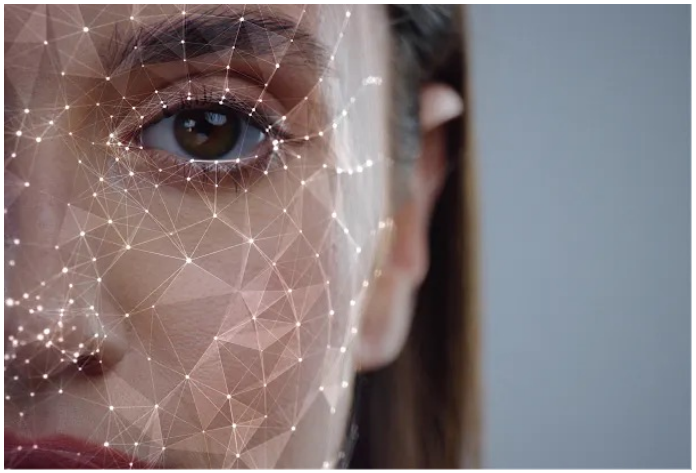
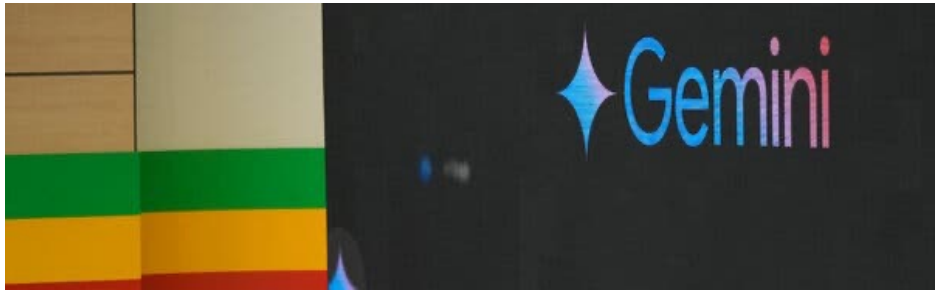
Provably robust artificial intelligence? A formal methods perspective

Prof. Marta Kwiatkowska
Department of Computer Science
University of Oxford

AI in everyday use…



A Tesla Model S



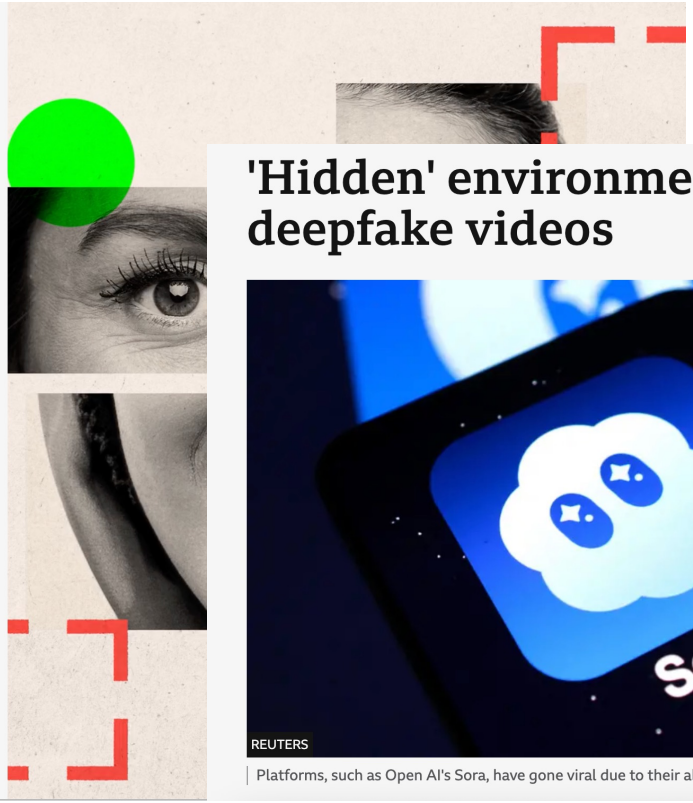
…and in the news

Animations off

BBC INDEPTH

The debate: Are facial recognition cameras in shops a step too far?

FALSE 24 October 2025



'Hidden' environmental impact of deepfake videos

Platforms, such as Open AI's Sora, have gone viral due to their ability to make hyper-realistic videos

Broad societal debate, seeking purpose and deployment opportunities

Safety and security risks in AI decision making

- AI decisions rely on neural network components
- Well known that neural networks are unstable to adversarial perturbations

Physical attack

Lightbeam attack

Patch attack

Real traffic sign

- For high-stakes applications, need provable guarantees on correctness
- Yet AI/ML community focuses on performance – formal verification to the rescue?

Safety Verification of Deep Neural Networks. CAV 2017 keynote

Michael Passingham
Senior researcher & writer

New research from Which? reveals that more than half of drivers are turning off safety tech in their cars with many finding the tech annoying, distracting or even dangerous. We explain why, and how to get the best from your current car or your next purchase.

