Explaining and Monitoring Models with Knowledge Graphs

AAAI 2023 Tutorial on Trustworthy and Responsible AI

Leilani H. Gilpin, PhD
Assistant Professor at UC Santa Cruz
lgilpin.com
lgilpin@ucsc.edu

Agenda

Motivate problem: Systems lack commonsense

Local sanity checks

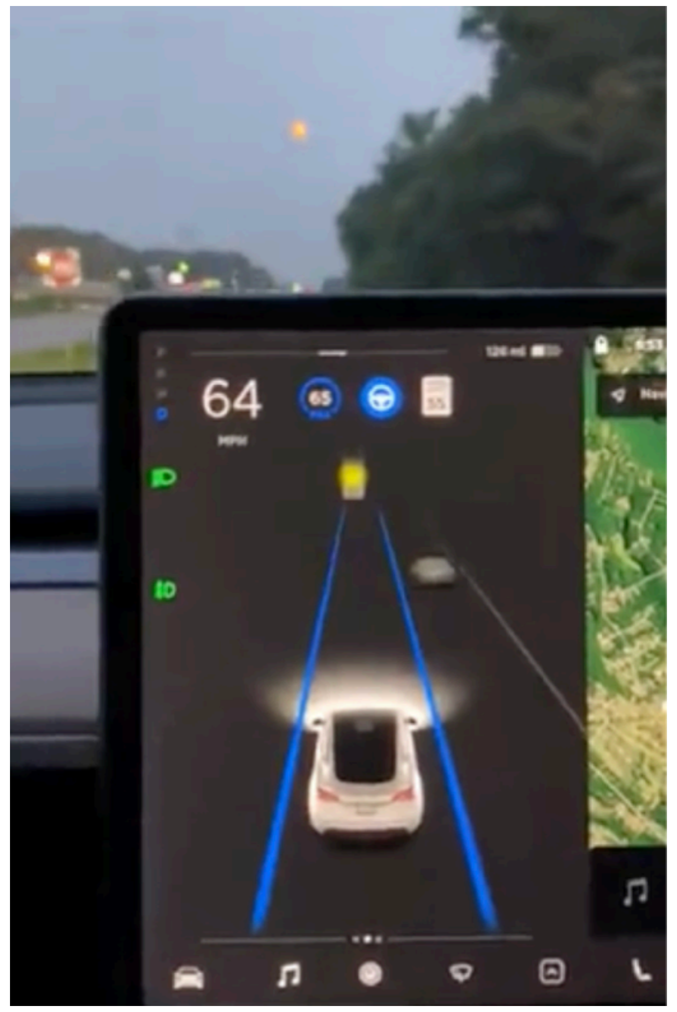
Using XAI + commonsense to “stress test” critical systems.

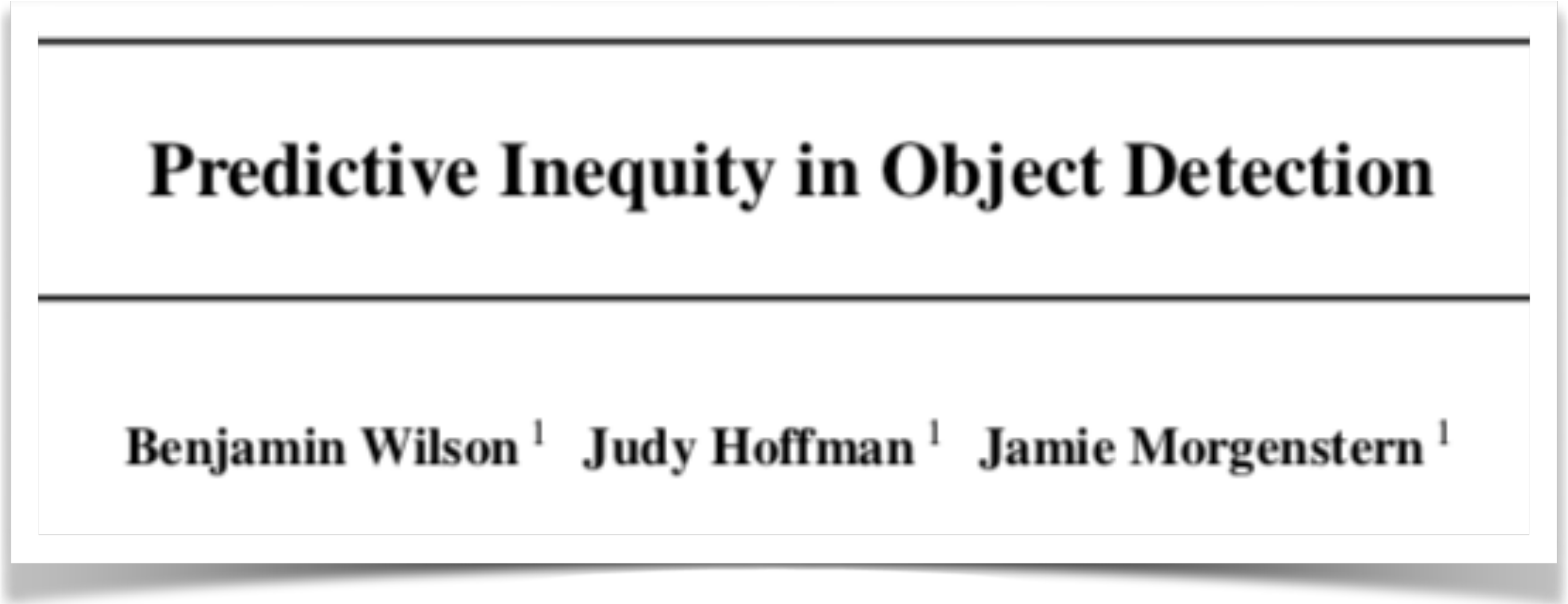
Open Challenges: Articulate systems by design.

Question: How to develop self-explaining architectures for system monitoring in critical domains?

Autonomous Vehicles Lack Common Sense

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





Predictive Inequity in Object Detection

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

Autonomous Vehicle Solutions are at Two Extremes

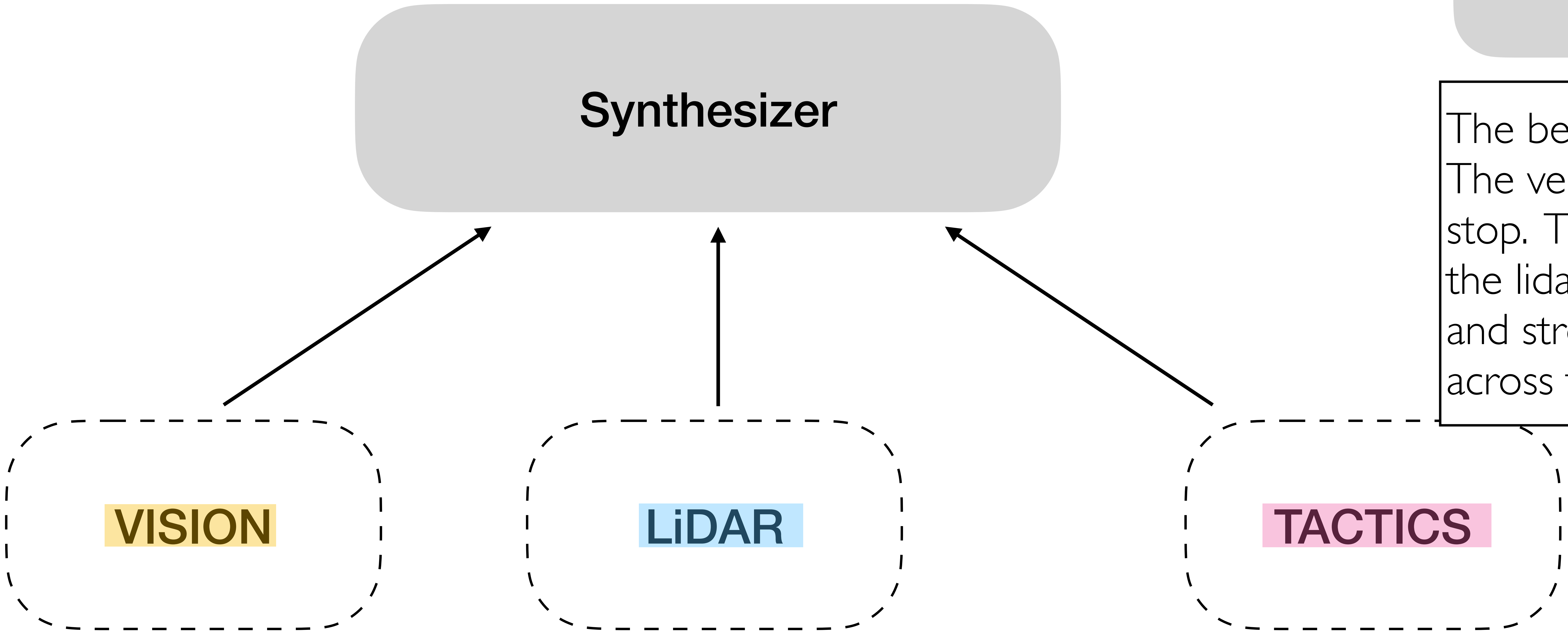
An Existing Problem

The Uber Accident

Solution: Internal Communication

Anomaly Detection through Explanations

Synthesizer to reconcile inconsistencies between monitor outputs.



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.

L..H. Gilpin. "Anomaly Detection Through Explanations." PhD Thesis, 2020.

L.H. Gilpin, V. Penubarthi, and L. Kagal. "Explaining Multimodal Errors in Autonomous Vehicles." 2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA). IEEE, 2021.

Agenda

Motivate problem: Systems lack commonsense

Local sanity checks

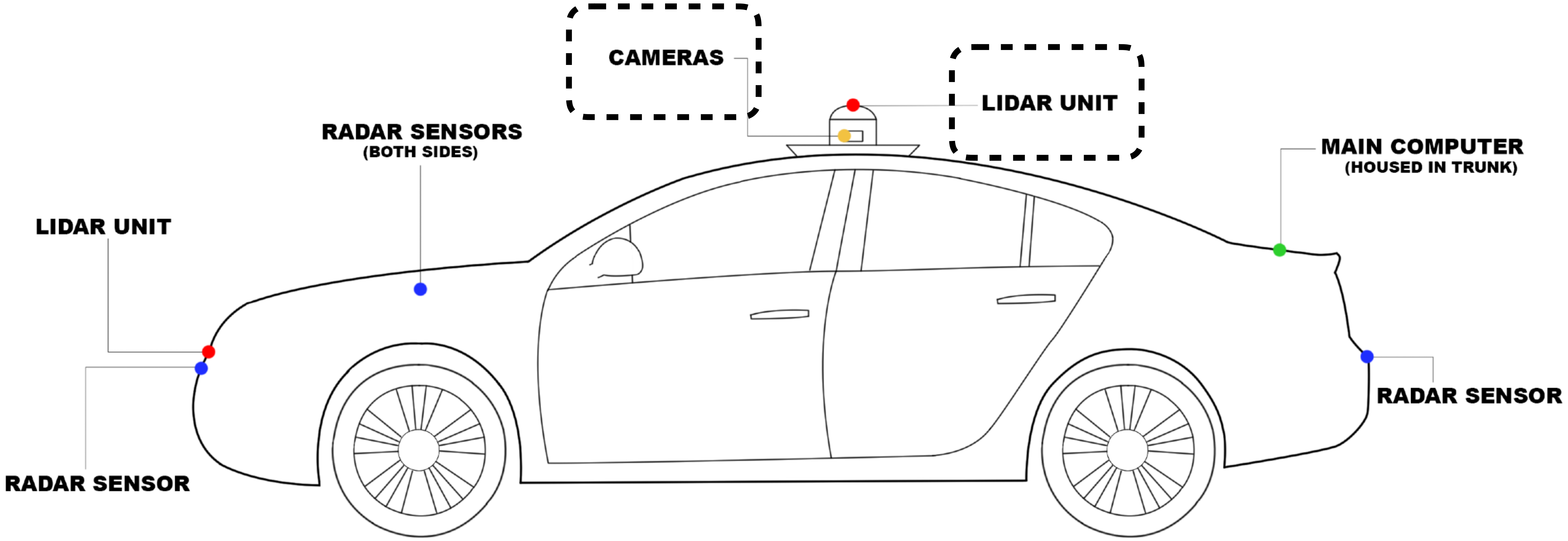
Using KG to “stress test” critical systems.

