Towards Trustworthy Autonomous Systems

Leilani Gilpin, **Vishnu Penubarthi** and Lalana Kagal

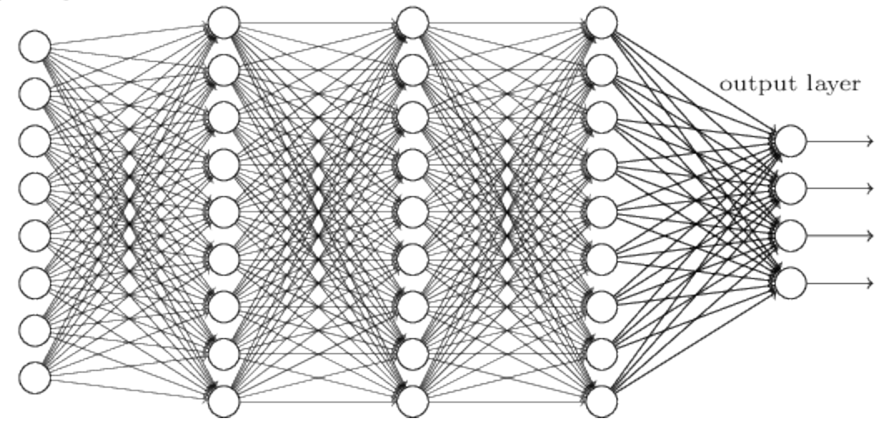
Motivation

- Autonomous systems are responsible for decisions previously entrusted to humans.
- The failure of these systems can have catastrophic consequences with significant loss of life and property.
- It is essential that these systems perform reliably and that their decisions are ***trustworthy*** even in the presence of anomalies and cyber attacks.
- ***Explanations*** can help ensure that these systems are working in our best interest and to help identify attacks and anomalies.
- Applications: self driving cars, adversarial ML (with Dr. Bhargava's group), IoT, disaster management, etc.

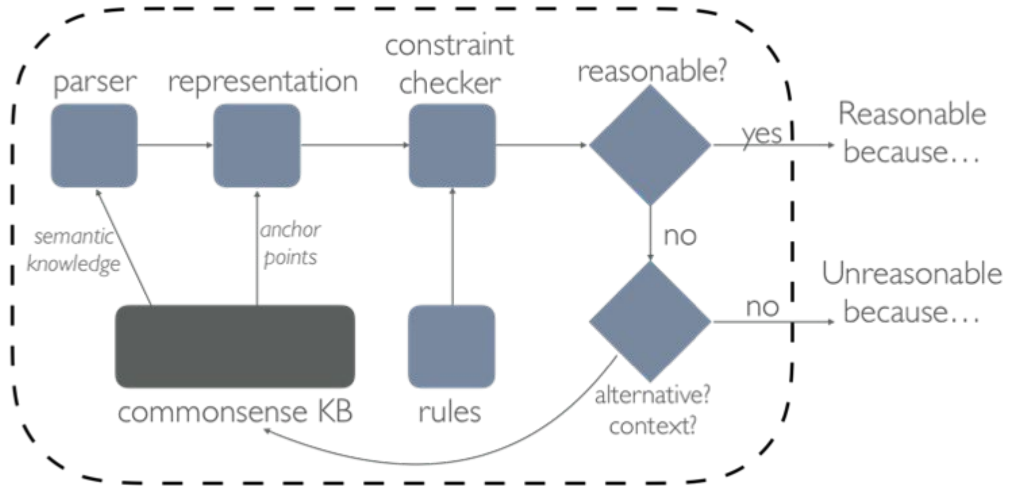
Vision: Articulate Systems that can Coherently Communicate to Resolve Issues

With Other Systems

Learning system



Symbolic system



Common language to complete tasks.

- Redundancy: systems solve problems in multiple ways.
- Hybrid processes: systems that learn from each other.

With Humans

humans

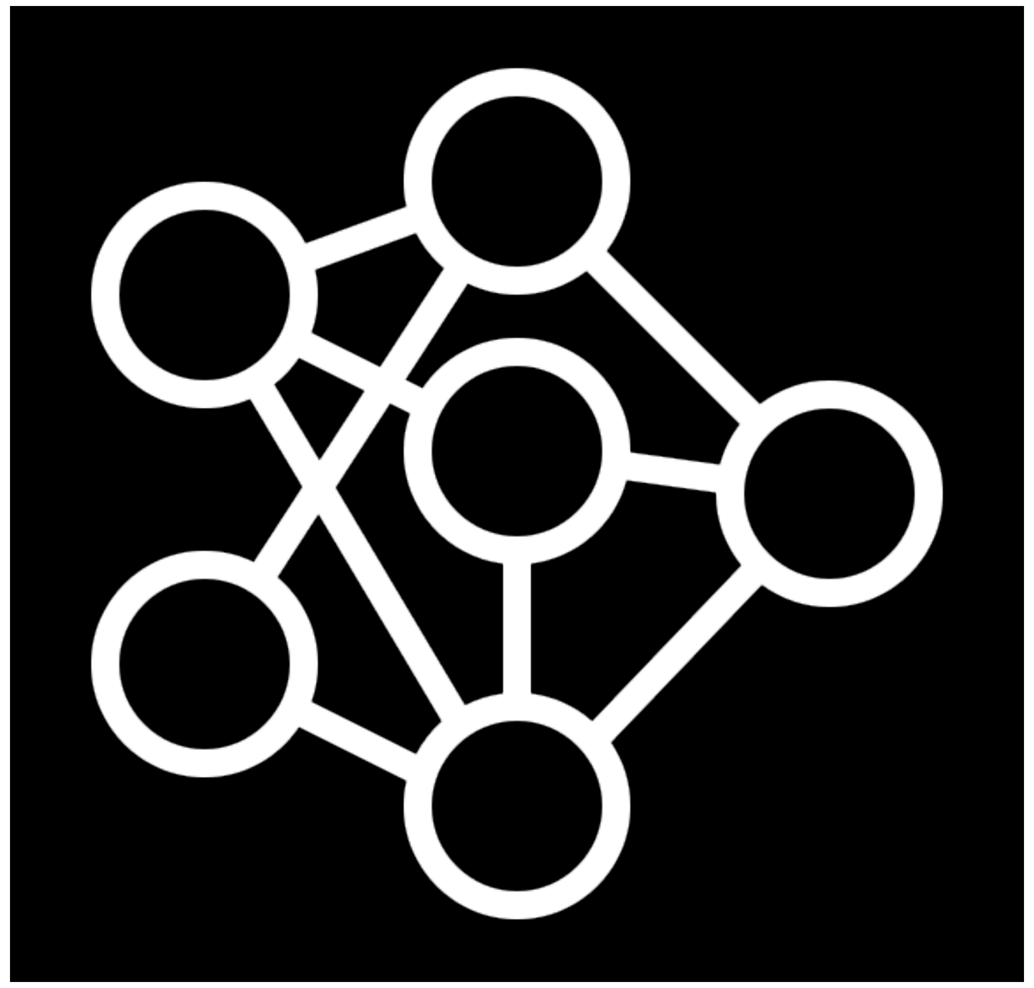
complex system

Explanations are a debugging language.

- Debugging: humans can improve complex systems
- Education: complex systems can “improve” or teach humans.

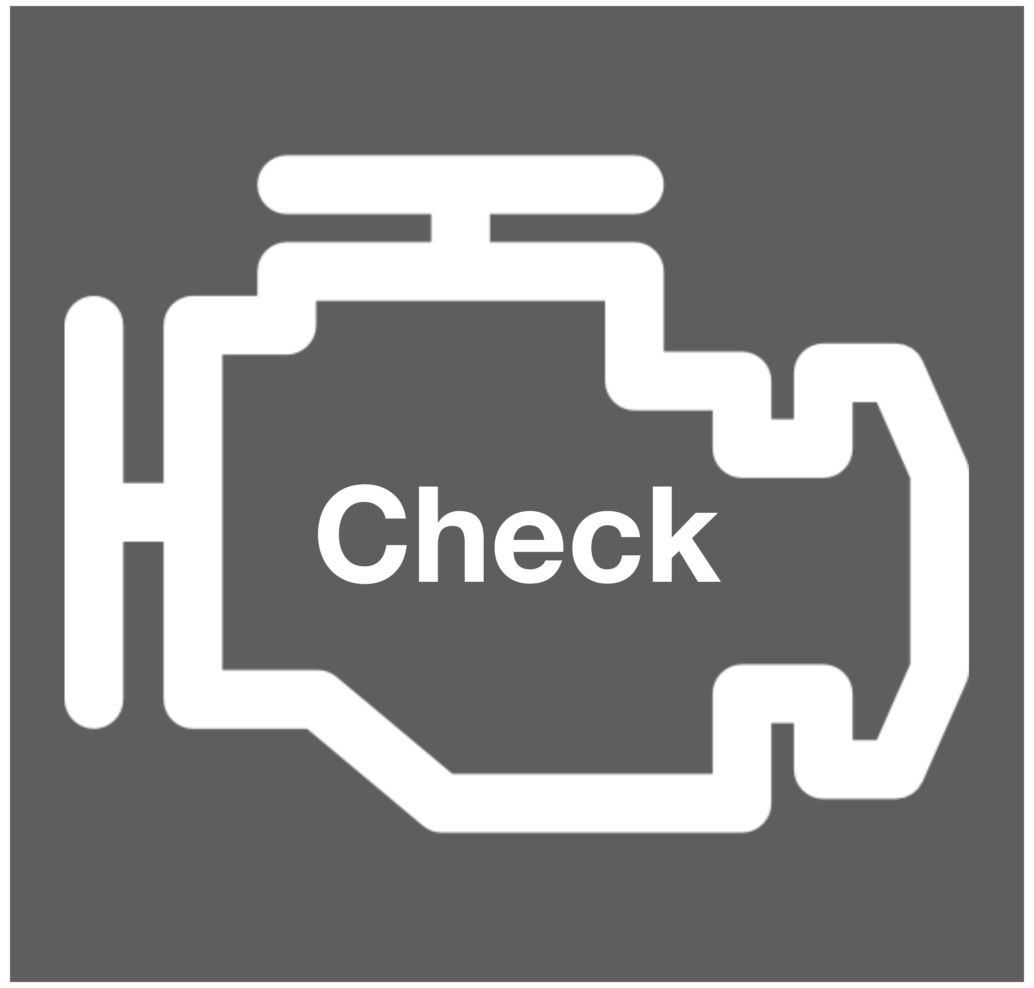
How can we leverage Explanations for Anomaly Detection

Black-box



Decisions supported with commonsense.

Imprecise



Localize errors with reasons.

System-level

Common language for debugging.