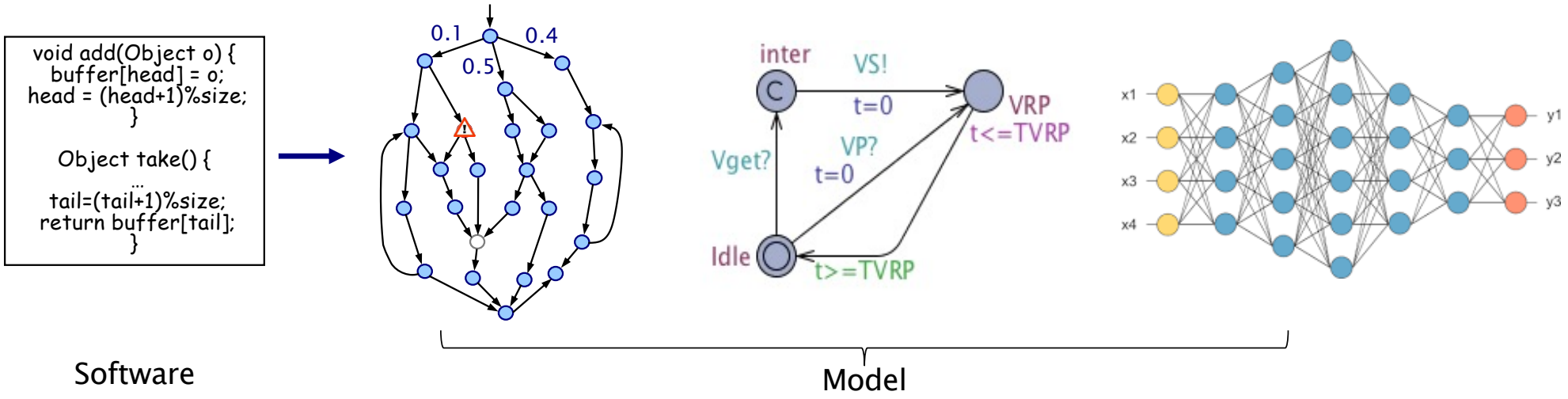
Like airbags and crumple zones, various car safety technologies are mandatory on new cars. While airbags are considered 'passive' safety tech – they only activate when you crash – Advanced Driver-Assistance Systems (ADAS) are 'active' and are intended to prevent you from having an accident in the first place.

With driver error a leading cause of road accidents in the UK, the best case scenario with ADAS features is that they prevent avoidable accidents. These include accidents where the driver unintentionally leaves their lane, is driving too fast or hasn't spotted an obstacle ahead of them.

However, Which? has found evidence that these features are being habitually turned off by drivers, with just over half of drivers who have an ADAS feature on their car reporting they turn at least one feature off at least some of the time. And when the tech is off, it isn't protecting anybody.

This highlights that there's a lot of room for improvement in the way these systems are implemented and explained.

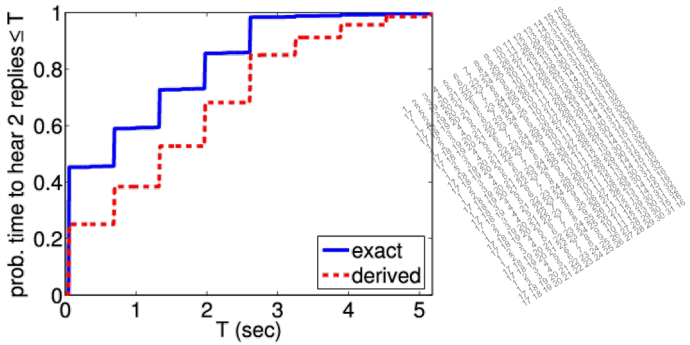
Formal verification provides provable guarantees



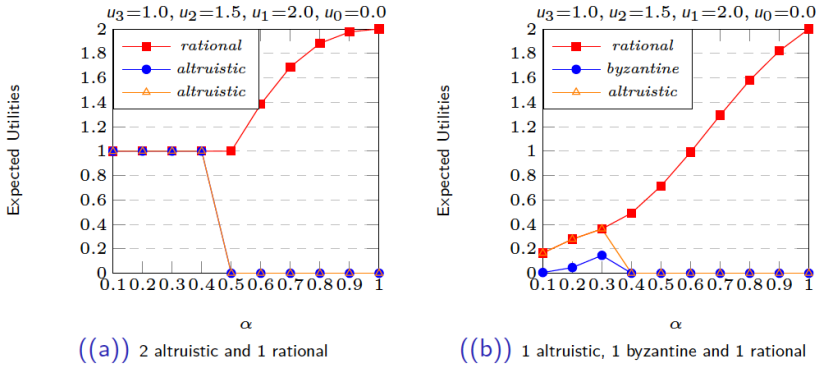
- Modelling = rigorous, mathematical abstraction
- Verification = proof that the model satisfies specification
- Synthesis = correct-by-construction model/policy from specification
- Automated = algorithmic, implemented in software

Probabilistic Model Checking in Autonomy. Kwiatkowska *et al*, Ann Rev of Control, Robotics and Aut. Sys. (2022).

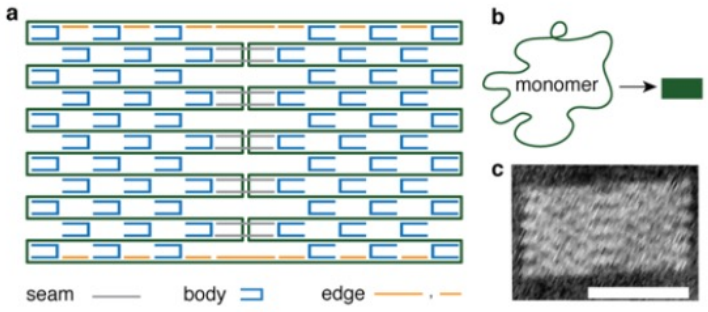
Multiple applications and use cases!



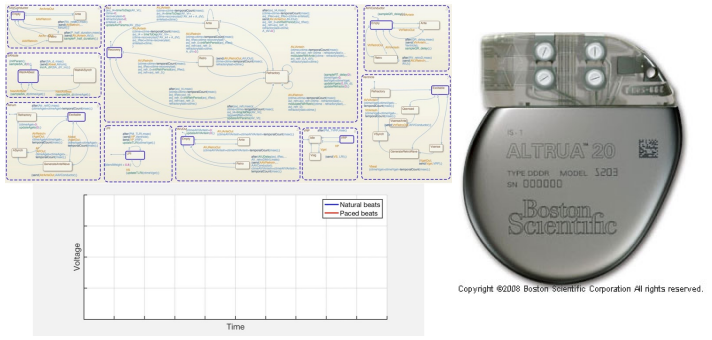
Protocol debugging



Protocol verification



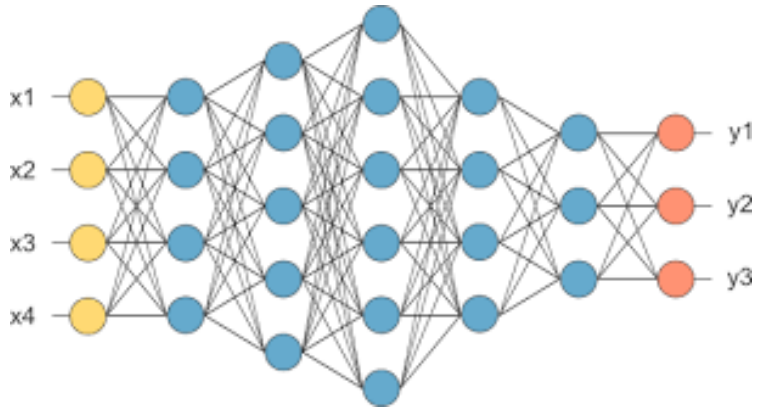
Prediction of DNA folding



Optimal controller synthesis

Formal verification for neural networks (NNs)