Open Challenges: Articulate systems by design.

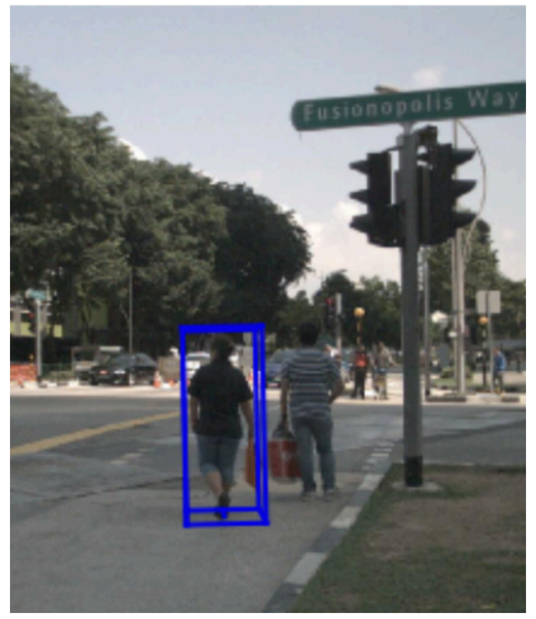
Complex Systems Fail in Two Ways

1. Failure *local* to a specific subsystem.

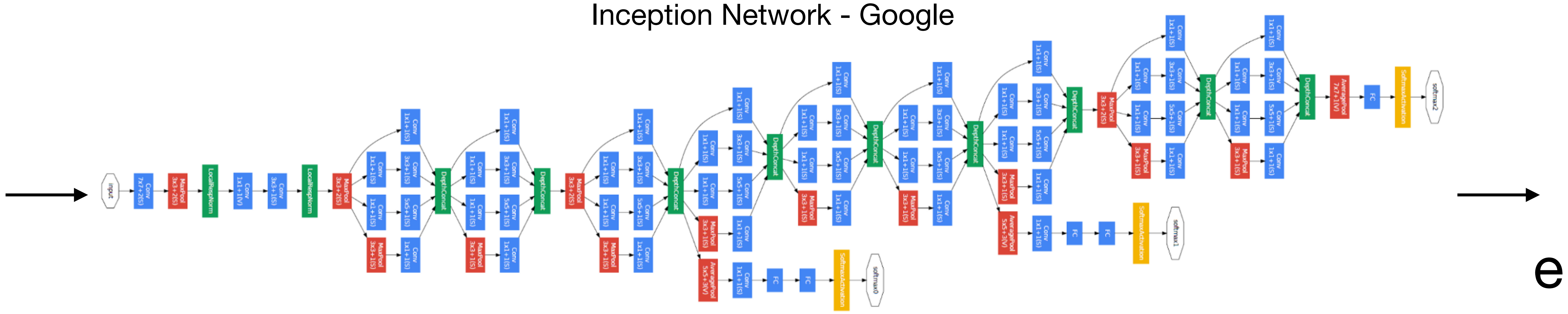
2. A failed *cooperation* amongst subsystems.



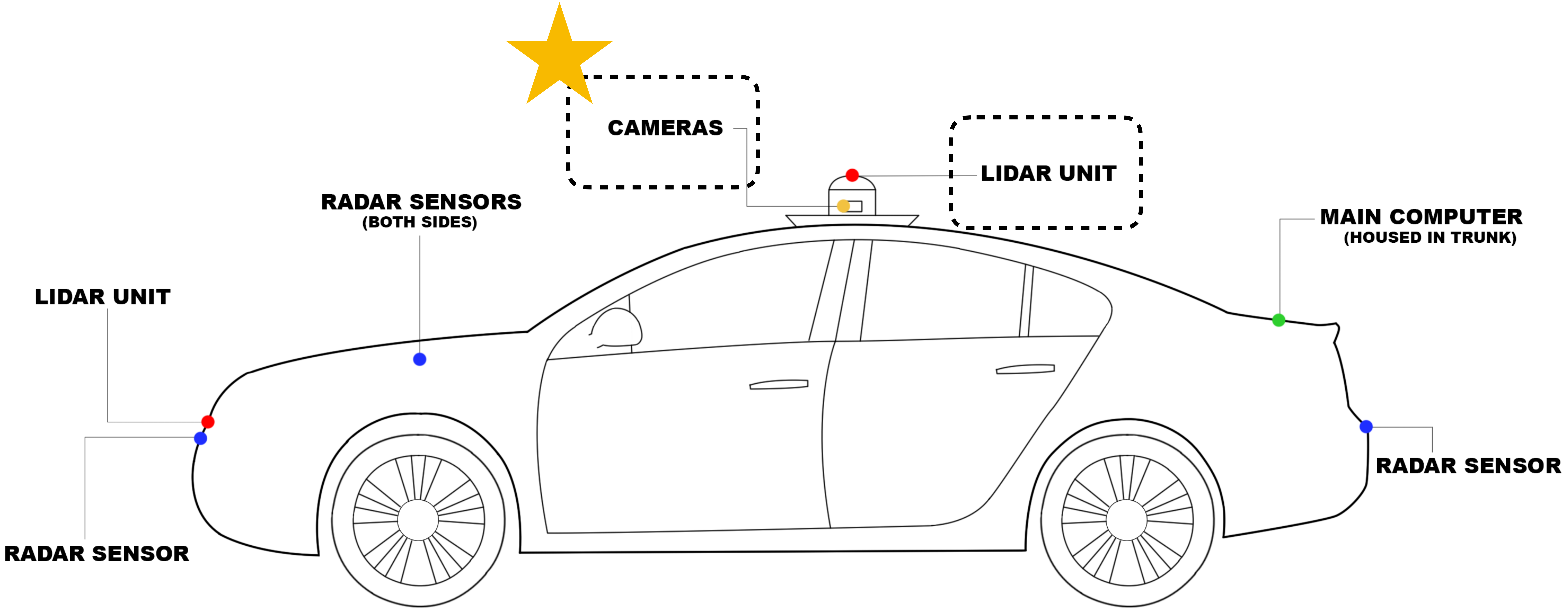
A Neural Network Labels Camera Data



Inception Network - Google

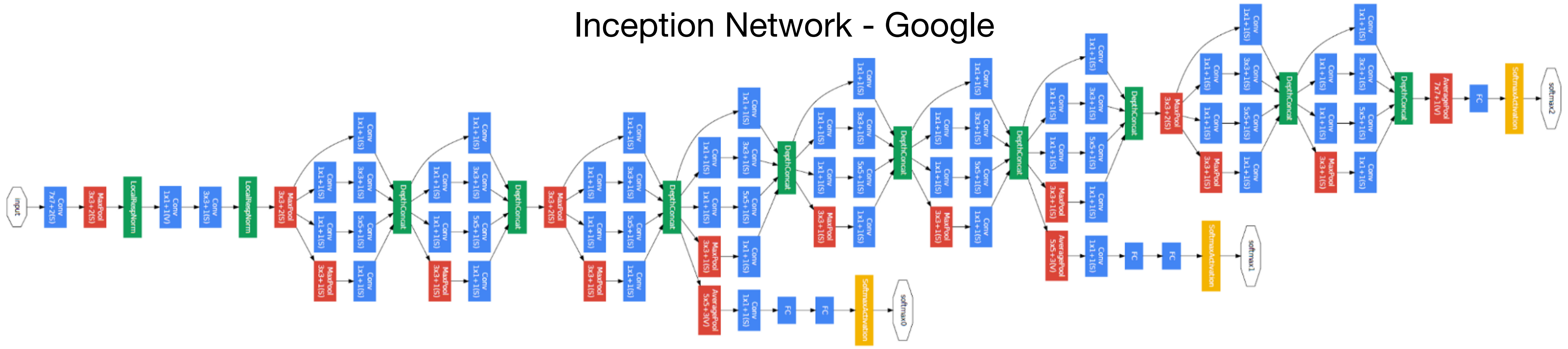


Label
e.g. pedestrian



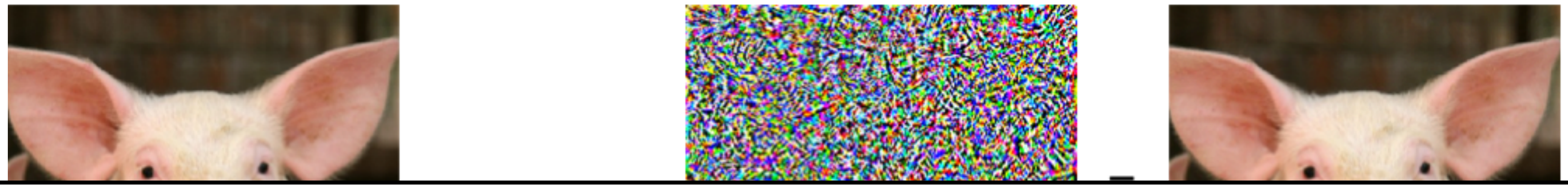
Problem: Neural Networks are Brittle

Inception Network - Google



Label
e.g. pedestrian

“pig” “airliner”



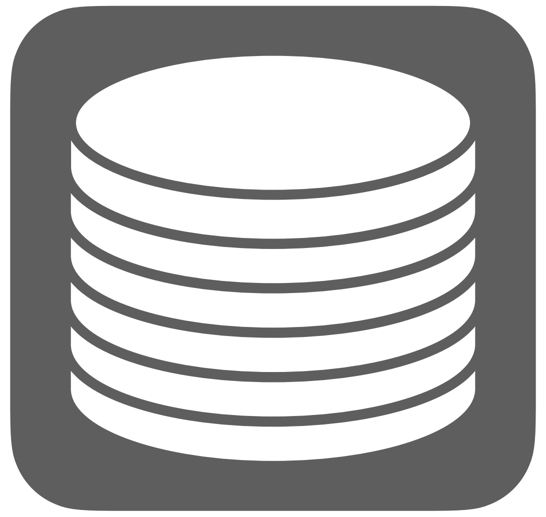
For self-driving, and other mission-critical, safety-critical applications, these mistakes have CONSEQUENCES.

