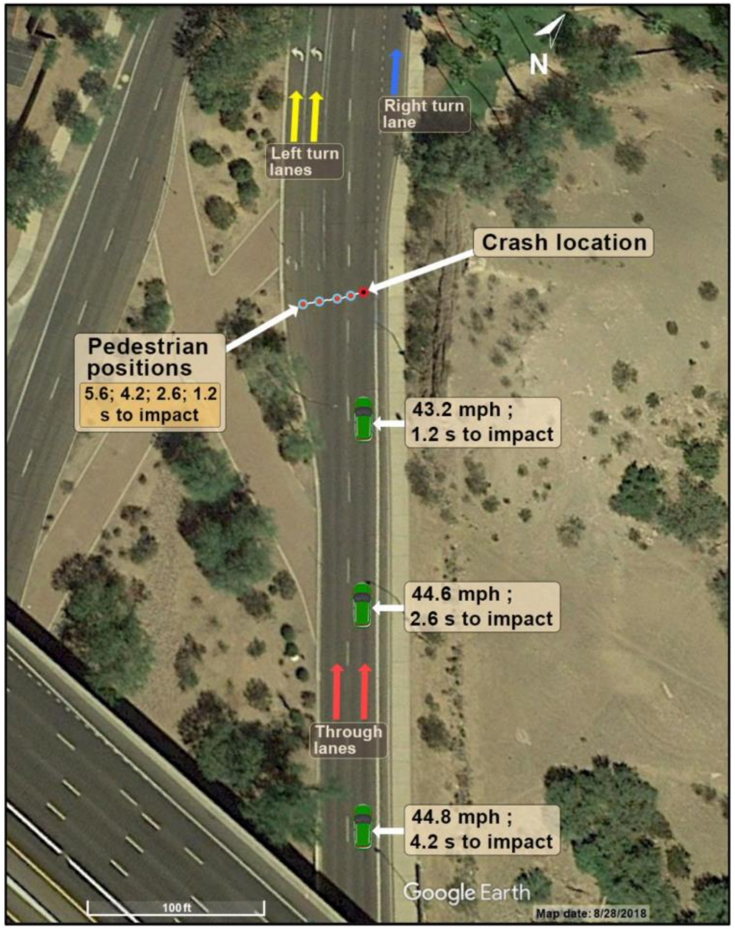
Fundamental Challenges

AI/ML Safety

- Safety processes assume
  - Know system boundary and it is fixed
  - Know (can specify precisely) system behaviour
  - Know system environment and can assess hazards
  - Life-cycle progressively adds detail so can analyse easily
- With AI/ML
  - Functional boundary unknown and may change
  - Behaviour not known precisely (learnt not specified)
    - Models can be opaque
  - Environment extremely complex (unpredictable)
  - Life-cycle highly iterative

Fundamental Challenges

Perception, Planning and More



Failure to regulate accountability for safety of automated driving

Inadequate engineering processes and lack of oversight of operators

Failure of operator to detect that system was not operating correctly

Failure of system to correctly detect pedestrian and avoid collision

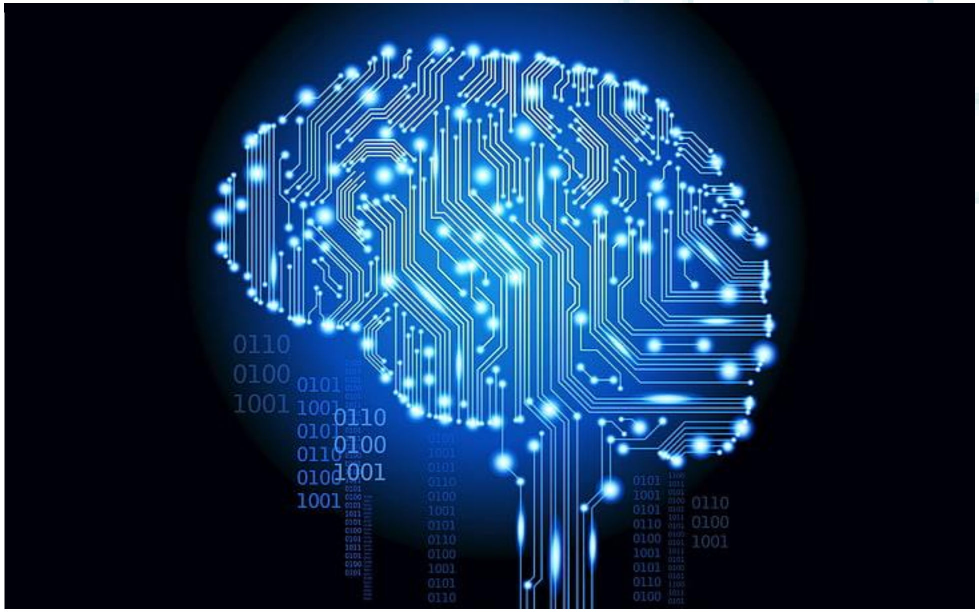
Agenda

Key Topics

- Challenges of AI and ML
  - Fundamentals
- Approaches
  - Life cycle models
  - Assurance of AI and ML
- Wider Issues
  - Analysis of perception and decision-making
  - Ethical Issues
- Conclusions