Domain: Self driving cars

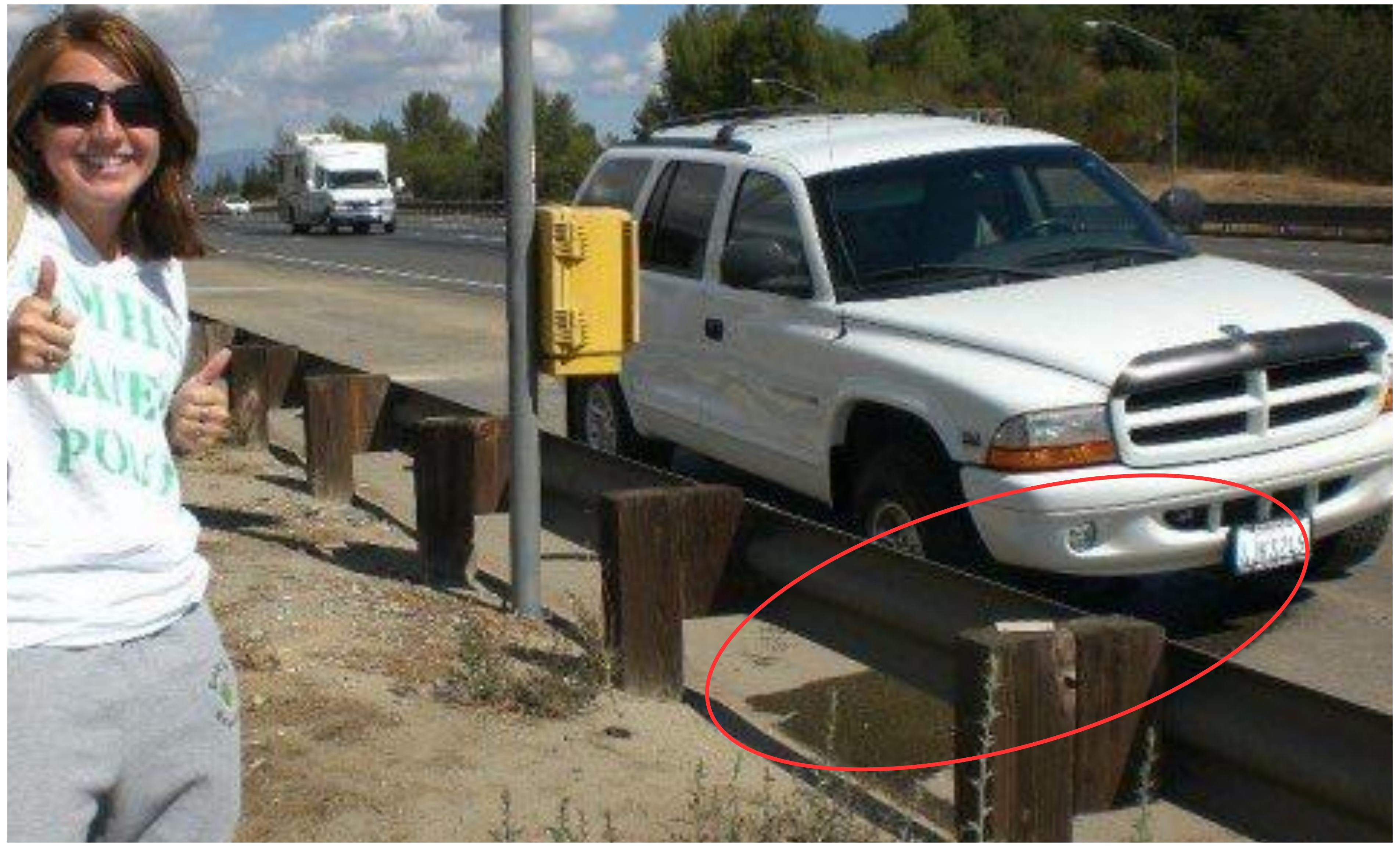
EXCLUSIVE AUTONOMOUS VEHICLES UBER/LYFT

Uber Finds Deadly Accident Likely Caused By Software Set to Ignore Objects On Road

By Amir Efrati May 07, 2018 9:48 AM PDT · Comments by Noah David, Michael D. Geer and 4 others

Subscribe now

Failure of Complex Systems

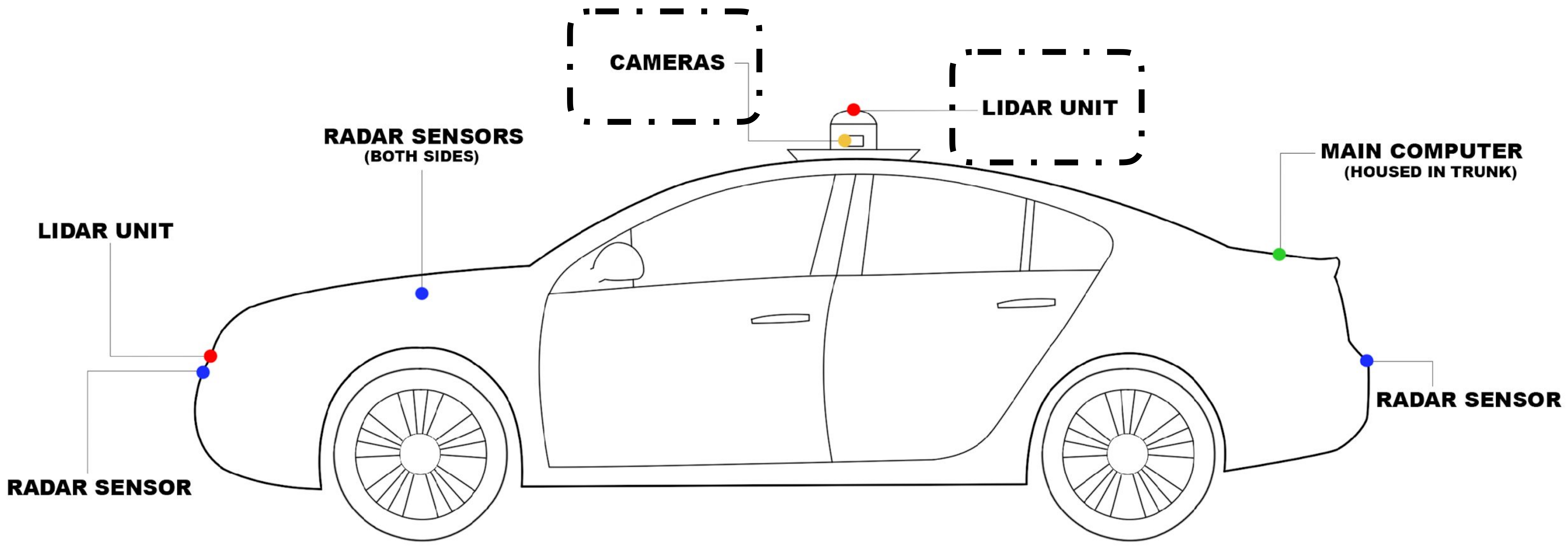



AI Mistakes Bus-Side Ad for Famous CEO, Charges Her With Jaywalking

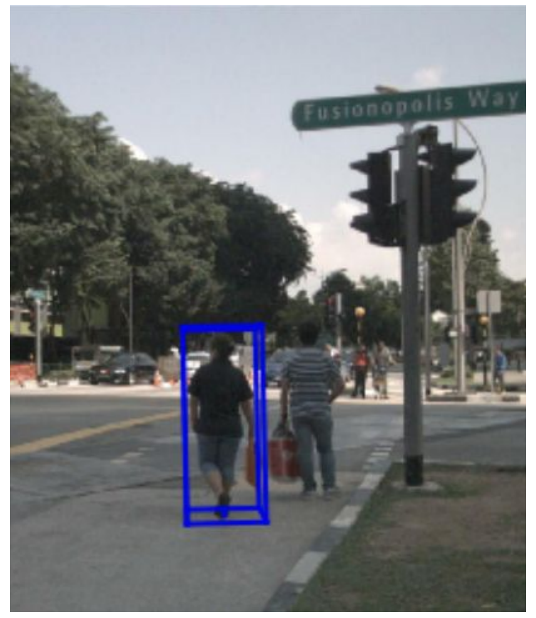
By Tang Ziyi / Nov 22, 2018 04:17 PM / Society & Culture

Complex Systems Fail in Two Ways

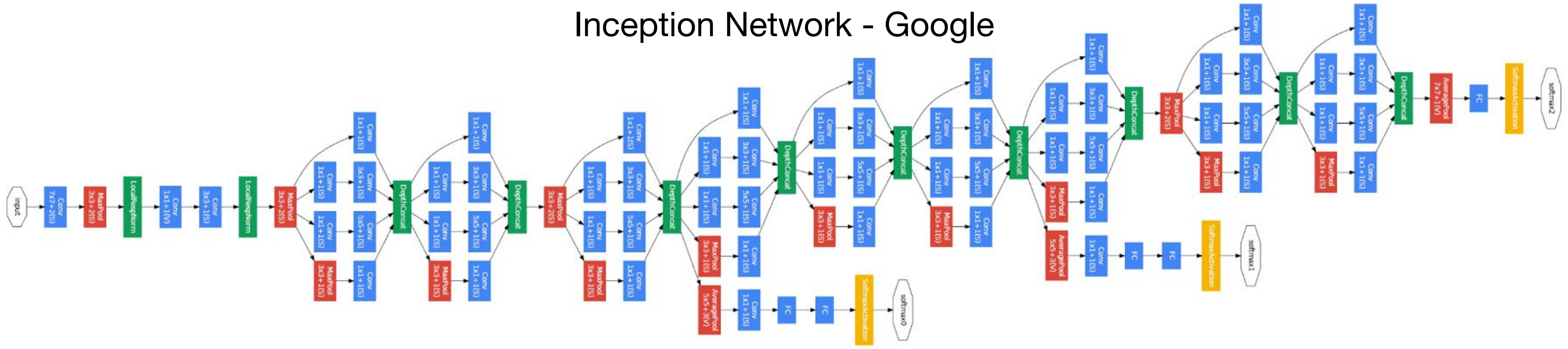
1. Failure *local* to a specific subsystem.
2. A failed *cooperation* amongst subsystems.



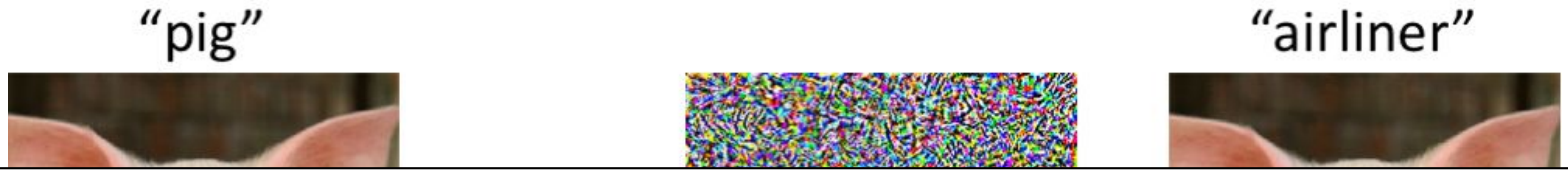
Local Problem: Neural Networks are Brittle and Biased



