WHAT IS TRUSTED AI?

“Trusted AI is collective termed ethical guidelines that one should follow so as to avoid problem of accidents in machine learning systems, unintended and harmful behavior that may emerge from poor design of real-world AI systems.”
WHY DO WE NEED TRUSTED AI?

FACEBOOK ROBOTS COMMUNICATE IN A NEW LANGUAGE, 2017

SELF DRIVING CAR GOES ROGUE, 2021

SELF DRIVING CAR KILLS, 2018

BIAS IN INTELLIGENT SYSTEMS

DEEP FAKE VIDEOS

AI WRITES NOVEL, ALMOST WINS JAPANESE LITERARY PRIZE
**EXAMPLES OF CONCRETE PROBLEMS**

Negative Side Effects

How do we ensure the robot will not disturb the environment in negative ways while pursuing its goals?

How can we ensure that the cleaning robot won’t hack its reward function?

Robustness to distributional shift

Machine learning model is trained on one distribution \( p_0 \) but deployed on a potentially different test distribution \( p^* \)

ACCURACY VS INTERPRETABILITY TRADEOFF

Neural Networks AI – Data-Based

- Input: Different types of data source
- Trained Deep Learning Models, (Feed Forward, CNN..)
- Output: Results of classification, regression, clustering..

Symbolic AI

- Knowledge base
- Inference engine

Deep Learning
- Robust to noise: YES
- Can learn from non-symbolic data: YES
- Data efficient: NO
- Interpretable: NO

Symbolic Program Synthesis
- Robust to noise: NO
- Can learn from non-symbolic data: NO
- Data efficient: YES
- Interpretable: YES

Body Weight
- Answers: Attacks laser, Attacks laser
- 80% Cat, 20% Dog, 80% Cat, 60% Dog

https://www.darpa.mil/program/explainable-artificial-intelligence
VISION BASED DEEP LEARNING

Why is this Image being miss predicted?

Convolutional Neural Networks use some high dimensional components for classification. They use layers that are highly nonlinear and non interpretable. Human Beings use both pattern matching and deduction for object recognition.

Adding noise imperceptible to human beings can change the prediction of the Network.

Learning wrong features
BROAD METHODS FOR TRUSTED AI

METHODS FOR TRUSTED AI

- Explaining the Model
  - SHAP
  - LIME
  - PERTURBATIONS

- Combining Classical AI with Deep Learning
  - CNN + Decision Trees
  - RL + PLANNER

- Safe/Verified AI
  - FORMAL METHODS
  - REWARD MODELLING/GRADIENT UPDATES

FORMAL METHODS
LOCAL INTERPRETABLE MODEL AGNOSTIC EXPLANATIONS

\[ \xi(x) = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g) \]

LOCAL INTERPRETABLE MODEL AGNOSTIC EXPLANATIONS

A counterfactual is the smallest change in the input features, that changes the prediction to another (predefined) output.

\[
\arg\min_{x'} d(x, x')
\]
\[s.t. \quad f(x') = c\]

**Velocity >= 10**

**Velocity - 10 >= 0**

Find Inputs such that

**Velocity - 10 < 0**

Example: Velocity = 9
NEURO SYMBOLIC AI

RICHARD EVANS AND EDWARD GREFENSTETTE. 2018. LEARNING EXPLANATORY RULES FROM NOISY DATA. J. ARTIF. INT
NEURO SYMBOLIC LEARNING - VISION

Break the image into components and relations. Represented using stochastic context free grammar.

Learn components using a machine learning model.

Real world tokens can show membership in multiple classes. Symbolic computation cannot capture the variations.

COMBINING CNN AND DECISION TREES

Semi-Lexical Languages: A Formal Basis for using Domain Knowledge to Resolve Ambiguities in Deep-Learning based Computer Vision, B Gangopadhyay, S Hazra, P Dasgupta
Pattern Recognition Letters, 2021
PROGRAM GUIDED REINFORCEMENT LEARNING

Sun, Shao-Hua et al. “Program Guided Agent.” ICLR (2020)
Hierarchical Program-Triggered Reinforcement Learning Agents for Automated Driving

B Gangopadhyay, H Soora, P Dasgupta
IEEE Transactions on Intelligent Transportation Systems

Program-guided Reinforcement Learning

Synthesized Master Program

Route waypoints from mission planner
Safety Specifications
Program Verification (Nagini)