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

Monitor Opaque Subsystems for Reasonableness

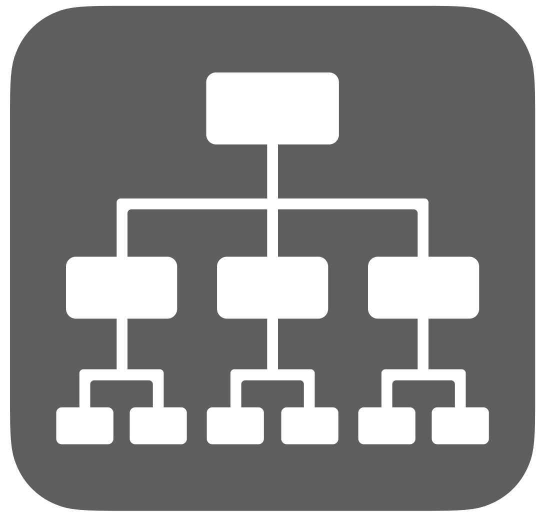
Label
e.g. pedestrian

Opaque
Mechanism



Commonsense
Knowledge Base

+



Flexible
Representation

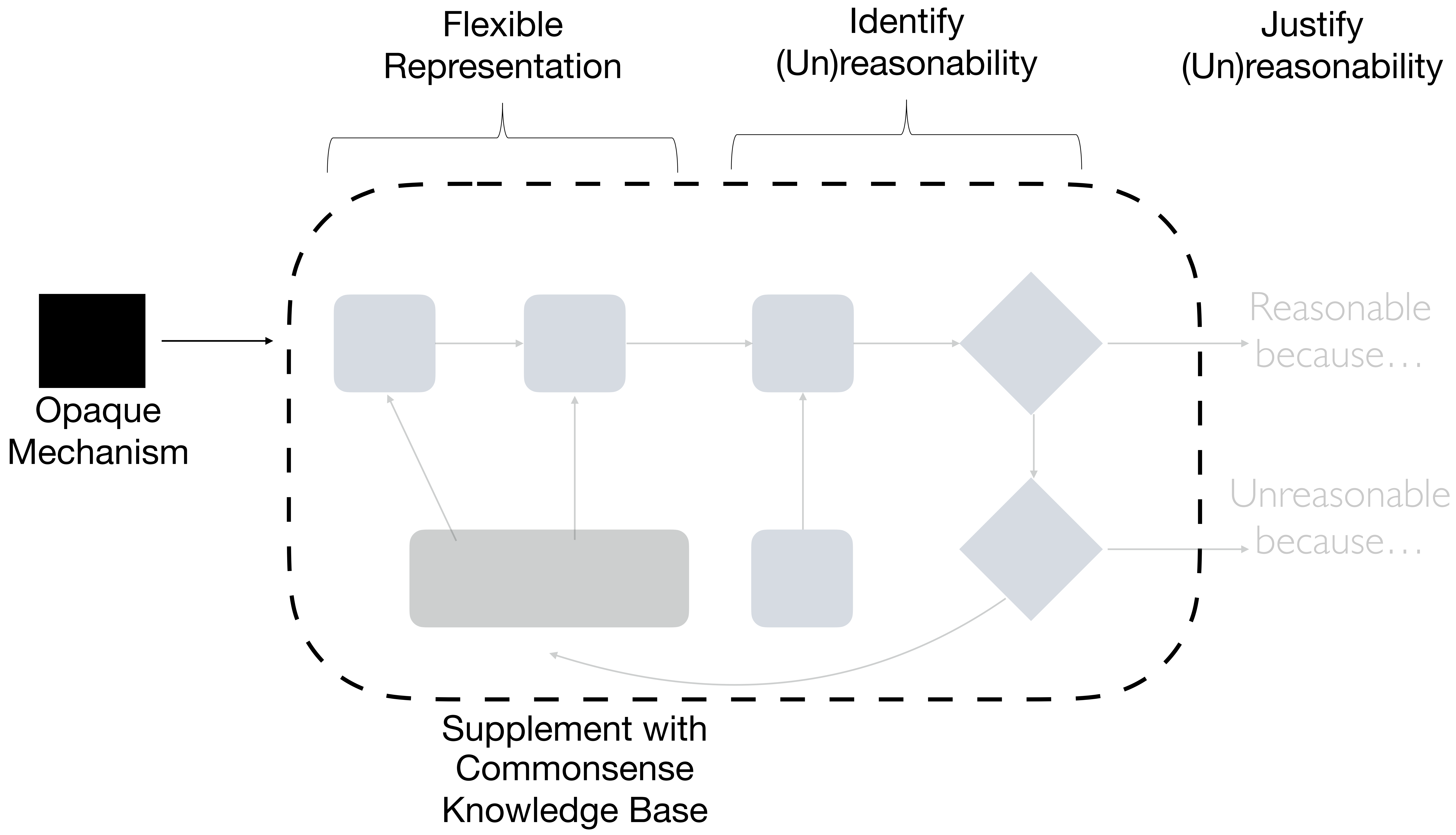
+

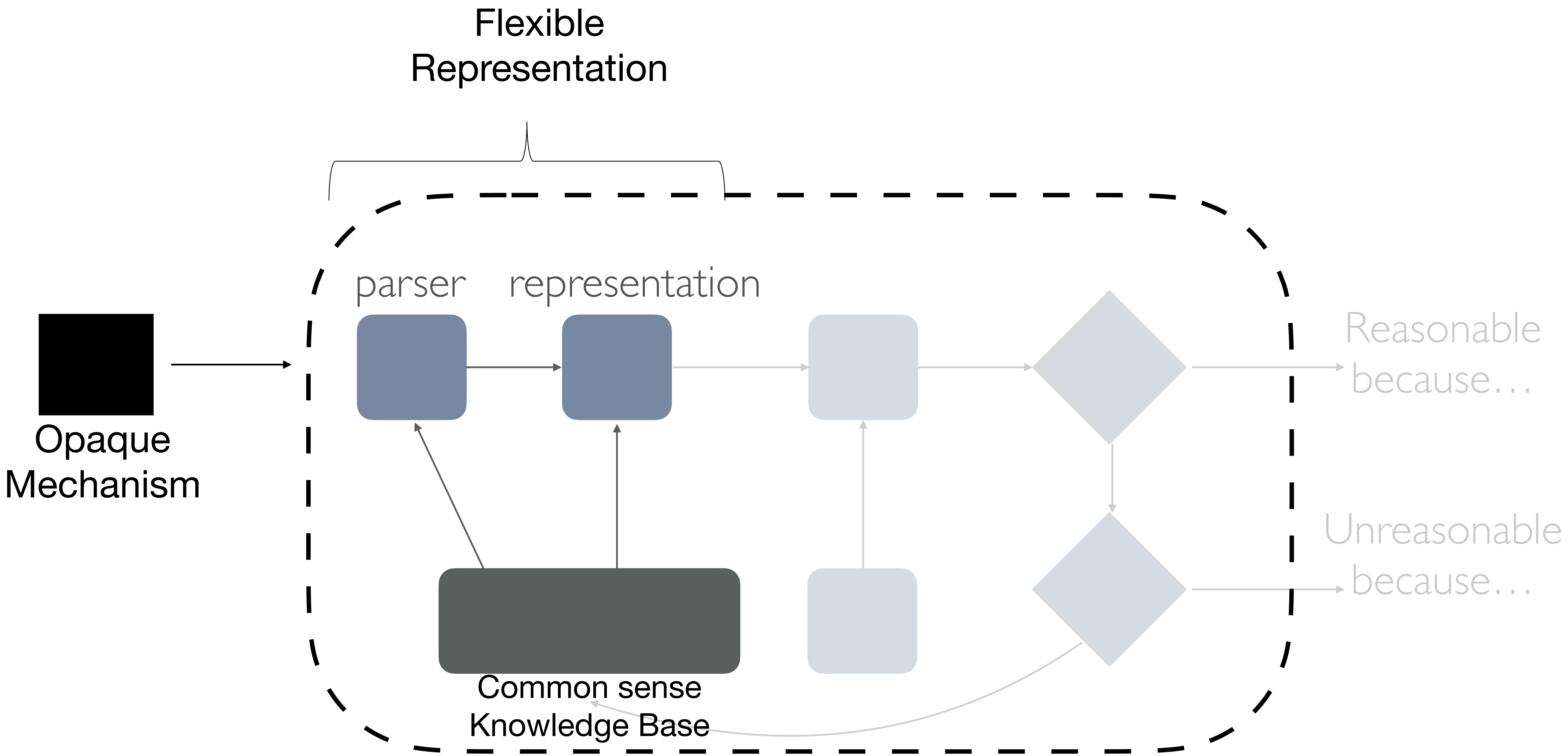
Identify
(Un)reasonability

+

Justify
(Un)reasonability

1. Judgement of reasonableness
2. Justification of reasonableness


Flexible
Representation
Identify
(Un)reasonability
Justify
(Un)reasonability
Opaque
Mechanism
Reasonable
because…
Unreasonable
because…
Supplement with
Commonsense
Knowledge Base



Flexible Representation
Opaque Mechanism
parser
representation
Reasonable because…
Unreasonable because…
Common sense Knowledge Base

Primitive Representations

Encode Understanding

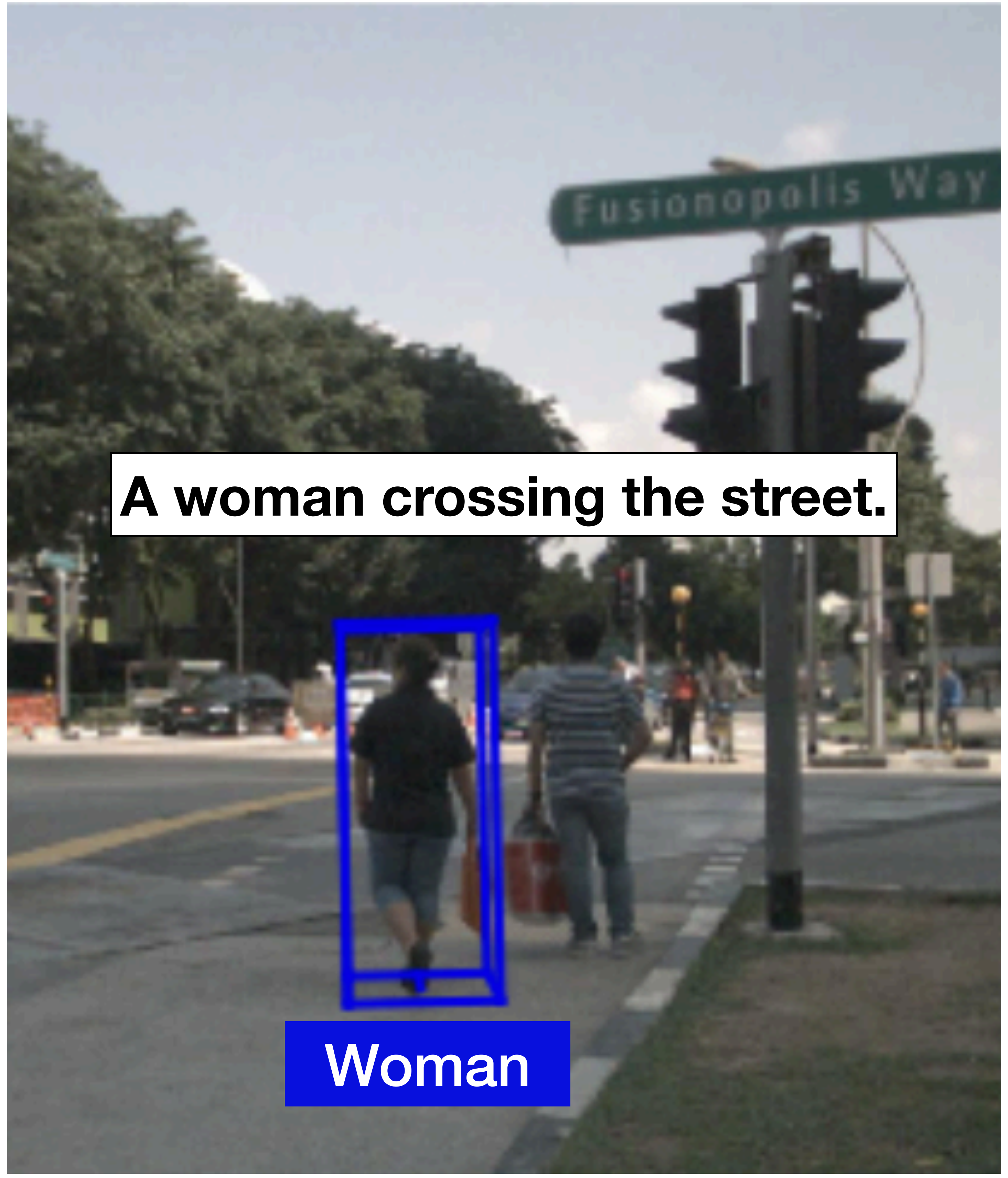
Conceptual Dependency Theory (CD), Schank 1975

11 primitives to account for *most* actions:

