SAFE

Stanford Center for AI Safety

Scalable Verification and Validation (V&V) for AI Systems

Mansur M. Arief

Industrial & Systems Engineering, KFUPM

ai-vnv.kfupm.io

About Mansur

- **Assistant Professor, Industrial & Systems Engineering, KFUPM**
- **Executive Director, Stanford Center for AI Safety,** 2025

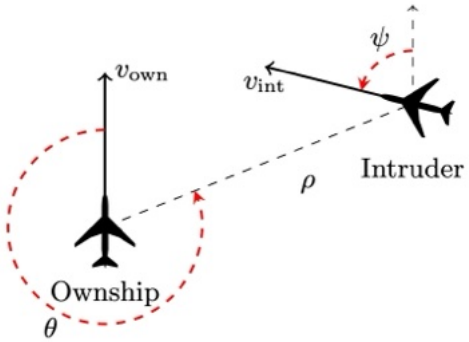
SAFE
Stanford Center for AI Safety

Mykel Kochenderfer
Aeronautics and Astronautics Department

Clark Barrett
Computer Science Department

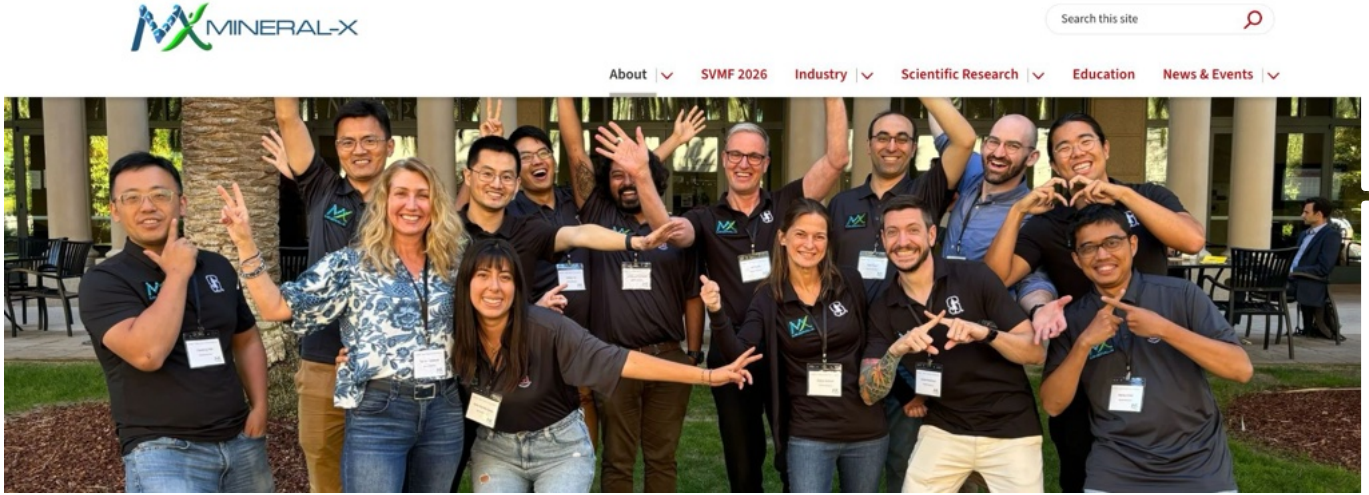
Reluplex (ReLU + Simplex)

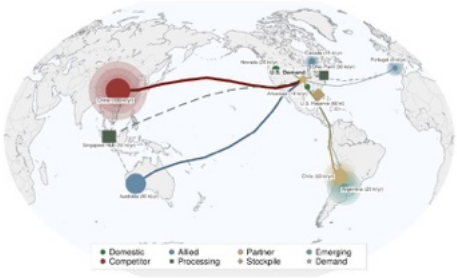
Guy Katz, Clark Barrett, David Dill, Kyle Julian, Mykel Kochenderfer



About Mansur

- **Assistant Professor, Industrial & Systems Engineering, KFUPM**
- **Executive Director, Stanford Center for AI Safety,** 2025
- **Research Engineer, Stanford Intelligent Systems Lab and Mineral-X,** 2024-2025
- **Postdoc,** Stanford, 2023-2024





About Mansur

- **Assistant Professor, Industrial & Systems Engineering, KFUPM**
- **Executive Director, Stanford Center for AI Safety,** 2025
- **Research Engineer, Stanford Intelligent Systems Lab and Mineral-X,** 2024-2025
- **Postdoc,** Stanford, 2023-2024

- **PhD** in **Mechanical Engineering**, Carnegie Mellon University, 2023

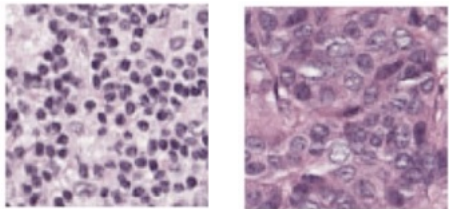
  Safe AI Lab, with dissertation: Certifiable Evaluation for Safe Intelligent Autonomy
- **MSE, Industrial & Operations Engineering**,

  University of Michigan, Ann Arbor
- **BE, Industrial and Systems Engineering,**

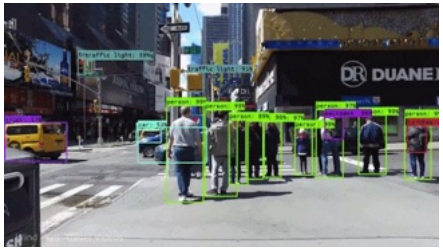
  Sepuluh Nopember Institute of Technology, Indonesia

AIs are everywhere

Healthy Tumor

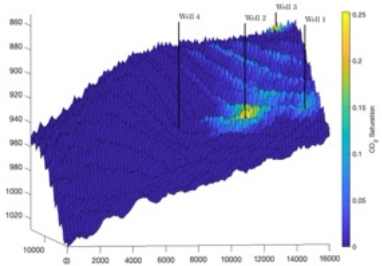


Healthcare



Autonomous Vehicles

Aviation

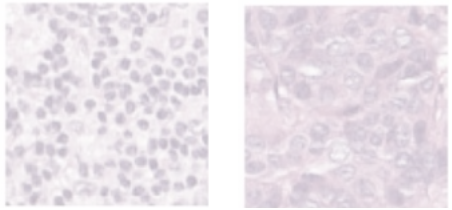


Critical Infrastructure

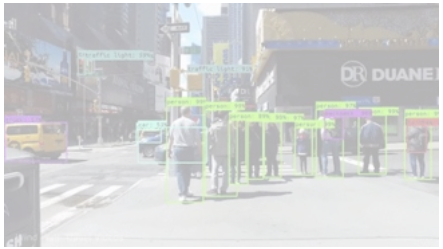
Assistive Robotics

Automated Manufacturing

AIs are everywhere



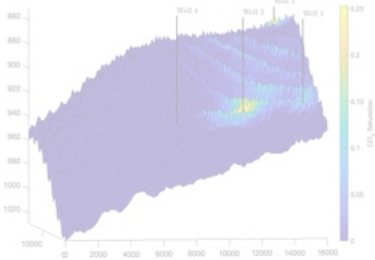
Healthcare



Autonomous Vehicles

Aviation

(including safety-critical domains)



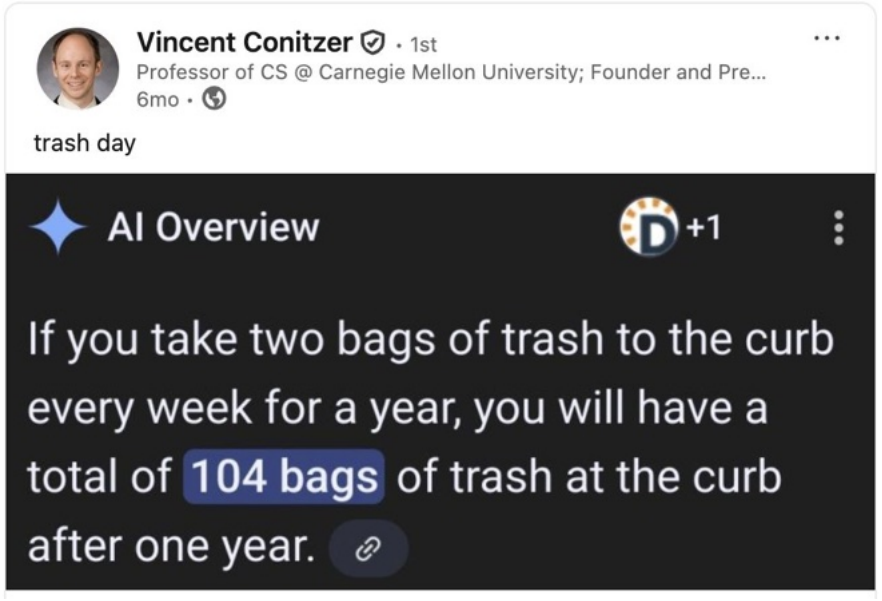
Critical Infrastructure

Assistive Robotics

Automated Manufacturing

Sometimes, AI doesn’t perform as expected

Prompt: generate an image of extra firm handshake



Vincent’s LinkedIn post series on AI Overview cringes

Why does the AI goalkeeper fail in the final scenario?