- Rigorous formal verification
  - can provide provable guarantees, e.g. that no adversarial examples exist
  - enables safety/security certification and correct-by-construction synthesis
  - crucial part of safety assurance



- Neural network models more challenging
  - black box, lack interpretability
  - high-dimensional function
  - interplay between architecture and training (non-linear optimization)

Image classifier is a function f: $R^n \rightarrow \{c_1,\ldots c_k\}$
Learnable weights and bias

Approximates human perception from M training examples

- Much progress since 2017: Reluplex, DeepPoly, ReluVal, CROWN, …

Safety Verification of Deep Neural Networks. CAV 2017 keynote

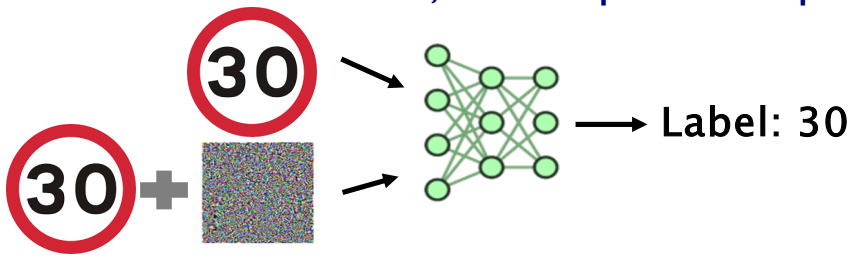
This talk: provable guarantees via formal verification

- AI as normal technology
  - a tool, used for automation and augmentation of human capabilities
  - beyond AI methods and benchmarks, towards real applications
- Brief recap of progress in (local) adversarial robustness certification
- Beyond (test-time) adversarial robustness
  - robustness to (training-time) poisoning attacks
  - robustness to strategic manipulations
  - robust decision making
  - robust collaboration and coordination with humans
- Conclusions and future directions

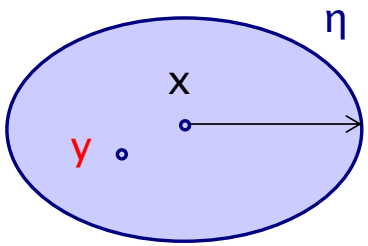
AI as Normal Technology. Narayanan & Kapoor, Knight First Amendment Institute, 2025

Recap of adversarial robustness

- Consider local adversarial robustness, for a specific input

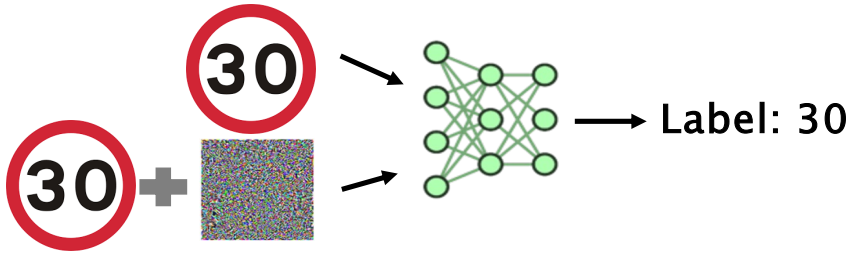


- Informally, no perturbation results in a misclassification
- More formally, assume given
  - trained neural network classifier $f : R^m \rightarrow \{c_1,\ldots c_k\}$
  - region η centred at x wrt distance function, e.g. L^2, L^∞
- Define local robustness at x wrt η by (SAT friendly)
  - $\nexists y \in \eta$ such that $f(x) \neq f(y)$
- Here, focus on computing provable guarantees on correctness, rather than constructing defences

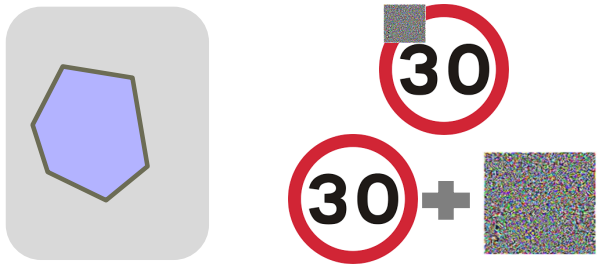


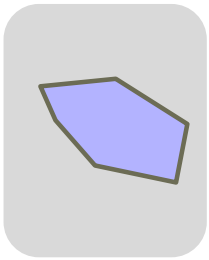
Neural network verification

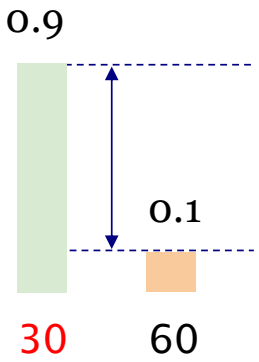
- Given a neural network $f: R^n \rightarrow R^m$, the NN verification problem is defined as $(\varphi_{pre}, \varphi_{post})$ requiring that
  - $\forall x \in R^n . x \vdash \varphi_{pre} \longrightarrow f(x) \vdash \varphi_{post}$



- Simplification to polyhedral input and output sets







$f_1(x) - f_2(x) \geq 0$

(for logits, m=2)

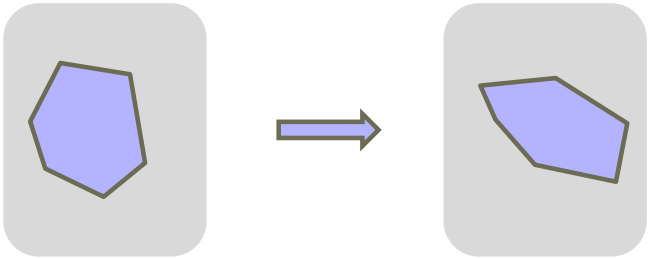
confidence gap

- Typically, exact verification intractable, focus on computing lower/upper bounds