Take-Away 1

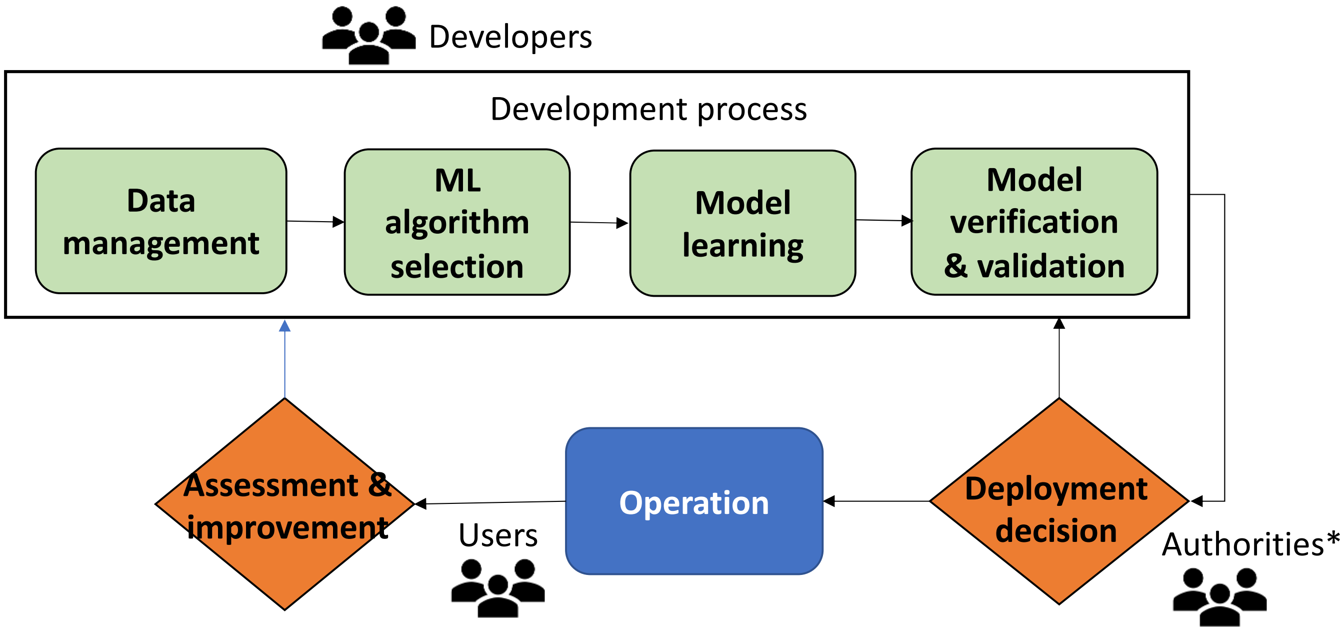
Safety Must Embrace ML



Safety must adopt ML models and methods to assure ML

ML Life-Cycle Model

Learning and Deployment



* may involve regulators, hospital managers, developers, insurers

Data Management

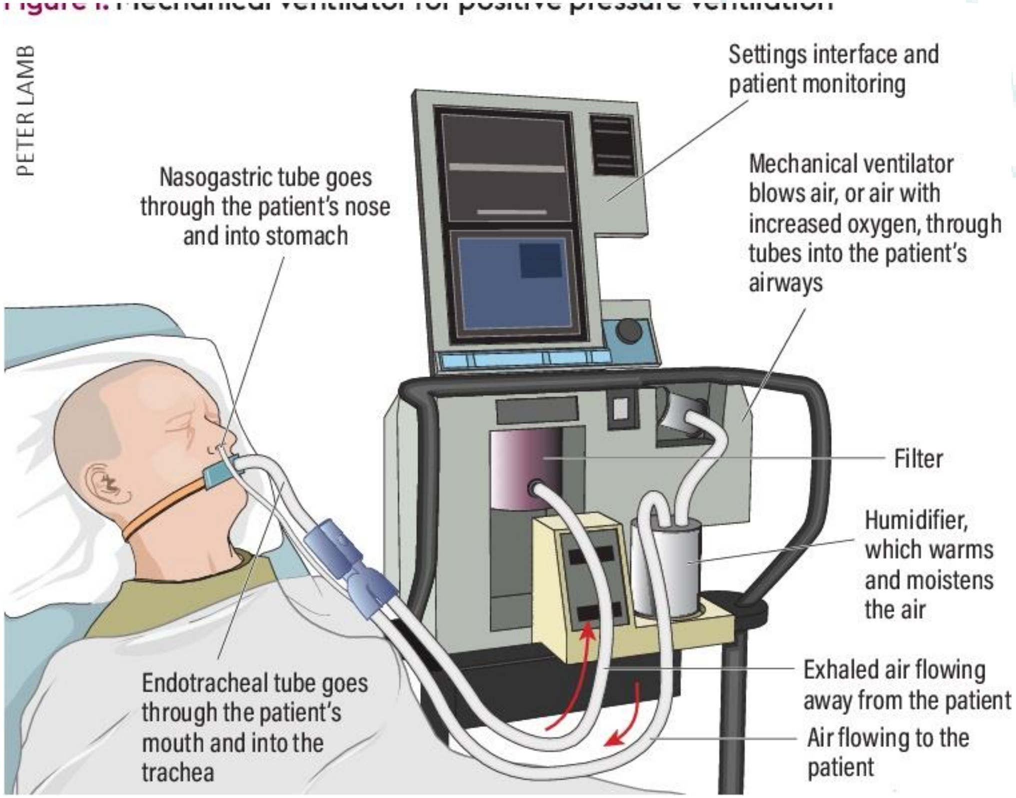
Garbage In – Garbage Out

- Data Management is critical – GIGO
  - Need to assure properties of data
- Criteria for data management
  - Conformant – data formats, units, etc. respected
  - Complete – all elements of records included
  - Accurate – reflects "ground truth"
  - Balanced – reflects the real-world distribution
  - Relevant – to the problem at hand, e.g. class of patient, road types, etc.

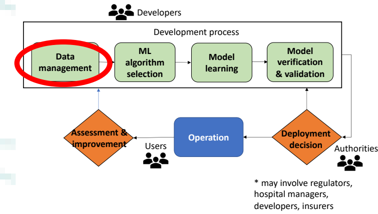
Illustrative Example

Weaning from Mechanical Ventilation

- Time of weaning from mechanical ventilation is critical
  - Too early, may lead to an emergency or reintubation
  - Too late, can lead to long-term effects, e.g. muscle damage
  - Clinically difficult judgment

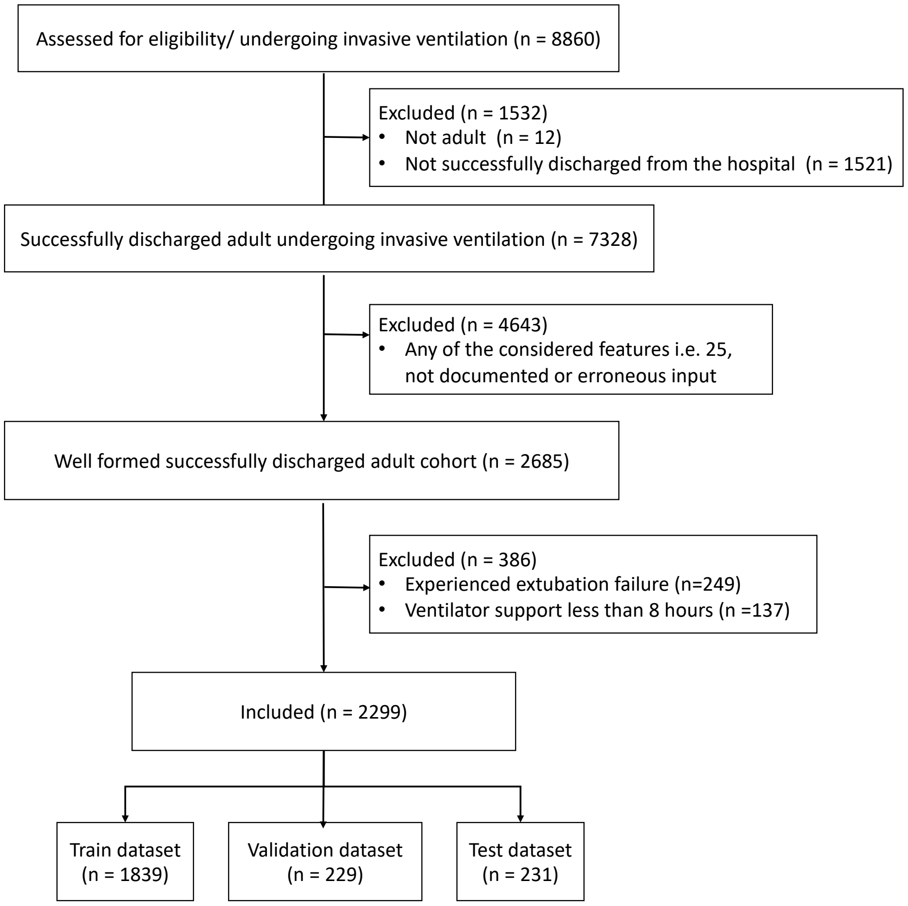


Illustrative Example



Data Selection

- Data selection shown diagrammatically
  - Data for training, verification and testing
- Shows data excluded if it is not conformant, complete, accurate or relevant
  - Balance depends on data sources