ATRANS
ATTEND
INGEST
EXPEL
GRASP
MBUILD
MTRANS
MOVE
PROPEL
PTRANS
SPEAK

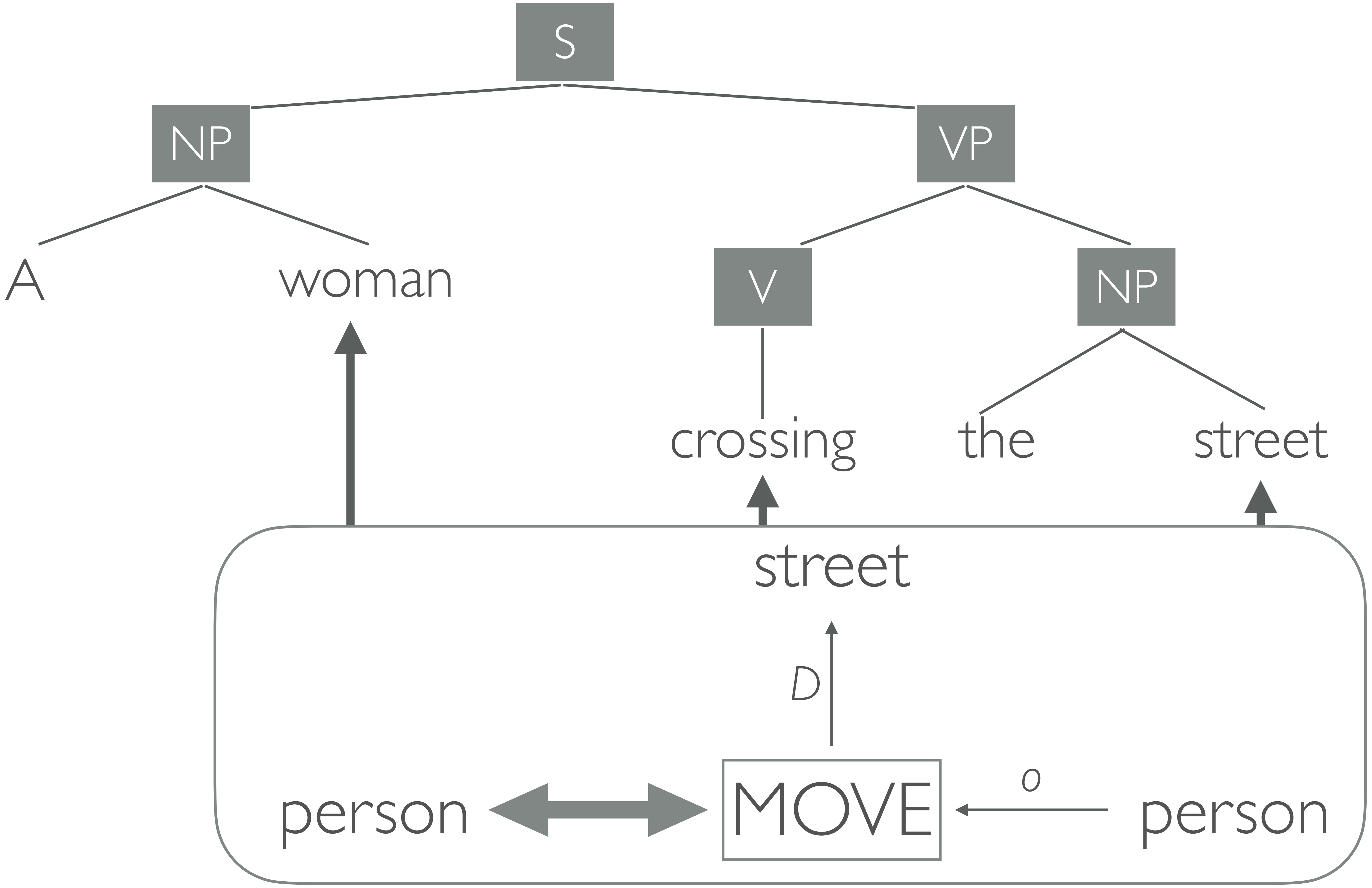
5 for physical actions
Extended to vehicle primitives

Parse Natural Language into Representation

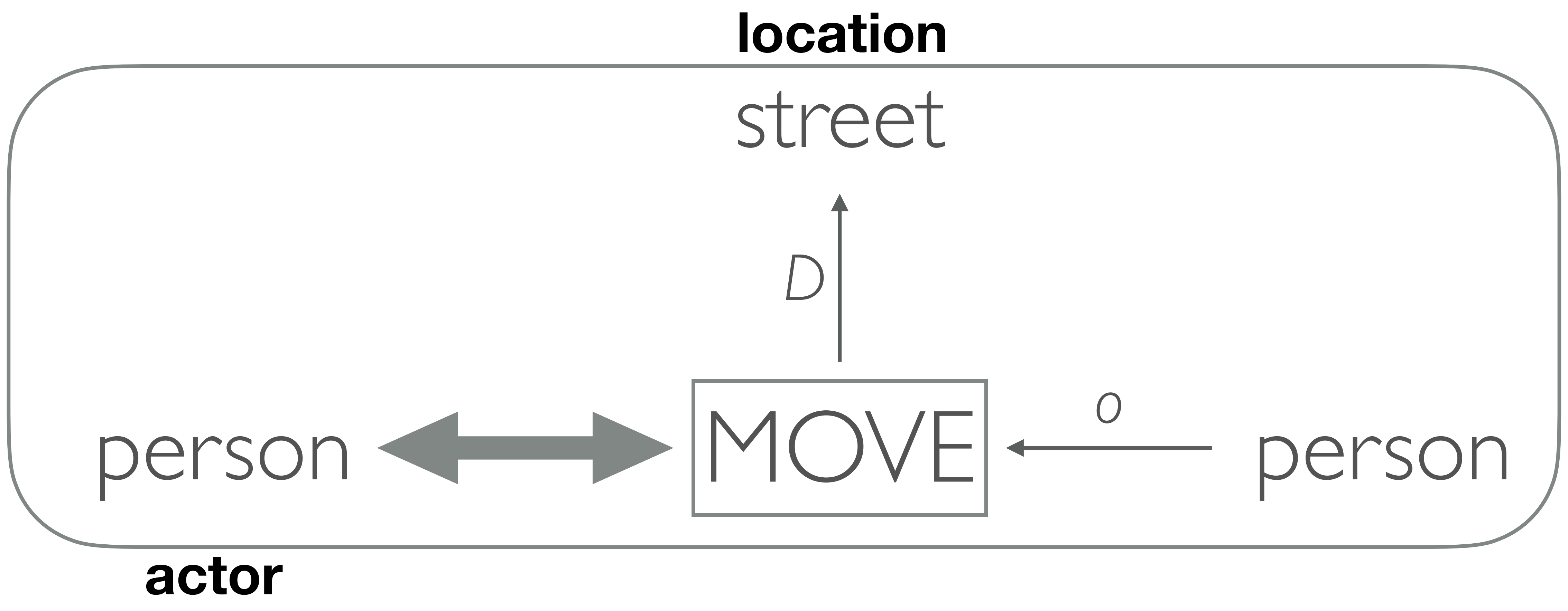


Data from Nuscenes

Parser



Representations with Implicit Rules



A perceived frame is REASONABLE

$$\big((x_1, p_1, y_1), \textbf{isA, REASONABLE}\big) \wedge$$
$$\big((x_2, p_2, y_2), \textbf{isA, REASONABLE}\big) \wedge$$
$$\ldots \wedge$$
$$\big((x_n, p_n, y_n), \textbf{isA, REASONABLE}\big)$$

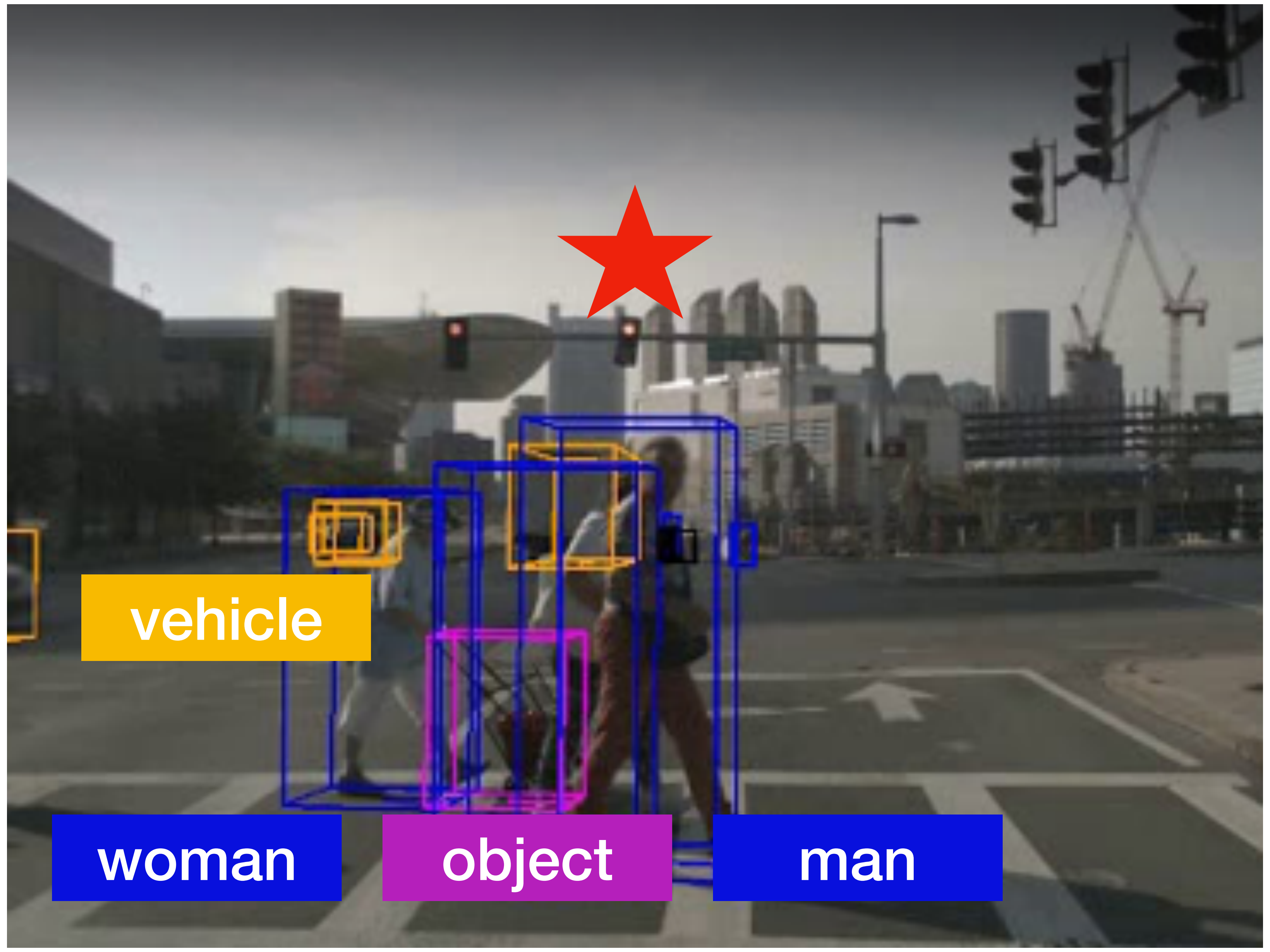
<u>Move Primitive Reasonability</u>

$$(x, hasProperty, animate) \wedge (x, locatedNear, y) \Rightarrow \big((x, MOVE, y)\ \textbf{isA, REASONABLE}\big)$$

