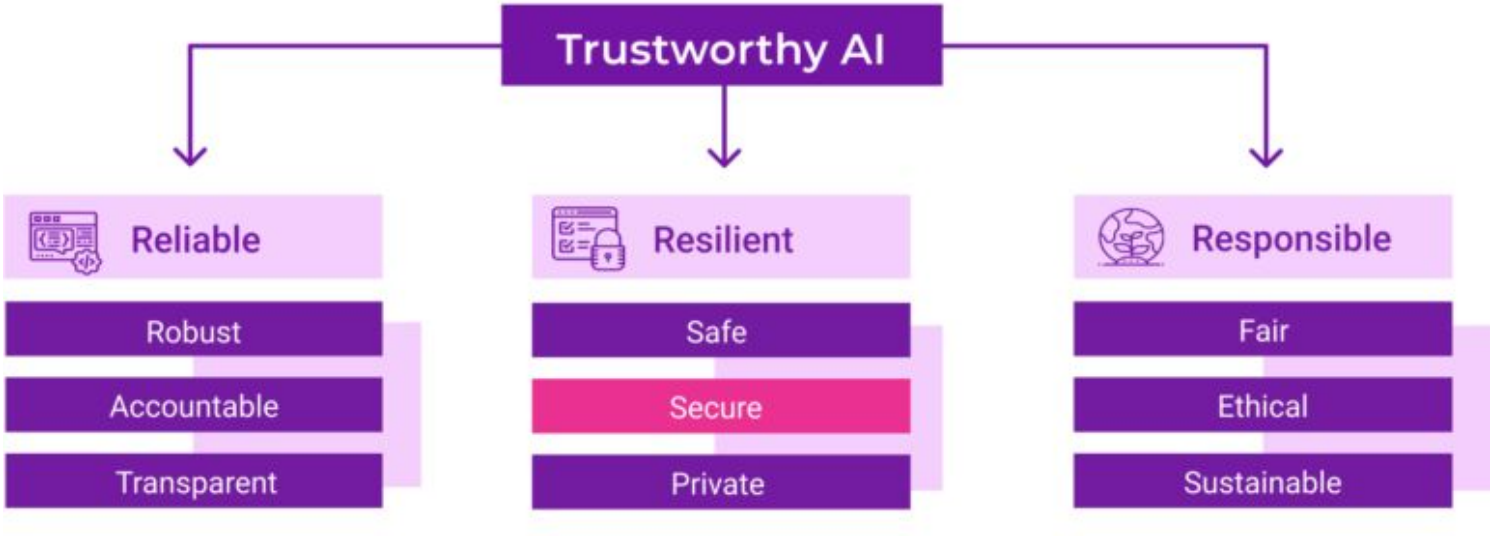
BUILDING TRUSTED AI

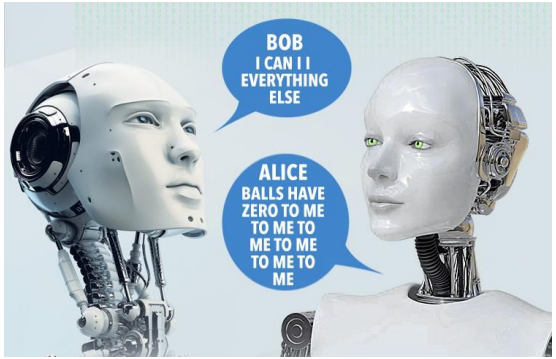
PRESENTED BY: BRITI GANGOPADHYAY

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?



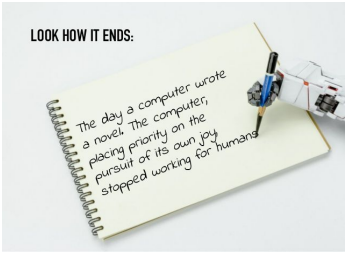
FACEBOOK ROBOTS COMMUNICATE IN A NEW LANGUAGE, 2017

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?

How can we ensure that the cleaning robot respects aspects of the objective that are too expensive to be frequently evaluated during training?

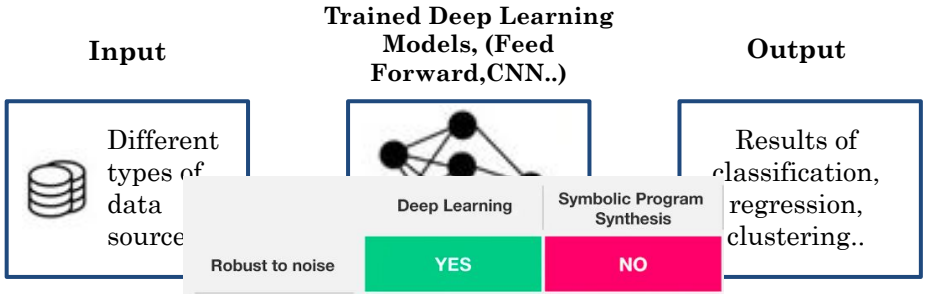
Robustness to distributional shift

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

Amodei, Dario et al. "Concrete Problems in AI Safety." ArXiv abs/1606.06565 (2016): n. pag.