Model Selection

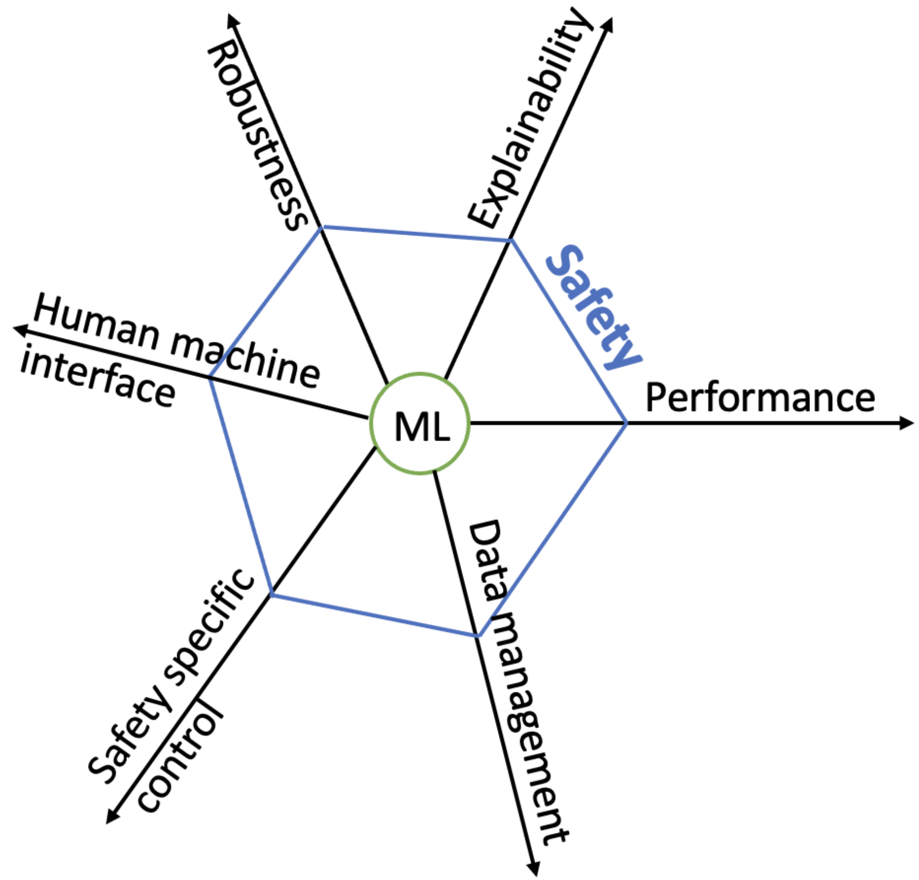
Performant and Assurable

- We need *ensure* the ML component works well
  - And to *assure* that it does so
- Some ML models are intrinsically explainable
  - Can interrogate the design to ascertain how decisions were made, e.g. classifying inputs
    - May be challenges with model size
- Some models are not “opaque”
  - But explainable AI (XAI) methods which can illuminate
  - Approximations to model behaviour

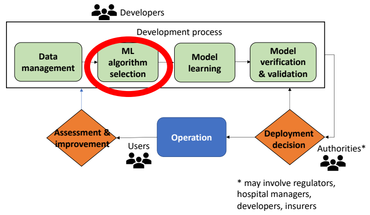
Take Away 2

A Trade-Space

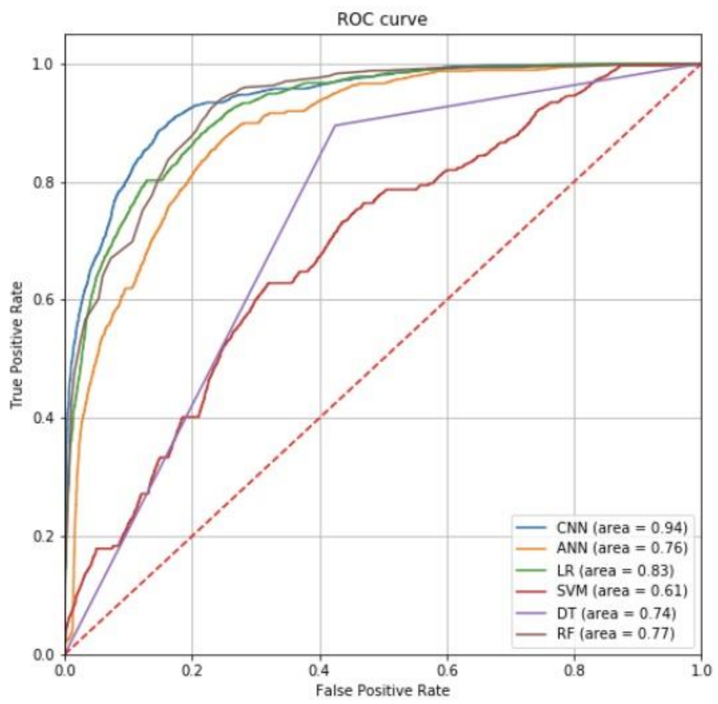
- Assurance is multi-faceted
  - Need to balance different facets
- Performance and explainability both important
  - Neither is over-riding
  - Model selection should consider both



Illustrative Example



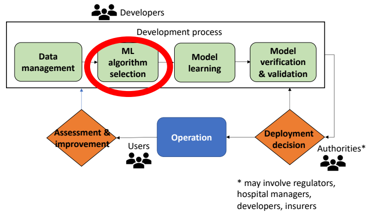
Comparing Model Performance



CNN	Convolutional Neural Network
ANN	Artificial Neural Network
LR	Logistic Regression
SVM	Support Vector Machine
DT	Decision Tree
RF	Random Forest

Area under the receiver-operator curve (AUC-ROC)

Illustrative Example



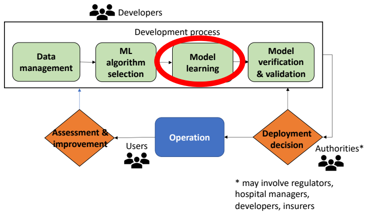
Performance Measures

- Logistic regression is the best performing of the intrinsically explainable models
  - But CNNs significantly better and there are XAI methods that can be used for CNNs

Methods	Accuracy	Precision	Recall	F1-Score	AUC
CNN	86%	82%	86%	84%	0.94
ANN	85%	84%	76%	79%	0.76
Logistic Regression	82%	78%	84%	79%	0.83
Support Vector Machine	70%	61%	61%	61%	0.61
Decision Tree	81%	76%	74%	74%	0.74
Random Forest Tree	87%	90%	77%	80%	0.77

Model Learning

Learning Safe Behaviour



- Models learn from the training dataset
- Performance is key
  - The model learning process focuses on meeting performance criteria
- Safety is also key
  - Performance criteria need to reflect safety constraints for the application
- Safety can influence model learning directly
  - Loss function “shaped” by safety considerations
  - Use of ML methods to improve robustness, etc.