Neural network verification: forward analysis

- Given a neural network $f: R^n \rightarrow R^m$, the NN verification problem is defined as $(\varphi_{pre}, \varphi_{post})$ requiring that
  - $\forall x \in R^n . x \vdash \varphi_{pre} \longrightarrow f(x) \vdash \varphi_{post}$

- Typical approach: forward analysis
  - start from $X = \{x \in R^n \mid x \vdash \varphi_{pre}\}$
  - bound the worst case on each layer
  - propagate bounds through layers
  - check whether the predicted labels are preserved

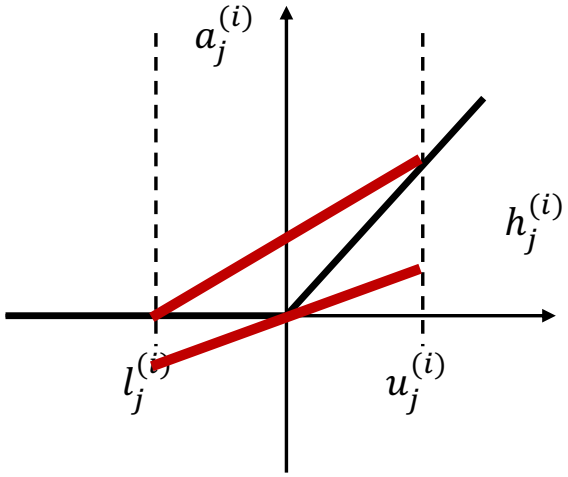


- Computes over-approximation of output set
- Note may result in loose bounds…

Progress in neural network verification

- Compute provable guarantees by lower/upper bounding the reachable values

- Methods include exact/approximate
  - search-based/Lipschitz, e.g. DLV
  - constraint solving/SMT/MIP, e.g., Reluplex
  - convex relaxation, e.g., interval/linear bound propagation, as in CROWN
  - abstract interpretation, e.g., DeepPoly
  - global optimisation, under assumption of Lipschitz continuity, e.g., DeepGO

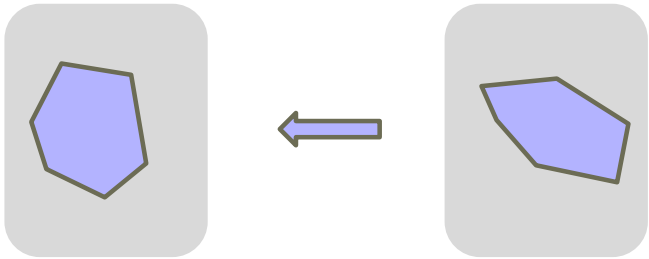


Linear bounding of ReLU activations
ReLU(x) := max(0, x)

- Hard problems, typically NP-completeness
- Convex relaxation best performers, see VNN-Comp
- Scaling, loose bounds and complex architectures an issue…

Neural network verification: backward analysis

- Given the NN verification problem $(\varphi_{pre}, \varphi_{post})$ for a neural network $f: R^n \rightarrow R^m$, requiring that
  - $\forall x \in R^n . x \vdash \varphi_{pre} \longrightarrow f(x) \vdash \varphi_{post}$
- Focus instead on backward analysis
- Characterize the inputs for output constraints $Y = \{y \in R^m \mid y \vdash \varphi_{post}\}$

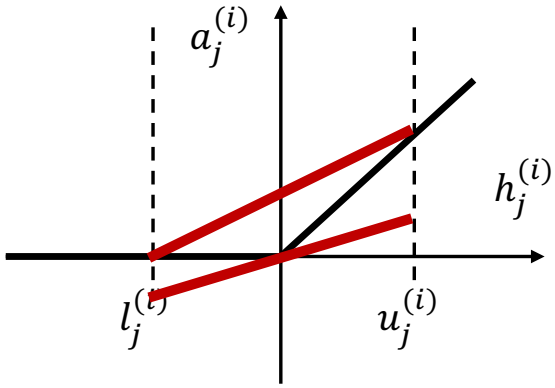


- Advantages
  - more precise correctness guarantees, particularly under-approximation
- but
  - exact preimage computation is intractable at scale, $O(2^n)$ for n unstable ReLU neurons

Provably bounding neural network preimages. Koha *et al*, In Proc. NeurIPS 2023.
Provable Preimage Under-Approximation for Neural Networks. Zhang *et al*, In Proc. TACAS 2024.

Preimage approximation

- Work backwards to generate preimage approximation via convex relaxation in terms of disjoint union of polytopes
- Given output specification $\mathrm{y} = f(x) \geq 0$ (any polyhedral property)
- Compute symbolic lower/upper bounding functions for activations from output layer to input:
  - $\underline{A}x + \underline{b} \leq f(x) \leq \overline{A}x + \overline{b}$
- Preimage under-approximation as a polytope:
  - $\{\mathrm{x} \mid \underline{A}x + \underline{b} \geq 0\} \longrightarrow \{\mathrm{x} \mid f(x) \geq 0\}$
- Also preimage over-approximation
- Method relies on
  - backward propagation
  - preimage refinement through input/ReLU splitting planes
  - heuristics and optimisations, to deal with exponential growth in constraints