actor **location**

Implementing a 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" .
```

Implementing Reasonableness Monitors
For Real-world Error Detection

- End-to-end prototype
  - Machine perception
  - Represented with 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.

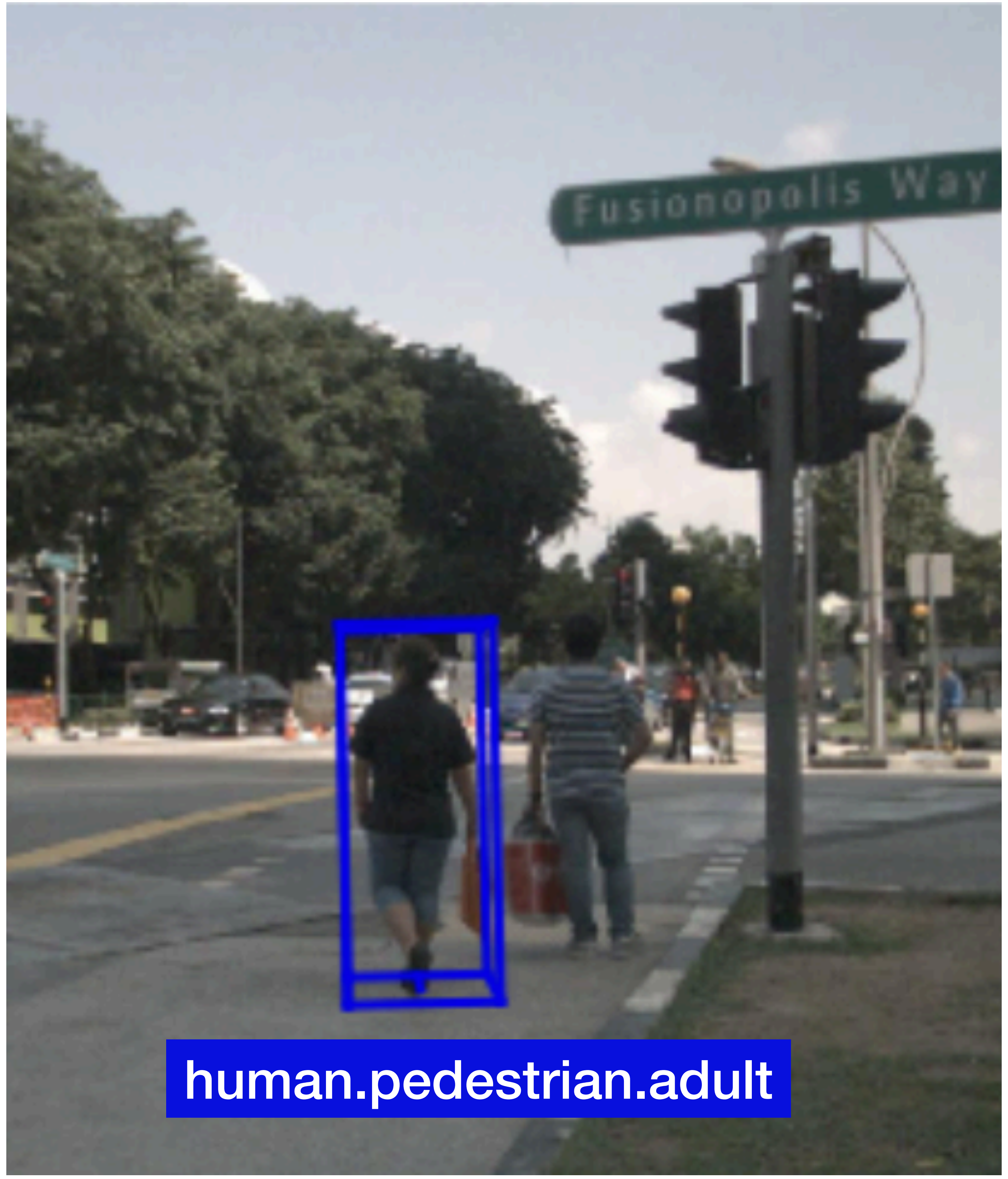
- Generalized framework
  - Reusable web standards
  - Extended Schank representations

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

Reasonableness Monitoring on Real Data

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

Commonsense is Unorganized

ConceptNet

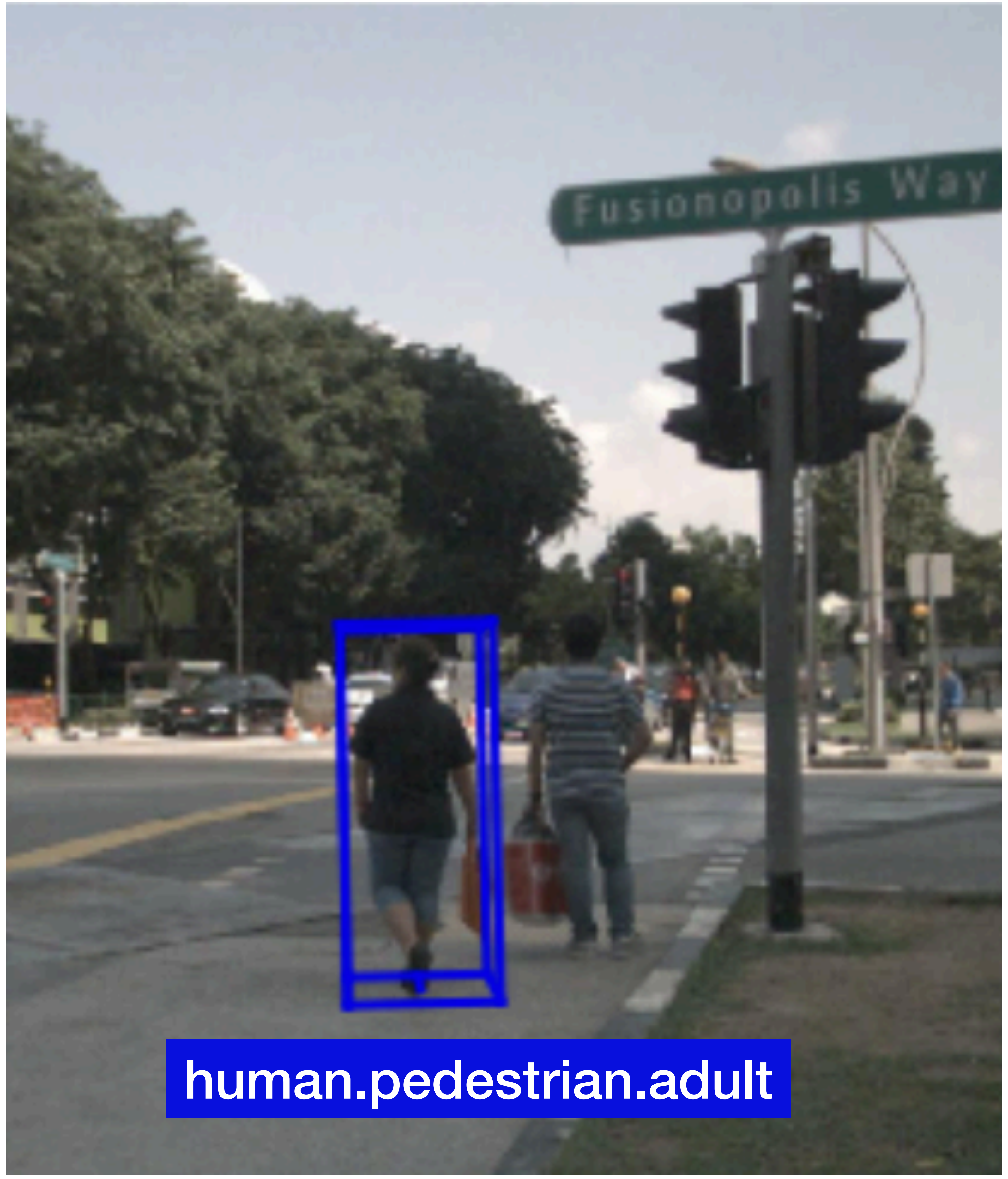
adult is a type of...

- en animal (n, wn)
- en person (n, wn)
- en animal (n)

adult is capable of...

- en help a child
- en dress herself
- en sign a contract
- en drink beer
- en work
- en act like a child
- en dress himself
- en drive a car
- en drive a train
- en explain the rules to a child