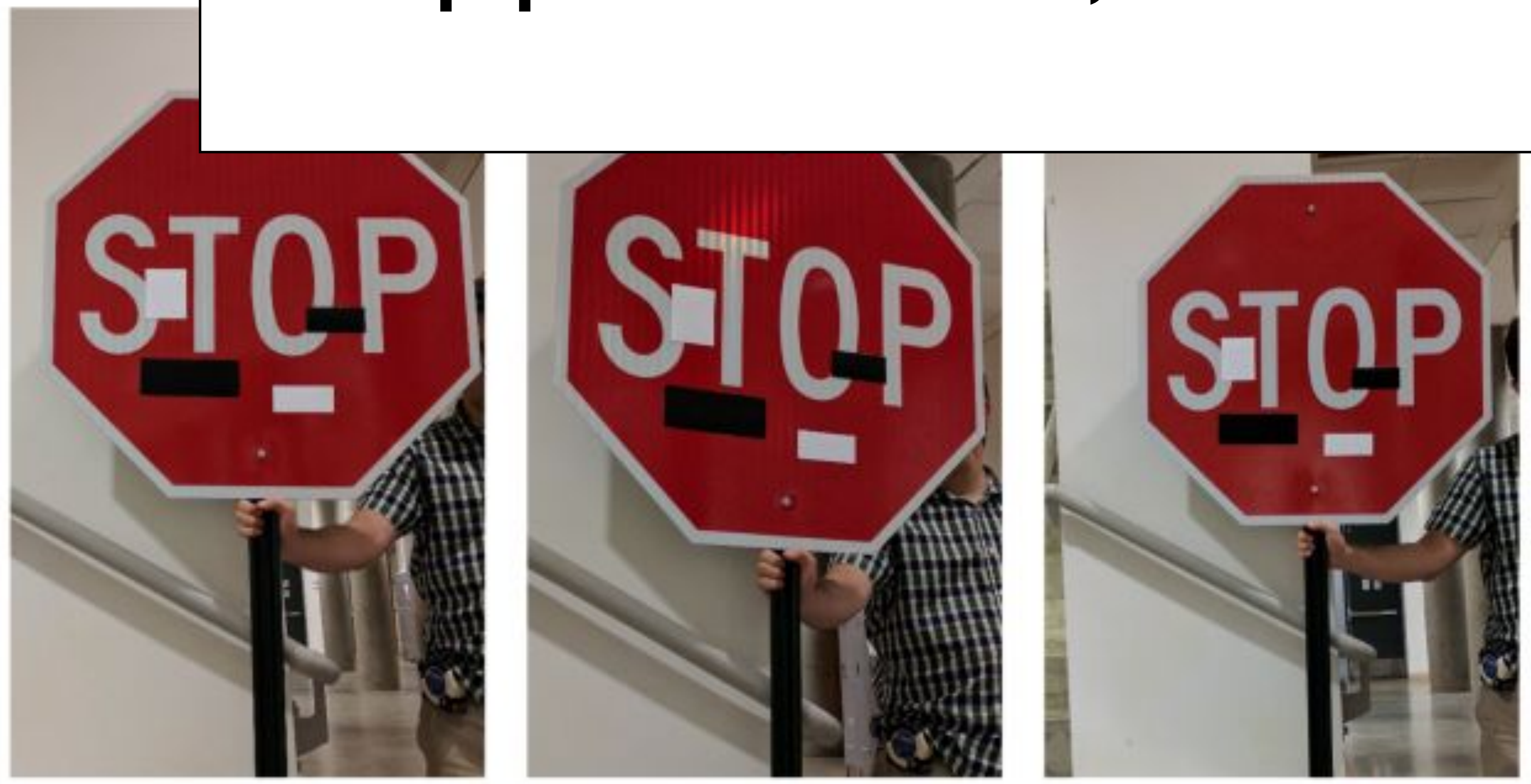
Inception Network - Google

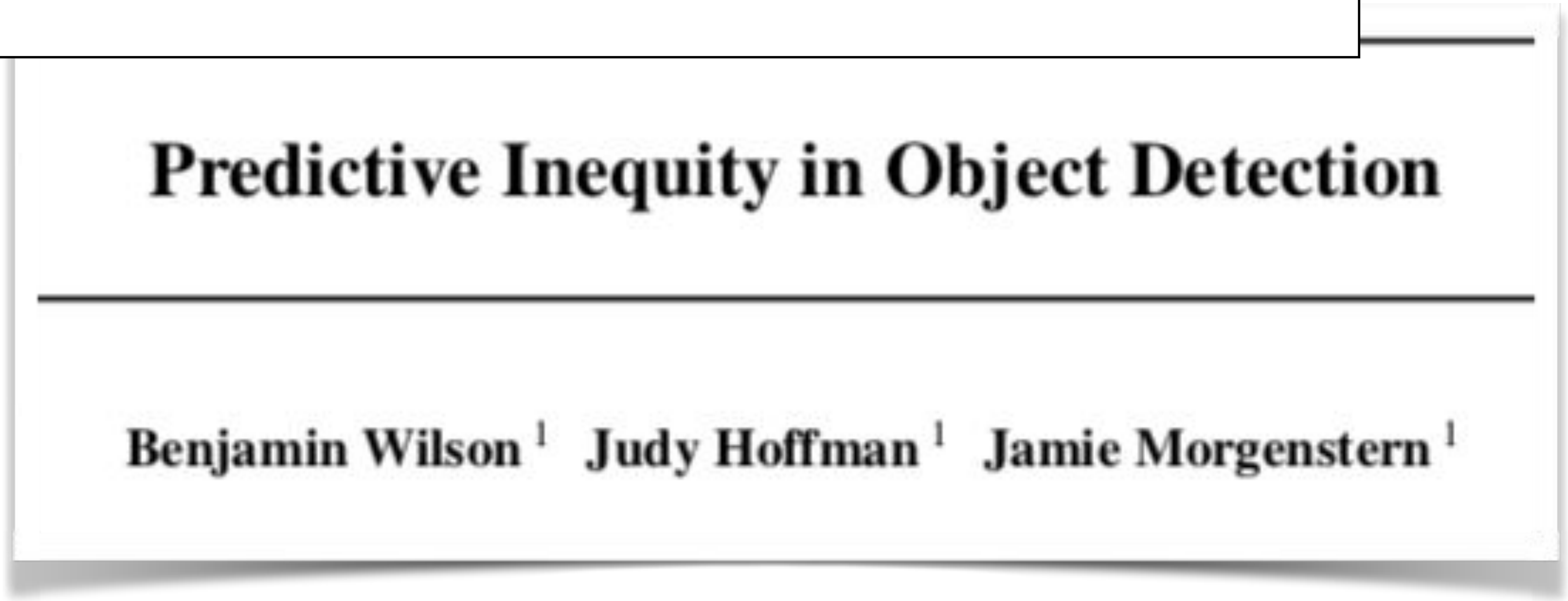


Label
e.g. pedestrian



For self-driving, and other mission-critical, safety-critical applications, these mistakes have CONSEQUENCES.





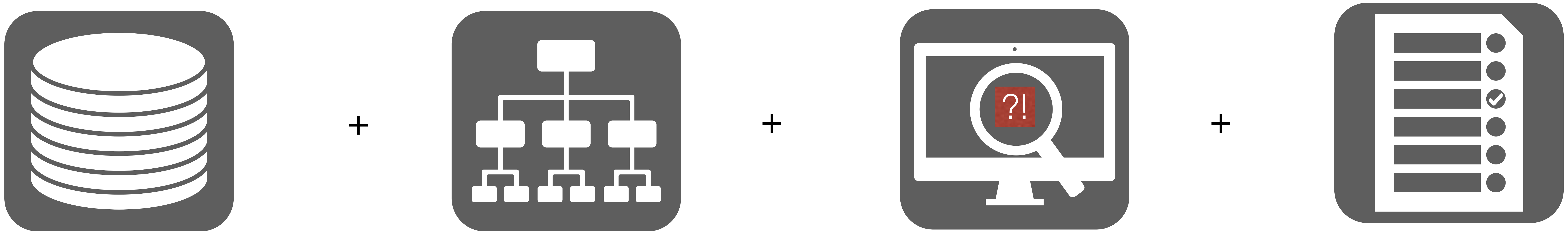
Predictive Inequity in Object Detection

Benjamin Wilson [1] Judy Hoffman [1] Jamie Morgenstern [1]

K. Eykholt et al. “Robust Physical-World Attacks on Deep Learning Visual Classification.”

Monitor Opaque Subsystems for Reasonableness

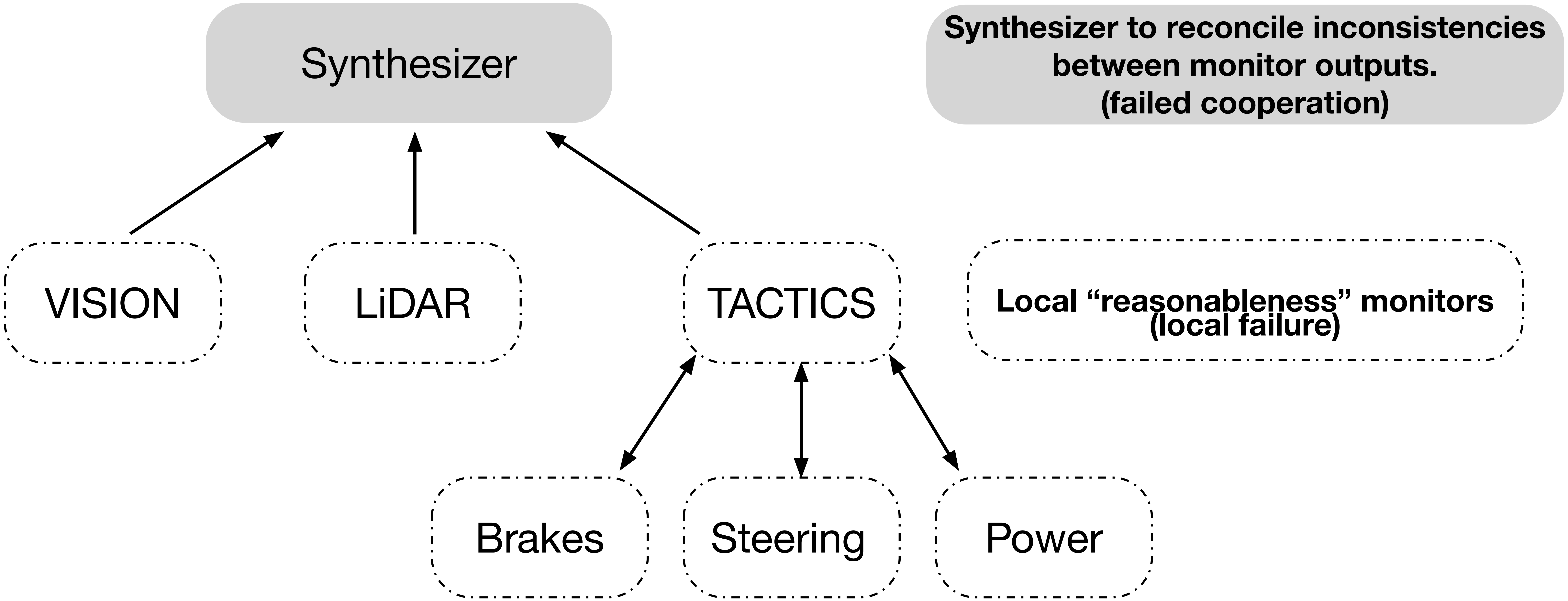
Opaque
Mechanism



Commonsense Knowledge Base + Flexible Representation + Identify (Un)reasonability + Justify (Un)reasonability

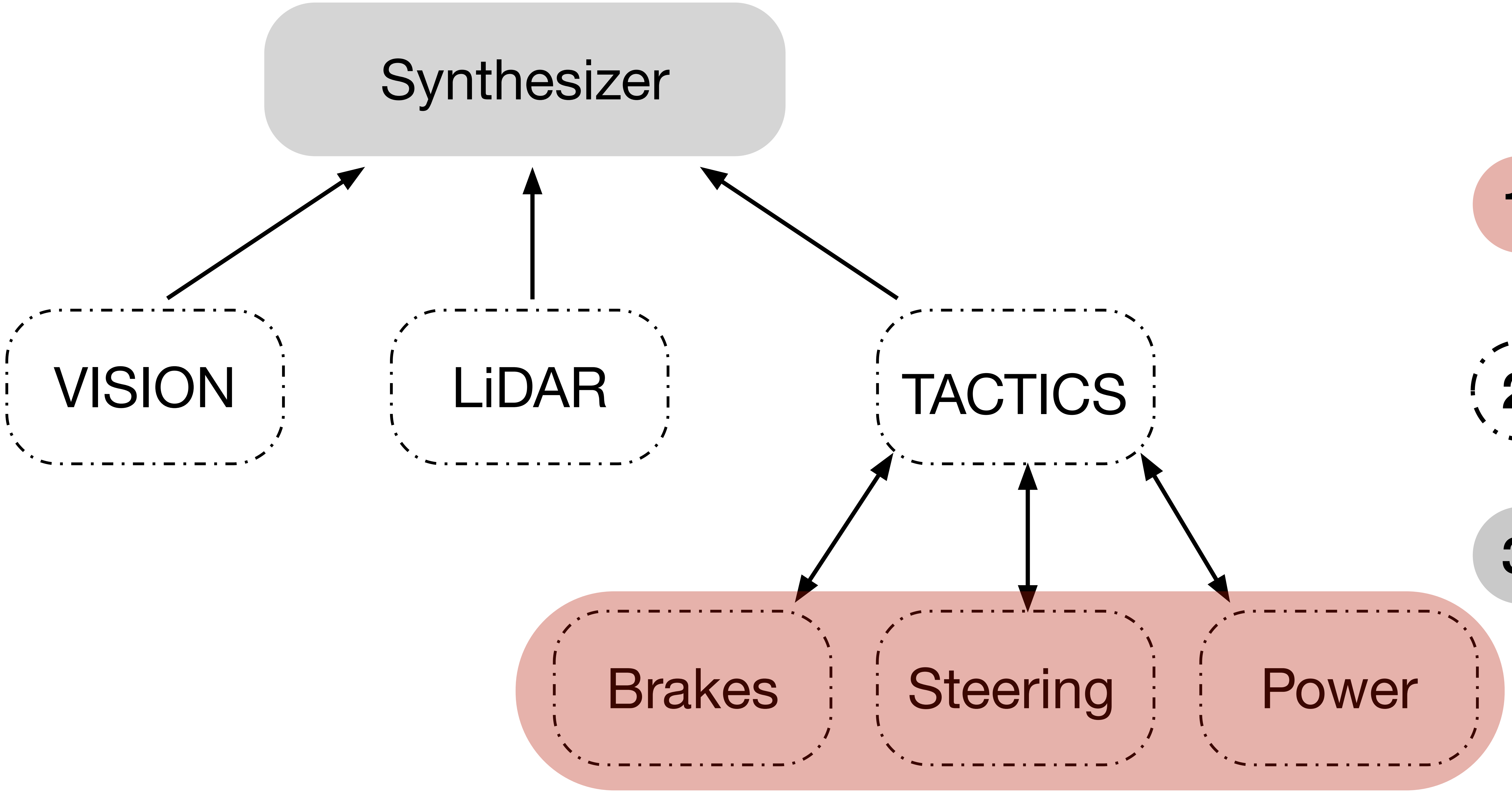
1. Judgement of reasonableness
2. Justification of reasonableness

System Architecture for Self-Driving Cars



L.H. Gilpin. Explaining possible futures for robust autonomous decision-making. Proceedings of the AAAI Fall Symposium on Anticipatory Thinking, 2019.