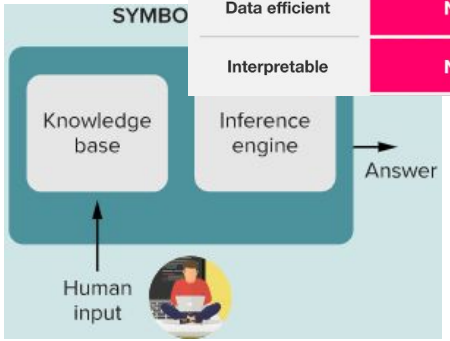
ACCURACY VS INTERPRETABILITY TRADEOFF

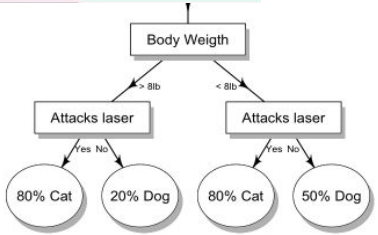
Neural Networks AI – Data-Based

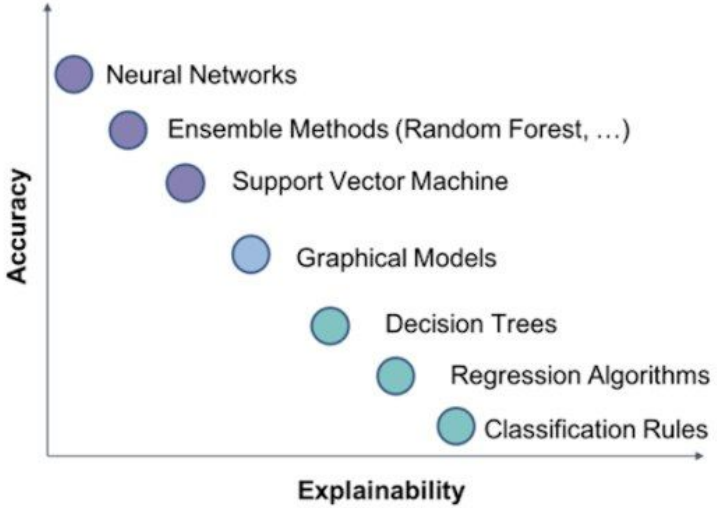


	Deep Learning	Symbolic Program Synthesis
Robust to noise	YES	NO
Can learn from non-symbolic data	YES	NO
Data efficient	NO	YES
Interpretable	NO	YES

Symbolic A







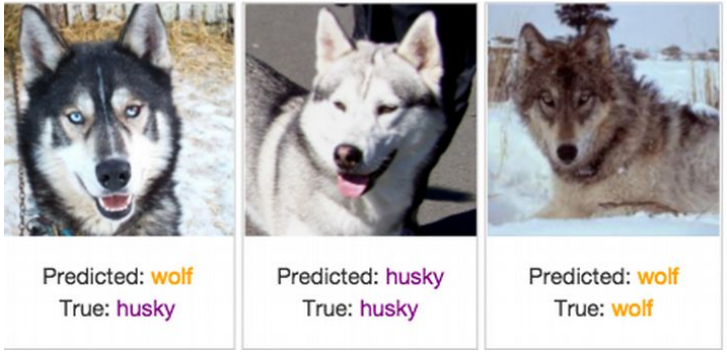
https://www.darpa.mil/program/explainable-artificial-intelligence

VISION BASED DEEP LEARNING

Why is this Image being miss predicted?

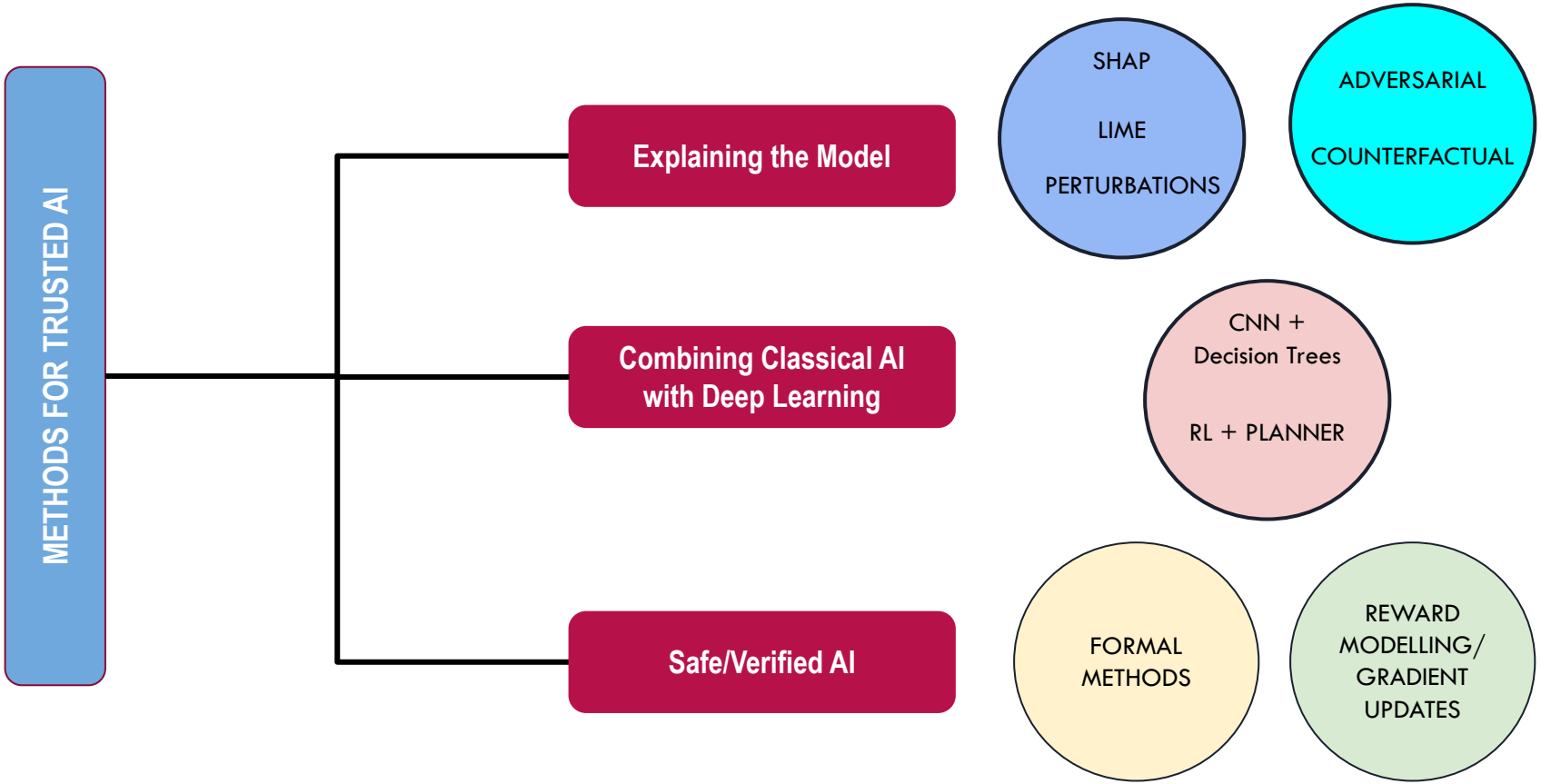
Convolutional Neural Networks use some high dimensional components for classification. They use layers that are highly nonlinear and non interpretable. Human Beings use both pattern matching and deduction for object recognition.