```
('adult, 'typeOf, 'animal)
('adult, 'isA, 'bigger than a child')
…
```

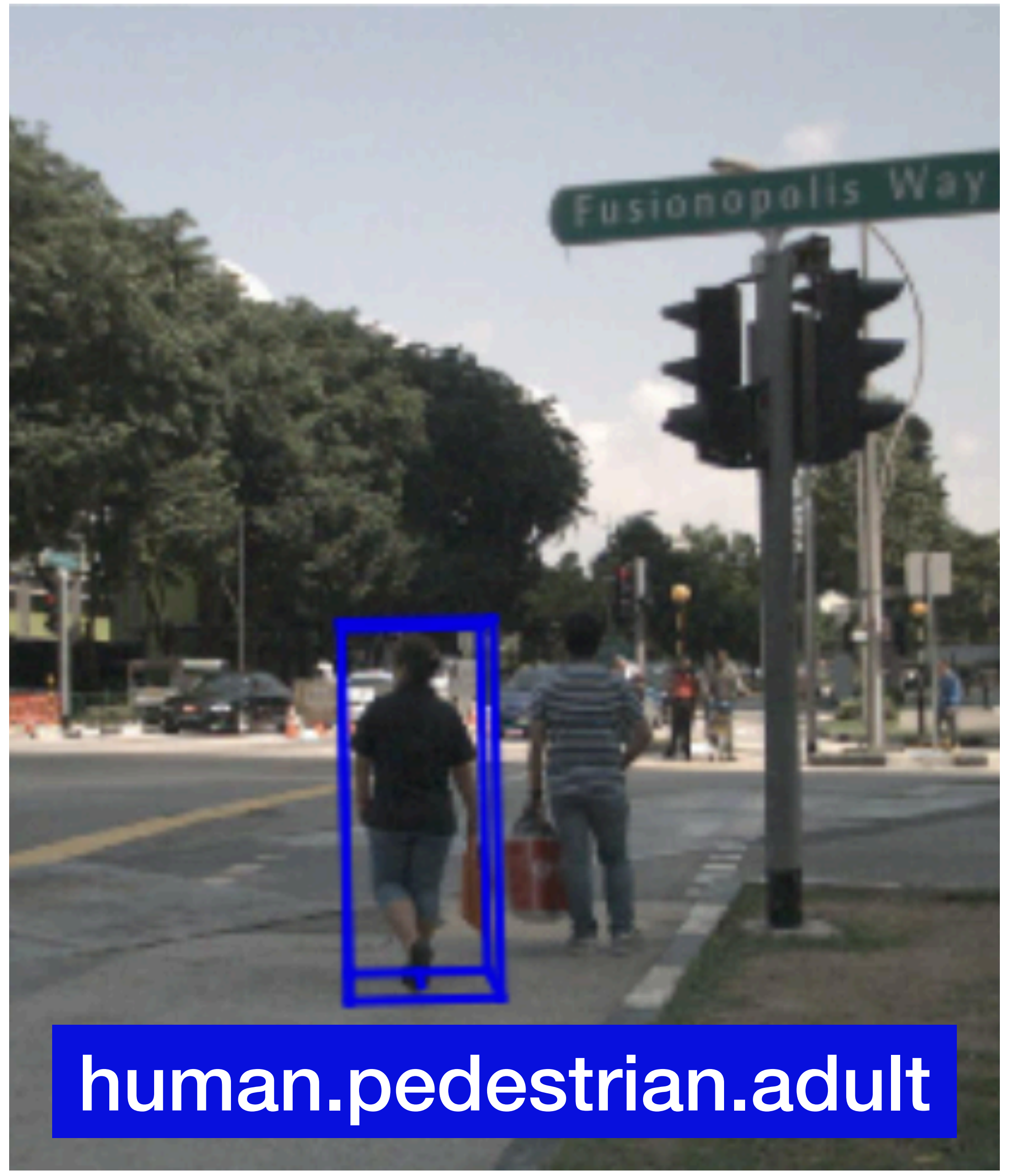


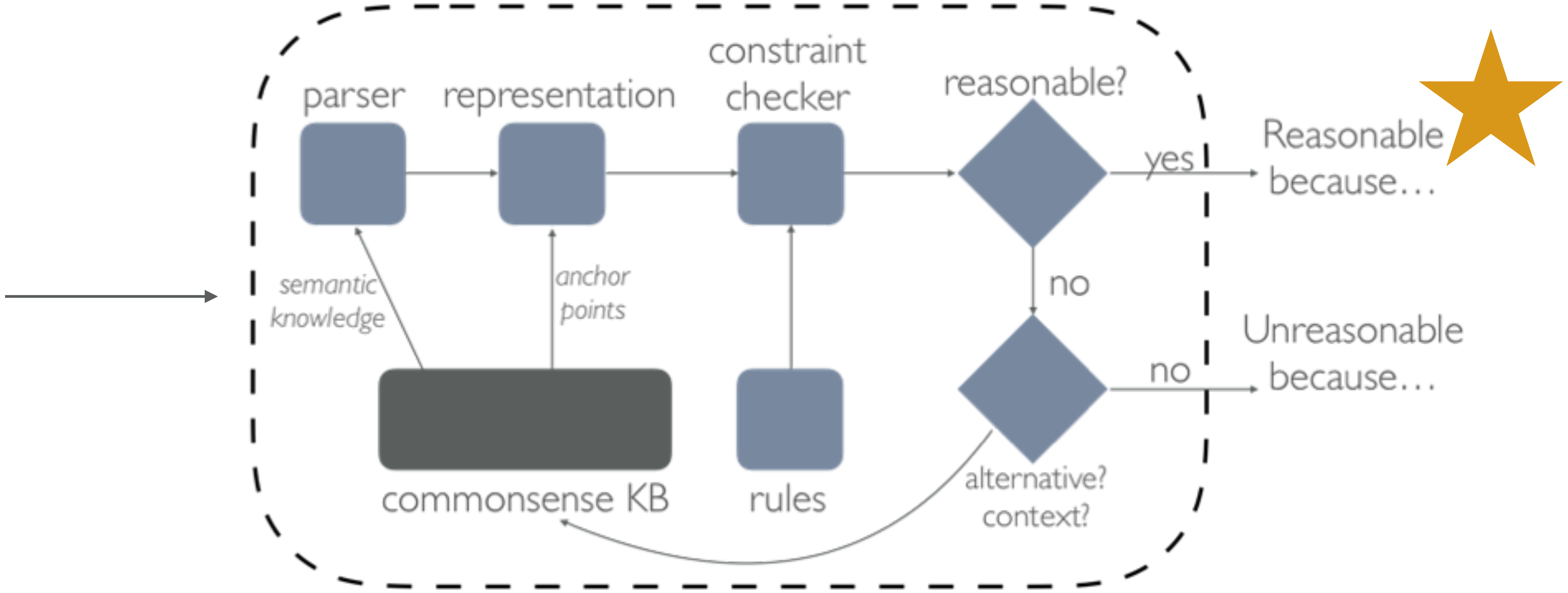
Data from NuScenes

Pain Points of Commonsense Knowledge

1. Organization of commonsense knowledge
   1. Top-down vs bottom-up - what is the sweet spot?
   2. Linguistic flexibility vs semantic expressivity
2. Flexible generalization with little data
   1. Reasoning by analogy seems promising
   2. Difficult and we don’t seem to have the right knowledge in the right form
3. Realistic evaluation tasks and datasets
   1. We tend to hack the tasks, and the language models are an excellent helper for it
   2. Embodied, multi-modal, explainable, open-ended tasks are all great efforts
   3. How to evaluate them at scale is not obvious

Monitor Outputs a Judgement and Justification





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.

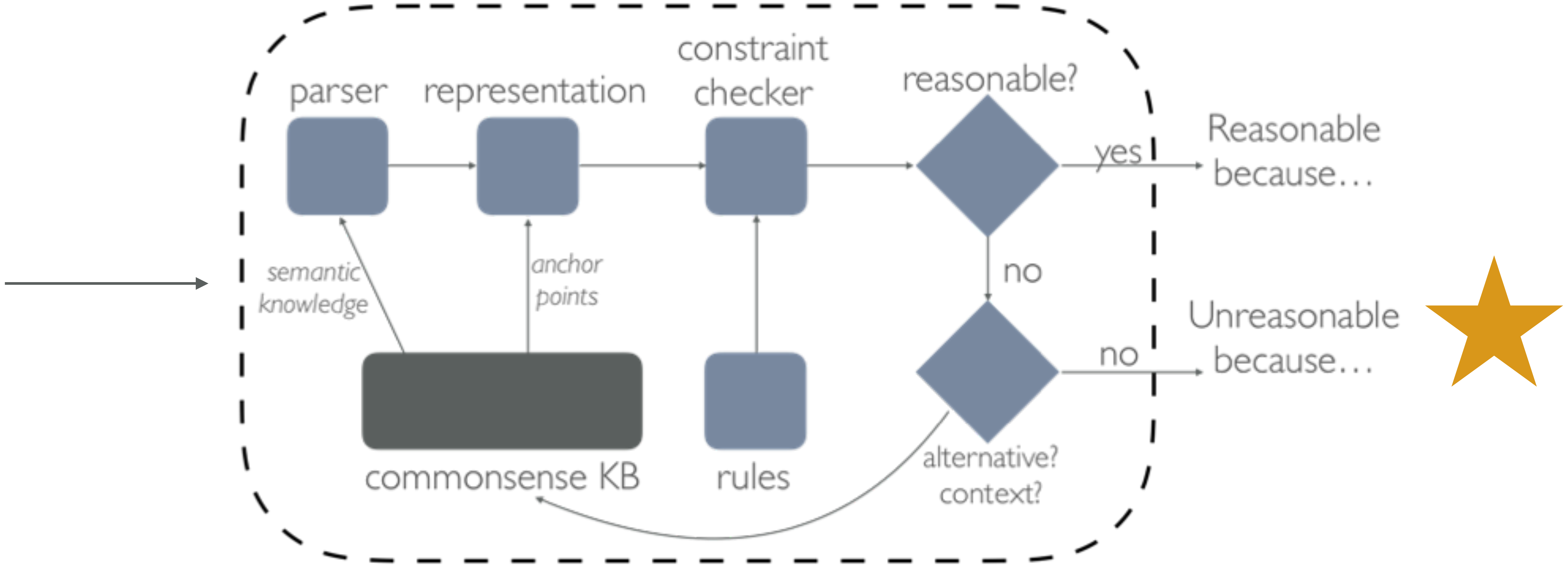
Evaluating Reasonableness Monitors

Building Errors

- Built an “unreasonable” image description dataset.
  - 100 descriptions.
  - Average of 4.47 words, with 57 unique words.
  - 14 verbs, 35 nouns, 8 articles/auxiliary verbs, prepositions.
  - 23 of the 100 had prepositional phrases.

- Self-driving image processing errors:
  - Real-time evaluation with Carla.
  - Added errors on existing datasets (NuScenes).
  - Examining errors on the validation dataset of NuScenes leaderboard.
  - Building challenge problems and scenarios.

Adding 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 common sense supporting this judgement. Discounting objects detected in the same region.

Insights from Misclassifications

Commonsense Assumptions

- Built an “unreasonable” image description dataset.
  - 100 descriptions.
  - Average of 4.47 words, with 57 unique words.
  - 14 verbs, 35 nouns, 8 articles/auxiliary verbs, prepositions.
  - 23 of the 100 had prepositional phrases.

		Classify as:	
		Reasonable	Unreasonable
Label as:	Reasonable		Parser: 2 ConceptNet: 8
	Unreasonable	Parser: 2 ConceptNet: 6	

Agenda

Motivate problem: Systems lack commonsense