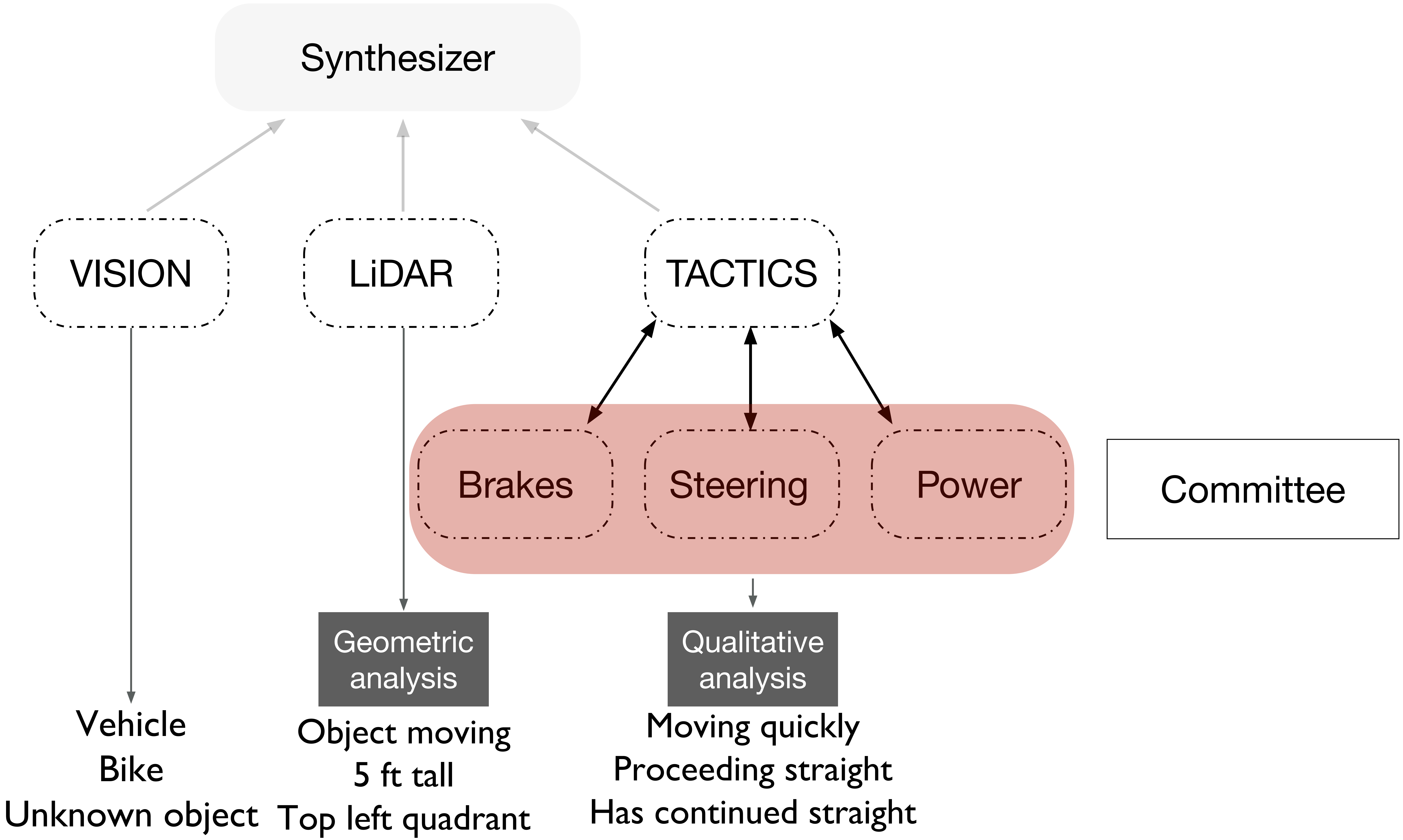
Anomaly Detection Through Explanations in Three Steps



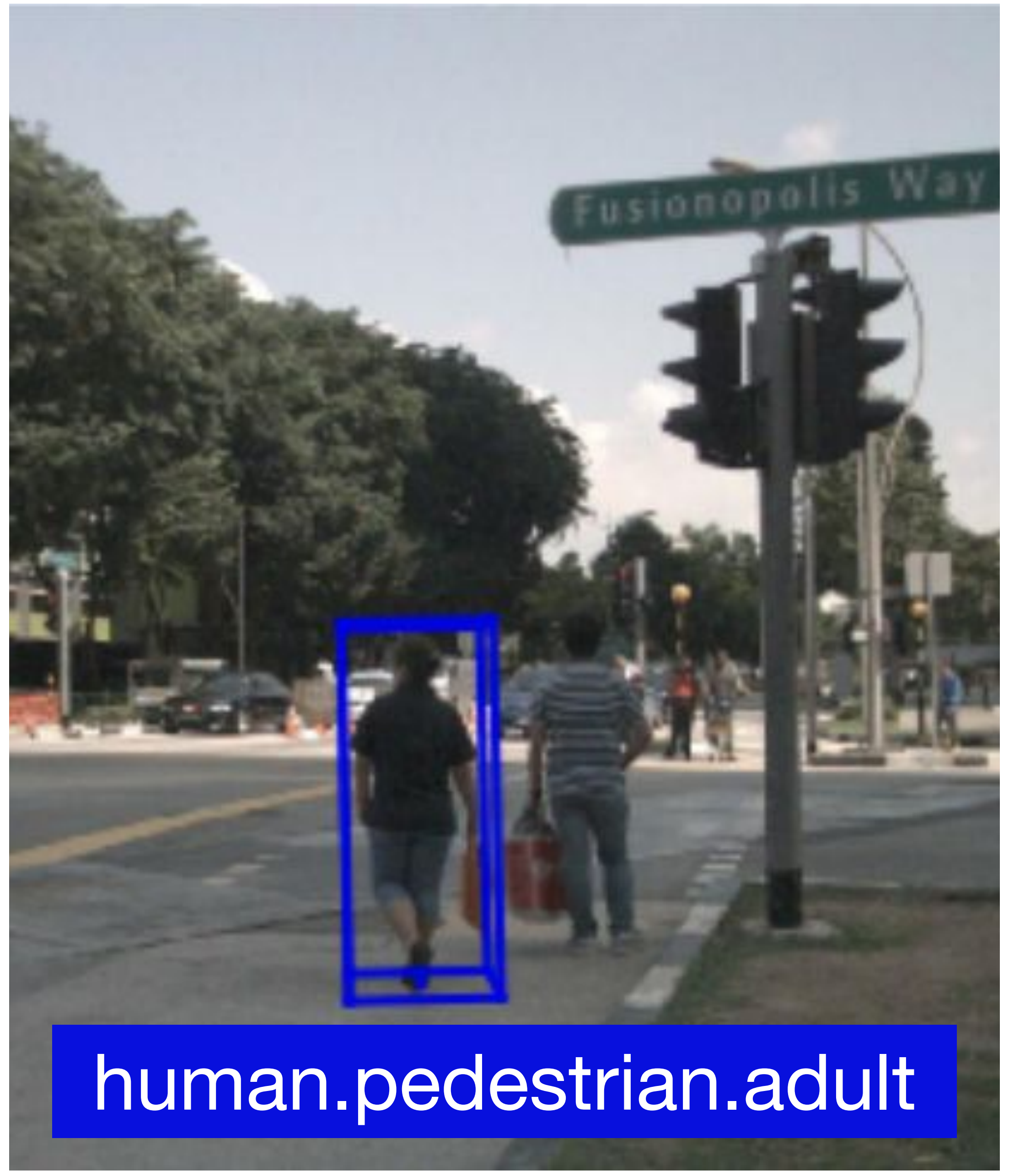
1. Generate Symbolic Qualitative Descriptions for each committee.
2. Input qualitative descriptions into local “reasonableness” monitors.
3. Use a synthesizer to reconcile inconsistencies between monitors.

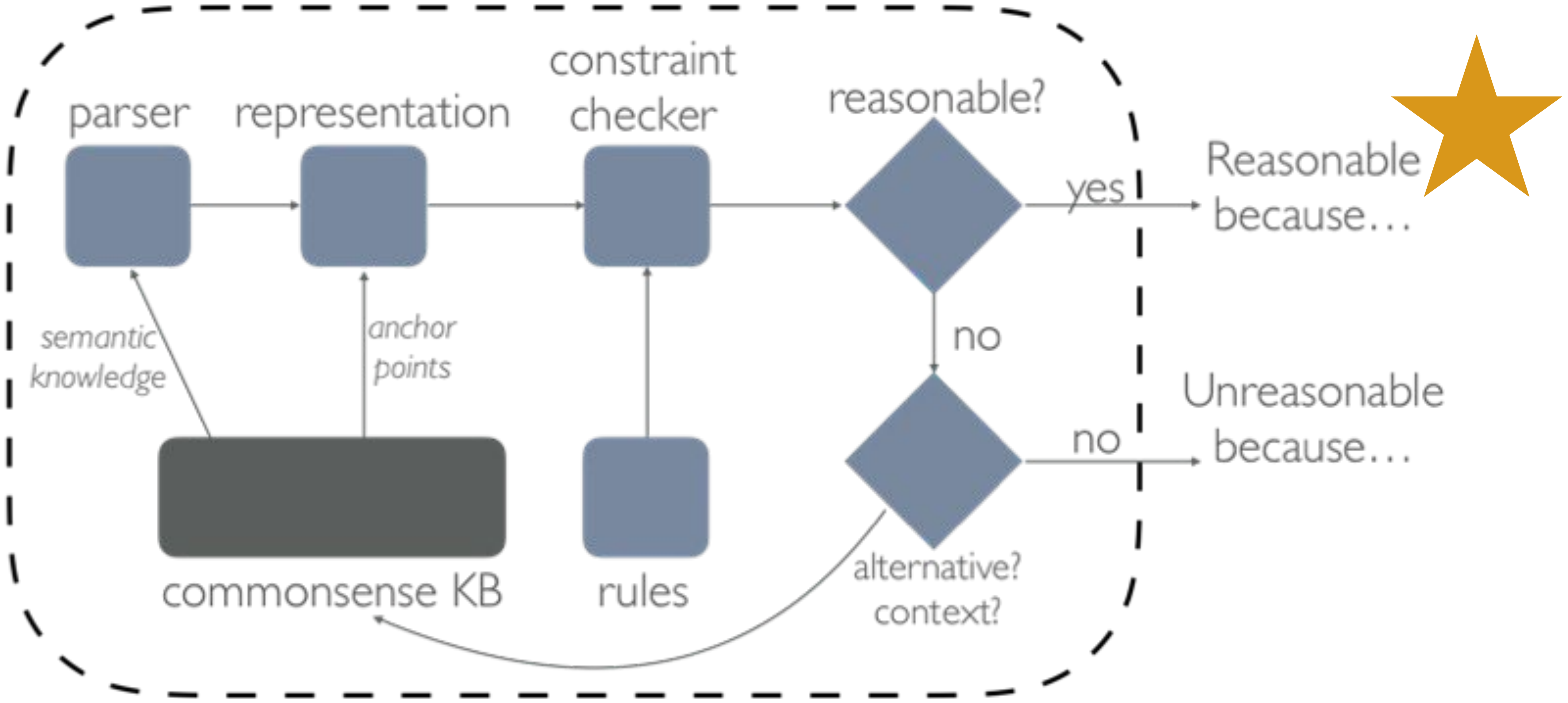
L.H. Gilpin, V. Penubarthi, L. Kagal. “Anomaly Detection through Explanations.” To be submitted.