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

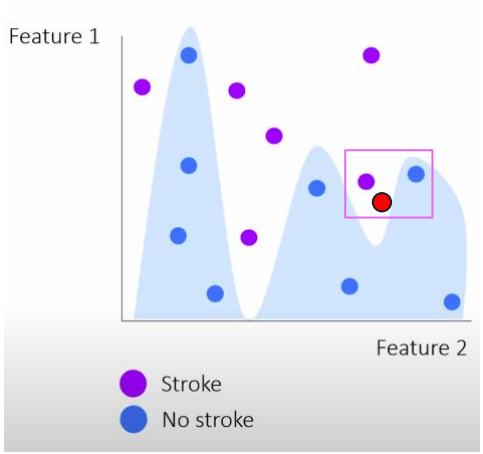


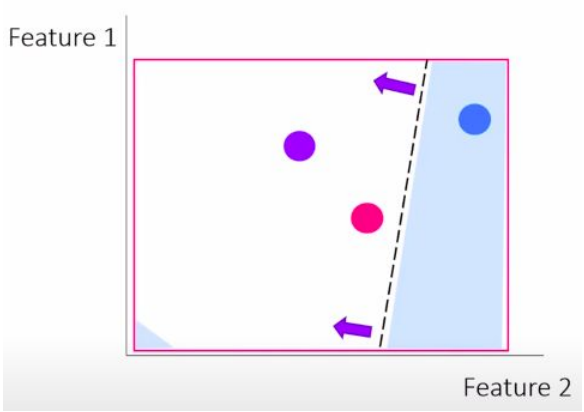
Learning wrong features

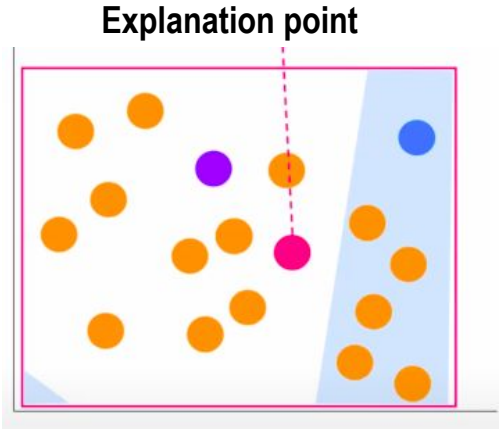
BROAD METHODS FOR TRUSTED AI



LOCAL INTERPRETABLE MODEL AGNOSTIC EXPLANATIONS







Complex model

Simple model

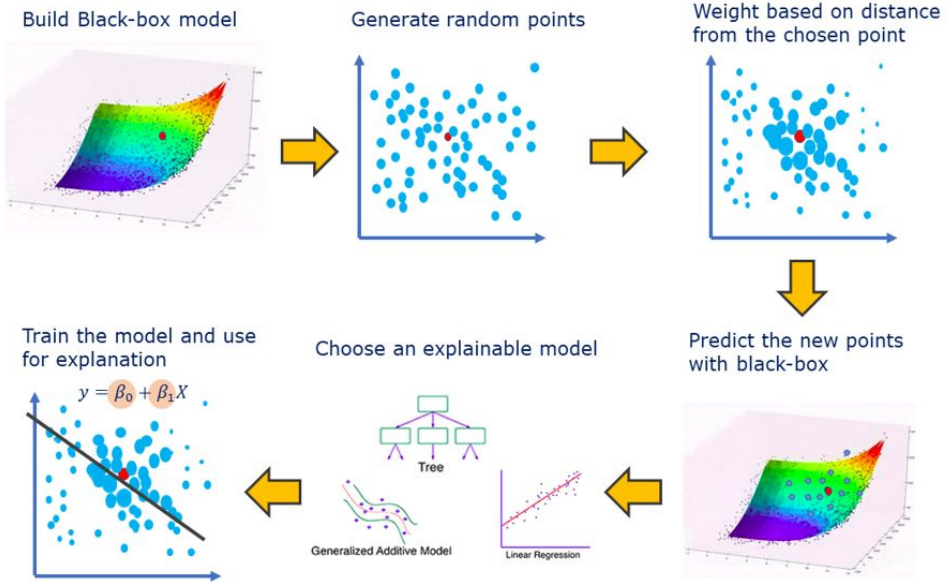
$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \; \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

Input Features

Complexity of the interpretable model

Ribeiro, Marco Tulio et al. "'Why Should I Trust You?": Explaining the Predictions of Any Classifier." Proceedings of the 22nd ACM SIGKDD (2016)

LOCAL INTERPRETABLE MODEL AGNOSTIC EXPLANATIONS



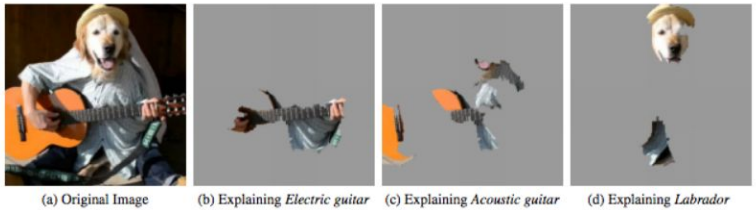
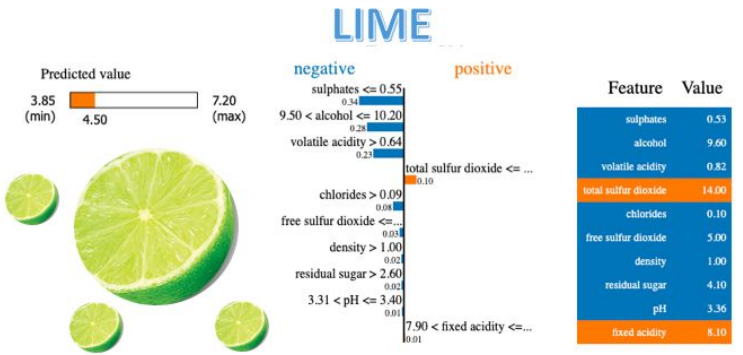