Local sanity checks

Using XAI + commonsense to “stress test” critical systems.

Open Challenges: Articulate systems by design.

Vision: Real World Adversarial Examples

“Realistic” Adversarial examples

L. H. Gilpin, A. Amos-Binks, "Close Syntax but Far Semantics: A Risk Management Problem for Autonomous Vehicles.” *To Appear in Abstracts of the AAAI Fall Symposium on Cognitive Systems for Anticipatory Thinking.*

Vision: Real World Adversarial Examples

Anticipatory Thinking Layer for Error Detection

"Realistic" Adversarial examples

The traffic lights are on top of the truck. The lights are not illuminated. The lights are moving at the same rate as the truck, therefore this is not a "regular" traffic light for slowing down and stopping at.

Testing Framework in Two Parts

The traffic lights are on top of the truck. The lights are not illuminated. The lights are moving at the same rate as the truck, therefore this is not a "regular" traffic light for slowing down and stopping at.

Explanatory Error Detection

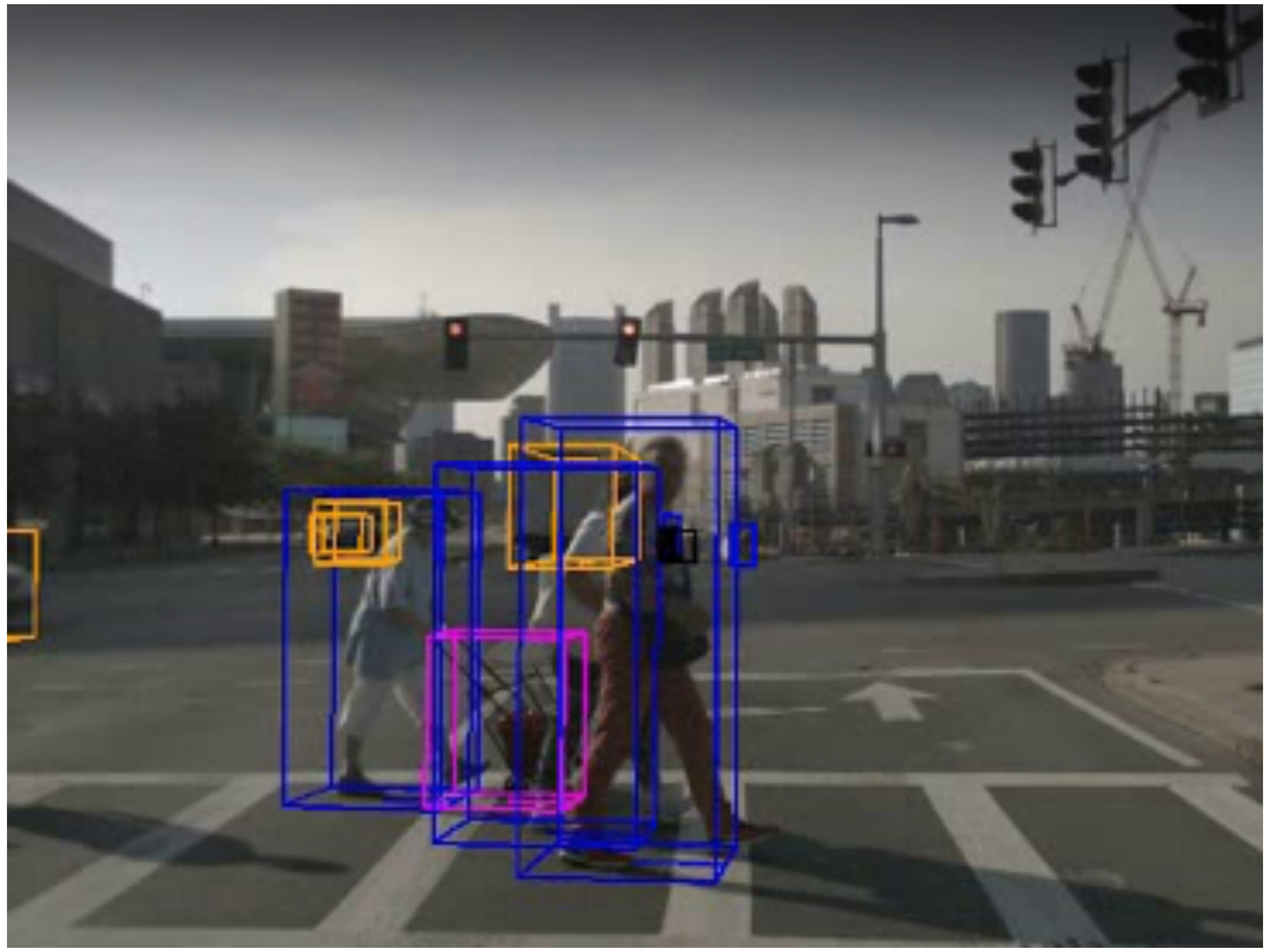
Content generation

!

Deploy

Lack of Data and Challenges for AVs

- Existing Challenges
  - Targeted as optimizing a mission or trajectory and not safety.
  - Data is hand-curated.
- Failure data is not available
  - Unethical to get it (cannot just drive into bad situations).
  - Want the data to be realistic (usually difficult in simulation).



Data from NuScenes

Need for Context and Explanation

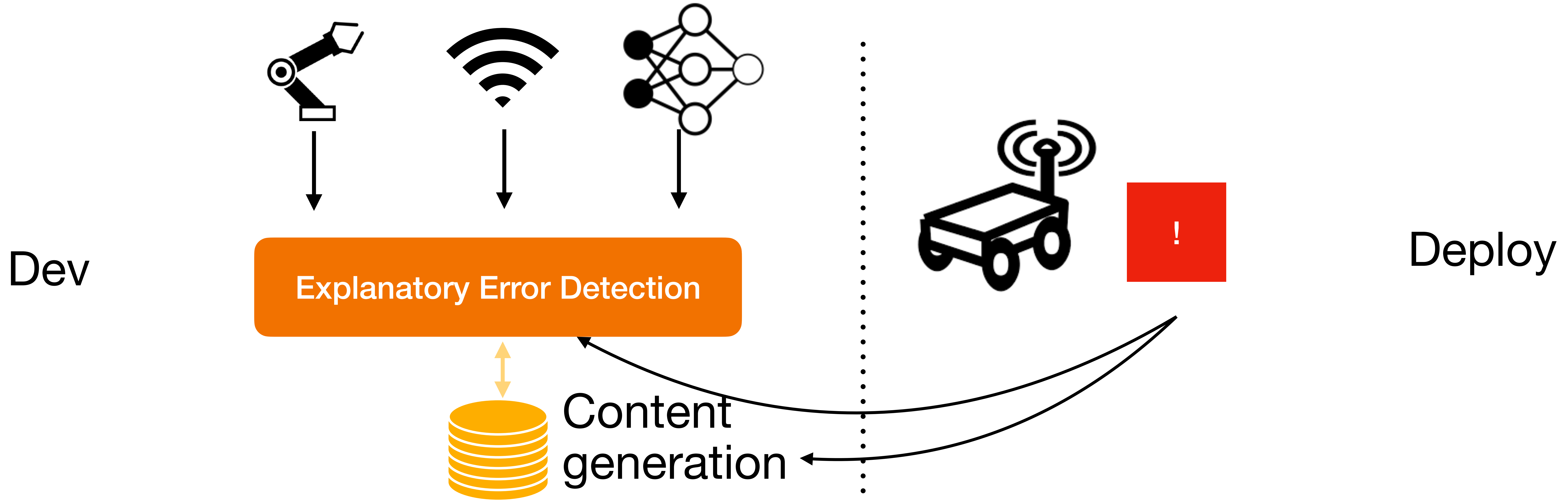
“Realistic” Adversarial

Approach: How it Works

Use Adversarial Images in Dev Testing

- Solution: Use a cognitive architecture that helps to anticipate and understand these failure cases.

- Assess autonomous vehicles for their risk management capabilities **before** being deployed and provide incident level risk management explanations in human readable form.



Agenda

Motivate problem: Systems lack commonsense

Local sanity checks

Using XAI + commonsense to “stress test” critical systems.

Open challenges: Articulate systems by design.

Wrap Up Discussion: Open Challenges