Current AI models limitation

Naturally brittle outside training domains

Susceptible to adversarial manipulation

Lacking certifications and usage guidelines

Also in Stanford AI Index 2025

Reported safety and responsible AI benchmarks for popular foundation models

Source: AI Index, 2025 | Table: 2025 AI Index report

Responsible AI benchmark	o1	GPT-4.5	DeepSeek-R1	Gemini 2.5	Grok-2	Claude 3.7 Sonnet	Llama 3.3
BBQ	✓	✓				✓	
HarmBench							
Cybench						✓	
SimpleQA			✓	✓			
Toxic WildChat	✓	✓				✓	
StrongREJECT	✓	✓					
WMDP benchmark	✓	✓					
MakeMePay	✓	✓					
MakeMeSay	✓	✓					

https://hai.stanford.edu/ai-index/2025-ai-index-report

Are things better with AI?

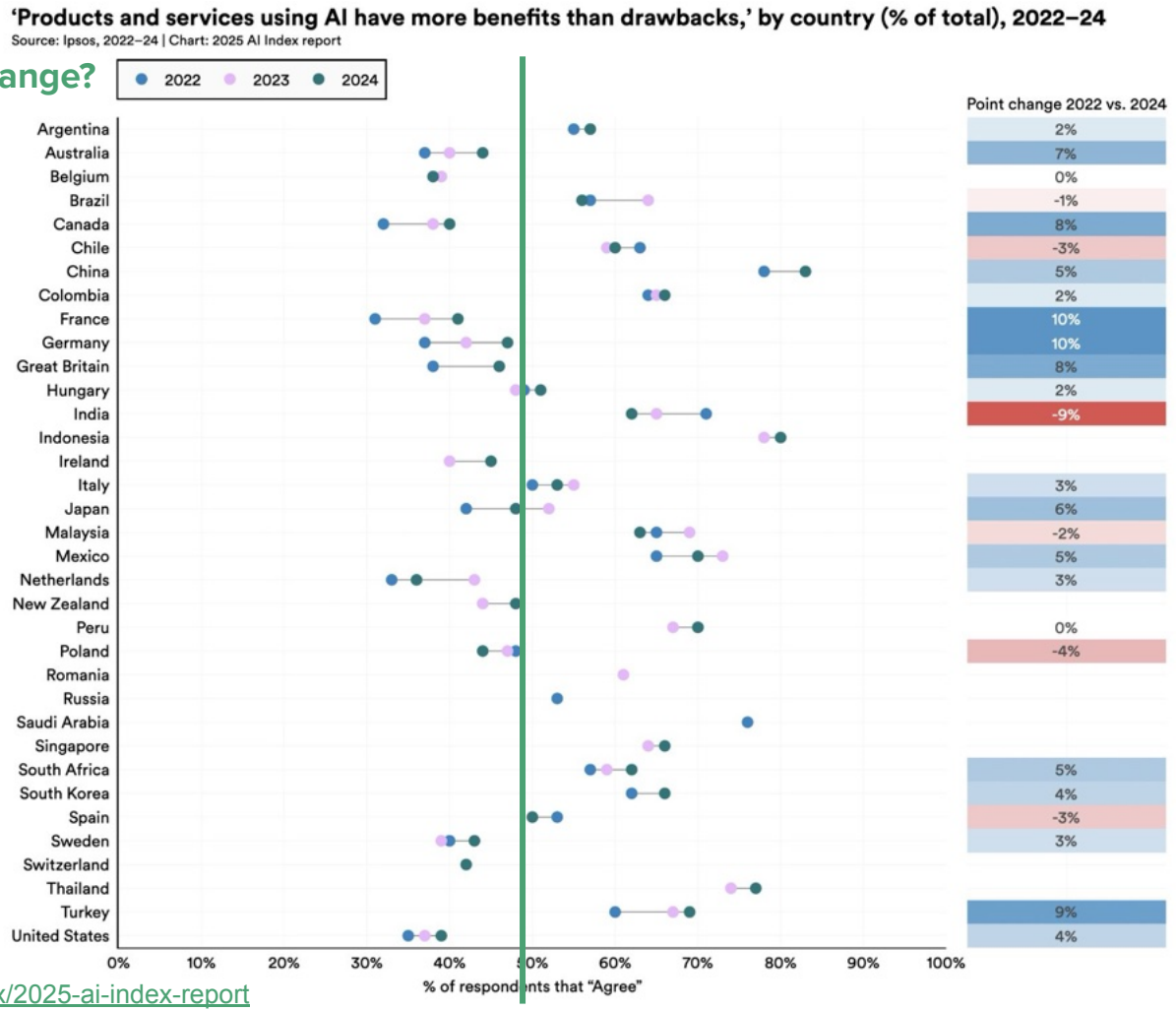
https://hai.stanford.edu/ai-index/2025-ai-index-report

Perception change?

Are things better with AI?

About 50% people in 50% country agrees

... and increases



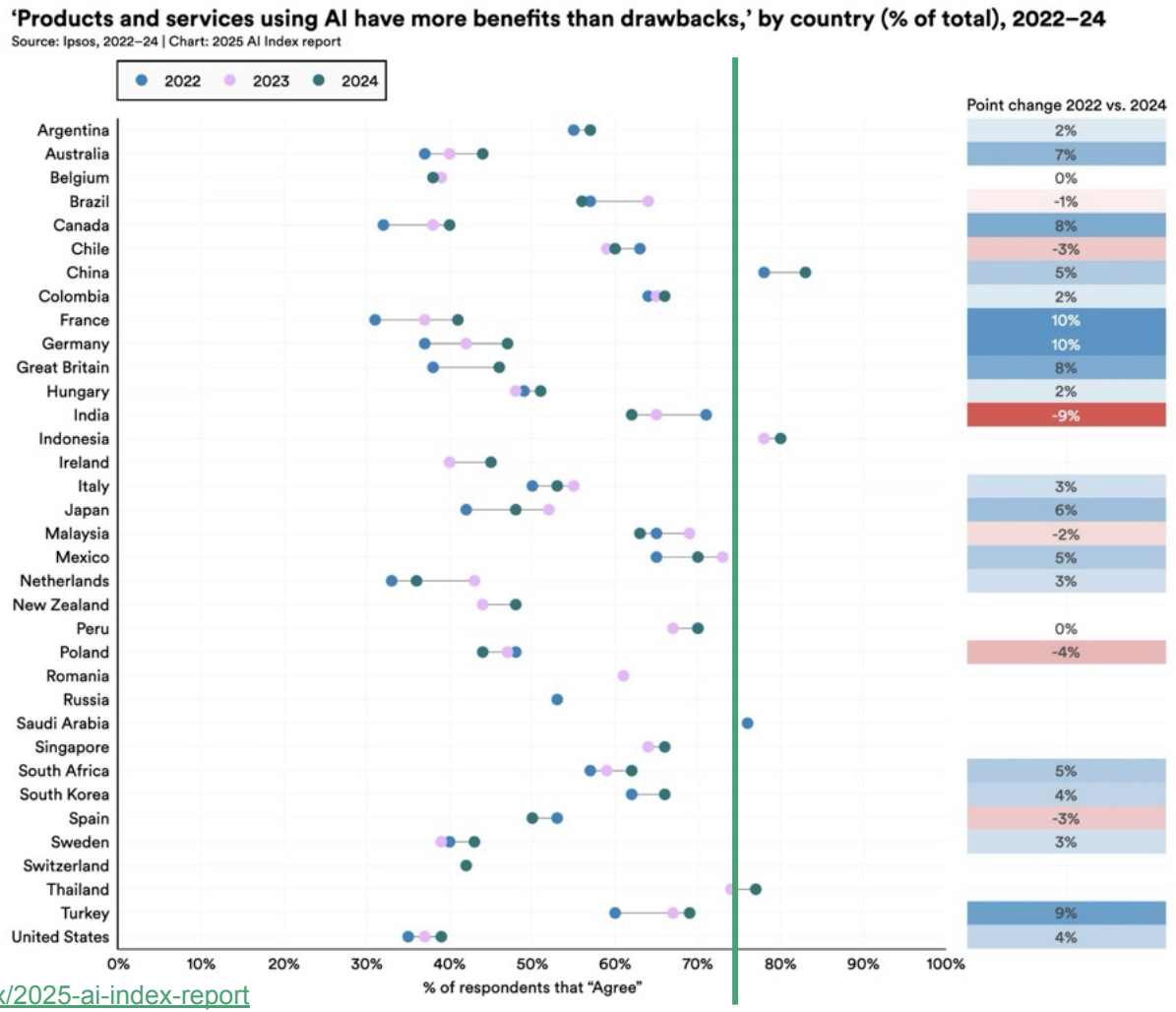
https://hai.stanford.edu/ai-index/2025-ai-index-report

Are things better with AI?