Performance Requirements

Example 1.

Consider an autonomous driving application in which a subsystem may be required to identify pedestrians at a crossing. A component within the perception pipeline may have a requirement of the form “When Ego is 50 metres from the crossing, the object detection component shall identify pedestrians that are on or close to the crossing in their correct position.”

Take Away 3

Safety should Drive Design

- Good safety engineering ***improves*** design
  - Principle still applies with ML
  - Use classical safety methods, e.g. HAZOP, adapted if necessary to produce safety requirements

HAZOP

Table 1. Fragment of SHARD analysis showing a single hazard

Guide word	Deviation (Hazards)	Possible Causes	Effects	Severity
Incorrect	Sudden change of vasopressor dose is administered (concerns two consecutive doses)	1 Kink of line 2 The pump fails, e.g. due to electrical problem or bag/syringe not installed correctly 3 The delivery line might not be connected to patient's central line, e.g. due to the patient pulling out the central line 4 The drug might not be added to the diluent, so the syringe/bag just contains saline (a problem when bags/syringes are being changed over) 5 Initial recommendation by doctor has a sharp change in dose and doctor carried through the recommendation (not considered in this paper) 6 RL agent recommends a sharp change in dose and doctor accepts the advice, e.g. due to automation bias 7 Inappropriate titration of dose by nurse 8 Doctor fails to check current dose 9 Features in state space of the RL model are not sufficient to represent the patient conditions for sepsis decision making 10 Reward function used for RL model is coarse 11 Cost function used for RL model development is not appropriate 12 Hyperparameters used for RL model development are not optimised 13 Training data for RL model development is not appropriate 14 Nurse prepared wrong dose (e.g. due to calculation error) 15 Data corruption (e.g. invalid or wrong data produced by over-writing patient's features) 16 Features for wrong patient entered 17 Wrong patient feature values entered (e.g. due to unit difference) 18 Test results for wrong patient received 19 Incorrect test results received	Acute Hypotension, Strokes, Renal failure, Heart attack could occur from a sharp drop in the dose Hypertension, Cardiac Arrhythmia, Strokes, Raised intracranial pressure, Pulmonary oedema could occur from a sharp rise in the dose	Major/ considerable

DSRs

Table 4. Major changes in the modified RL model

	Features in state space (R1)	Cost Function(R3)
RL model in [32]	48	$L(\theta) = E[(Q_{double-target} - Q(s,a;\theta))^2] + \lambda_1 max(\lvert Q(s,a;\theta)\rvert - Q_{thresh}, 0)$
Modified RL model	48 (Removed one feature – timestep, added an extra one – relative dose change)	$L(\theta) = E[(Q_{double-target} - Q(s,a;\theta))^2] + \lambda_1 max(\lvert Q(s,a;\theta)\rvert - Q_{thresh}, 0) + \lambda_2 max(\lvert V_{change}\rvert - 0.75, 0)$ V_{change} is the agent recommended dose (argmax of $Q(s,a;\theta)$) minus the vasopressor dose in the previous step; λ_1 and λ_2 are the tuning parameters that decide how much to penalise the flexibility of the model.

ML Performance

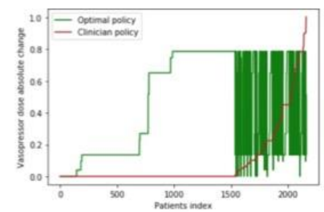


Figure 4. Original Policy: Comparison of max absolute vasopressor dose change in one step for each patient in the test data set between the clinician and the learnt optimal policy

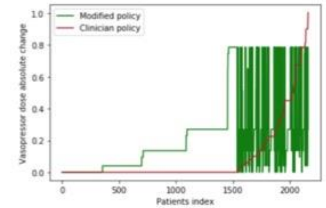


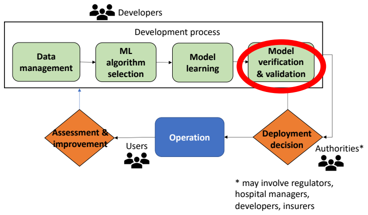
Figure 5. Modified Policy: Comparison of max absolute vasopressor dose change in one step for each patient in the test data set between the clinician and the learnt modified policy

Model V&V

Verification and Validation

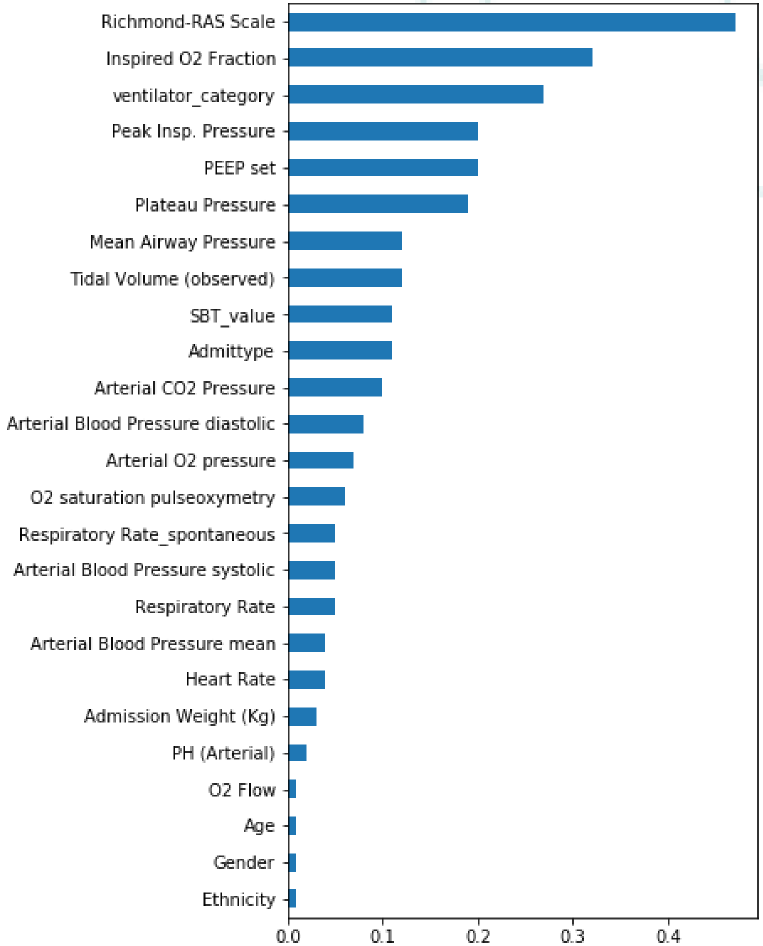
- Verification a core part of ML development
  - Undertaken as part of model development
  - Tested using separate dataset (recall three-way split of data in data management)
- Validation is concerned with how well the models work in the real world
  - On the road, in the clinical setting, etc.
  - Hard to evaluate prior to deployment
  - Explanations (XAI Methods) have a role to play in making the “black box” models open for validation