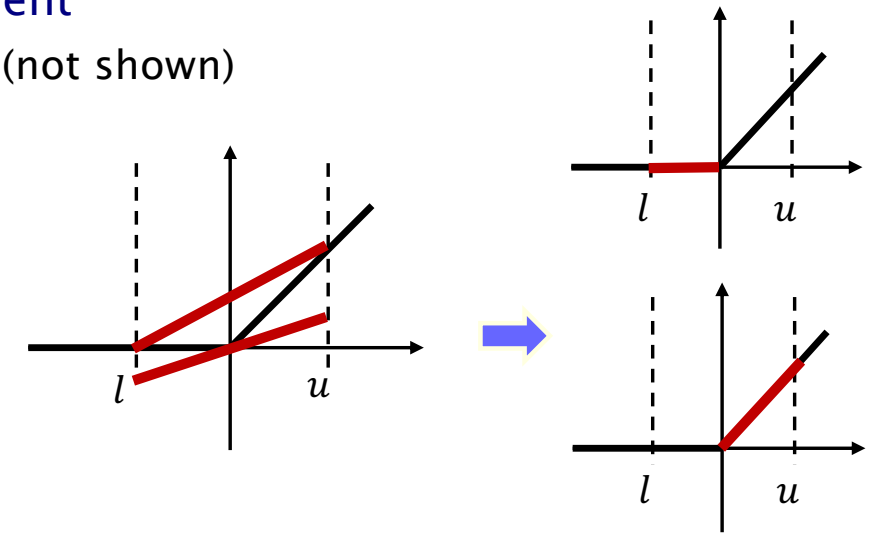


Linear bounding of ReLU activations

PREMAP: A Unifying PREiMage APproximation Framework for Neural Networks, to appear in JMRL 2025, arXiv:2408.09262

Preimage under/over-approximation

- Anytime algorithm
- Preimage refinement to handle approximation loss
  - parallel processing of split regions
  - tightening of approximation by optimizing relaxation parameters
  - (novel differential objective)
- Two types of (sound) preimage refinement
  - via input-feature-aligned cutting plane (not shown)
  - via ReLU-aligned cutting plane
  - (unstable ReLU neuron into two stable cases: approximation becomes exact)
- Volume-estimated prioritization of splitting subregions
- Exact volume for final verification



PREMAP: A Unifying PREiMage APproximation Framework for Neural Networks, to appear in JMLR, arXiv:2408.09262

Experimental results: preimage under-approximation

- Method scales to high-dimensional tasks
  - first method to scale to l_∞ attack (noise in all image pixels) and patch attack
  - recently improved and extended to CNNs

  - evaluated on MNIST, GTSRB and SVHN with varied size and position of the patch, indicating areas of vulnerability
  - provides quantitative coverage results for larger perturbation bounds

L_∞ attack	#Poly	Cov(%)	Time(s)	Patch attack	#Poly	Cov(%)	Time(s)
0.05	2	100.0	3.107	3×3(center)	1	100.0	2.611
0.07	247	75.2	121.661	4×4(center)	678	38.2	455.988
0.08	522	75.1	305.867	6×6(corner)	2	100.0	9.065
0.09	733	16.5	507.116	7×7(corner)	7	84.2	10.128

Efficient Preimage Approximation for Neural Network Certification. Bjorklund *et al*, arxiv.org/abs/2505.22798

Quantitative neural network verification

- Preimage under-approximation enables quantitative verification
  - i.e. estimating proportion of inputs that satisfy φ_{post}
  - sound and complete
- Useful in cases when verification fails
- Complementary to robustness verifiers, benchmarked against winner of VNN-Comp 2023

Task	α, β-CROWN		Our		
	Result	Time(s)	Cov(%)	#Poly	Time(s)
Cartpole ($\dot{\theta} \in [-1.642, -1.546]$)	yes	3.349	100.0	1	1.137
Cartpole ($\dot{\theta} \in [-1.642, 0]$)	no	6.927	94.9	2	3.632
MNIST (L_∞ 0.026)	yes	3.415	100.0	1	2.649
MNIST (L_∞ 0.04)	unknown	267.139	100.0	2	3.019

Provable Preimage Under-Approximation for Neural Networks. Zhang *et al*, In Proc. TACAS 2024.

Beyond adversarial robustness: poisoning attacks

- Test-time adversarial robustness not sufficient
- Training of neural networks exposed to
  - poisoning attacks by injecting malicious training data
  - data prone to corruption, such as missing data or biases
  - critical for sensitive domains, e.g. healthcare, finance, etc

- Defences against poisoning lack formal guarantees
  - e.g. robust training, randomized smoothing

- Provable guarantees for certifiable training?
  - adaptation of test-time (interval/polyhedral) certification important first step
  - approximates training dynamics layer by layer
  - suffers from over-approximation and divergence

MIBP-Cert: Certified Training against Data Perturbations with Mixed-Integer Bilinear Programs, Lorenz *et al*, NeurIPS 2025

Certified training against data perturbations

- Aim to bound the error introduced by that perturbations of training data
- Intractable in general
- Formulate a mixed-integer bilinear programming problem
  - compute exact bounds for a single training step
  - bounding parameters at each step for tractability
  - ensure soundness of bounds
- Compute certified accuracy on UCI datasets, for complex perturbations

Precondition	Certification Rate	Certified Accuracy
Assuming accurate health data	(100.0%)	56.3%
Modeling missing mental health values	98.6%	56.3%
Modeling missing values across all features	95.8%	53.5%
Modeling mental health over-reporting	91.5%	50.7%

  - can guarantee correct prediction for all test points even for large perturbations

MIBP-Cert: Certified Training against Data Perturbations with Mixed-Integer Bilinear Programs, Lorenz *et al*, NeurIPS 2025

Beyond adversaries: strategyproof robustness

- So far, consider only adversarial robustness to individual perturbations,
  - but AI agents can behave strategically
- Can we instead devise strategyproof policy learning? (correctness by design)

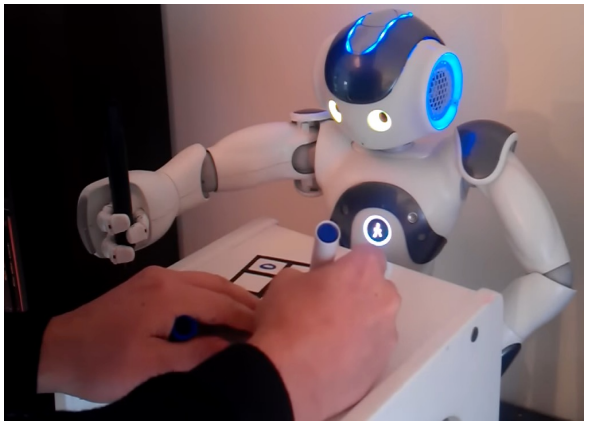
- Consider RLHF (reinforcement learning from human feedback)
  - multiple agents, diverse preferences, leading to potential bias in learnt policy decisions
  - but agents can also strategically manipulate the decisions in their favour by misreporting their preferences
  - existing RLHF methods not strategyproof…