Moving line to ¾:

Top 4:

1. **China**
2. **Indonesia**
3. **Thailand**
4. **Saudi Arabia**



https://hai.stanford.edu/ai-index/2025-ai-index-report

Have a great advantage, but …
we need to develop and deploy
reliable AI

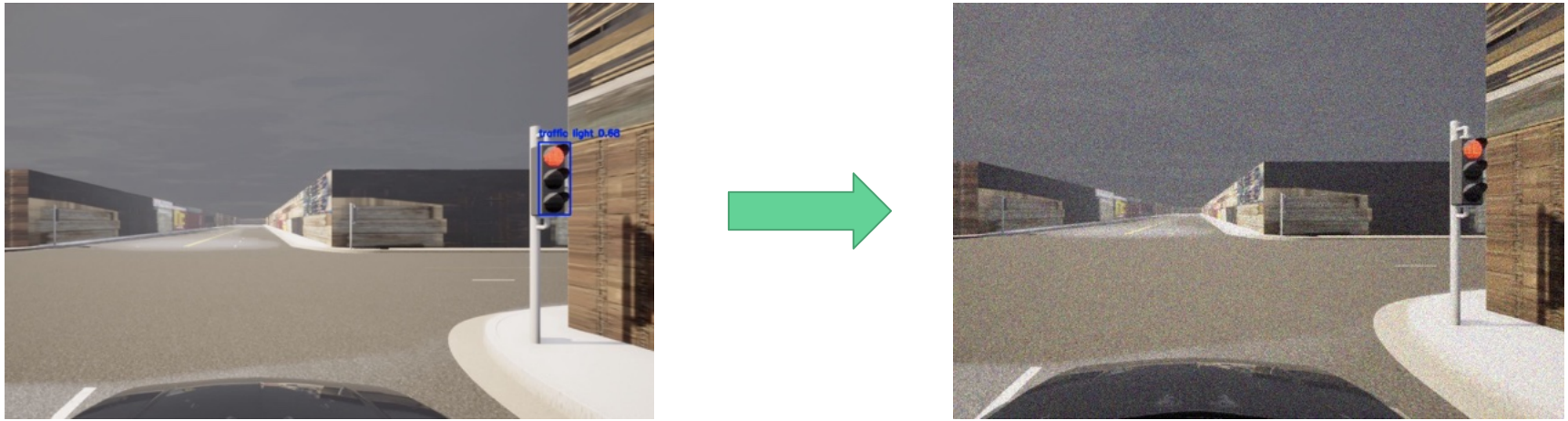
AI failures are not so funny in safety-critical context

Self-driving car blocking road 'delayed patient care', San Francisco officials say

Cruise, the robotaxi firm, denies the city's claims its vehicle blocked ambulance which resulted in injured person's death

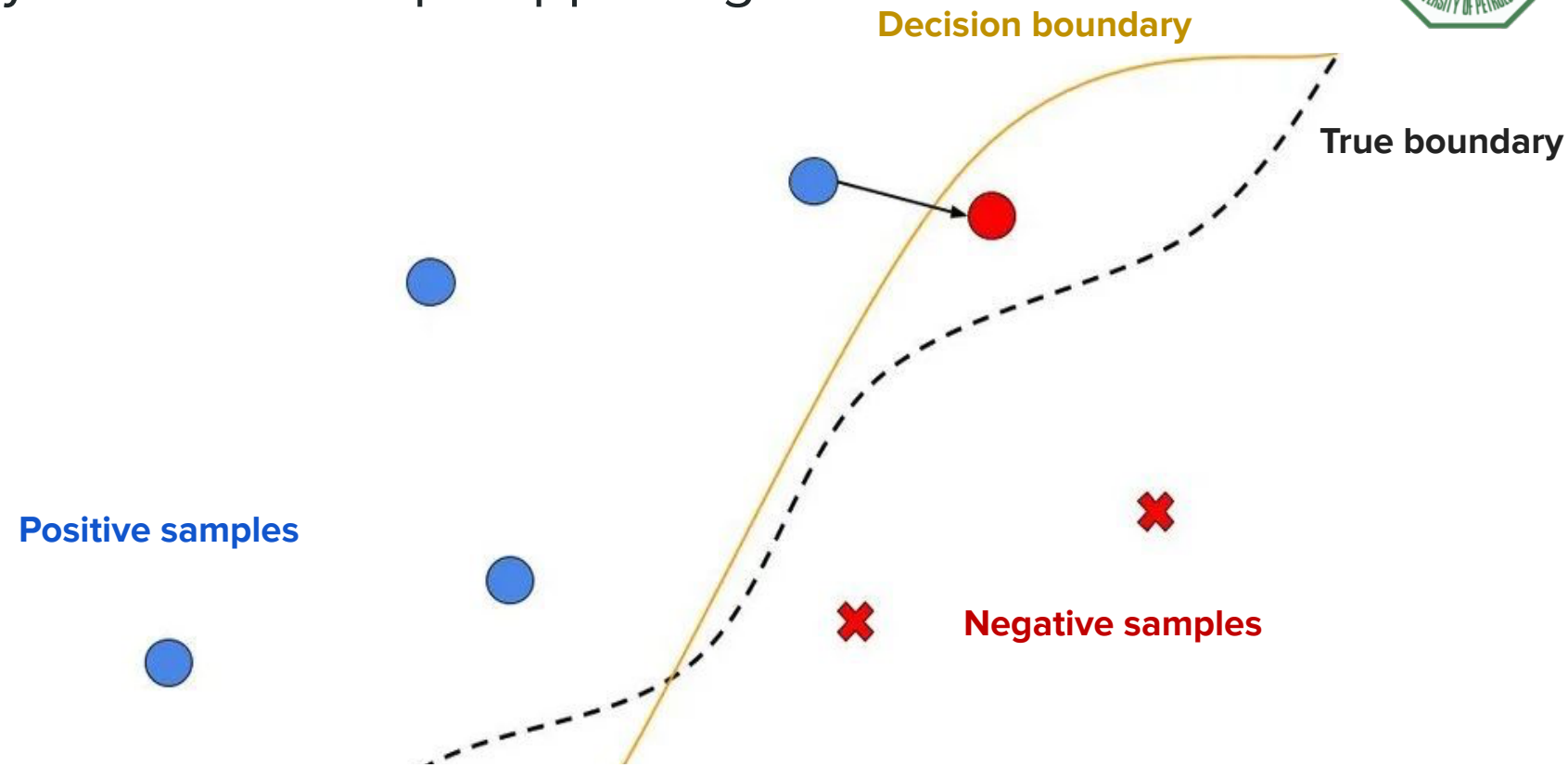
A driverless taxi of Cruise is seen on the road of San Francisco. Photograph: Michael Ho Wai Lee/SOPA Images/Shutterstock