1. Generate Symbolic Qualitative Descriptions for each committee.



Local Monitoring

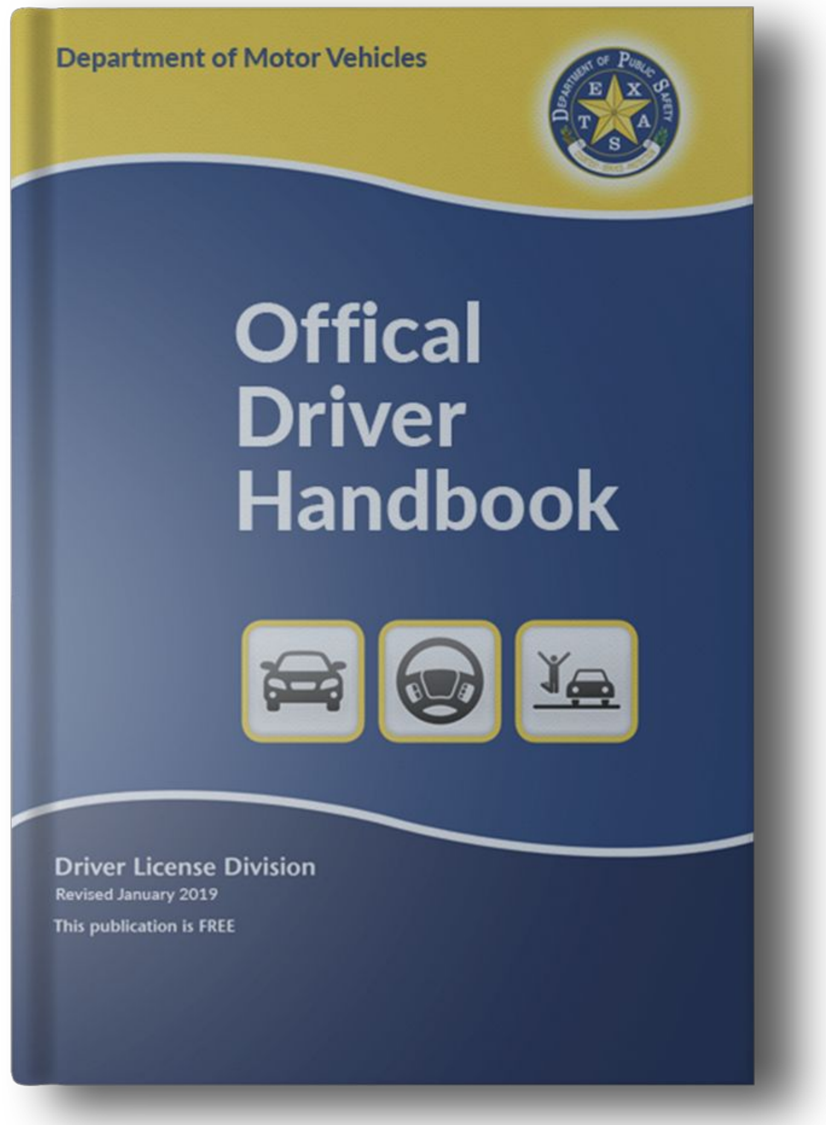




This perception is reasonable. An adult is typically a large person. They are usually located walking on the street. Its approximate dimensions of [0.621, 0.669, 1.642] is approximately the correct size in meters.

Start with Baseline Rules

Identify (Un)reasonability



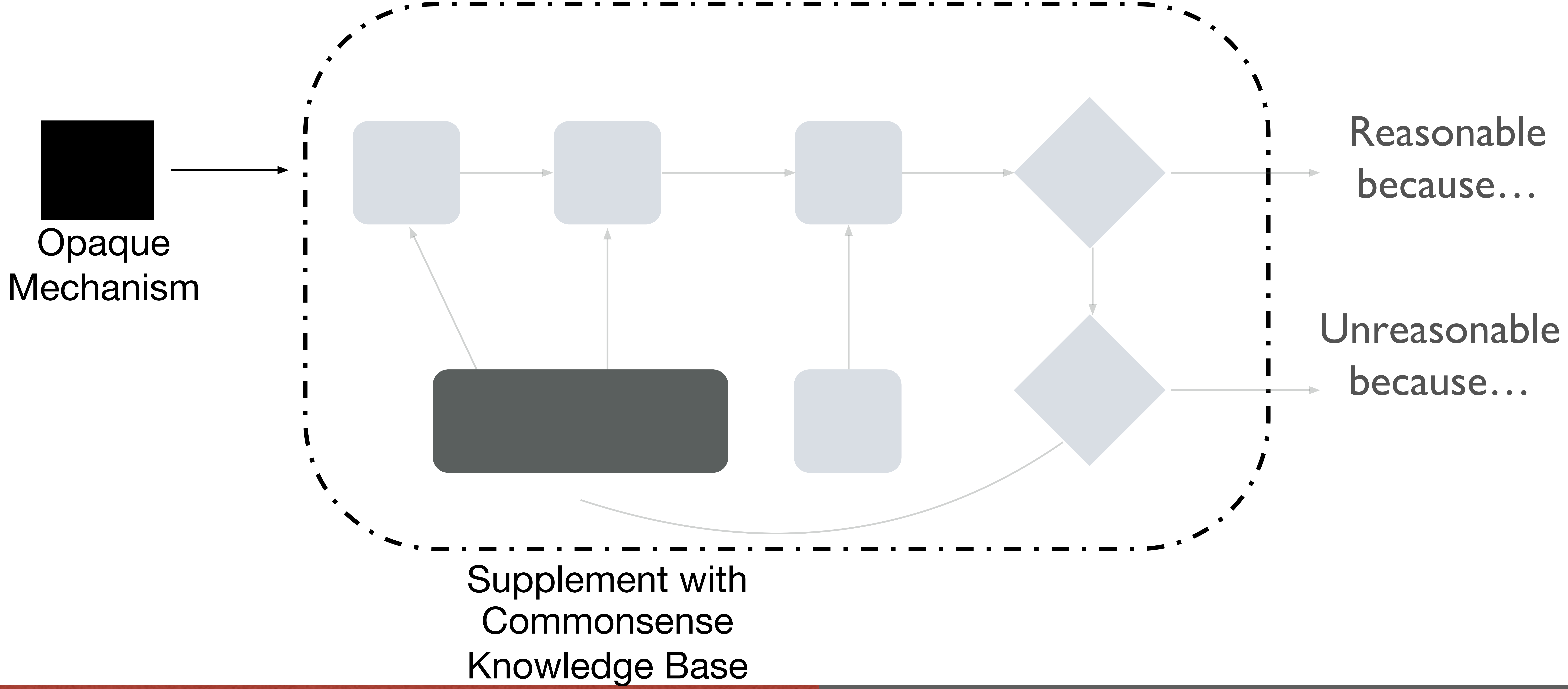

```
:safe_car_policy a air:Policy;
        air:rule :light-rule;
        air:rule :pedestrian-rule;
        air:rile :speed-rule;
        rdfs:comment "Safe driving tactics";
        rdfs:label "Safe driving tactics by the state of MA."

:pedestrian-rule a air:Belif-rule;
        rdfs:comment "Ensure that pedestrians are safe.";
        air:if {
            :EVENT a :V;
                car_ont:InPathOf :V.
        };
        air:then [
            air:description ("There is a pedestrian");
            air:assert [air:statement{:Event
                    air:compliant-with :safe_car_policy .}]] .
        air:else [
            air:description ("There is not a pedestrian");
            air:assert [air:statement{:Event
                    air:non-compliant-with :safe_car_policy .}]] .
```

+ reasoner

http://dig.csail.mit.edu/2009/AIR/

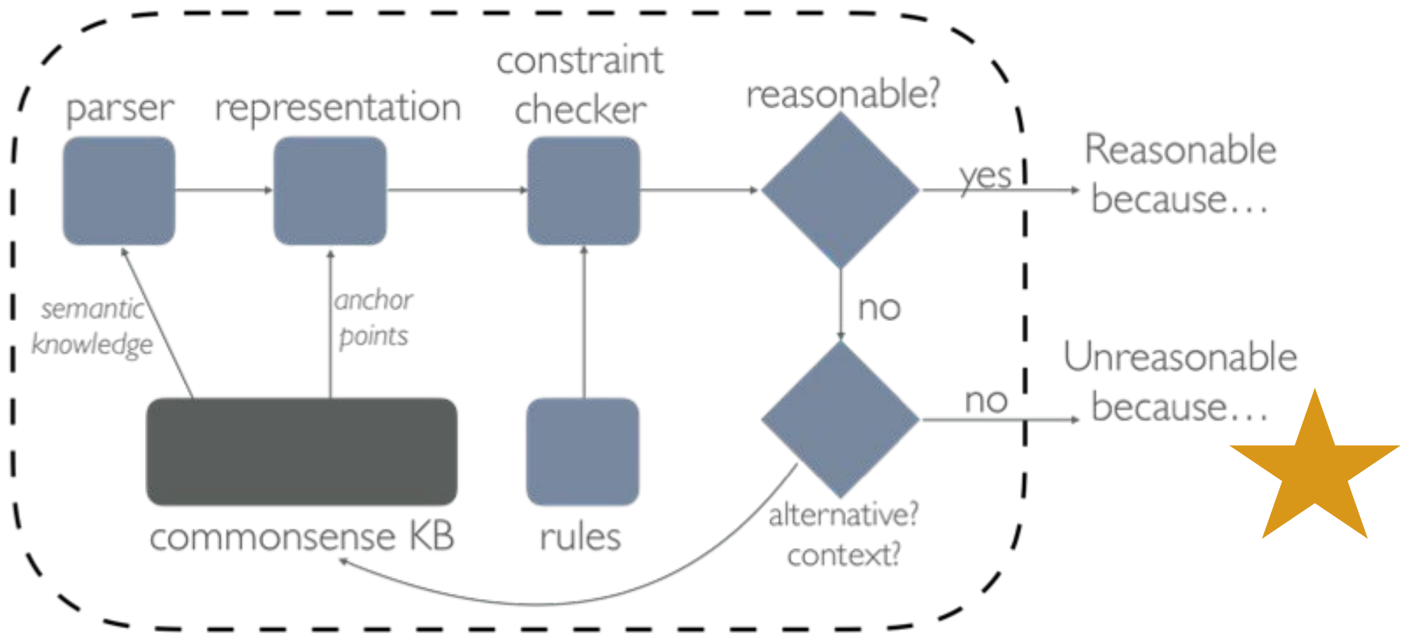
L.H. Gilpin and L. Kagal. "An Adaptable Self-Monitoring Framework for Opaque Machines." AAMAS 2019.

Semantic Knowledge Bases Provide Commonsense



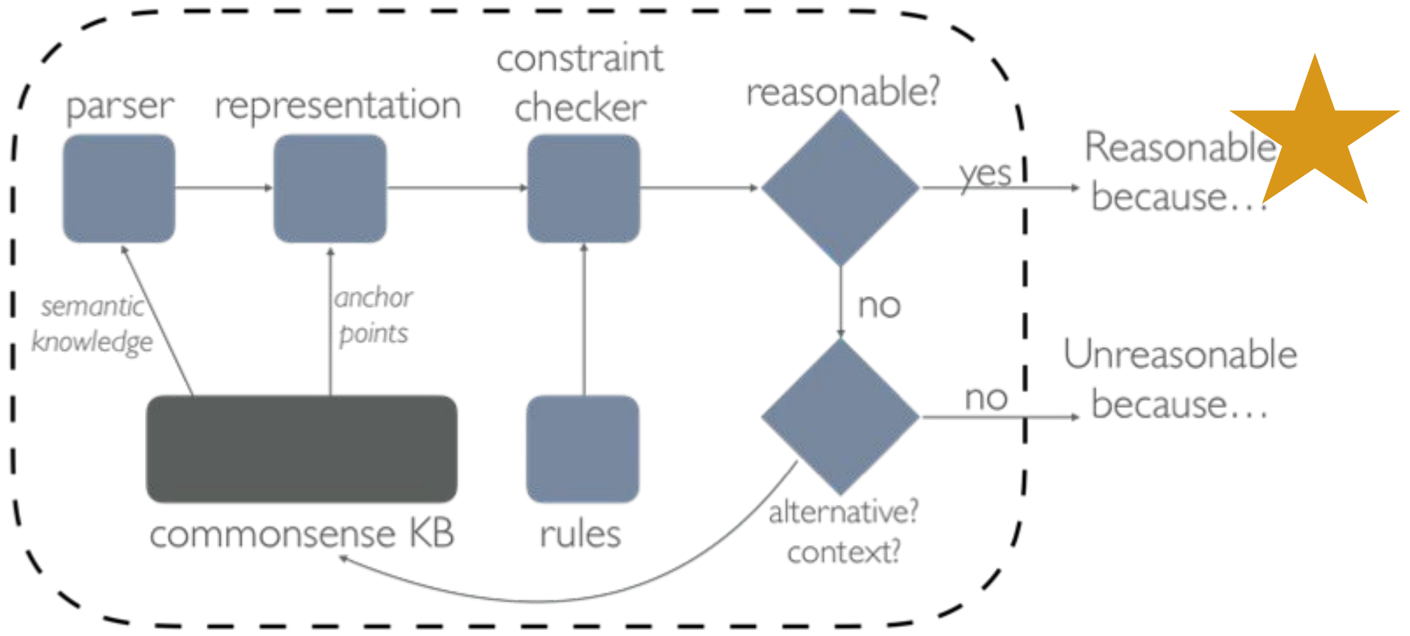
2. Input qualitative descriptions into local "reasonableness" monitors.