Illustrative Example



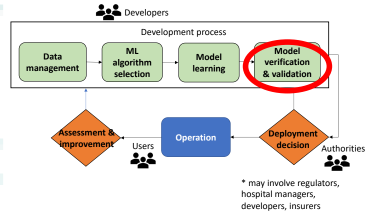
Explanations for Validation

- Validation needs to be carried out by clinicians
  - Example illustrates feature importance
  - Clinicians can judge if the ranking is plausible
    - Age, gender, ethnicity not relevant here (NB ethics)
    - No "absolute" but can refer to clinical literature and compare different models (CNNs "better" than ANNs)



Illustrative Example

Explanations and Robustness

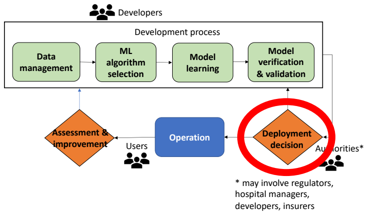


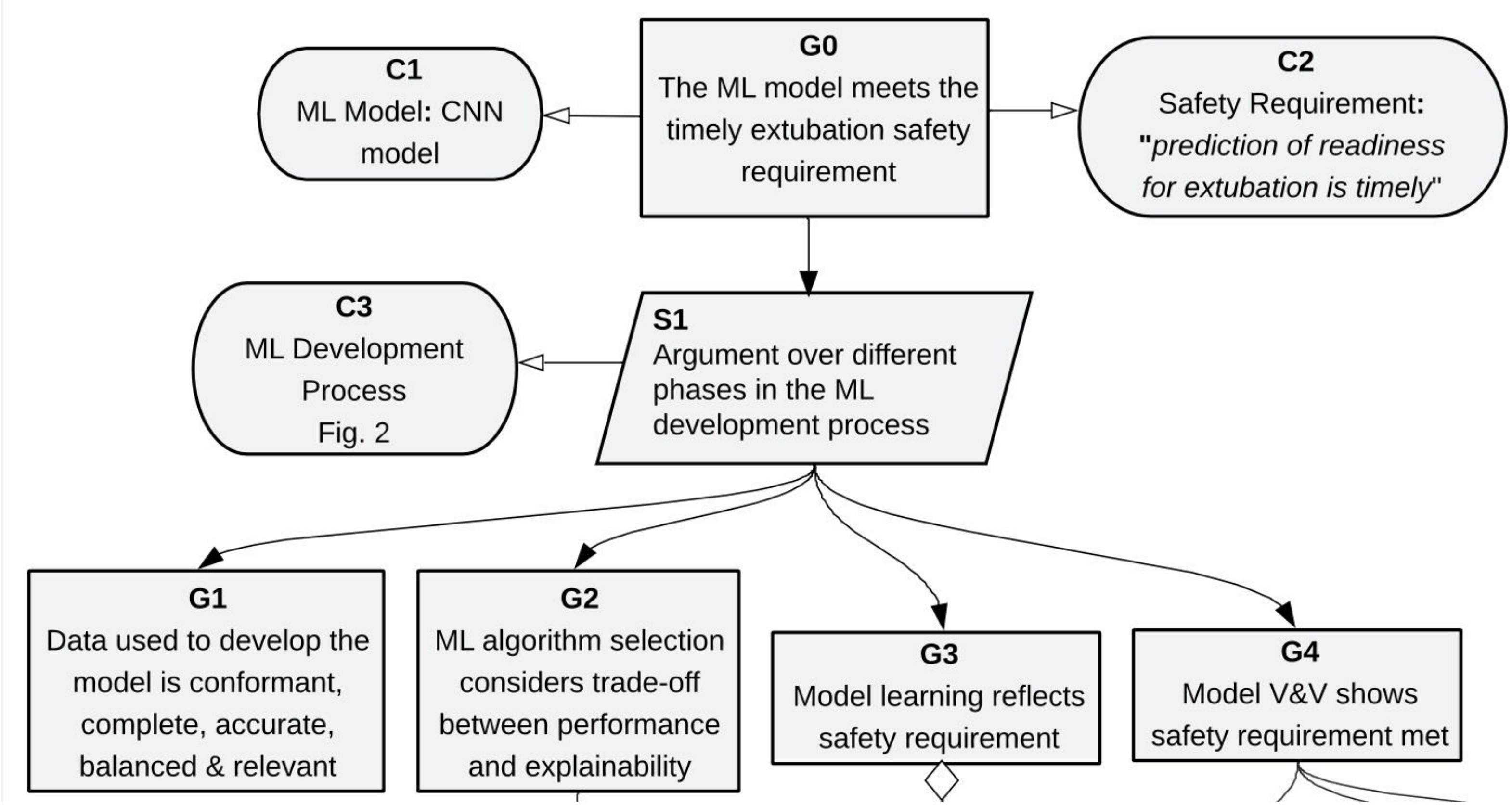
- Robustness is also important
  - Counterfactuals – input change to change output
  - How well the models cope with changes in inputs
  - Some similarity with ”no single point of failure criterion”

Features	Original instance	Counterfactual Examples 1	2	3
Admit Type	Emergency	—	—	—
Ethnicity	White	—	—	—
Gender	Female	—	—	—
Age	78.2	—	—	—
Admission Weight	86.5	—	—	—
Heart Rate	119	—	110	—
Respiratory Rate	24	26	—	—
SpO2	98	—	—	96
Inspired O2 Fraction	100%	—	40%	—
PEEP set	10	5	5	5
Mean Airway Pressure	14	—	10	—
Tidal Volume (observed)	541	—	—	560
PH (Arterial)	7.46	—	—	—
Respiratory Rate(Spont)	0	—	24	—
Richmond-RAS Scale	-1	—	0	—
Peak Insp. Pressure	21	—	—	—
O2 Flow	5	—	—	—
Plateau Pressure	19	—	—	—
Arterial O2 pressure	124	108	118	—
Arterial CO2 Pressure	33	—	—	—
Blood Pressure (systolic)	101	—	—	—
Blood Pressure (diastolic)	65	—	—	—
Blood Pressure (mean)	76	—	—	—
Spontaneous breathing trials	No result	Successfully Completed	Successfully Completed	Successfully Completed
Ventilator Mode	CMV/ASSIST/ AutoFlow	PCV+	SIMV/PSV	SIMV/PSV
Predicted outcome	0.93	0.44	0.17	0.36