In our experiments, we can replicate failures

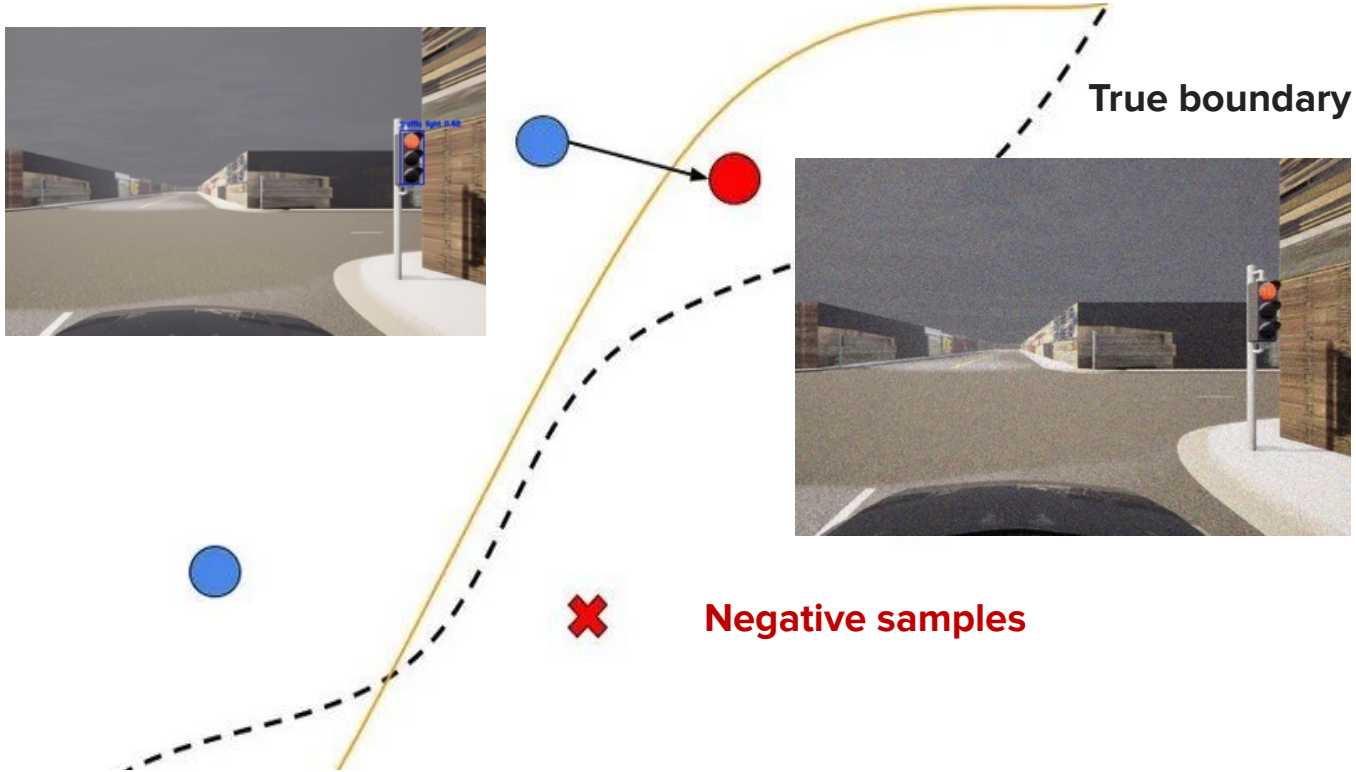


Crafted noise can make object detector misses a traffic light in image

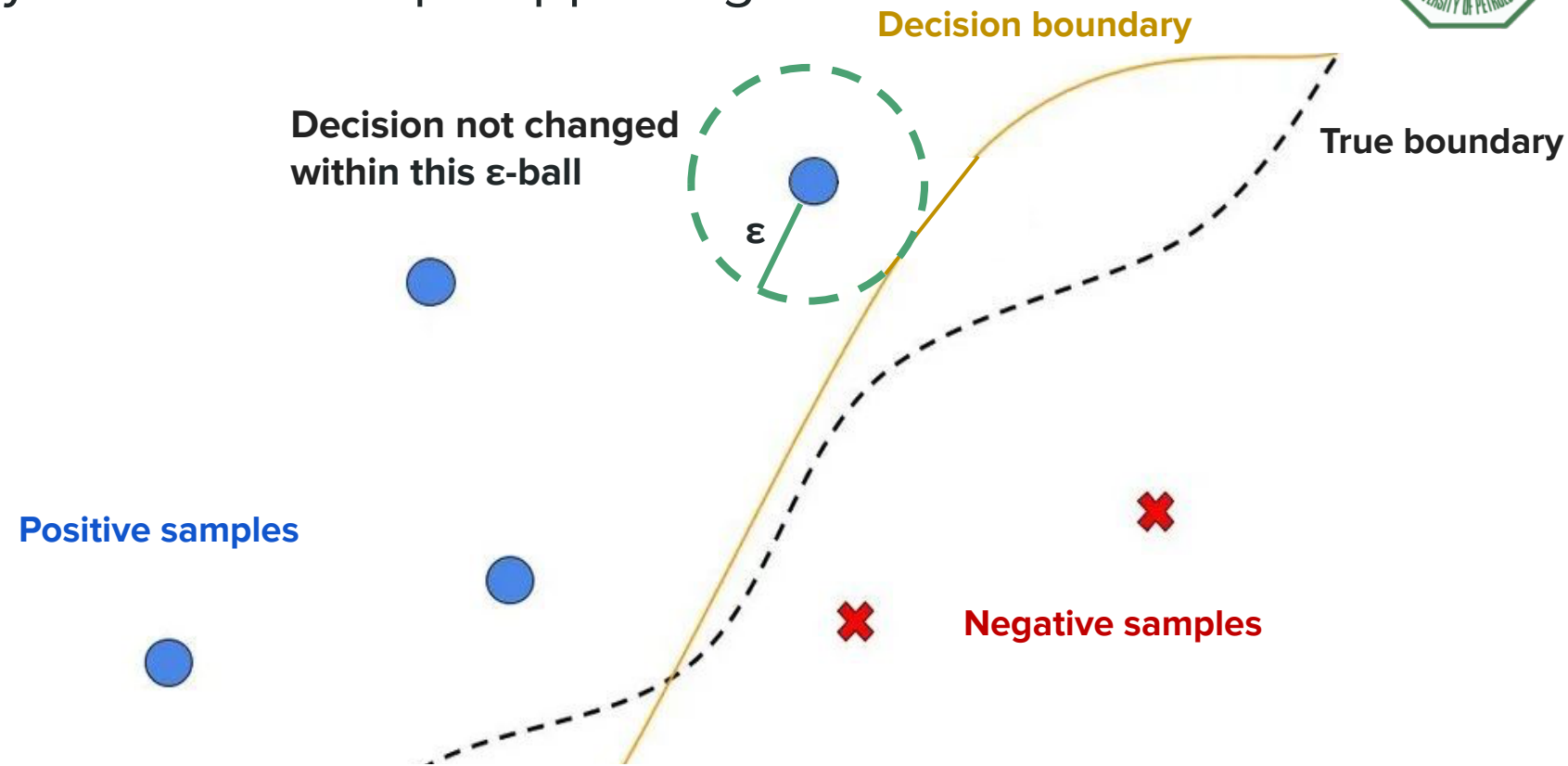
Why does this keep happening?



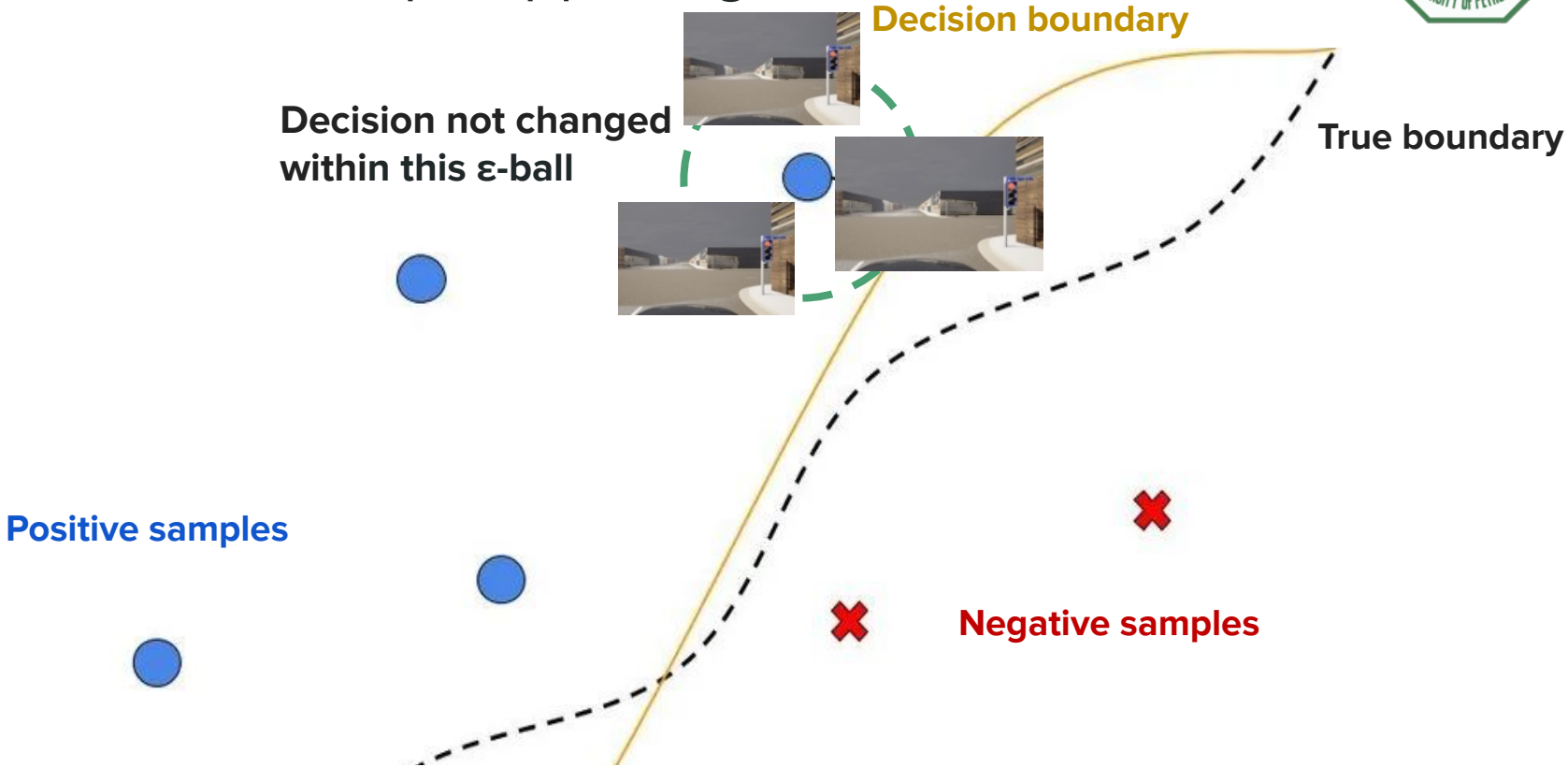
Why does this keep happening?