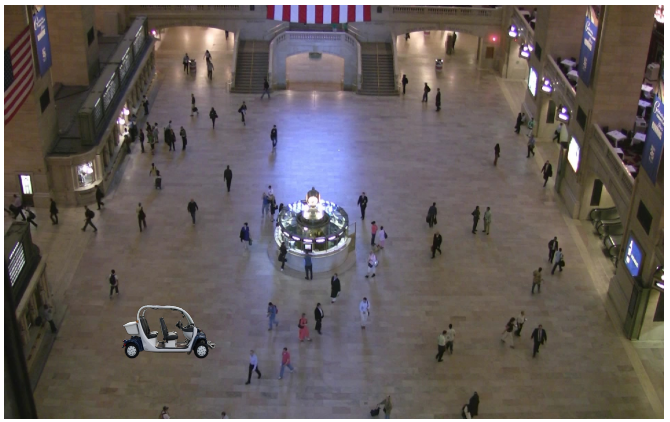
- Aim to devise strategyproof RLHF through mechanism design
  - how? incentivise truthful reporting
  - can provide an algorithm that is approximately strategyproof and converges to the optimal policy as the number of individuals and samples increases

Strategyproof Reinforcement Learning from Human Feedback. Kleine Buening *et al*, NeurIPS 2025, arXiv:2503.09561v1

Human–AI collaborations

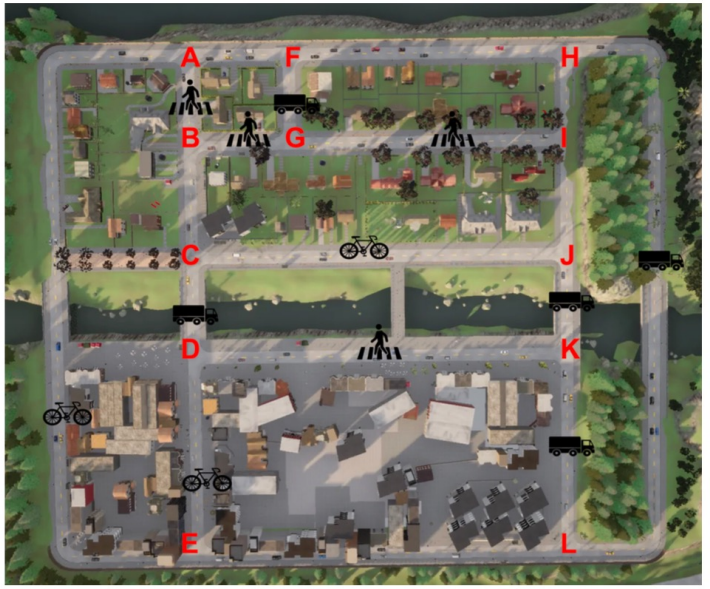
- Human–robot interactive systems: important yet challenging
  - different interaction patterns (collaborative, adversarial)
  - human cognitive states (trust, intention) difficult to predict
  - uncertainty, dynamic environments
  - complex specifications (beyond reachability, reward maximisation)
  - can we obtain rigorous guarantees?

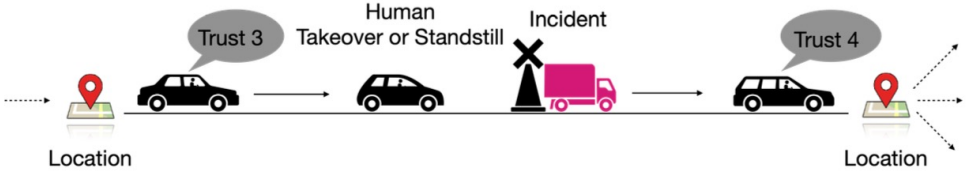




Trust-aware human-robot collaboration

- Model trust-based human-robot collaborations as POMDPs (partially observable Markov decision processes)
  - human trust is the hidden state
  - robot performance affects trust
  - trust informs human decision
- Temporal logic specifications
  - "visit locations G, J and L (in this order) from A"
  - "the trust level must not fall below a certain threshold"
- Optimal policy synthesis
- Correctness guarantees, subject to measurement precision

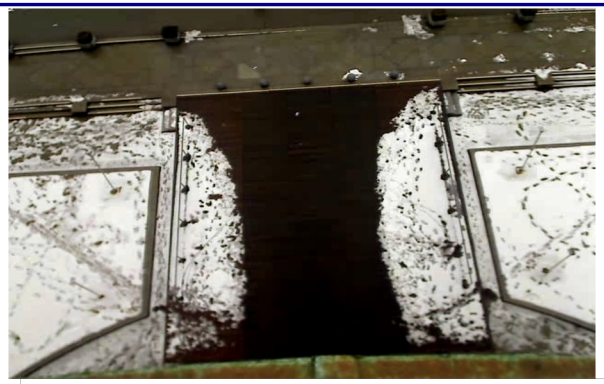


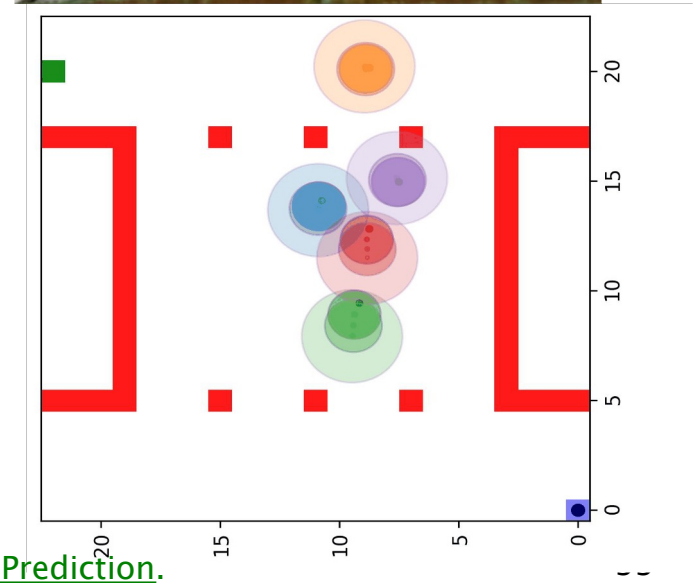


Trust-Aware Motion Planning for Human-Robot Collaboration under Distribution Temporal Logic Specifications.
Yu *et al*, In Proc. ICRA 2024.