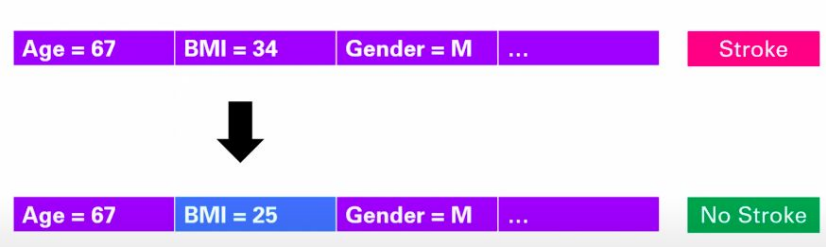
Figure 4: Explaining an image classification prediction made by Google's Inception network, highlighting positive pixels. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)



Ribeiro, Marco Tulio et al. "'Why Should I Trust You?": Explaining the Predictions of Any Classifier." Proceedings of the 22nd ACM SIGKDD (2016)

COUNTERFACTUAL AND ADVERSARIAL EXAMPLES

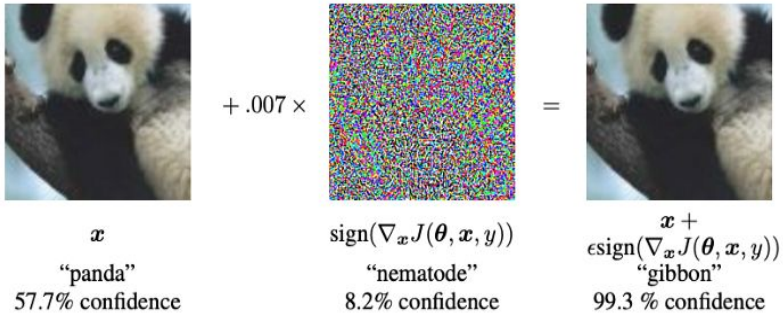
A counterfactual is the smallest change in the input features, that changes the prediction to another (predefined) output.