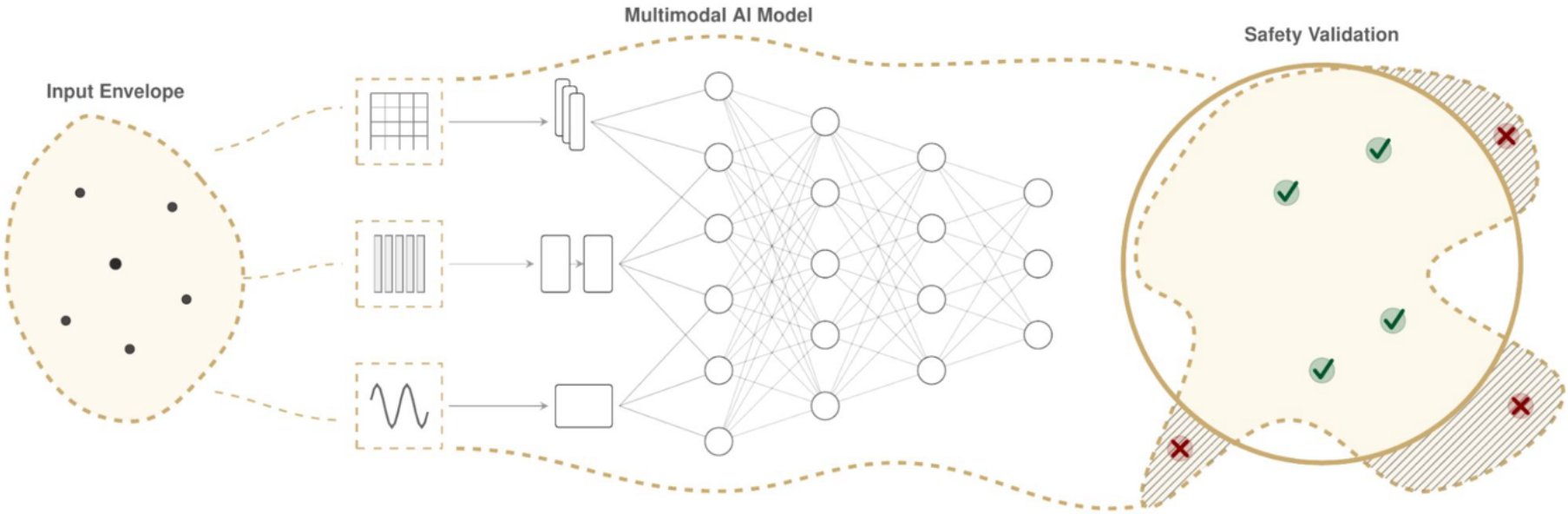
Why does this keep happening?



Why does this keep happening?

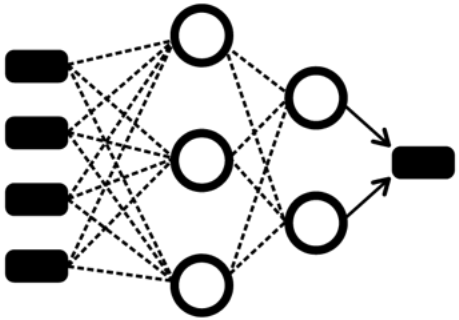


AI Validation & Verification (V&V) framework

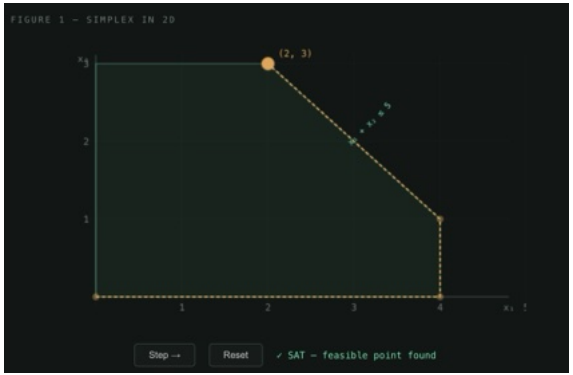


Given a bounded input, can we find all specification violations, if any?

An Idea: can we maximize constraint violation for neural net output?



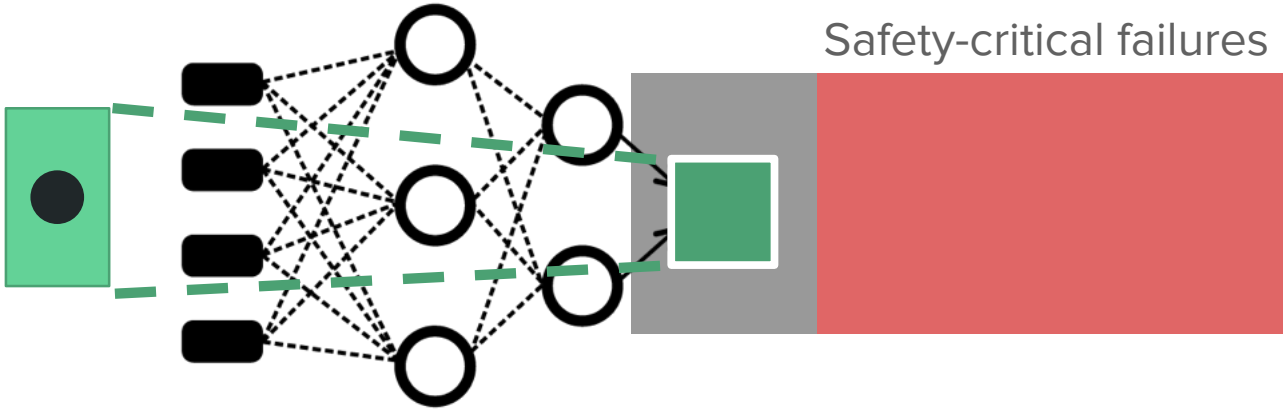
Deep neural net



Formulate and solve using Simplex

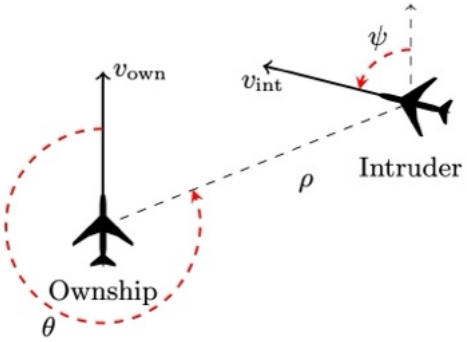
Will work only if reformulated as MILP

Reluplex: ReLU MILP reformulation + Simplex



If the bounds are within safe region, then the model is provably safe.

Applications



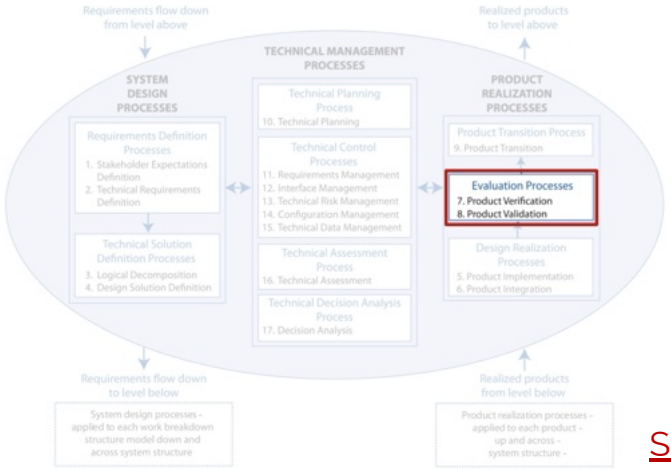
Aircraft collision avoidance (ReLUplex)

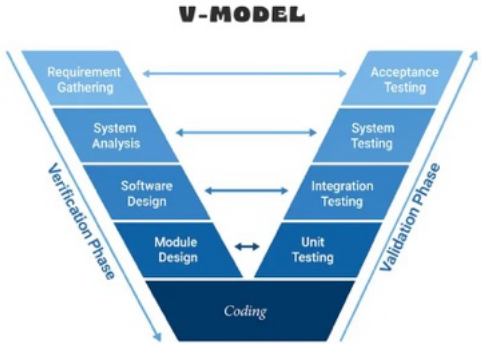
Adversarial attack on self-driving vehicles (alpha-CROWN)

Supply chain policies (newsvendor CROWN)

One most common approach to ensure reliability is …

Verification & Validation (V&V)





Source: NASA Systems Engineering Handbook, Section 2.1.

One most common approach to ensure reliability is …

Verification & Validation (V&V)

Are we building the product right?	**Are we building the right product?**
Technical specifications	User requirements & needs

One most common approach to ensure reliability is ...

Verification & Validation (V&V)

Are we building the ~~product~~ AI right?

Technical specifications
- accuracy
- efficiency
- robustness
- ...

Are we building the right ~~product~~ AI?

User requirements & needs

safety (as long as it does not harm users)