Vehicle
Bike
Unknown object



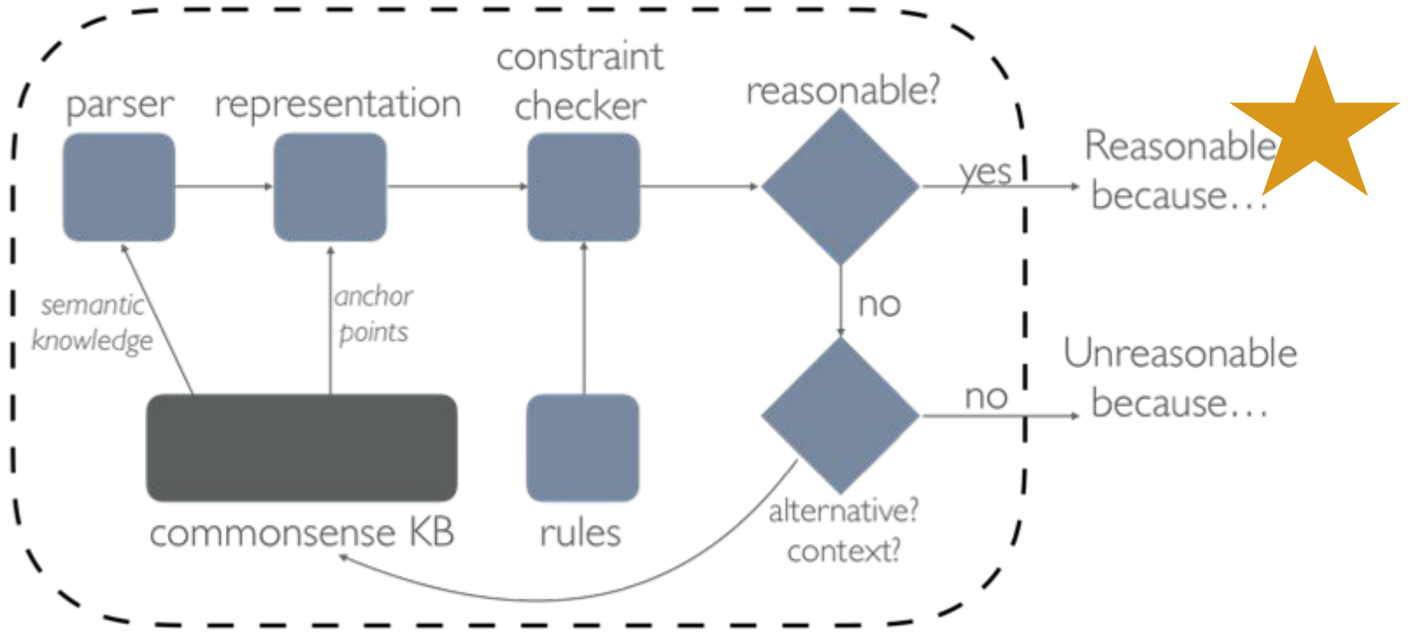
This vision perception is unreasonable. There is no commonsense data supporting the similarity between a vehicle, bike and unknown object except that they can be located at the same location. This component should be ignored.

Object moving
5 ft tall
Top left quadrant



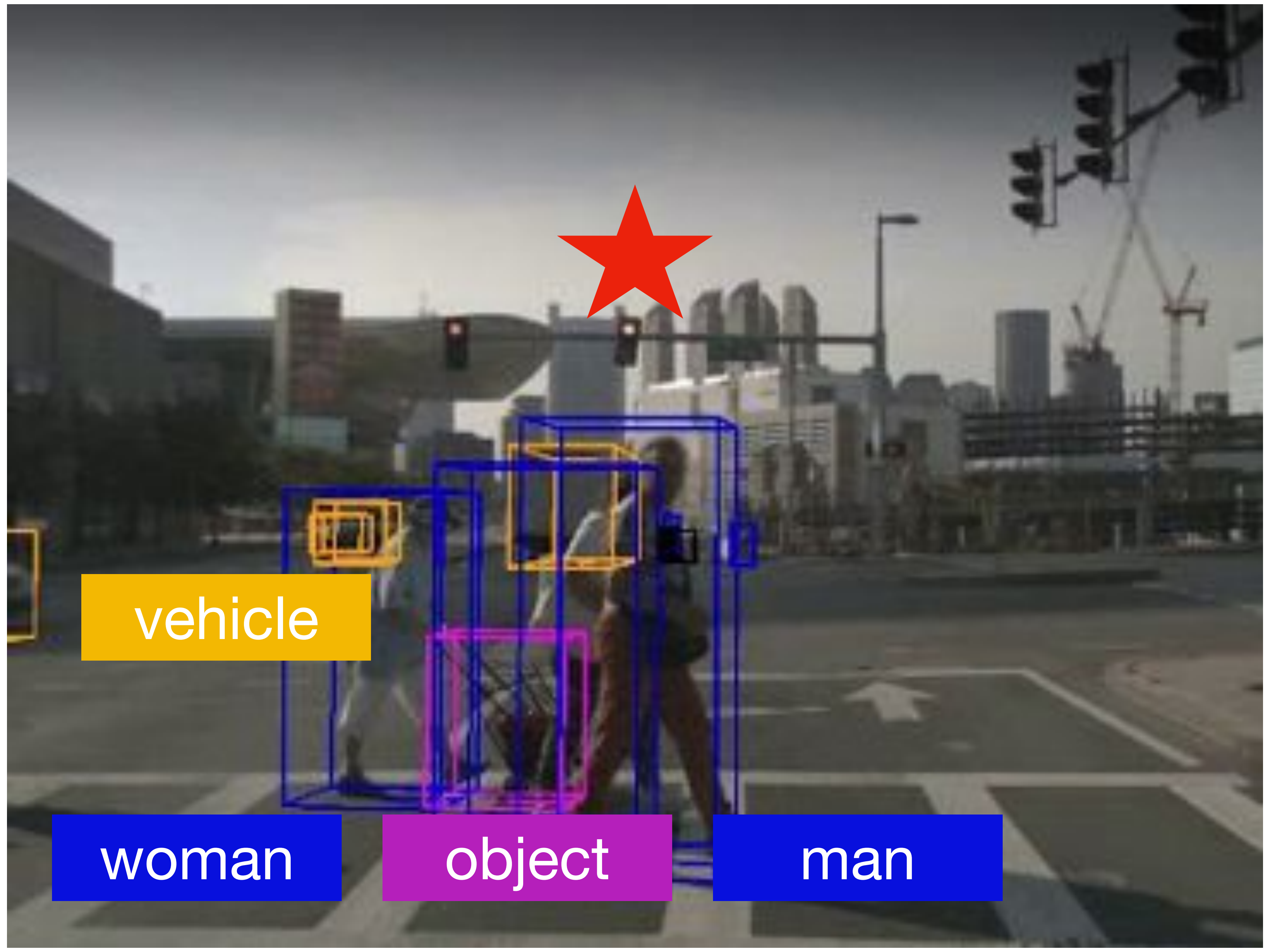
This lidar perception is reasonable. An object moving of this size is a large moving object that should be avoided.

Moving quickly
Proceeding straight
Has continued straight



This system state is reasonable given that the vehicle has been moving quickly and proceeding straight for the last 10 second history.

Flexible Representation with Implicit Reasonableness Rules



Data from Nuscenes

actor

woman

man

object

direction

```
@prefix foo: <http://foo#>.
@prefix car_ont: <http://car_ont#>.

foo:my_car
    a  car_ont:Vehicle ;
    car_ont:LastState "stop" ;
    car_ont:CurrentState "stop" ;
    car_ont:direction foo:some_traffic_light .

foo:some_pedestrians
    a car_ont:Pedestrian ;
    car_ont:label woman ;
    car_ont:CurrentState "move" ;
    car_ont:propel foo:woman-object ;
    car_ont:InPathOf foo:my_car .

    a car_ont:Pedestrian ;
    car_ont:label man ;
    car_ont:CurrentState "move" ;
    car_ont:NextTo foo:woman-object ;
    car_ont:InPathOf foo:my_car .

foo:woman-object
    a car_ont:Object ;
    car_ont:CurrentState "propel" ;
    car_ont:InPathOf foo:my_car .

foo:some_traffic_light
    a car_ont:TrafficLight ;
    car_ont:LightColor “red" .
```

L.H. Gilpin and L. Kagal. “An Adaptable Self-Monitoring Framework for Opaque Machines.” AAMAS 2019.

Symbolic reasons

3. Use a synthesizer to reconcile inconsistencies between monitors.

```
(monitor, judgement, unreasonable)
(input, isType, labels)
(all_labels, inconsistent, negRel)
(isA, hasProperty, negRel)
…
(all_labels, notProperty, nearMiss)
(all_labels, locatedAt, consistent)
(monitor, recommend, discount)
```

```
(monitor, judgement, reasonable)
(input_data, isType, sensor)
…
(input_data[4], hasSize, large)
(input_data[4], IsA, large_object)
(input_data[4], moving, True)
(input_data[4], hasProperty, avoid)
…
```

```
(monitor, judgement, reasonable)
(input, isType, history)
(input_data, moving, True)
(input_data, direction, forward)
(input_data, speed, fast)
(input_data, consistent, True)
(monitor, recommend, proceed)
```

→ Synthesizer →

The best option is to veer and slow down. The vehicle is traveling too fast to suddenly stop. The vision system is inconsistent, but the lidar system has provided a reasonable and strong claim to avoid the object moving across the street.

3. Use a synthesizer to reconcile inconsistencies between monitors.

Synthesizer + Priority Hierarchy → Abstract Goals

- Explanation synthesizer to deal with *inconsistencies.*
  - Argument tree.
  - Queried for support or counterfactuals.

1. Passenger Safety
2. Passenger Perceived Safety
3. Passenger Comfort
4. Efficiency (e.g. Route efficiency)

→

A passenger is safe if:

- The vehicle proceeds at the same speed and direction.
- The vehicle avoids threatening objects.


Synthesizer to reconcile inconsistencies between monitor outputs.
Synthesizer
The best option is to veer and slow down. The vehicle is traveling too fast to suddenly stop. The vision system is inconsistent, but the lidar system has provided a reasonable and strong claim to avoid the object moving across the street.
VISION
LiDAR
TACTICS
(monitor, judgement, unreasonable)
(input, isType, labels)
(all_labels, inconsistent, negRel)
(isA, hasProperty, negRel)
…
(all_labels, notProperty, nearMiss)
(all_labels, locatedAt, consistent)
(monitor, recommend, ignore)
(monitor, judgement, reasonable)
(input_data, isType, sensor)
…
(input_data[4], hasSize, large)
(input_data[4], IsA, large_object)
(input_data[4], moving, True)
(input_data[4], hasProperty, avoid)
…
(monitor, recommend, avoid)
(monitor, judgement, reasonable)
(input, isType, history)
(input_data, moving, True)
(input_data, direction, forward)
(input_data, speed, fast)
(input_data, consistent, True)
(monitor, recommend, proceed)

Framework for Real World Error Detection

- End-to-end prototype
  - Machine perception
  - Represented with frame-based primitives (Schank conceptual dependency primitives).