$$argmin_{x'} \; d(x, x')$$

$$s.t. \quad f(x') = c$$

Counterfactual Explanations



Adversarial Attacks

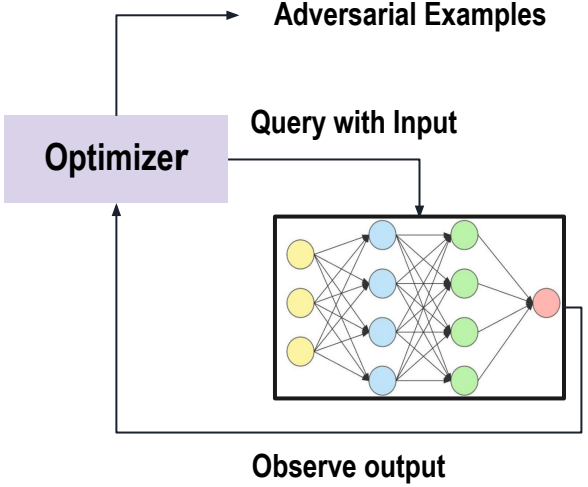
Velocity >= 10

Velocity - 10 >= 0

Find Inputs such that

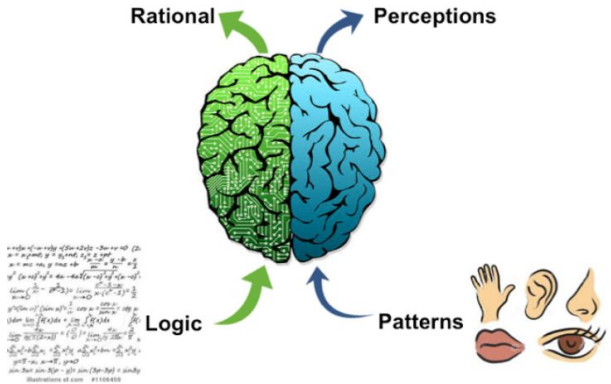
Velocity - 10 < 0

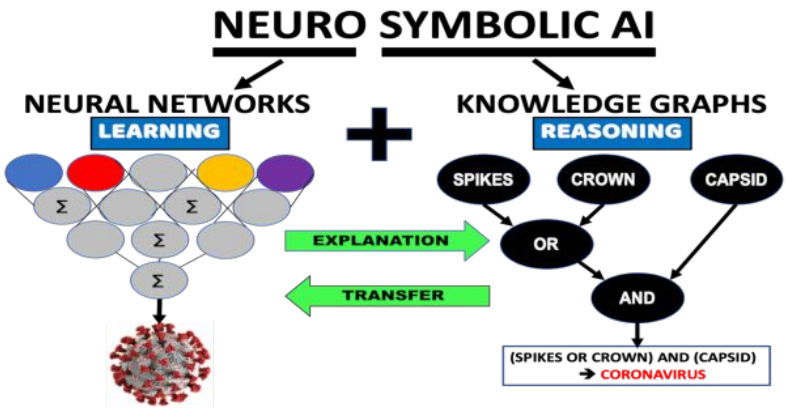
Example : Velocity = 9

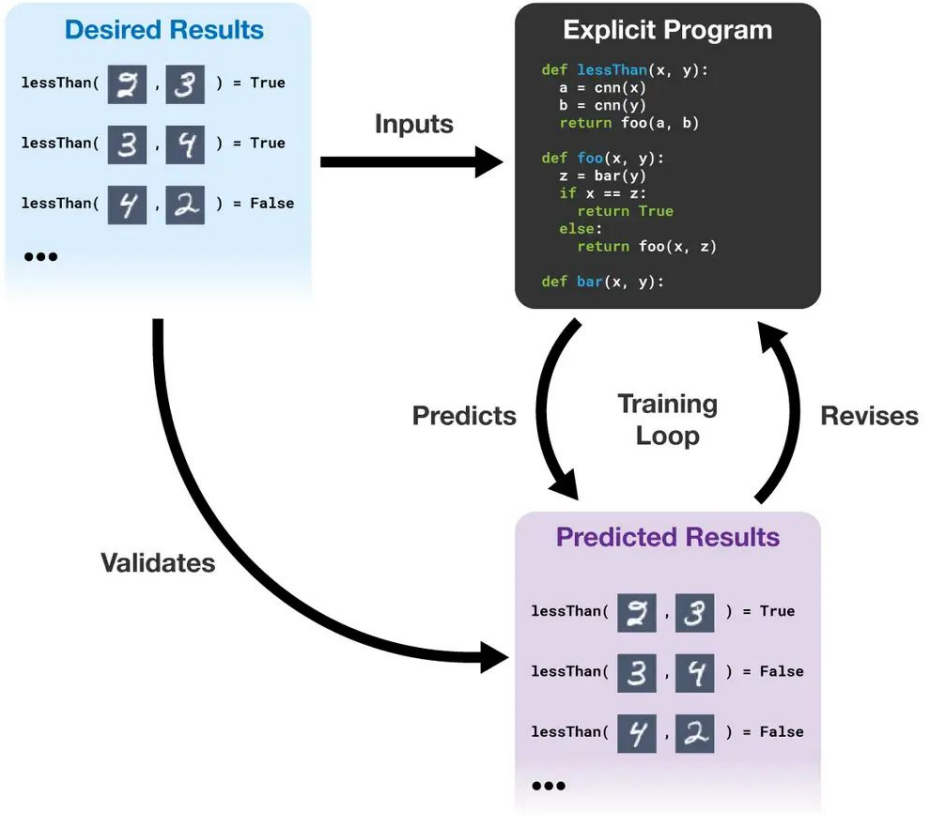


Gangopadhyay, Briti et al. "Identification of Test Cases for Automated Driving Systems Using Bayesian Optimization." 2019 IEEE Intelligent Transportation Systems Conference (2019)

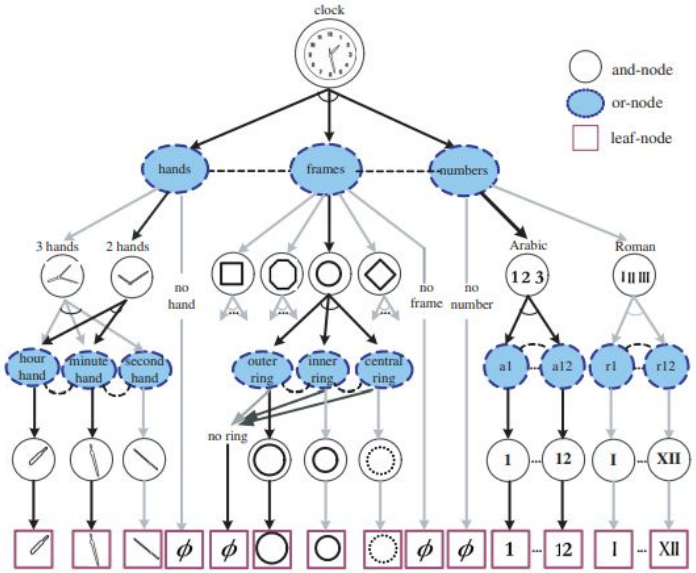
NEURO SYMBOLIC AI



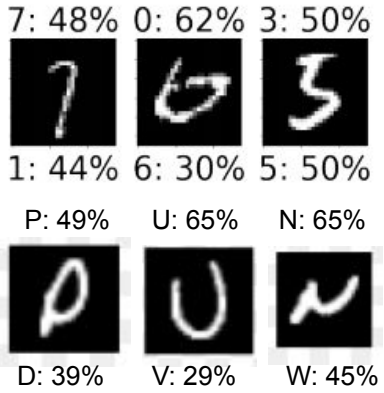




NEURO SYMBOLIC LEARNING - VISION

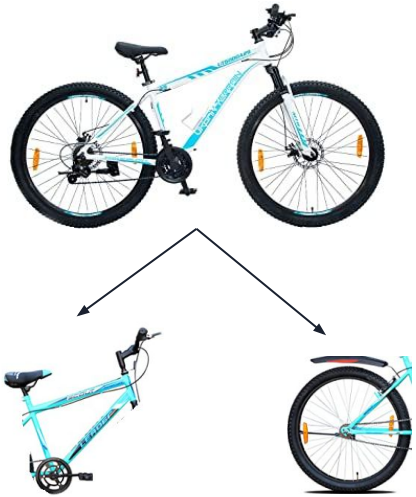


Break the image into components and relations.
Represented using stochastic context free grammar.



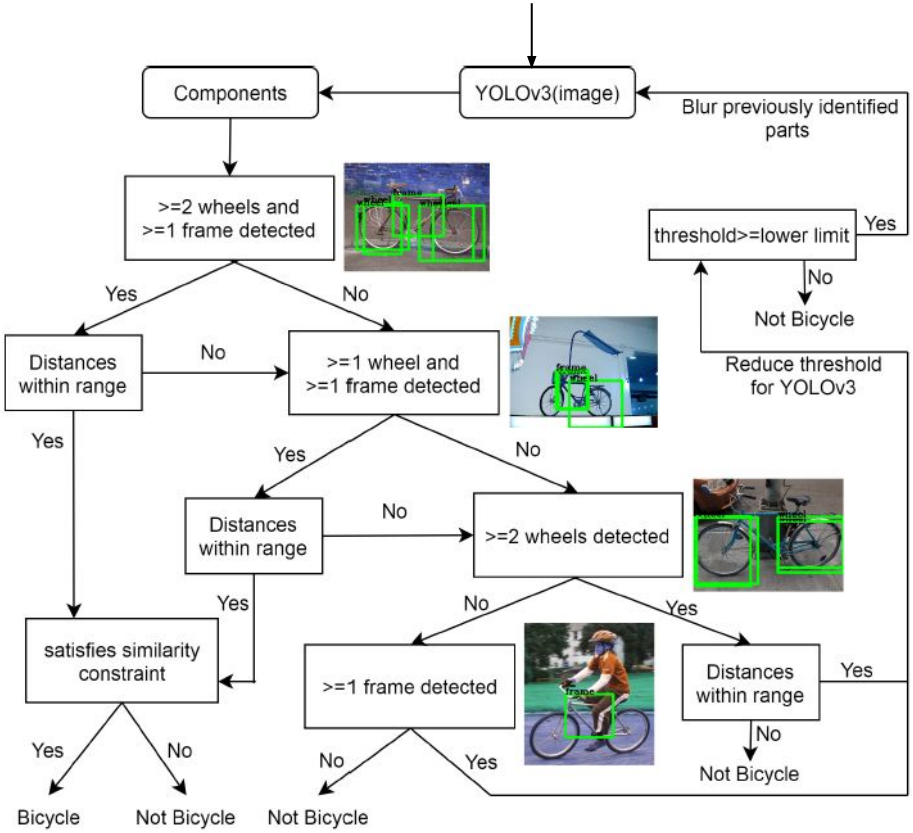
Real world tokens can show membership in multiple classes.
Symbolic computation cannot capture the variations.