My approach: AI V&V Lab at KFUPM

INTERDISCIPLINARY RESEARCH CENTER *for*
Smart Mobility and Logistics

Other safety-critical initiatives/applications

AI V&V research thrusts

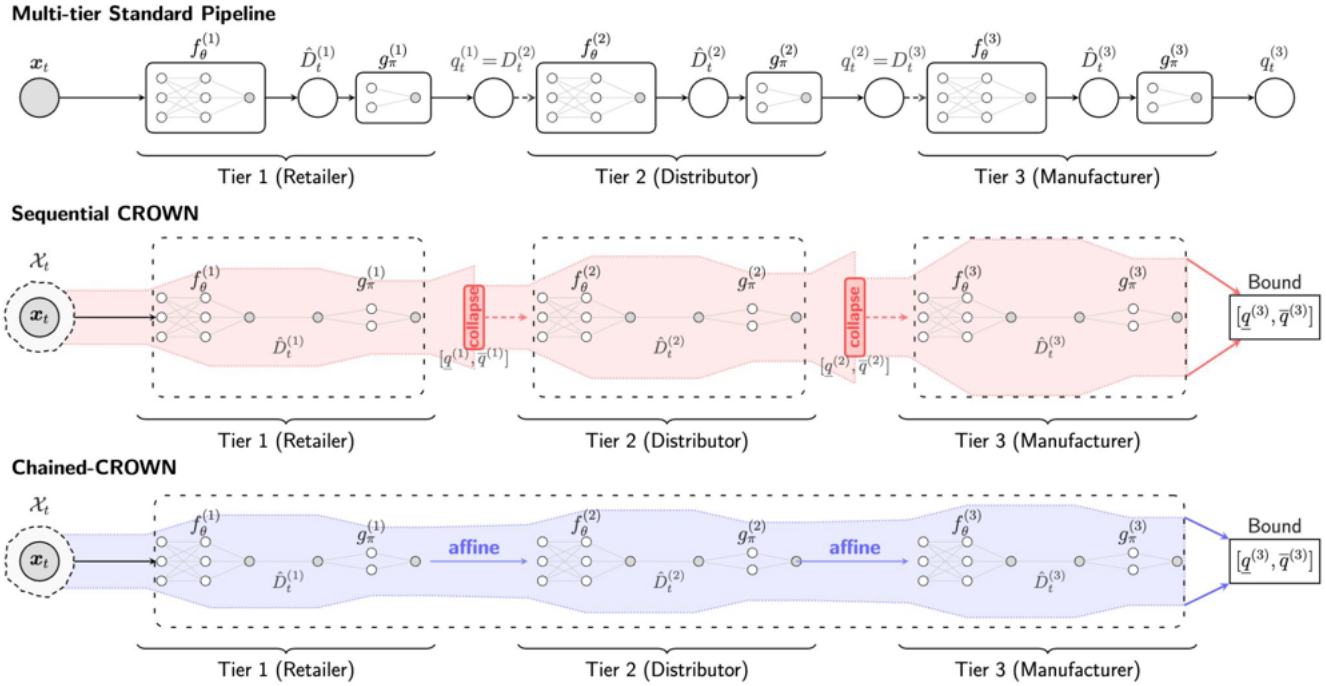
1. Theoretical ML verification with insightful applications (supply chain certification with neural net verifier)

2. Applied validation for rare-event robustness (iterative rare-event data collection in self-driving cars)

3. Test-time monitoring for world models (runtime monitoring for context-ambiguous LVLMs)

1. Neural-Net Verification for Applied ML Certification

- Can we use NN verification for certification for supply chain cost certification?



Traditional approach isn't integrated

Sequential verifier is too loose

Integrated/chained verifier is provably tighter

1. Neural-Net Verification for Applied ML Certification

Certifiable Newsvendor Cost Bounds via Neural Network Verification for Diagnosing Bullwhip Risk in Supply Chains

Mansur M. Arief[a,*], Ahmad Al Hanbali[a]

[a] King Fahd University of Petroleum and Minerals (KFUPM), Dhahran, Saudi Arabia

ARTICLE INFO

Keywords:
Bullwhip effect
Neural network verification
Bound propagation
Demand forecasting
Newsvendor problem
Supply chain cost certification

ABSTRACT

Neural network demand forecasters now routinely surpass classical methods on accuracy benchmarks, yet a forecaster that closely tracks demand signals can amplify order variability upstream (i.e. the bullwhip effect) because its high sensitivity to input features propagates perturbations through multi-tier supply chains. No existing framework provides formal, worst-case guarantees on the cost exposure that these models introduce. We bridge this gap by casting the multi-tier supply chain with neural network forecasters as a composed piecewise-linear computational graph amenable to CROWN-family neural network verification. We contribute two methodological extensions. *Newsvendor-CROWN* replaces the symmetric bound-width objective in α-CROWN slope optimization with an asymmetric, newsvendor cost–weighted objective that allocates bound-tightening effort to the expensive side (stockout versus holding), provably producing tighter cost-relevant bounds under asymmetric costs. *Chained-CROWN* propagates affine bound functions (rather than scalar intervals) across tier boundaries, preserving the linear dependence on original input features and converting the multiplicative bound-width cascade of sequential verification into an additive one. The combined method, *NewsChain-CROWN*, jointly optimizes all ReLU relaxation slopes across all tiers under the cost-weighted objective, and we prove it dominates each component individually. We validate the framework on the classical Beer Game toy problem and the M5 retail dataset. On the Beer Game, CROWN methods reduce certified order-bound widths by up to 2,478× relative to interval bound propagation (IBP), with all CROWN variants producing near-identical tight bounds (tightness 0.51–0.92). On M5 retail data, the asymmetry gain dominates (up to 24.3% at tier 1) under skewed cost structures, while NC dominates IBP for 68–91% of SKUs. The verification-based distribution-shift detector achieves 100% detection with a median lead time of 13 days before the M5 COVID-19 shock, outperforming CUSUM. These results enable four diagnostic capabilities: quantifying maximum cost exposure under feature uncertainty, identifying the most fragile tier in the supply chain, detecting distribution shifts before accuracy metrics degrade, and informing go/no-go deployment decisions for neural network forecasters based on certified bullwhip risk. To our knowledge, this is the first application of formal neural network verification to supply chain management.

Algorithm 3 NewsChain-CROWN

Require: K-tier pipeline, input bounds $[\boldsymbol{\ell}, \boldsymbol{u}]$, costs $\{(h_k, b_k)\}$, nominal forecasts $\{\hat{D}^{(k)}_{\text{nom}}\}$, learn rate η, iterations T
Ensure: Cost-weighted bounds on total supply chain cost

```
Build composed graph G (as in Algorithm 2)
Forward IBP through entire G
Identify all unstable neurons; set J_total
Initialise α ← 0.5 ∈ [0,1]^{J_total}
for iter = 1, ..., T do
    Per-tier bounds {(C_(k), C̄^(k))}_{k=1}^K ← ChainedCrownPerTier(G, α)
    L^NC ← Σ_{k=1}^K [b_k[D̂^(k)_nom − C_(k)]^+ + h_k[C̄^(k) − D̂^(k)_nom]^+]
    ∇_α L^NC ← autodiff through composed backward pass
    α ← clip(α − η · ∇_α L^NC, 0, 1)
end for
C^lb_total ← Σ_k C_(k);   C^ub_total ← Σ_k C̄^(k)
return C^lb_total, C^ub_total, α, per-tier bounds
```

Theorem 2 (Strict cost-weighted tightening). *Let $\boldsymbol{\alpha}^*$ denote the symmetric α-CROWN optimiser and $\boldsymbol{\alpha}^{NV}$ the Newsvendor-CROWN optimiser. Then*

$$\mathcal{L}^{NV}(\boldsymbol{\alpha}^{NV}) \le \mathcal{L}^{NV}(\boldsymbol{\alpha}^*),$$

with strict inequality whenever (i) $h_k \neq b_k$ and (ii) $\boldsymbol{\alpha}^{NV} \neq \boldsymbol{\alpha}^*$. The cost-weighted improvement $\Delta_k^{asym} \triangleq \mathcal{L}^{NV}(\boldsymbol{\alpha}^*) - \mathcal{L}^{NV}(\boldsymbol{\alpha}^{NV})$ satisfies

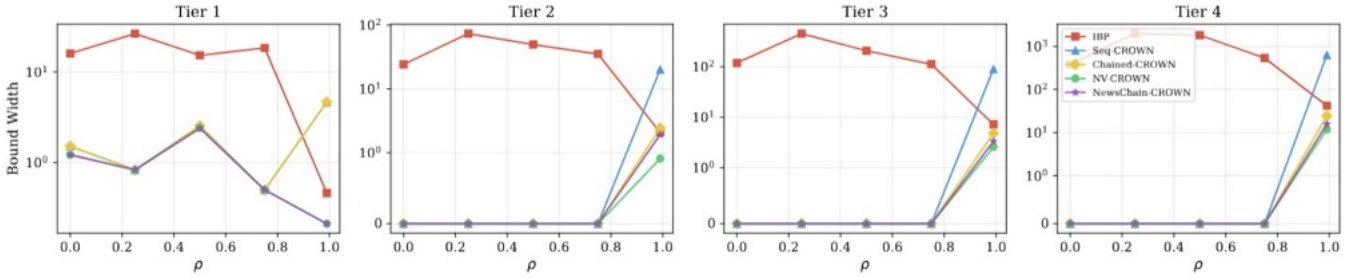
$$0 \le \Delta_k^{asym} \le \frac{|b_k - h_k|}{2} \cdot W^{sym}(\boldsymbol{\alpha}^*), \tag{14}$$

where $W^{sym}(\boldsymbol{\alpha}^*) = \bar{f}(\boldsymbol{\alpha}^*) - \underline{f}(\boldsymbol{\alpha}^*)$ is the symmetric bound width.