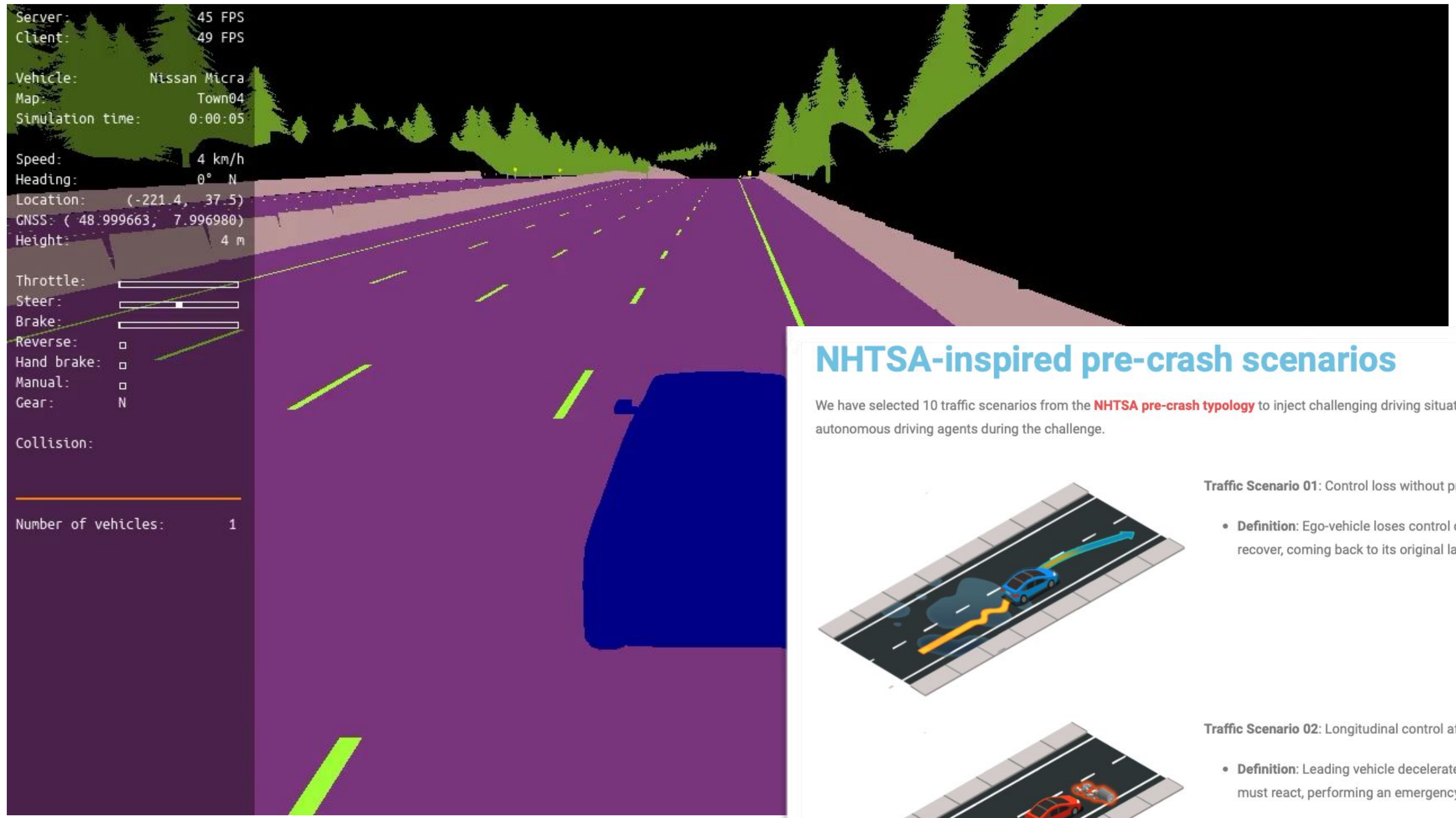
L.H. Gilpin, J.C. Macbeth and E. Florentine. "Monitoring scene understanders with conceptual primitive decomposition and commonsense knowledge." ACS 2018.

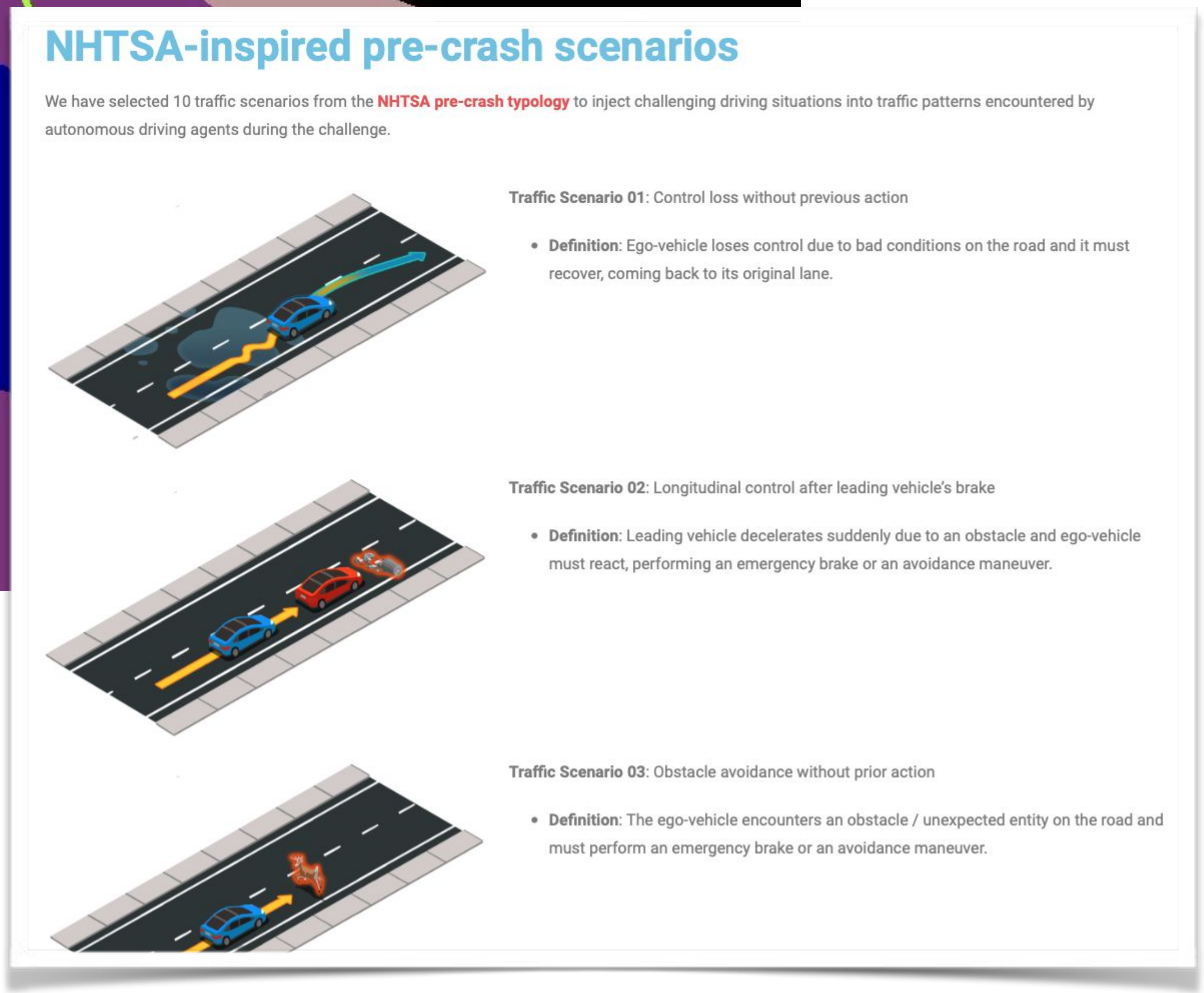
- Generalized framework
  - Reusable web standards
  - Extended primitive representations to apply to multiple applications.

L.H. Gilpin and L. Kagal. "An Adaptable Self-Monitoring Framework for Opaque Machines." AAMAS 2019.

System Evaluation

Carla Simulations - real-world inspired scenarios





NHTSA-inspired pre-crash scenarios

We have selected 10 traffic scenarios from the **NHTSA pre-crash typology** to inject challenging driving situations into traffic patterns encountered by autonomous driving agents during the challenge.

Traffic Scenario 01: Control loss without previous action

- **Definition**: Ego-vehicle loses control due to bad conditions on the road and it must recover, coming back to its original lane.

Traffic Scenario 02: Longitudinal control after leading vehicle's brake

- **Definition**: Leading vehicle decelerates suddenly due to an obstacle and ego-vehicle must react, performing an emergency brake or an avoidance maneuver.

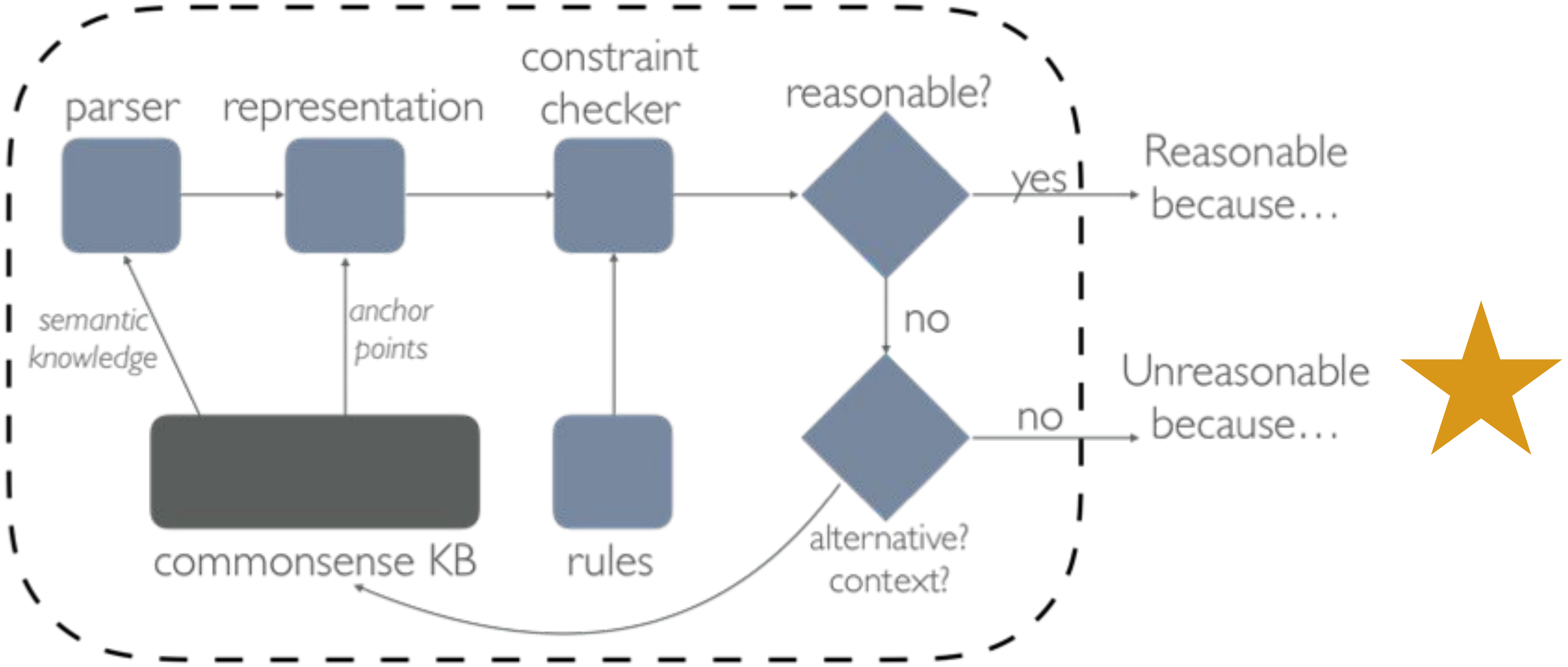
Traffic Scenario 03: Obstacle avoidance without prior action

- **Definition**: The ego-vehicle encounters an obstacle / unexpected entity on the road and must perform an emergency brake or an avoidance maneuver.

NuScenes dataset

- <u>Detection</u>: Generate logs from scenarios to detect failures.
- <u>Invoke errors</u>: Scrambling *multiple* labels on existing datasets.
- <u>Real errors</u>: Examining errors on the validation dataset of NuScenes leaderboard.

Invoking and Validating Errors



This perception is unreasonable. The movable_object.trafficcone located in the center region is not a reasonable size: it is too tall. There is no commonsense supporting this judgement. Discounting objects detected in the same region.