How to make systems that are articulate?

- How do we find the right common sense for specific tasks?
- What is the “right” representation (flexible but also specific).

How can systems communicate?

- Tackling the “interpretability” gap.
- How can we leverage KGs to help?

How can we detect (and explain) commonsense failures?

- What is the proper evaluation method or metrics?
- “Near misses” in commonsense reasoning.

Systems lack commonsense

Explanation

Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning

Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter and Lalana Kagal
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139
{lgilpin, davidbau, bzy, abajwa, specter, lkagal}@ mit.edu

Dynamic explanations, under uncertainty

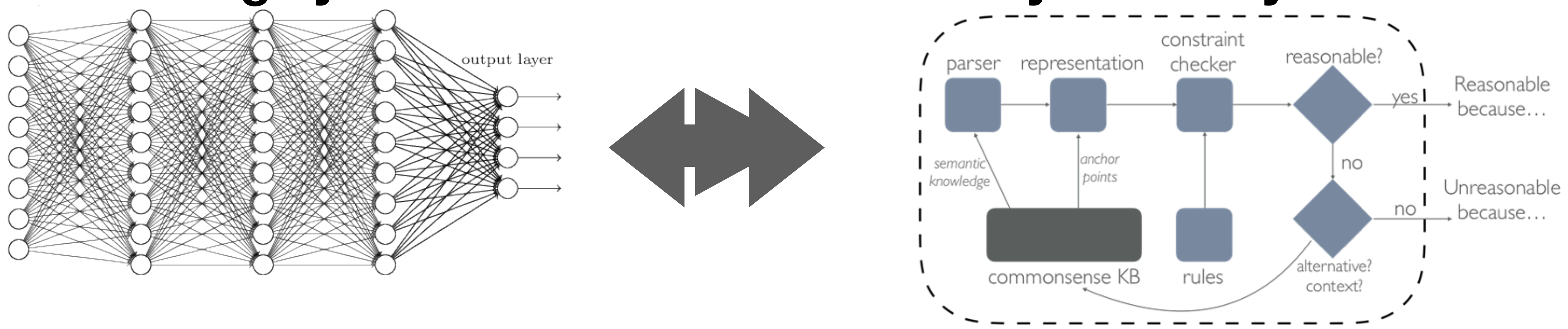
Self-explaining architectures

Vision: Articulate Machines

Coherent Communication

With Other Systems

Learning system **Symbolic system**

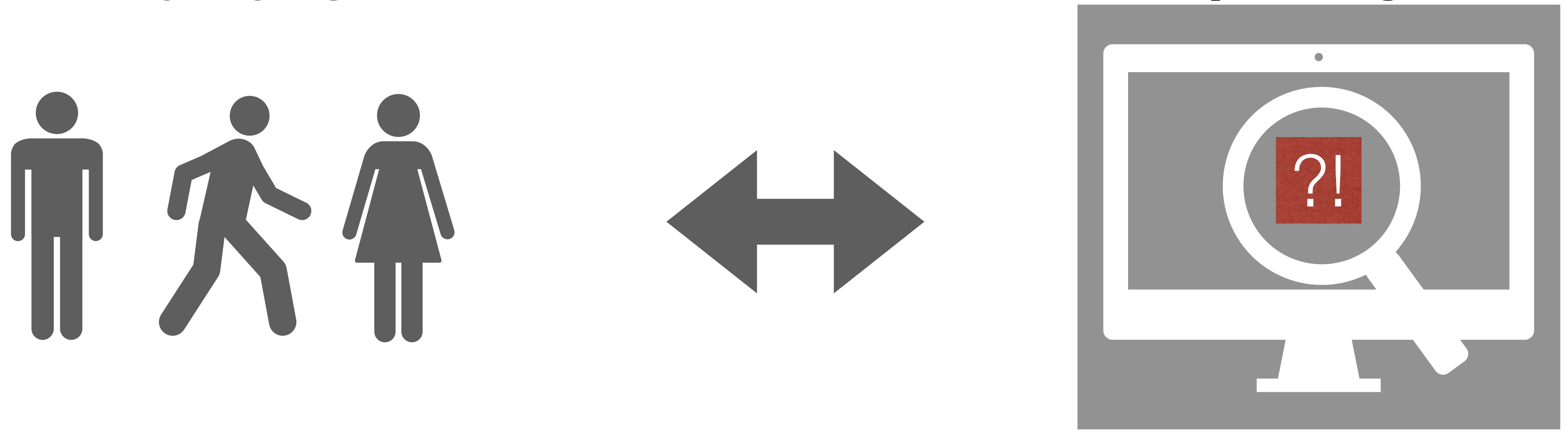


Common language to complete tasks.

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

With Humans

humans **complex system**

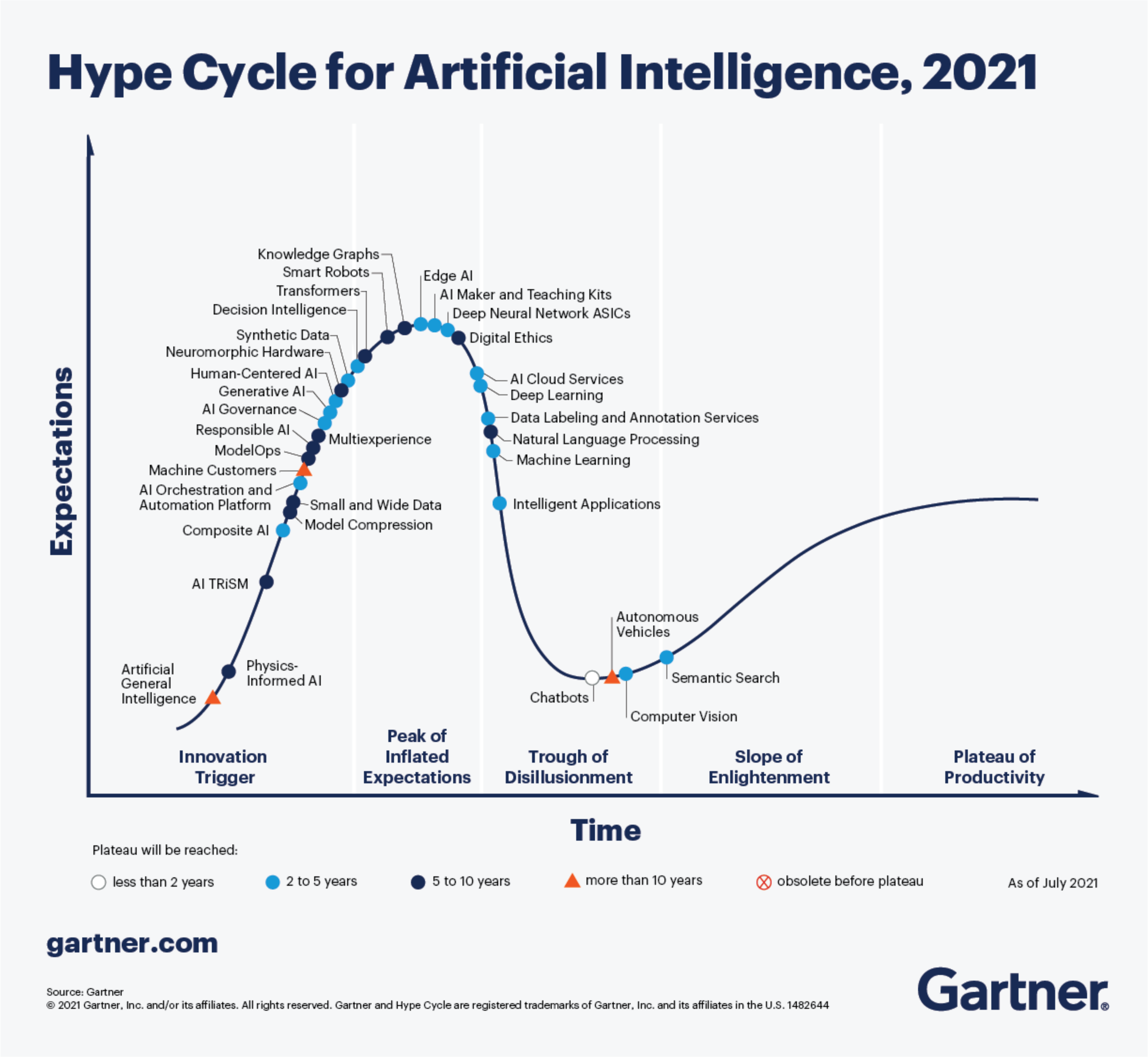


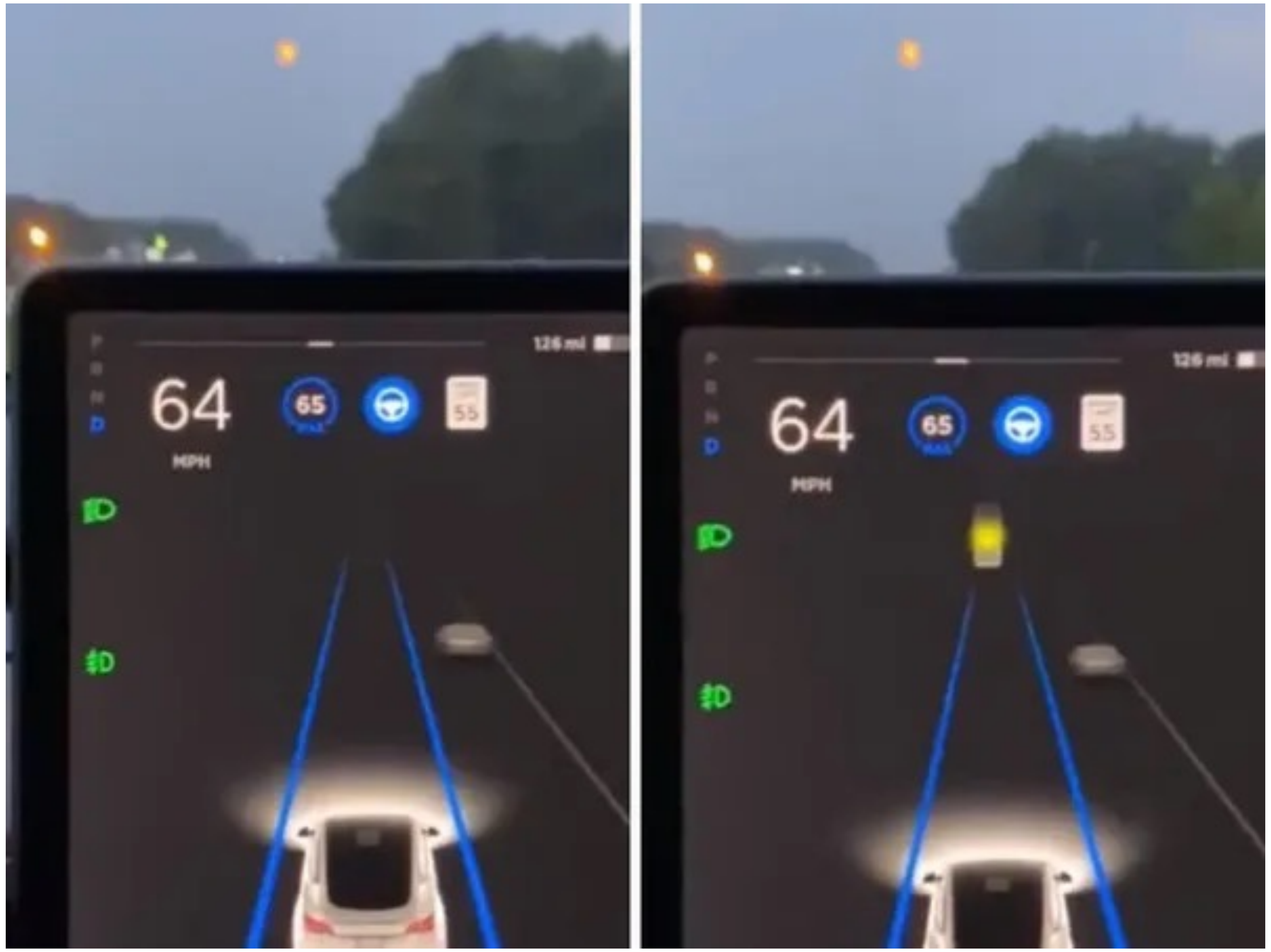
Explanations are a debugging language.

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

Vision: Articulate Machines

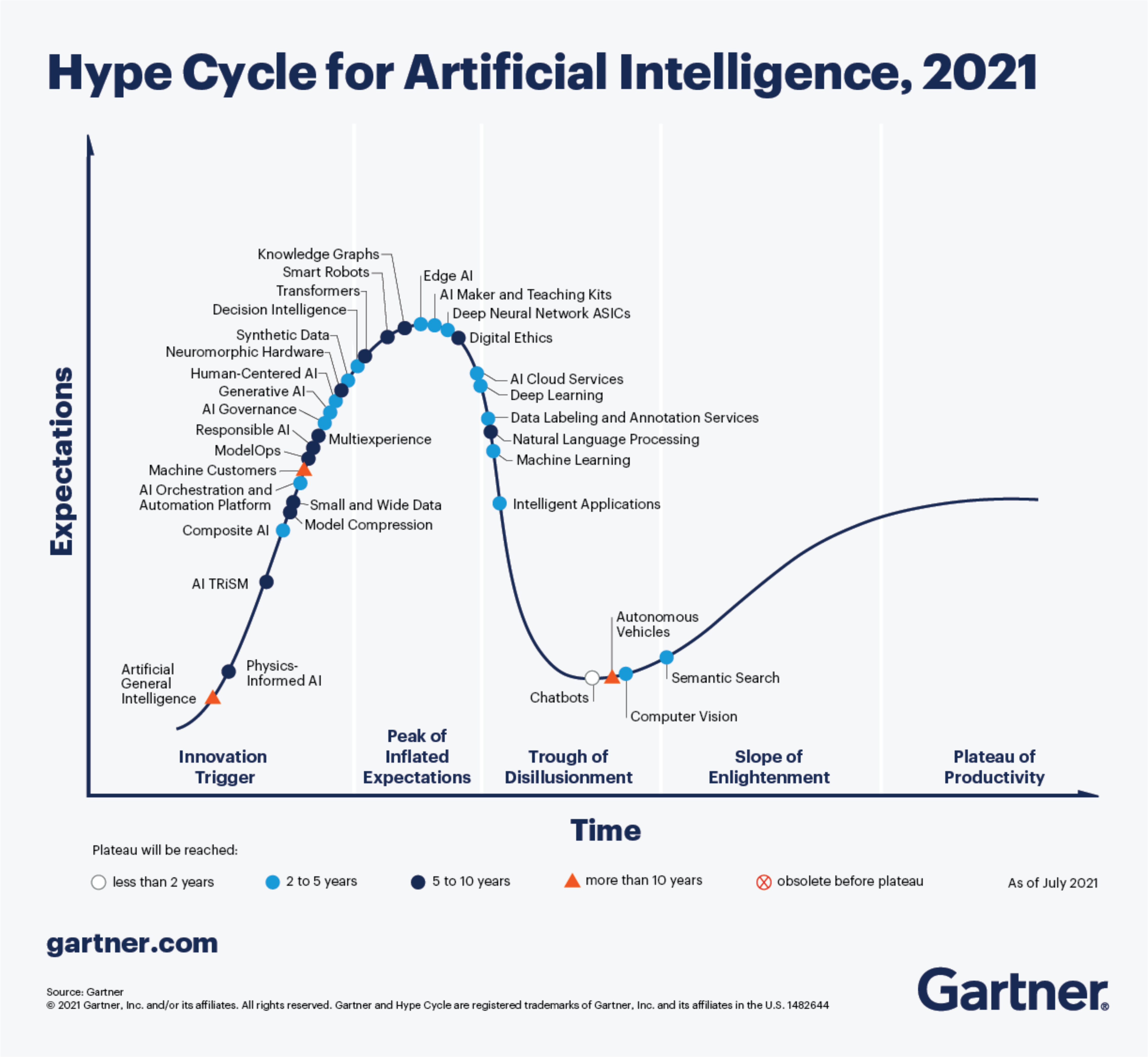
Using XAI+Commonsense

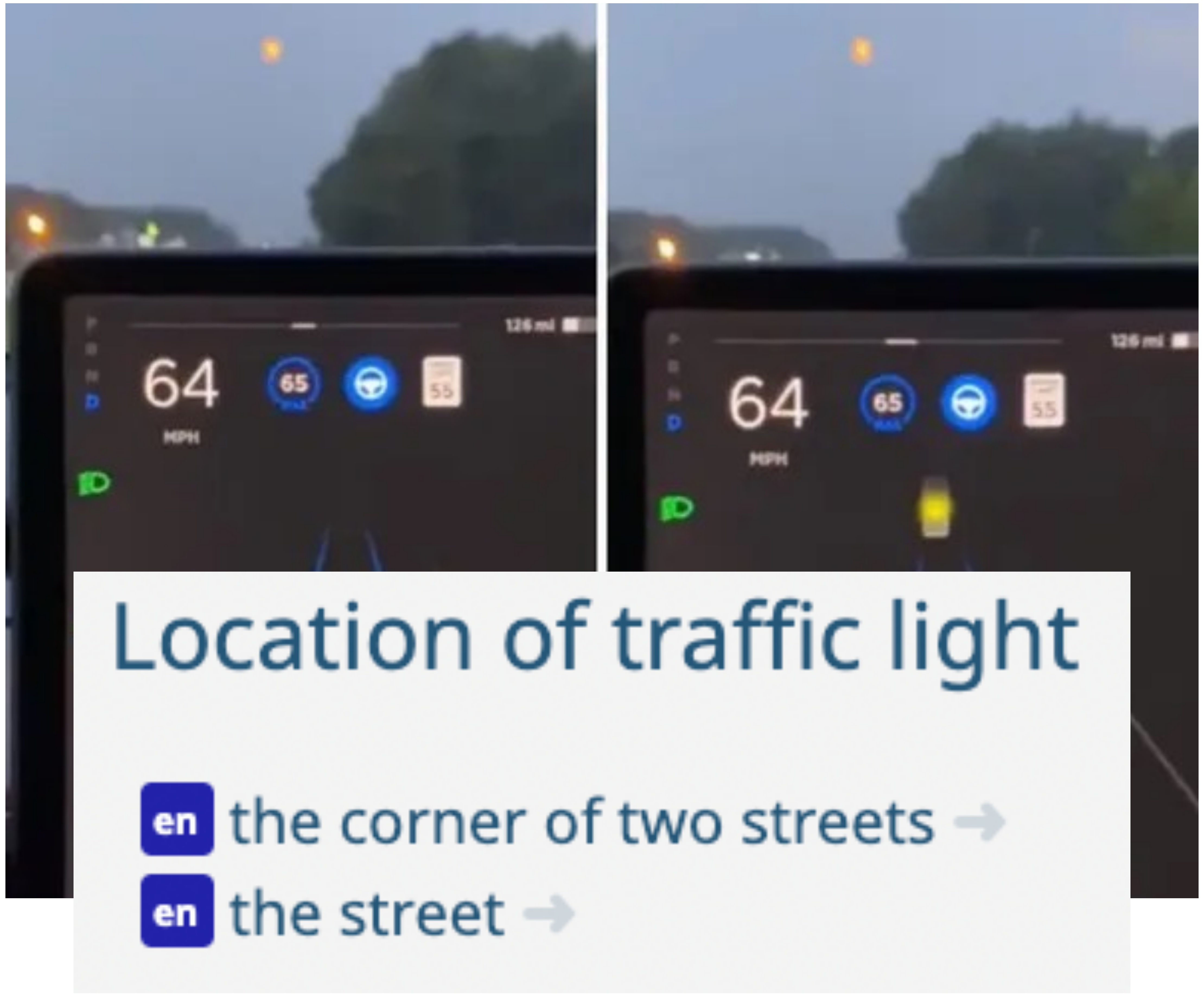




Vision: Articulate Machines

Using XAI+Commonsense





Resources and Future Reading

[1] Gilpin, Leilani. "Reasonableness monitors." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 32. No. 1. 2018.

[2] Gilpin, Leilani H., Jamie C. Macbeth, and Evelyn Florentine. "Monitoring scene understanders with conceptual primitive decomposition and commonsense knowledge." *Advances in Cognitive Systems* 6 (2018): 45-63.

[3] Gilpin, Leilani H., Vishnu Penubarthi, and Lalana Kagal. "Explaining multimodal errors in autonomous vehicles." *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 2021.

Tutorial website: https://yilunzhou.github.io/aaai2023tutorial/