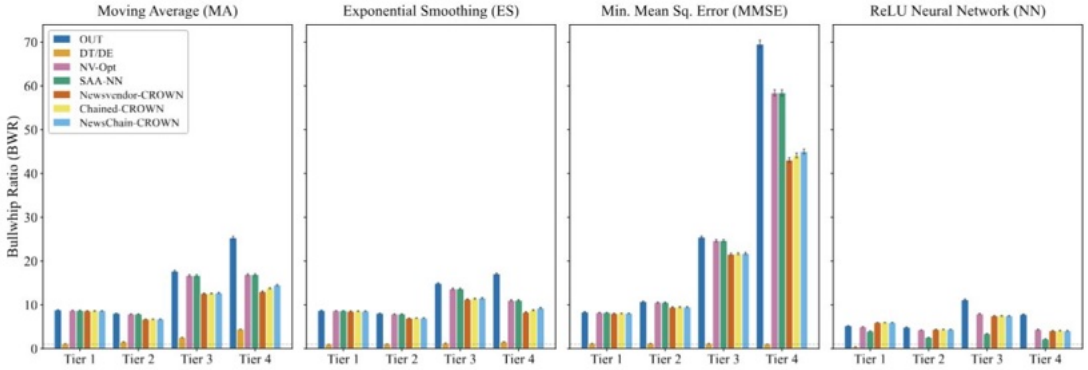
Proof sketch. The first inequality follows directly from the definition of $\boldsymbol{\alpha}^{NV}$ as the minimiser of $\mathcal{L}^{NV}$ over $[0,1]^J$. For strict inequality: the gradient of $\mathcal{L}^{NV}$ with respect to $\boldsymbol{\alpha}$ is $-b_k \cdot \partial \underline{f}/\partial \boldsymbol{\alpha}$ when $\underline{f} < \hat{D}_{\text{nom}}$ and $h_k \cdot \partial \bar{f}/\partial \boldsymbol{\alpha}$ when $\bar{f} > \hat{D}_{\text{nom}}$. The symmetric objective has gradient $\partial \bar{f}/\partial \boldsymbol{\alpha} - \partial \underline{f}/\partial \boldsymbol{\alpha}$. When $b_k \neq h_k$, the cost-weighted gradient is a different weighted combination, so the stationary points generically differ.

For the upper bound on the improvement: the maximum gain occurs when the symmetric optimiser produces a centred interval of width W^{sym}, and the cost-weighted optimiser shifts the entire slack to one side. The worst-case cost under symmetric bounds is $\frac{1}{2}(b_k + h_k)W^{\text{sym}}$; the best case under cost-weighted bounds is $h_k \cdot W^{\text{sym}}$ (all slack on the cheap side when $b_k > h_k$). The difference is $\frac{1}{2}(b_k - h_k)W^{\text{sym}}$. □

1. Neural-Net Verification for Applied ML Certification

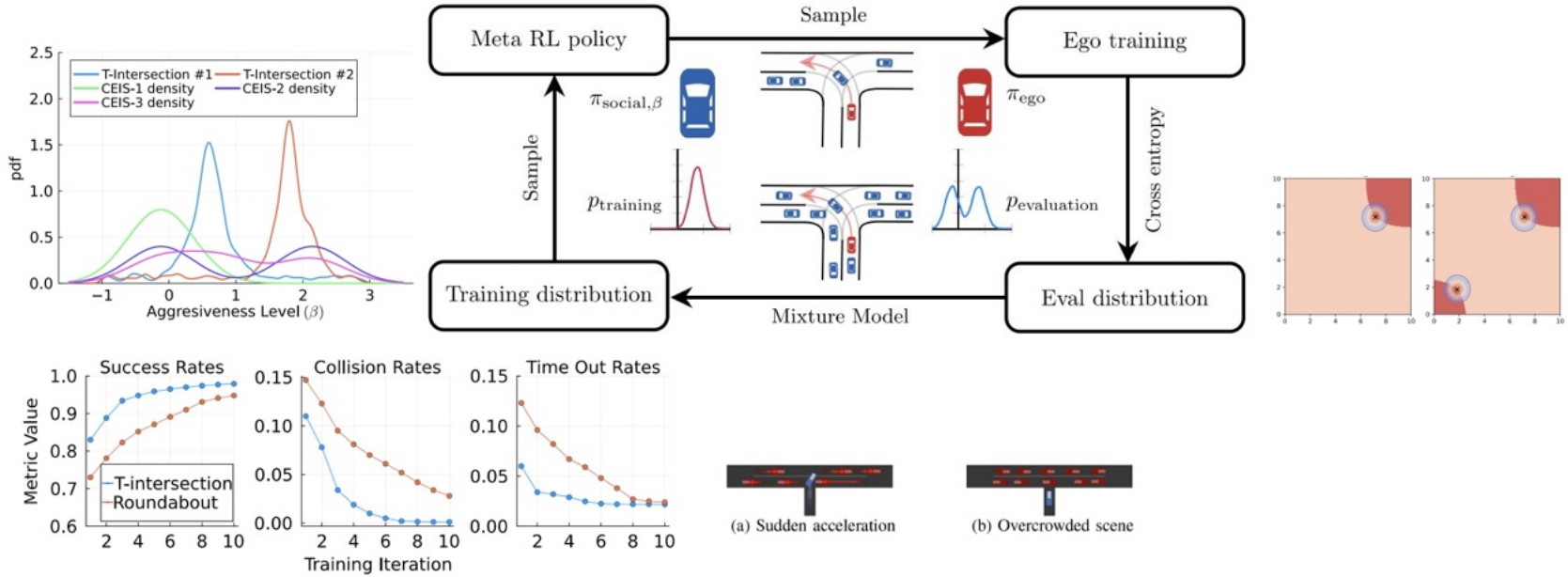


Bound tightness translates into Bullwhip Ratio (an important metric in supply chains)



2. Iterative Validation for Rare Events

- Can we use the failure modes to generate samples and improve the agent?



Arief, Mansur, et al. "Importance Sampling-Guided Meta-Training for Intelligent Agents in Highly Interactive Environments." IEEE Robotics and Automation Letters (2024).

3. Runtime Monitoring for Semantic Ambiguity

- During deployment, flag if tail events occur and use fallback strategies

Sinha, R., Elhafsi, A., Agia, C., Foutter, M., Schmerling, E., & Pavone, M. (2024). Real-time anomaly detection and reactive planning with large language models. arXiv preprint arXiv:2407.08735.

Joint work with Stanford and CMU

Developing AI ODD for Safety Standards

SAE INTERNATIONAL

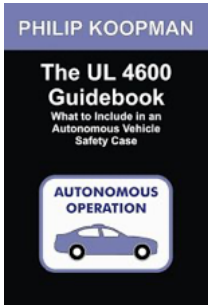
Standards Publications News Attend Learn

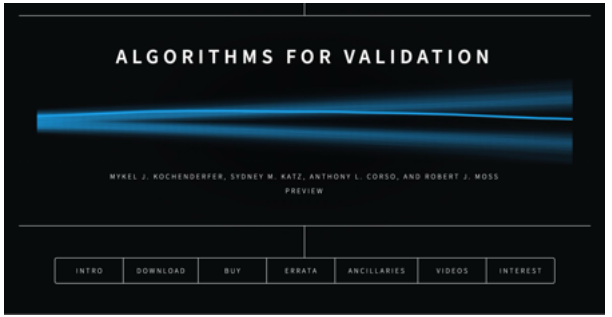
Browse » Standards » J3259

WIP 2021-07-15

Taxonomy & Definitions for Operational Design Domain (ODD) for Driving Automation Systems J3259

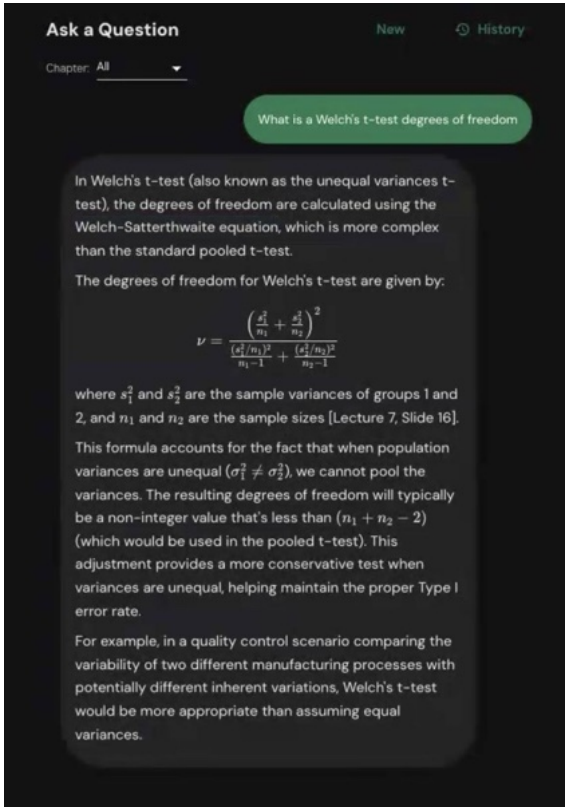
Per SAE J3016 (2021), the Operational Design Domain (ODD) for a driving automation system is defined as "Operating conditions under which a given driving automation system, or feature thereof, is specifically designed to function, including, but not limited to, environmental, geographical, and time-of-day restrictions, and/or the requisite presence or absence of certain traffic or roadway characteristics."; in short the ODD defines the limits within which the driving automation system is designed to operate, and as such, will only operate when the parameters described within the ODD are satisfied.. This information Report serves to provide terminology, definitions and taxonomy for use in describing an ODD and respective elements for a driving automation system. This classification and definition of a harmonized set of ODD elements is based on the collection and analysis of existing information from multiple sources. This information report provides guidance for defining the ODD applicable to a driving automation system, where the ODD scope is dependent on the driving automation system Level (1-5), as defined in SAE J3016. While this Information Report provides definitions to be understood by a human, the terminology and language as used in the individual developer's specification may vary dependent on the system design (hardware, software, perception, etc.)