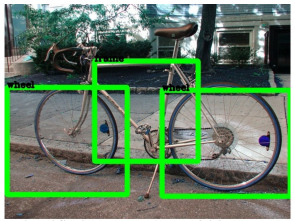
Learn components using a machine learning model.



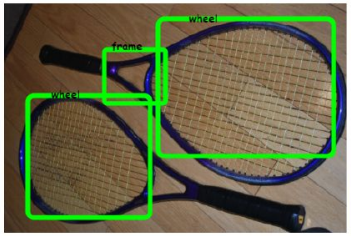
Song-Chun Zhu; David Mumford, A Stochastic Grammar of Images, 2007.

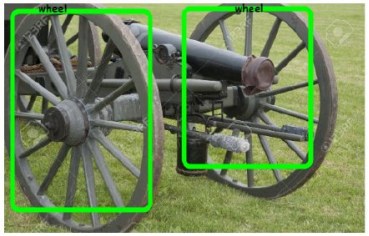
COMBINING CNN AND DECISION TREES

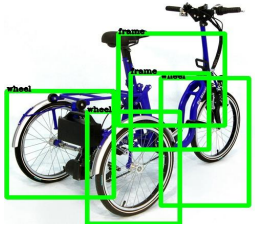




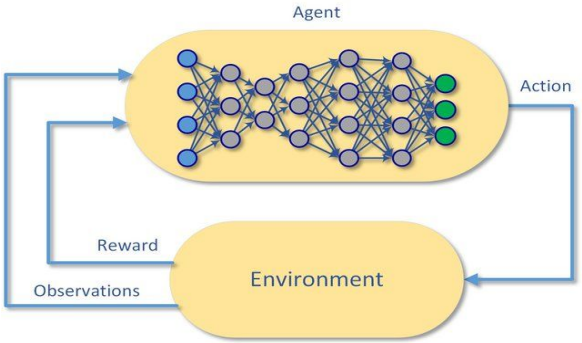
Identified components within range:
[2 wheels, 1 frame]

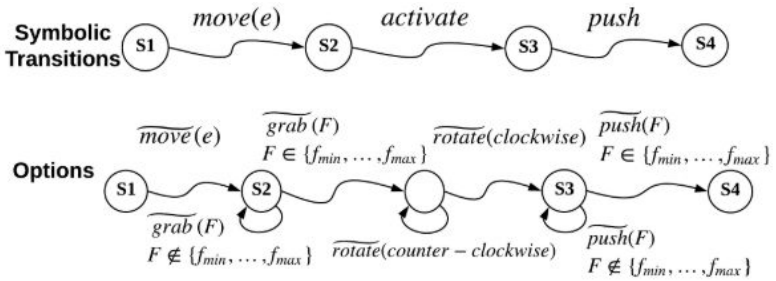


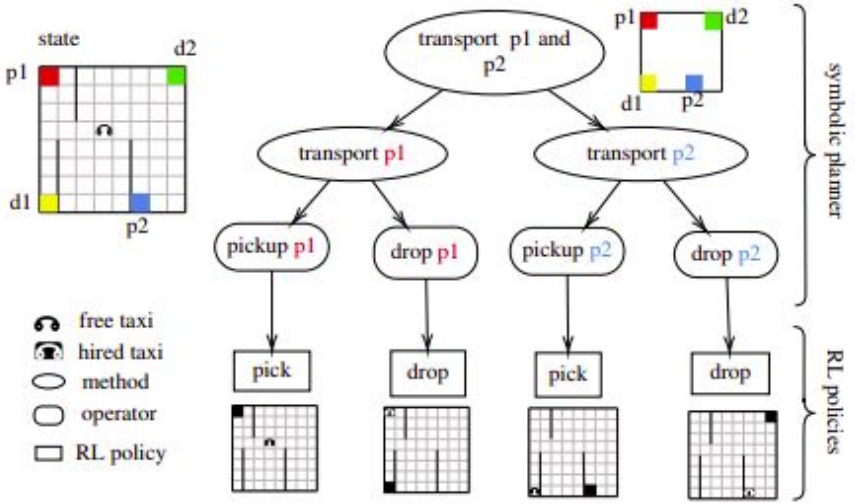




PLANNING AND REINFORCEMENT LEARNING







Yang, Fangkai et al. "PEORL: Integrating Symbolic Planning and Hierarchical Reinforcement Learning for Robust Decision-Making." IJCAI (2018).