Safe online planning in a crowd

- Robotic agent modelled as a POMDP
  - partial observability (e.g., perception inaccuracy)
- Environment is populated with pedestrians
- Pedestrian trajectory prediction
  - data-driven trajectory predictor
  - uncertainty quantification via adaptive conformal prediction (ACP)
- Safe online planning via shielding
  - on-the-fly safety shield construction
  - adapt shielding
- Safety guarantee, given any probability threshold

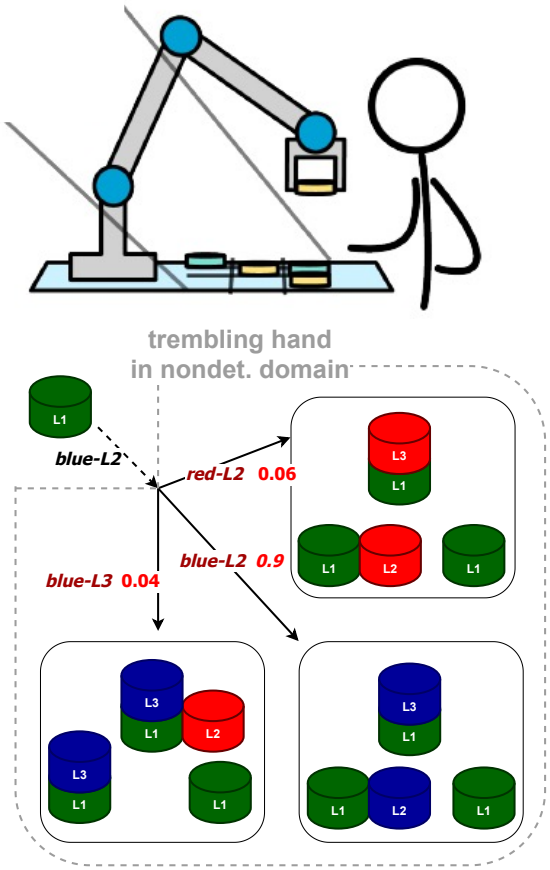




Safe POMDP Online Planning Among Dynamic Agents via Adaptive Comformal Prediction.
Sheng *et al*, RAL 2024.

Robot planning with human interventions

- Robot has a trembling hand
  - erroneously selects unintended actions with a small probability
- Human (collaborate/adversary) unpredictable
- Model collaboration as set-valued MDPs
  - set-valued probabilistic transitions capture both action instruction errors and human nondeterminism
- Temporal logic specifications
  - "eventually goal configuration and never undesired configuration"
- Sound verification and synthesis algorithms
  - via simplified Bellman equation
- Can generate plans



The Trembling Hand Problem For LTLf Planning. Yu *et al*, In Proc. IJCAI 2024.

Provable guarantees for RL decision policies

- Learning of optimal policies from temporal logic specifications
  - Sample Efficient Model-free Reinforcement Learning from LTL Specifications with Optimality Guarantees. Shao *et al*, Proc. IJCAI 2023
  - Converts LTL to limit-deterministic Buchi automata

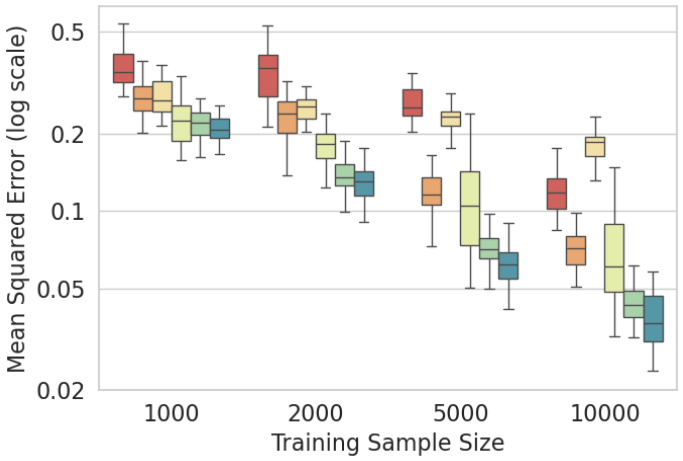
- Learning temporal logic specifications to debug/explain RL policies
  - Learning Probabilistic Temporal Logic Specifications for Stochastic Systems. Roy *et al*, Proc. IJCAI 2025
  - Learns concise probabilistic LTL from positive and negative examples

- Extension to imitation learning incorporating causal inference
  - A Unifying Framework for Causal Imitation Learning with Hidden Confounders. Shao *et al*, In ICLR 2025 Workshop on Spurious Correlation and Shortcut Learning: Foundations and Solutions
  - Learns causal effects using instrumental variables

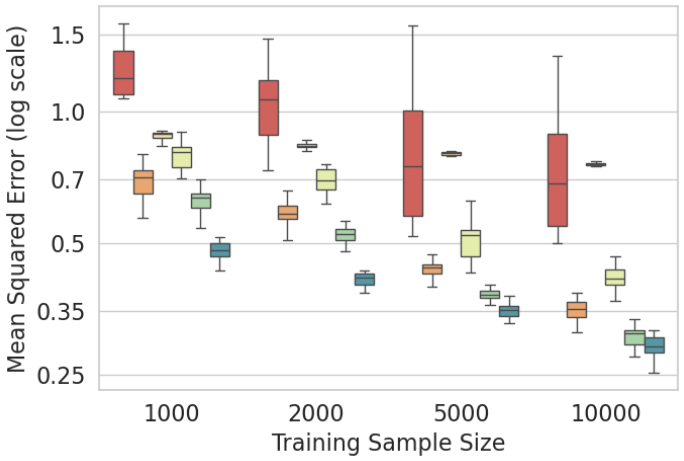
Experiments

- Evaluate DML-IV (ours) on benchmarks and semi-synthetic real-world datasets (infant development and cardiovascular mortality rate datasets)
- Compare the error of the learned causal effect of actions (lower is better):