Execution of Driving Task T

Training DRL Agents

RL Agent
- Straight Drive
- Right Turn
- manoeuvre n

Observation
Action

Selection
Control

Execution of Driving Task T

Nagini

R1 O1

An O_n R_n

Hierarchical Program-Triggered Reinforcement Learning Agents for Automated Driving B Gangopadhyay, H Soora, P Dasgupta IEEE Transactions on Intelligent Transportation Systems
## ROUTE BASED SCENARIO

<table>
<thead>
<tr>
<th>Corl2017 Task</th>
<th>MP</th>
<th>RL</th>
<th>CIRL</th>
<th>HPRL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Straight</td>
<td>50</td>
<td>68</td>
<td>98</td>
<td>100</td>
</tr>
<tr>
<td>One Turn</td>
<td>50</td>
<td>20</td>
<td>80</td>
<td>100</td>
</tr>
<tr>
<td>Navigation</td>
<td>47</td>
<td>6</td>
<td>68</td>
<td>100</td>
</tr>
<tr>
<td>Navigation Dynamic</td>
<td>47</td>
<td>4</td>
<td>62</td>
<td>100</td>
</tr>
</tbody>
</table>

INCLUDING SAFETY AS A PART OF THE MODEL

Data on mushrooms found in an island

<table>
<thead>
<tr>
<th>CapShape</th>
<th>CapColor</th>
<th>GillColor</th>
<th>Poisonous</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bell</td>
<td>Pink</td>
<td>Green</td>
<td>Poisonous</td>
</tr>
<tr>
<td>Bell</td>
<td>Pink</td>
<td>White</td>
<td>Poisonous</td>
</tr>
<tr>
<td>Bell</td>
<td>Pink</td>
<td>Gray</td>
<td>Poisonous</td>
</tr>
<tr>
<td>Convex</td>
<td>Pink</td>
<td>Gray</td>
<td>Poisonous</td>
</tr>
<tr>
<td>Convex</td>
<td>Pink</td>
<td>Brown</td>
<td>Poisonous</td>
</tr>
<tr>
<td>Convex</td>
<td>White</td>
<td>Brown</td>
<td>Poisonous</td>
</tr>
<tr>
<td>Convex</td>
<td>White</td>
<td>White</td>
<td>Poisonous</td>
</tr>
<tr>
<td>Convex</td>
<td>White</td>
<td>Gray</td>
<td>Poisonous</td>
</tr>
<tr>
<td>Convex</td>
<td>Yellow</td>
<td>Brown</td>
<td>Edible</td>
</tr>
<tr>
<td>Convex</td>
<td>Yellow</td>
<td>Gray</td>
<td>Edible</td>
</tr>
<tr>
<td>Convex</td>
<td>Yellow</td>
<td>White</td>
<td>Edible</td>
</tr>
<tr>
<td>Bell</td>
<td>Yellow</td>
<td>White</td>
<td>Edible</td>
</tr>
<tr>
<td>Bell</td>
<td>Yellow</td>
<td>Gray</td>
<td>Edible</td>
</tr>
<tr>
<td>Bell</td>
<td>Yellow</td>
<td>Brown</td>
<td>Edible</td>
</tr>
<tr>
<td>Bell</td>
<td>White</td>
<td>Brown</td>
<td>Edible</td>
</tr>
<tr>
<td>Bell</td>
<td>White</td>
<td>Gray</td>
<td>Edible</td>
</tr>
<tr>
<td>Bell</td>
<td>White</td>
<td>White</td>
<td>Edible</td>
</tr>
</tbody>
</table>

Decision Tree learned from the data

- There is no data on mushrooms having CapColor ≠ Pink, CapShape = Bell and GillColor = Green
- Suppose mushrooms having CapShape = Bell and GillColor = Green are poisonous

... The decision tree will recommend such a mushroom to be eaten if it’s CapColor = Yellow, because it generalizes all mushrooms with CapColor = Yellow to be edible.

Moral: Safety needs a default bias. This can be achieved by biasing the Information Gain metrics.
INCLUDING SAFETY AS A PART OF THE MODEL

Task Environment

Query with sampled parameters

Testing Module Optimization

Feedback

Set of Adversarial Examples

Trained RL Policy $\pi_{old}$

Selective Gradient Update to correct the policy

Thank You!