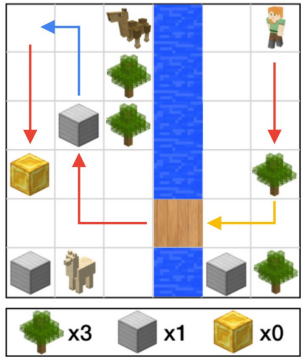
PROGRAM GUIDED REINFORCEMENT LEARNING

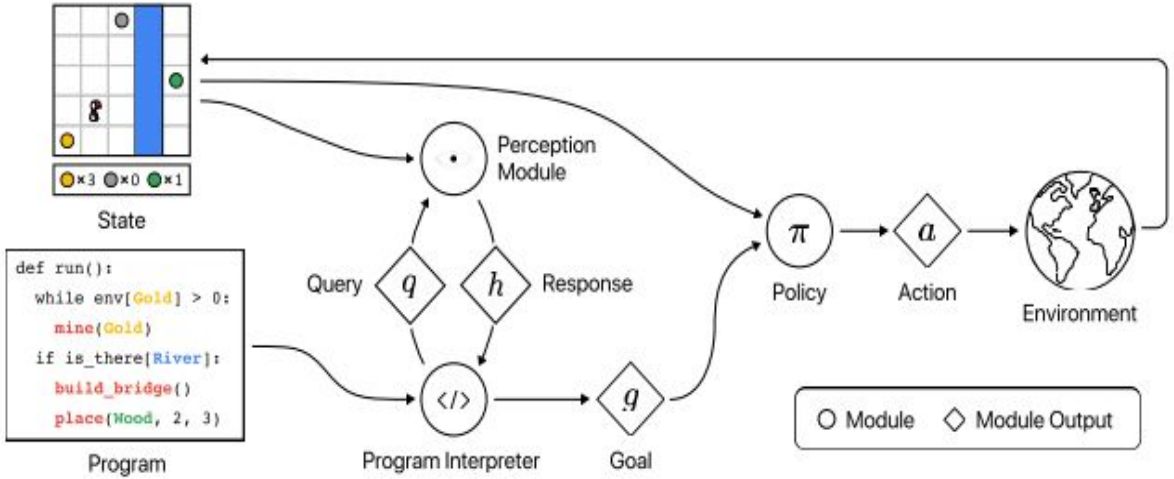
#	Language Instructions	Ground Truth Program	Alternative Interpretation

(a)	If there is a river, build a bridge. Repeat the followings 3 times: mine a gold, and if environment has no more than 8 gold, mine iron, and then sell an iron.	def run():
		  if is_there[River]:
		  build_bridge()
		  loop(3):
		  mine(Gold)
		  if env[Gold] <= 8:
		  mine(Gold)
		  sell(Iron)
			def run():
			  if is_there[River]:
			  build_bridge()
			  loop(3):
			  mine(Gold)
			  if env[Gold] <= 8:
			  mine(Gold)
			  sell(Iron)

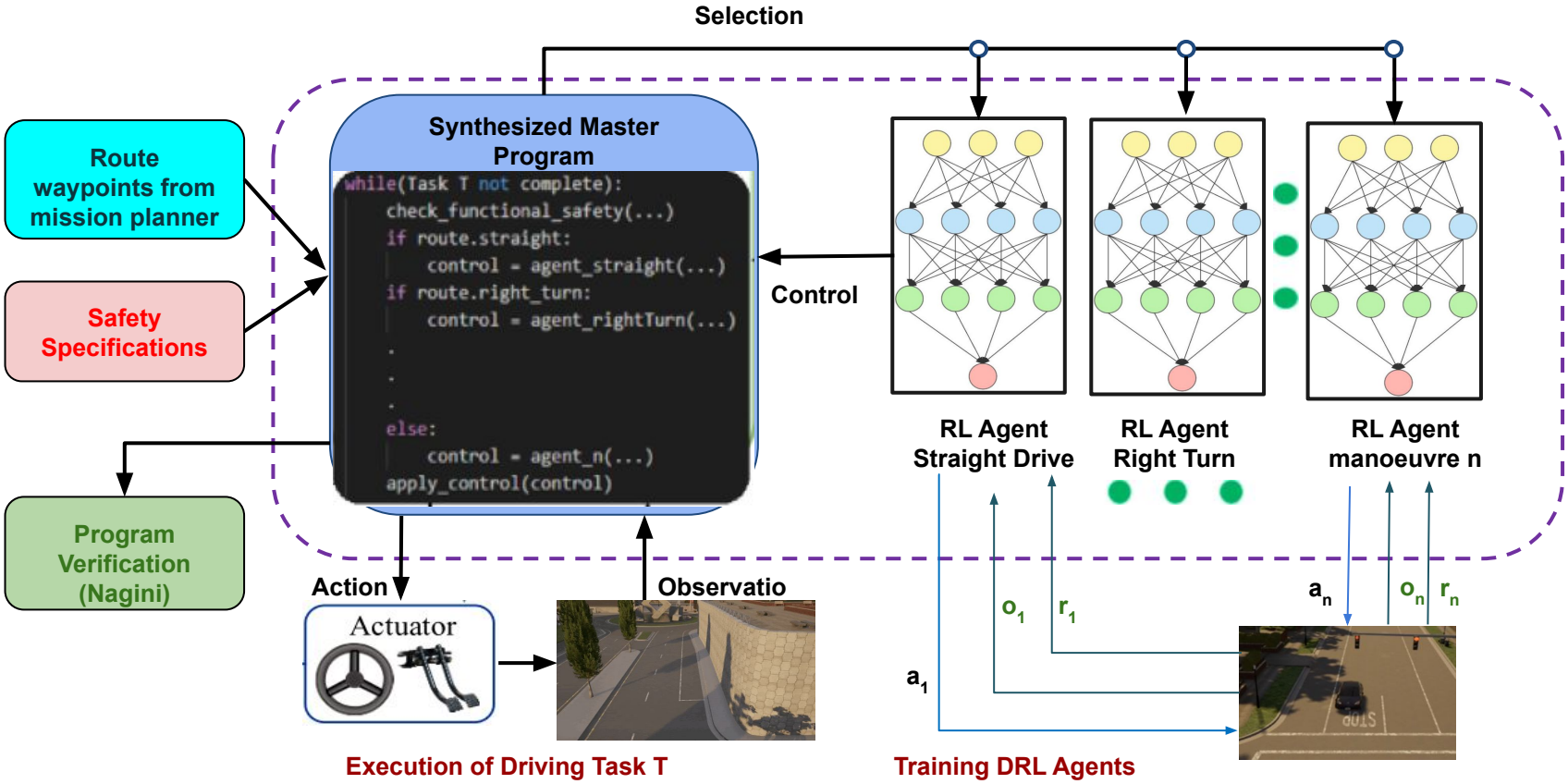
Program

```
def run():
  if is_there[River]:
    mine(Wood)
    build_bridge()
    if agent[Iron]<3:
      mine(Iron)
    place(Iron, 1, 1)
  else:
    goto(4, 2)
  while env[Gold]>0:
    mine(Gold)
```