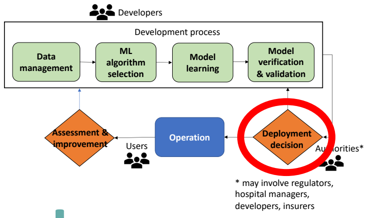
Illustrative Example

Safety Case

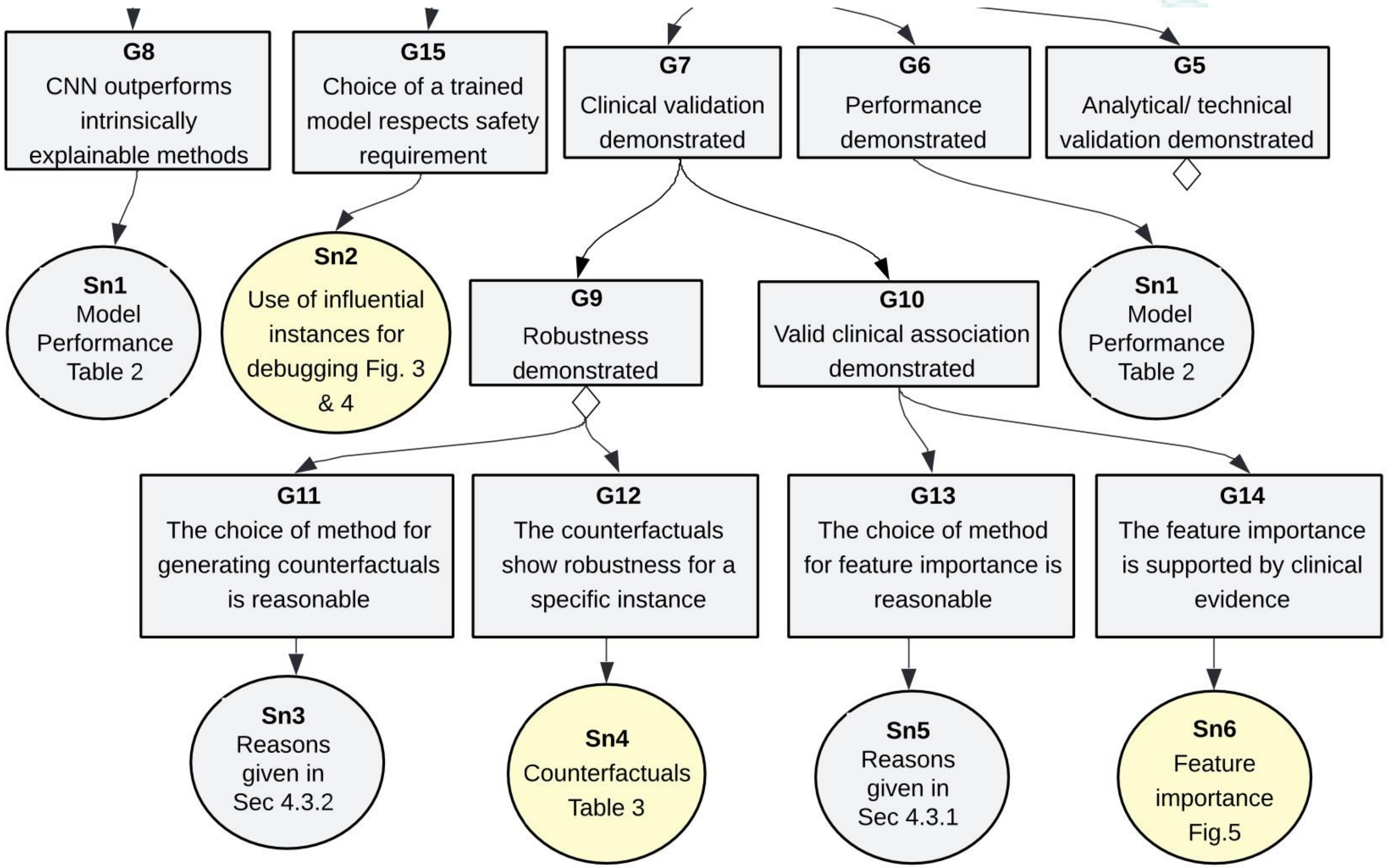




Illustrative Example



Role of Artefacts across Life Cycle



Agenda

Key Topics

- Challenges of AI and ML
  - Fundamentals
- Approaches
  - Life cycle models
  - Assurance of AI and ML
- Wider Issues
  - Analysis of perception and decision-making
  - Ethical Issues
- Conclusions

Wider Issues

ML is part of a Wider System

- Need to analyse the system as a whole
  - Socio-technical system, e.g. in healthcare
  - Technical system e.g. autonomous vehicles
  - Complex environment – technical, human, organisation
- AAIP is addressing the system issues
  - Assurance and safety analysis processes
- AAIP also considers applications across domains
  - Land, sea, air, healthcare, space, quarrying/mining, factory automation, solar farms … including tailoring

AAIP Research Strategy

Key Research Pillars

- Five pillars defining a safety and assurance process for robotics and autonomous systems
  - Societal Acceptability of Autonomous Systems (SOCA)
  - Safety of Autonomy in Complex Environments (SACE)
  - Safety Assurance of Understanding in AS (SAUS)
  - Safety Assurance of Decision-Making in AS (SADA)
  - Assurance of Machine Learning for AS (AMLAS)
- Producing 5 linked manuals/guides for use by engineers, developers and regulators
  - But ***generic***, so need ***tailoring*** to application domains …

AMLAS

ASSURING AUTONOMY
INTERNATIONAL PROGRAMME

Guidance on the Assurance of Machine Learning in Autonomous Systems (AMLAS)

Richard Hawkins, Colin Paterson, Chiara Picardi, Yan Jia, Radu Calinescu and Ibrahim Habli.

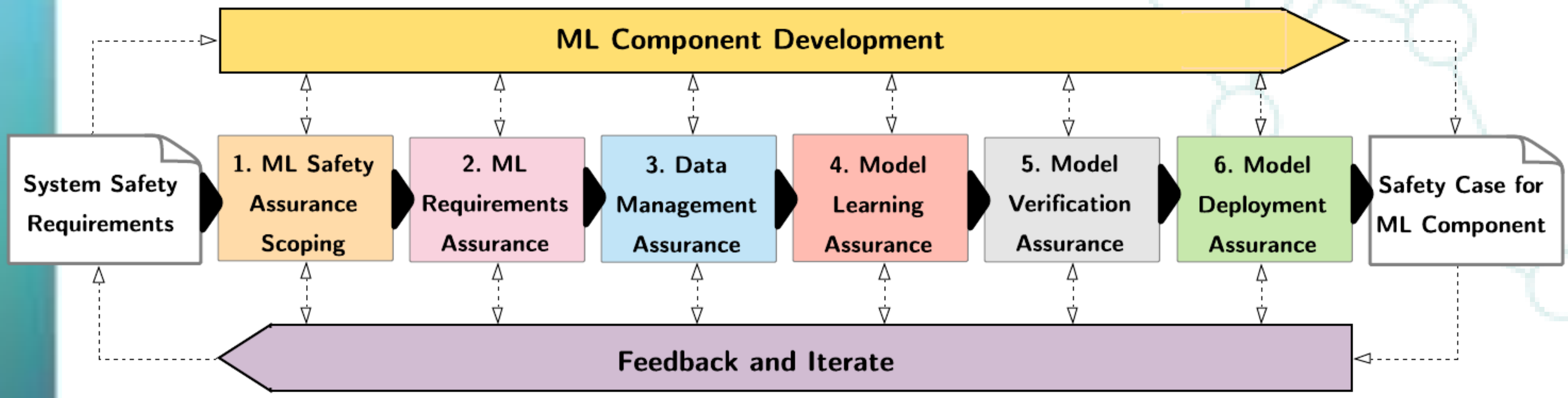
Assuring Autonomy International Programme (AAIP)
University of York