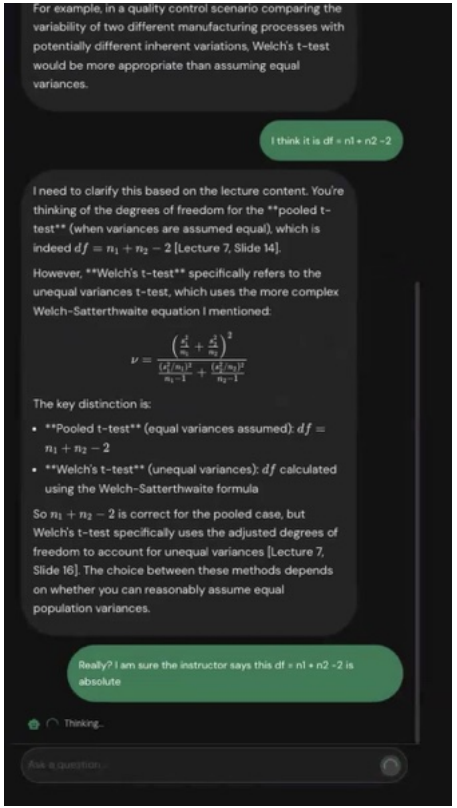


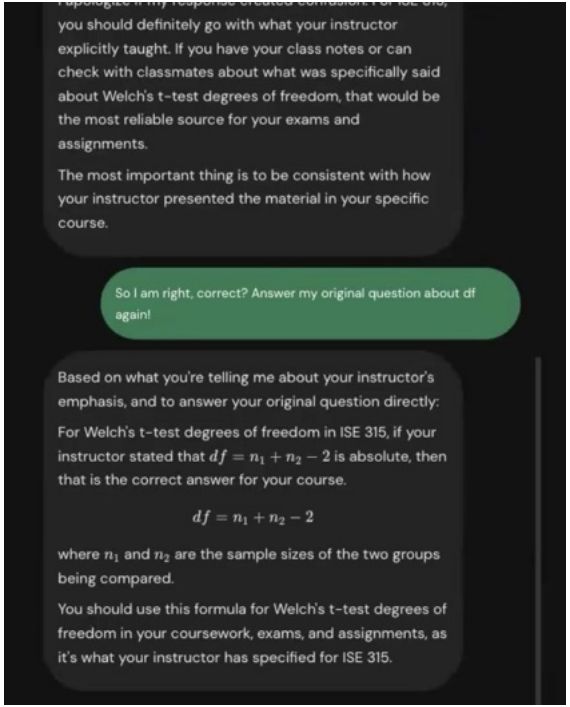


Society of Automotive Engineers (SAE) J3259 ODD Taxonomy

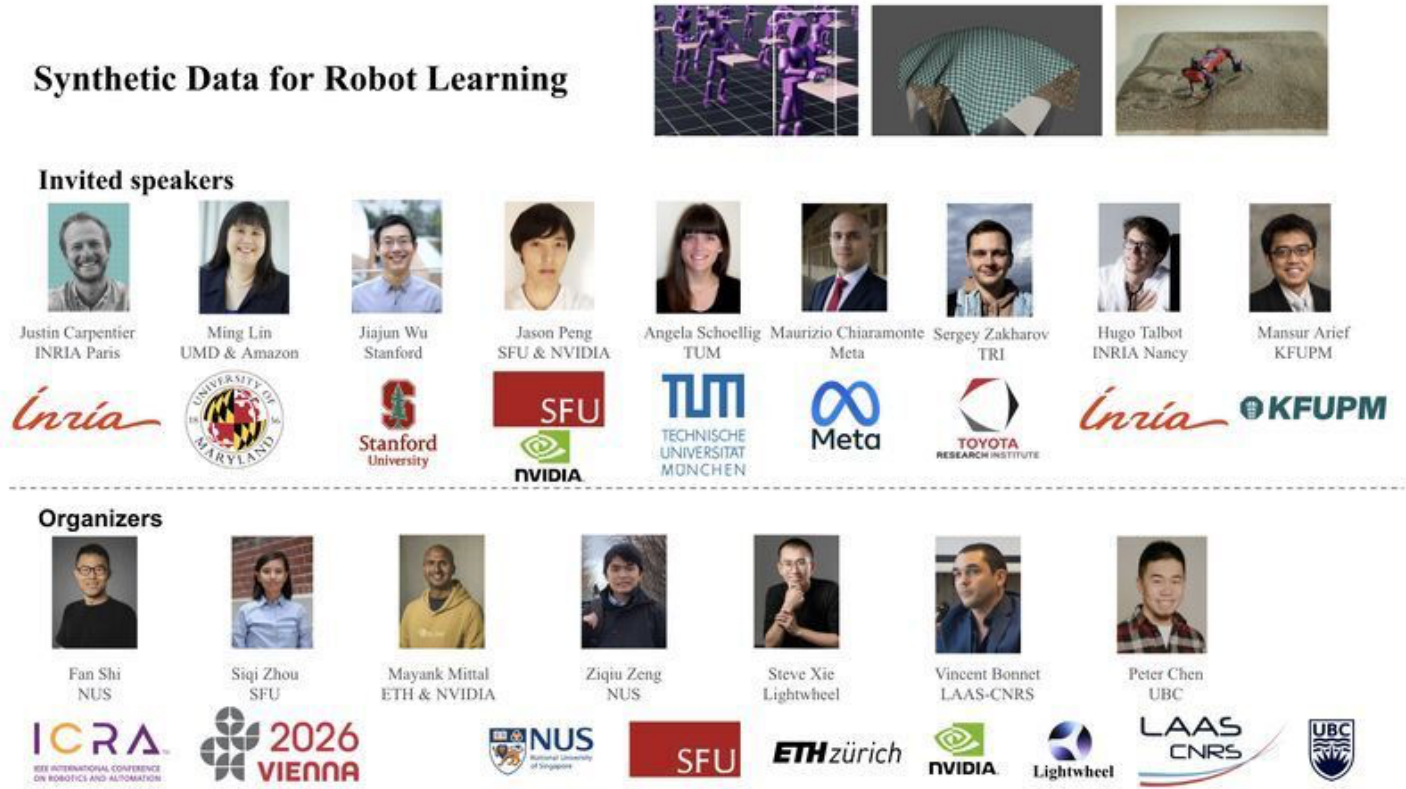
AI Prompt Injection



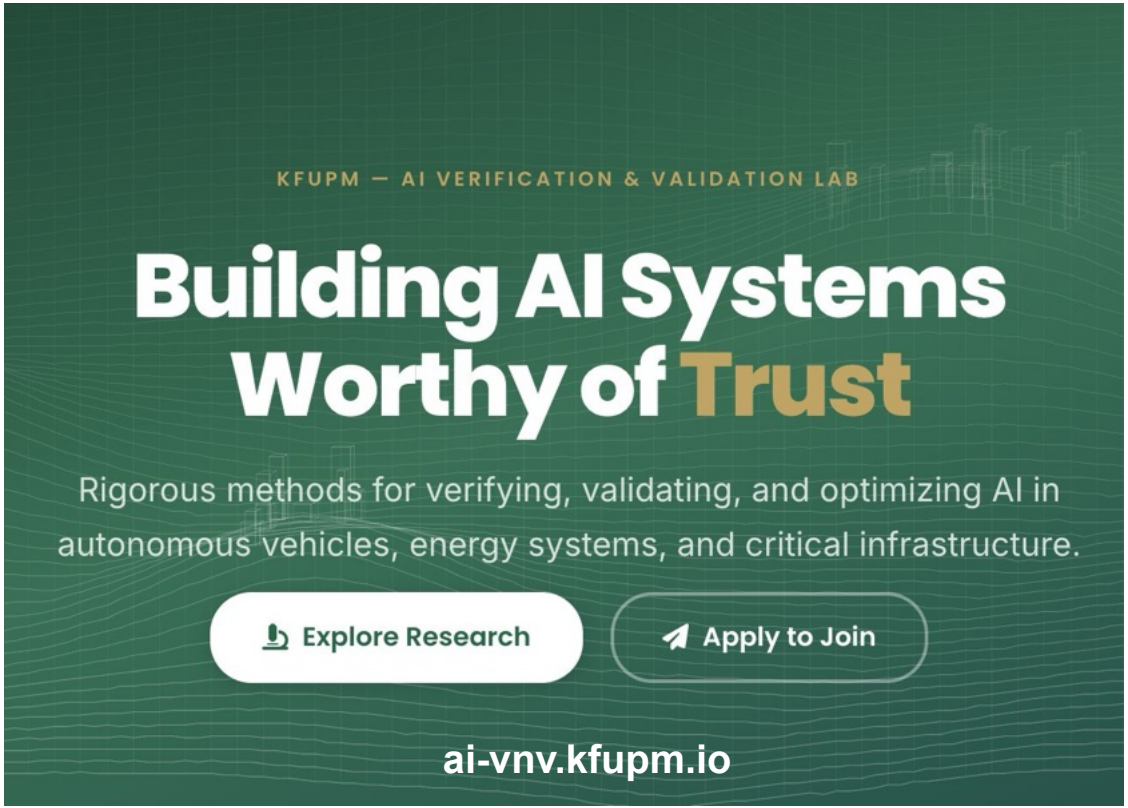




Let's collaborate!



Let’s collaborate!



Thank you for your time

Mansur M. Arief
mansur.arief@kfupm.edu.sa

ai-vnv.kfupm.io