PROGRAM GUIDED REINFORCEMENT LEARNING



ROUTE BASED SCENARIO

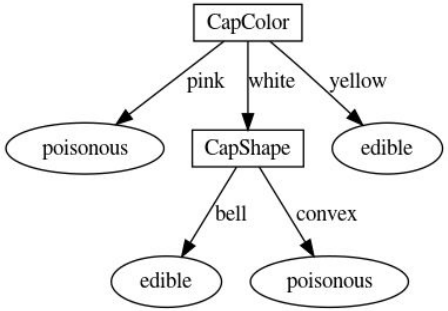
Corl2017 Task	MP	RL	CIRL	HPRL
Straight	50	68	98	**100**
One Turn	50	20	80	**100**
Navigation	47	6	68	**100**
Navigation Dynamic	47	4	62	**100**

A. DOSOVITSKIY, G. ROS, F. CODEVILLA, A. LOPEZ, AND V. KOLTUN, "CARLA: AN OPEN URBAN DRIVING SIMULATOR," IN PROCEEDINGS OF THE 1ST ANNUAL CONFERENCE ON ROBOT LEARNING, 2017, PP. 1–16.

INCLUDING SAFETY AS A PART OF THE MODEL

Data on mushrooms found in an island

CapShape	CapColor	GillColor	Poisonous
Bell	Pink	Green	Poisonous
Bell	Pink	White	Poisonous
Bell	Pink	Gray	Poisonous
Convex	Pink	Gray	Poisonous
Convex	Pink	Brown	Poisonous
Convex	White	Brown	Poisonous
Convex	White	White	Poisonous
Convex	White	Gray	Poisonous
Convex	Yellow	Brown	Edible
Convex	Yellow	Gray	Edible
Convex	Yellow	White	Edible
Bell	Yellow	White	Edible
Bell	Yellow	Gray	Edible
Bell	Yellow	Brown	Edible
Bell	White	Brown	Edible
Bell	White	Gray	Edible
Bell	White	White	Edible



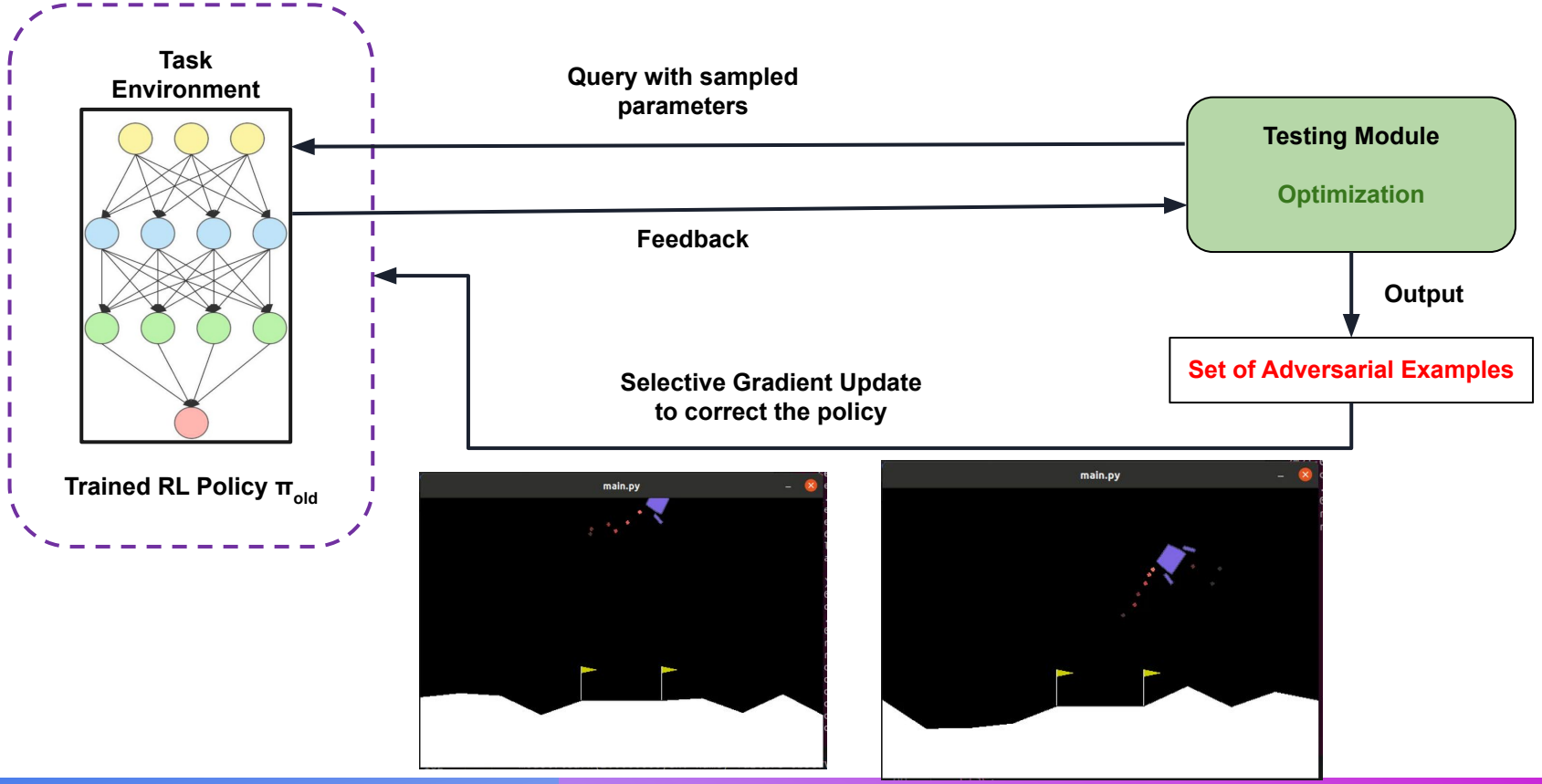
Decision Tree learned from the data

- 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



B. Gangopadhyay, P. Dasgupta, Counterexample Guided RL Policy Refinement Using Bayesian Optimization 35th Conference on Neural Information Processing Systems (NeurIPS 2021)


THANK
YOU!