Version 1, February 2021

- AMLAS provides
  - Defined **process**
  - Set of **safety case patterns**
- **AMLAS enables**
  1. Integration of safety assurance into development of ML components
  2. Generation of evidence base for justifying acceptable safety
- **Resulting in structured safety case for ML component**
  - Which will become part of the overall (AS/AV) safety case

https://www.york.ac.uk/assuring-autonomy/guidance/amlas/

AMLAS Overview

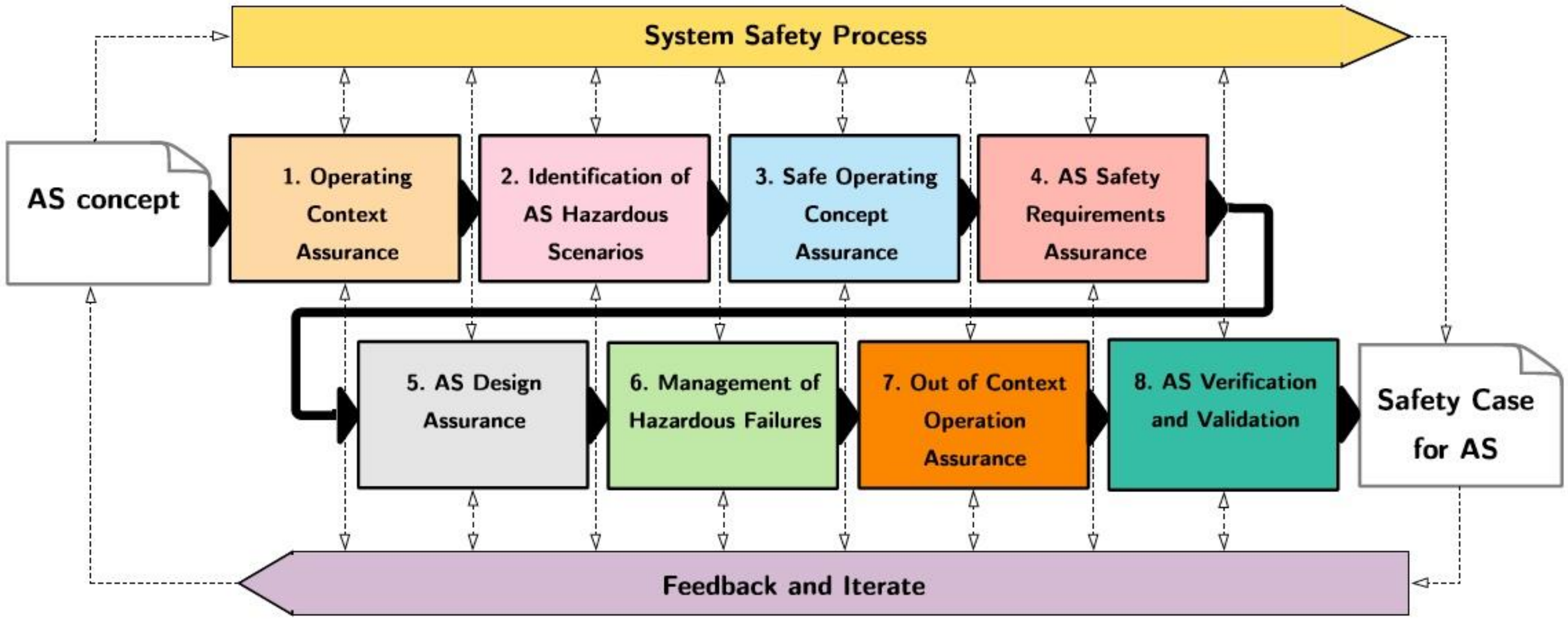


- For each stage AMLAS provides
  - Process description
    - Defined activities and artefacts (evidence)
  - Safety argument pattern

SACE

System in Context

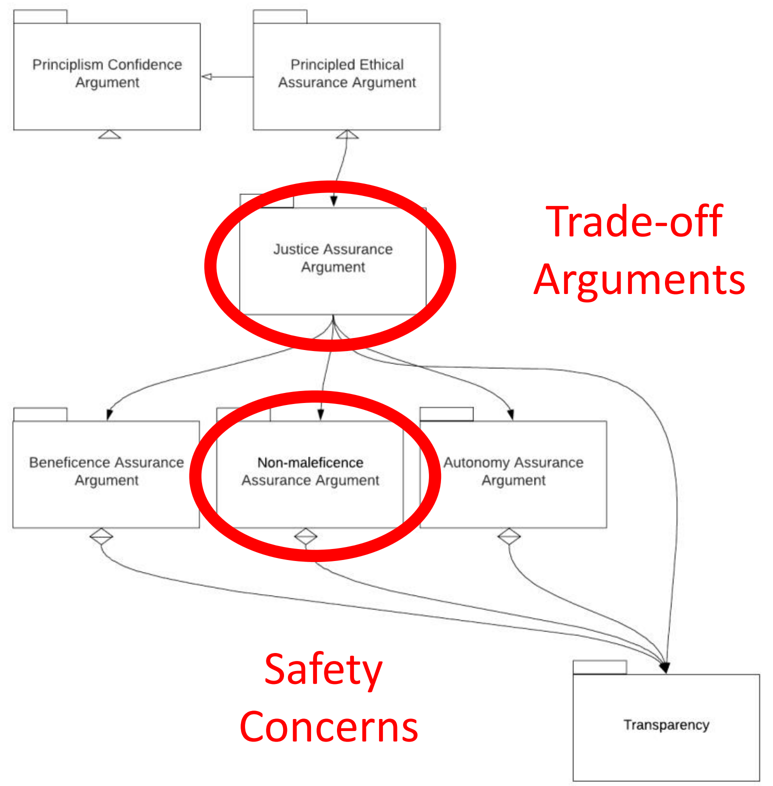
- Connects with AMLAS
  - Safety requirements flow down to ML components