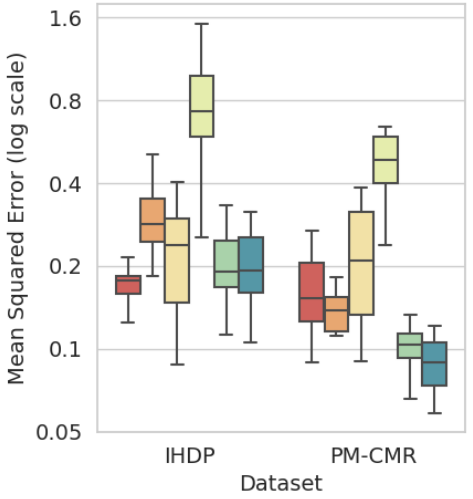
Low-dimensional



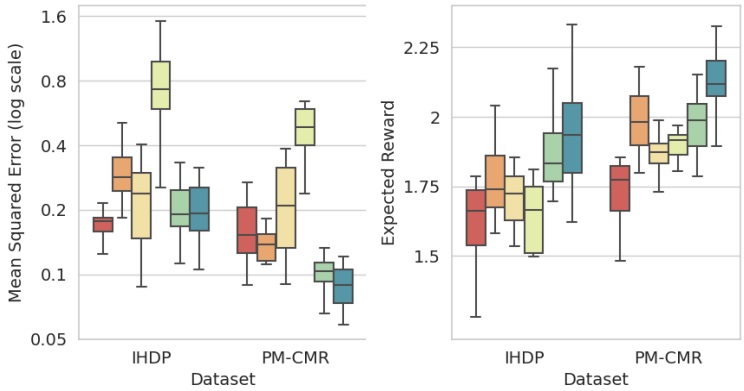
High-dimensional



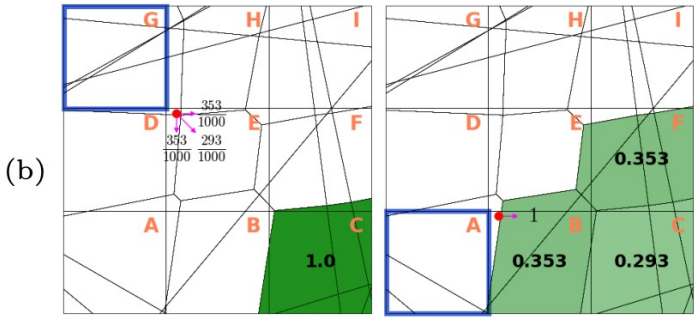
Real-World



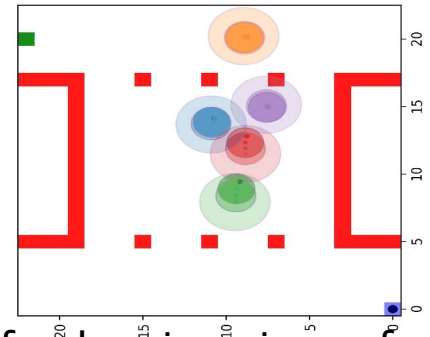
Multiple applications and NN verification use cases!



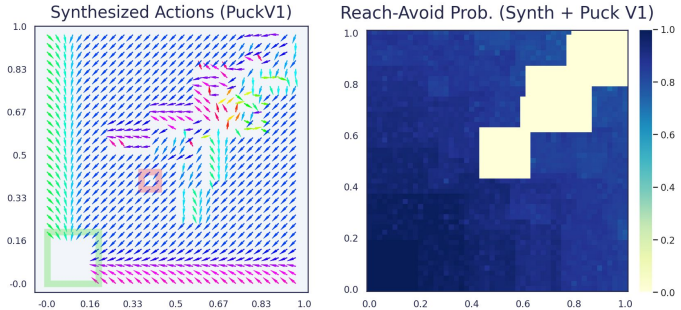
Sample-efficient policy learning



Protocol verification



Safe planning via conformal prediction



Optimal controller synthesis

http://fun2model.org/

Human-in-the-loop oversight

- Algorithmic decisions prone to bias
- Human oversight often required to approve or reject AI decisions
- Can human-in-the-loop (HITL) oversight reliably correct AI errors?

- Study modelling real-world welfare allocation scenarios
  - range of algorithmic biases in recommendations, different interventions

- Observe that HITL failed to reduce algorithmic errors
  - often amplified the errors, partisan differences
  - interventions lead to at best modest gains, outcomes depend on ideaology
  - e.g. financial incentives for correct decisions or inform decisions would retrain

- More research needed…

Human-in-the-loop Oversight of AI is Compromised by Political Preferences. Dores Cruz *et al*, In PsyArXiv. 2025. https://osf.io/preprints/psyarxiv/9ecms_v2

Concluding remarks

- Range of techniques developed in the AI/ML and formal methods communities
  - provable guarantees needed for high-stakes decisions
  - robustness to adversarial, poisoning and strategic manipulations
  - robustness and reliability of AI decisions desirable but challenging
  - human involvement in decisions proposed
  - but human oversight can exacerbate algorithmic errors
- Despite progress, major challenges remain
  - scalability to complex architectures and properties
  - foundational understanding needed
  - robust learning for correct-by-construction models and policies
  - need support for interactions with human decision makers
- Need integrated processes for validation and safety assurance

Acknowledgements

- My group and collaborators in this work

- Project funding
  - ERC Advanced Grant fun2model
  - EPSRC project FAIR: Framework for responsible adoption of artificial intelligence in the financial services industry, https://www.turing.ac.uk/research/research-projects/project-fair-framework-responsible-adoption-artificial-intelligence
  - ELSA European Lighthouse on Secure and Safe AI, https://www.elsa-ai.eu/
  - UKRI AI Hub on Mathematical foundations of intelligence: an 'Erlangen Programme' for AI

- See also
  - PRISM www.prismmodelchecker.org