Evaluating the UBER accident

```
(monitor, judgement, unreasonable)
(input, isType, labels)
(all_labels, inconsistent, negRel)


…
(all_labels, notProperty, nearMiss)
(all_labels, locatedAt, consistent)
(monitor, recommend, ignore)
```

```
(monitor, judgement, reasonable)
(input, isType, sensor)
…
(input_data[4], hasSize, large)
(input_data[4], IsA, large_object)
(input_data[4], moving, True)
(input_data[4], hasProperty, avoid)
…
(monitor, recommend, avoid)
```

!

```
(monitor, judgement, reasonable)
(input, isType, history)
(input_data, moving, True)
(input_data, direction, forward)
(input_data, speed, fast)
(input_data, consistent, True)
(monitor, recommend, proceed)
```

Abstract Goal Tree

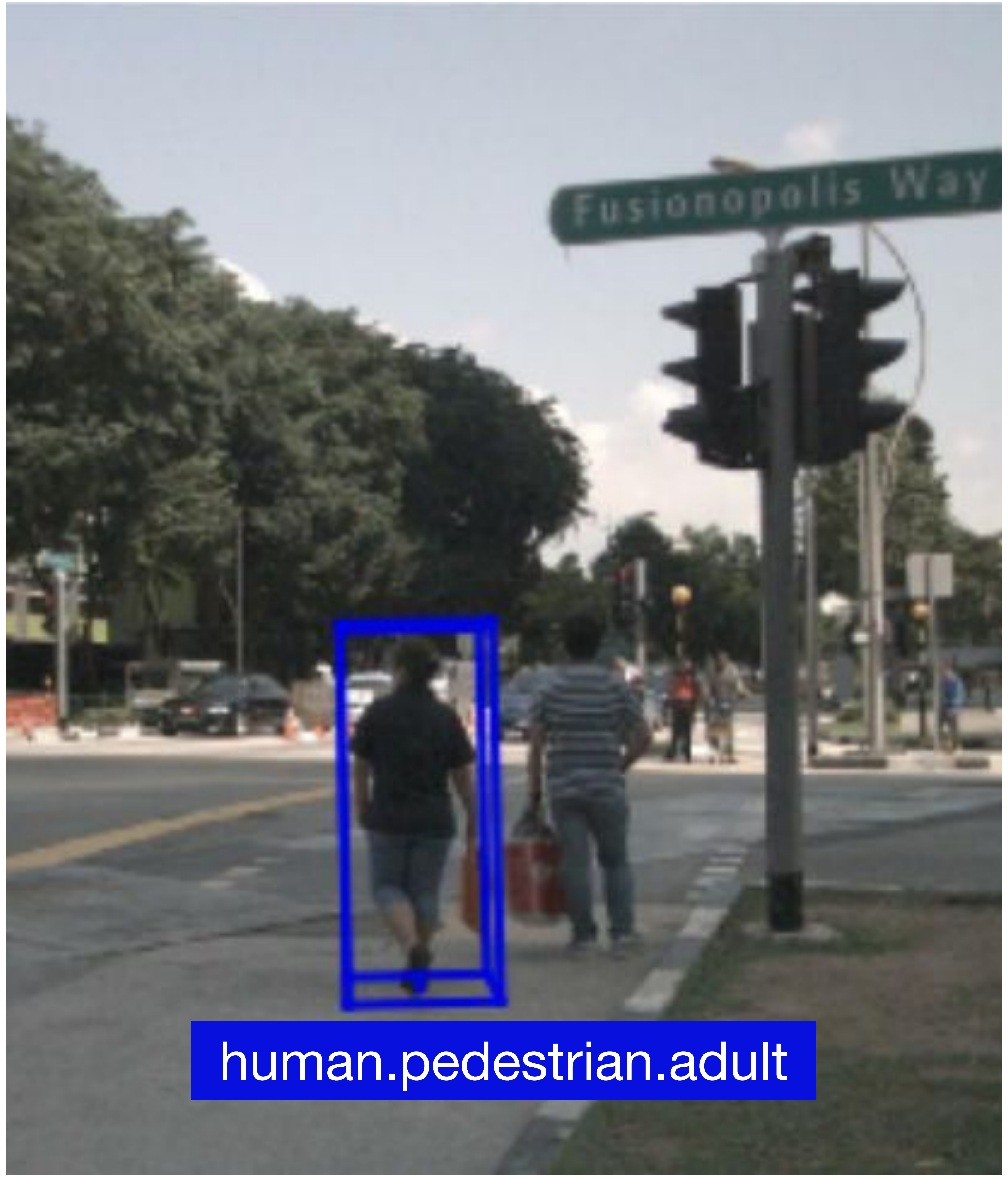
```
'passenger is safe',
AND(
  'safe transitions',
  NOT('threatening objects')
```

!

The best option is to veer and slow down. The vehicle is traveling too fast to suddenly stop. The vision system is inconsistent, but the lidar system has provided a reasonable and strong claim to avoid the object moving across the street.

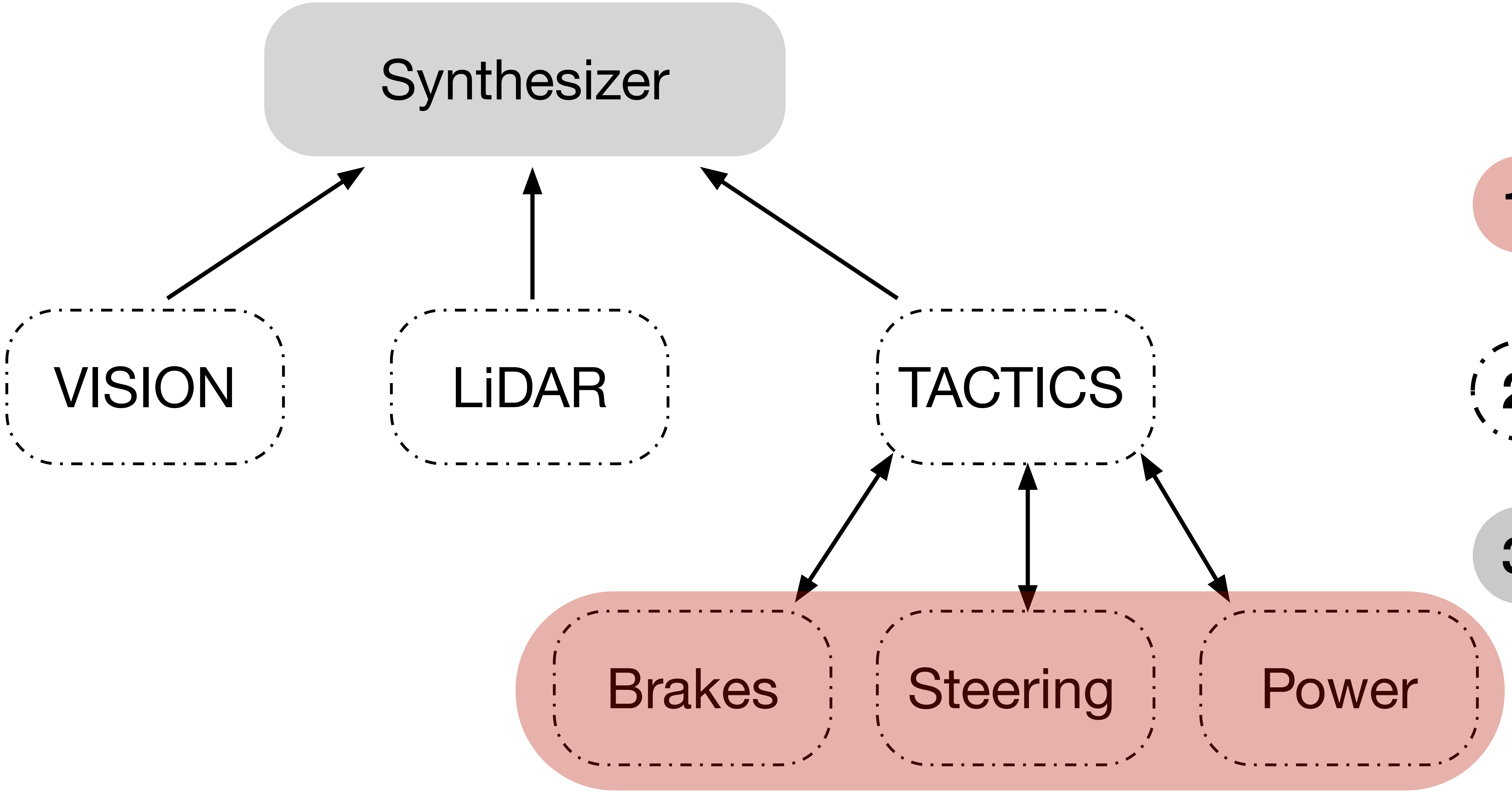
Evaluation of Reasonableness on NuScenes


```
{'token': '70aecbe9b64f4722ab3c230391a3beb8',
 'sample_token': 'cd21dbfc3bd749c7b10a5c42562e0c42',
 'instance_token': '6dd2cbf4c24b4caeb625035869bca7b5',
 'visibility_token': '4',
 'attribute_tokens': ['4d8821270b4a47e3a8a300cbec48188e'],
 'translation': [373.214, 1130.48, 1.25],
 'size': [0.621, 0.669, 1.642],
 'rotation': [0.9831098797903927, 0.0, 0.0, -0.18301629506281616],
 'prev': 'a1721876c0944cdd92ebc3c75d55d693',
 'next': '1e8e35d365a441a18dd5503a0ee1c208',
 'num_lidar_pts': 5,
 'num_radar_pts': 0,
 'category_name': 'human.pedestrian.adult'}
```

Data from NuScenes

Summary

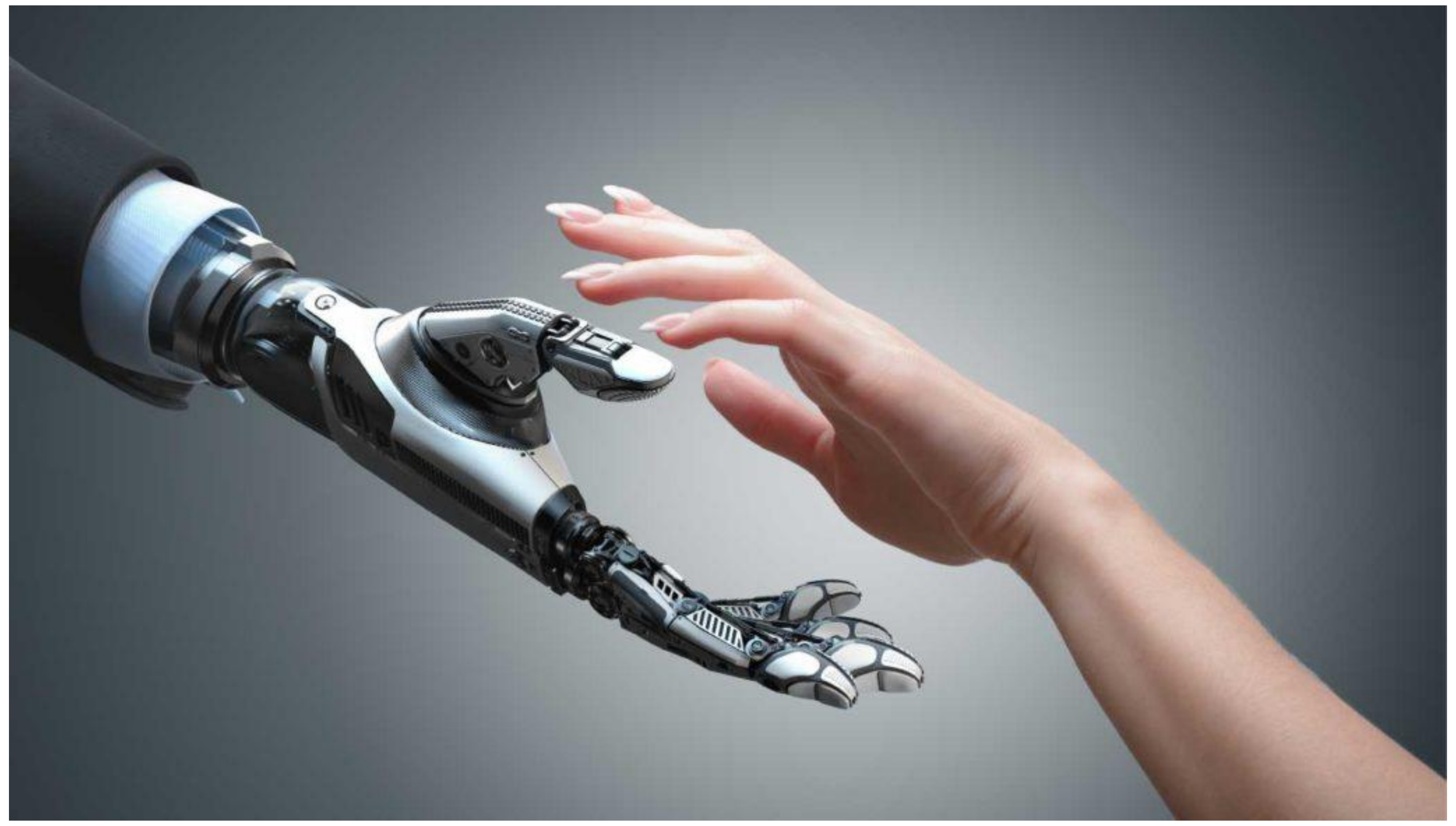


1. Generate Symbolic Qualitative Descriptions for each committee.
2. Input qualitative descriptions into local “reasonableness” monitors.
3. Use a synthesizer to reconcile inconsistencies between monitors.

L.H. Gilpin, V. Penubarthi, L. Kagal. “Anomaly Detection through Explanations.” To be submitted.

Applications

Society



Systems that articulately communicate with humans on shared tasks.

Liability

Systems that can testify, answer questions, and provide insights.

Robustness

Dynamic detection of failure and intrusion with precise mitigation.