Ethical Assurance

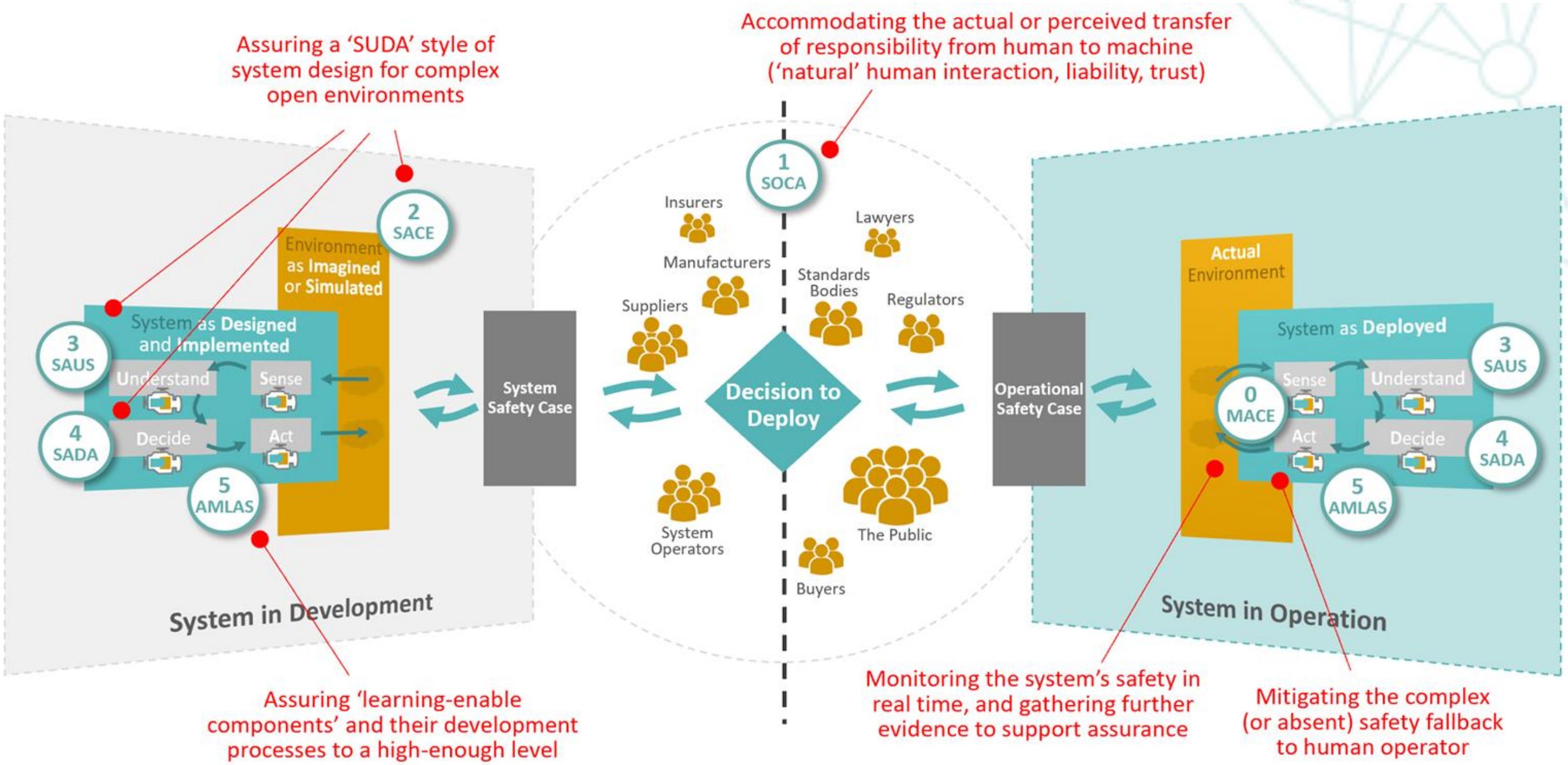
Drawing on Biomedical Ethics

- Can adapt safety arguments to include ethical issues
  - Central argument relates to beneficence (do good), maleficence (do no harm) and (human) autonomy
    - Supported by transparency
  - Principles are defeasible, so admit trade-offs
    - Would be reflected in the justice argument



Operational Monitoring

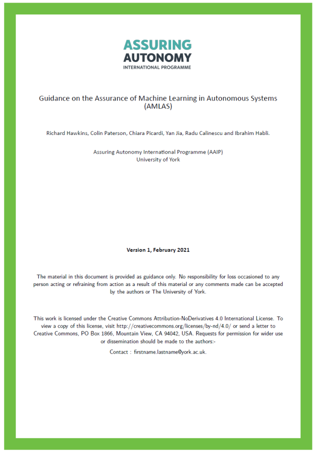
Monitoring to "Close the Loop"



Take Away 4

Specific and Generic

- Generic is ***valuable***
  - Identify all the dimensions of interest
    - Safety requirements, data management ...
  - Reusable across domains



ASSURING AUTONOMY
INTERNATIONAL PROGRAMME

Guidance on the Assurance of Machine Learning in Autonomous Systems (AMLAS)

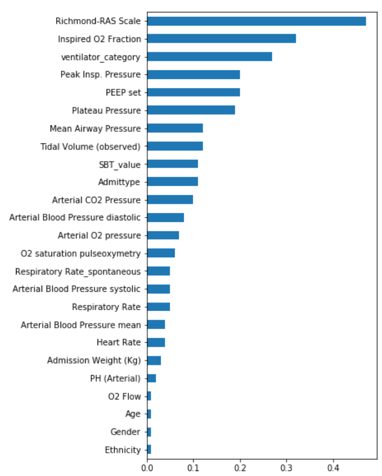
Richard Hawkins, Colin Paterson, Chiara Picardi, Yan Jia, Radu Calinescu and Ibrahim Habli.

Assuring Autonomy International Programme (AAIP)
University of York

Version 1, February 2021

The material in this document is provided as guidance only. No responsibility for loss occasioned to any person acting or refraining from action as a result of this material or any comments made can be accepted by the authors or The University of York.

Contact : firstname.lastname@york.ac.uk.

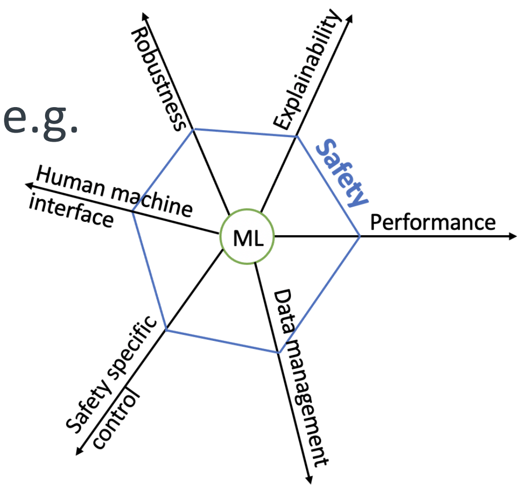


- Tailoring to a domain is ***essential***
  - Adopts its language
  - Addresses particular concerns, e.g. explainability
  - Illustrative example reflects AMLAS, but particularised to healthcare

Conclusions

Plus ça change, plus c'est la même chose

- Assurance of AI/ML-based systems poses unique challenges
  - ML developed iteratively, not via a conventional life-cycle, opacity of learnt models, etc. – plus ça change
- Must adopt & adapt established safety engineering methods
  - Hazard analysis, derived safety requirements, etc – plus c'est la même chose
- Safety and ML need to "embrace each other", e.g.
  - Apply ML methods to assuring safety of ML
- Recognise that assurance is multi-faceted
  - A lot to do, but a much already done
- International collaboration needed to solve these challenging problems



References

Where to learn more

- AAIP: https://www.york.ac.uk/assuring-autonomy/
- AMLAS: https://www.assuringautonomy.com/amlas
- Illustrative (weaning) example: https://ieeexplore.ieee.org/abstract/document/9769937
- Safety-driven design in healthcare: https://www.sciencedirect.com/science/article/pii/S1532046421000915 (example on slide 19)
- Ethical assurance argument: https://arxiv.org/abs/2203.15370

UNIVERSITY of York

ASSURING AUTONOMY
INTERNATIONAL PROGRAMME

Addressing **global challenges** in **assuring the safety** of robotics and autonomous systems