Reflections on Safety and Artificial Intelligence

Eric Horvitz

Exploratory Technical Workshop on Safety and Control for AI
Carnegie Mellon University

Pittsburgh, PA
June 27, 2016

AI & Safety

Constellation of methods referred to as Artificial Intelligence will touch our lives more closely and intimately

AI moving into high-stakes applications

Healthcare

Transportation

Finance

Public policy

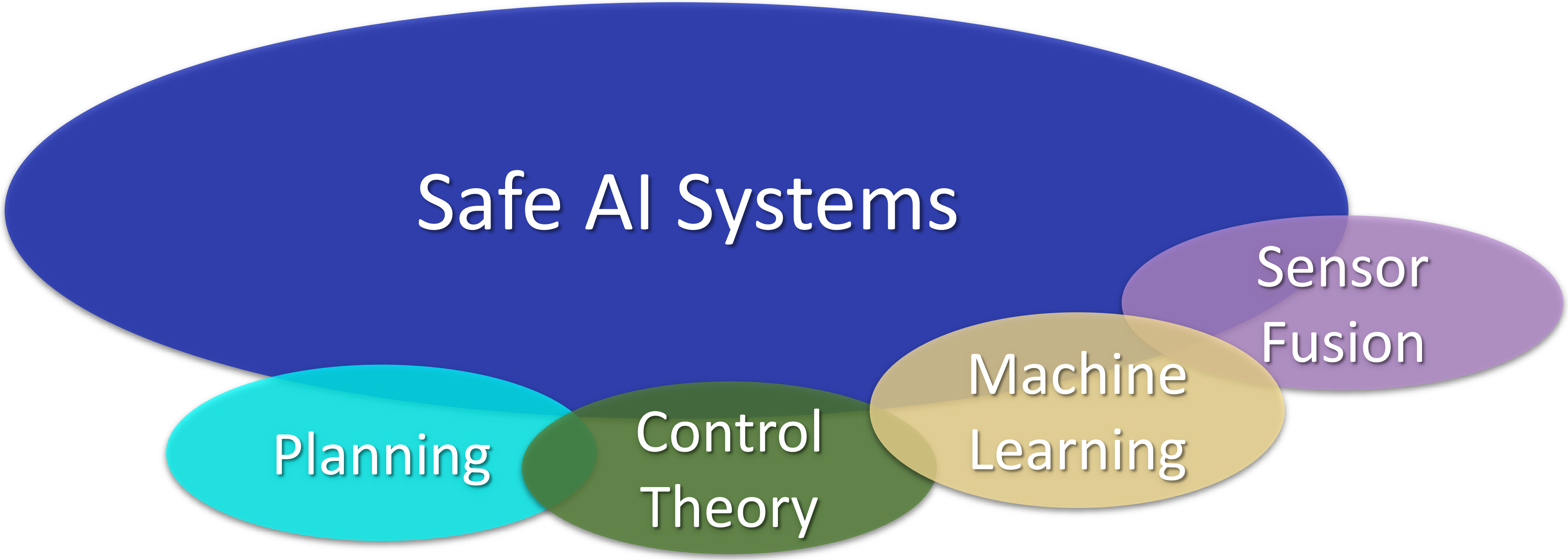
Defense

→ Much to do on principles, methods, and best practices

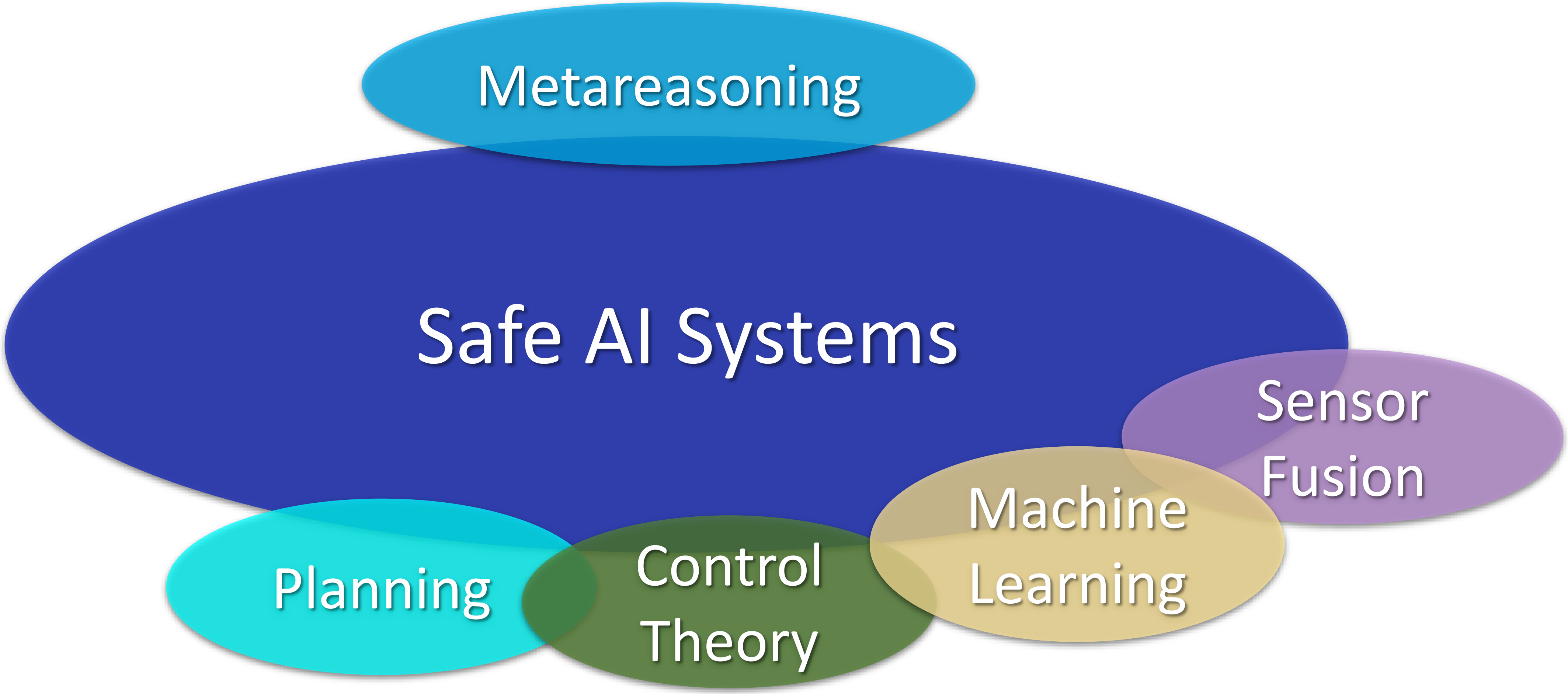
Relevance of Multiple Subdisciplines

Safe AI Systems

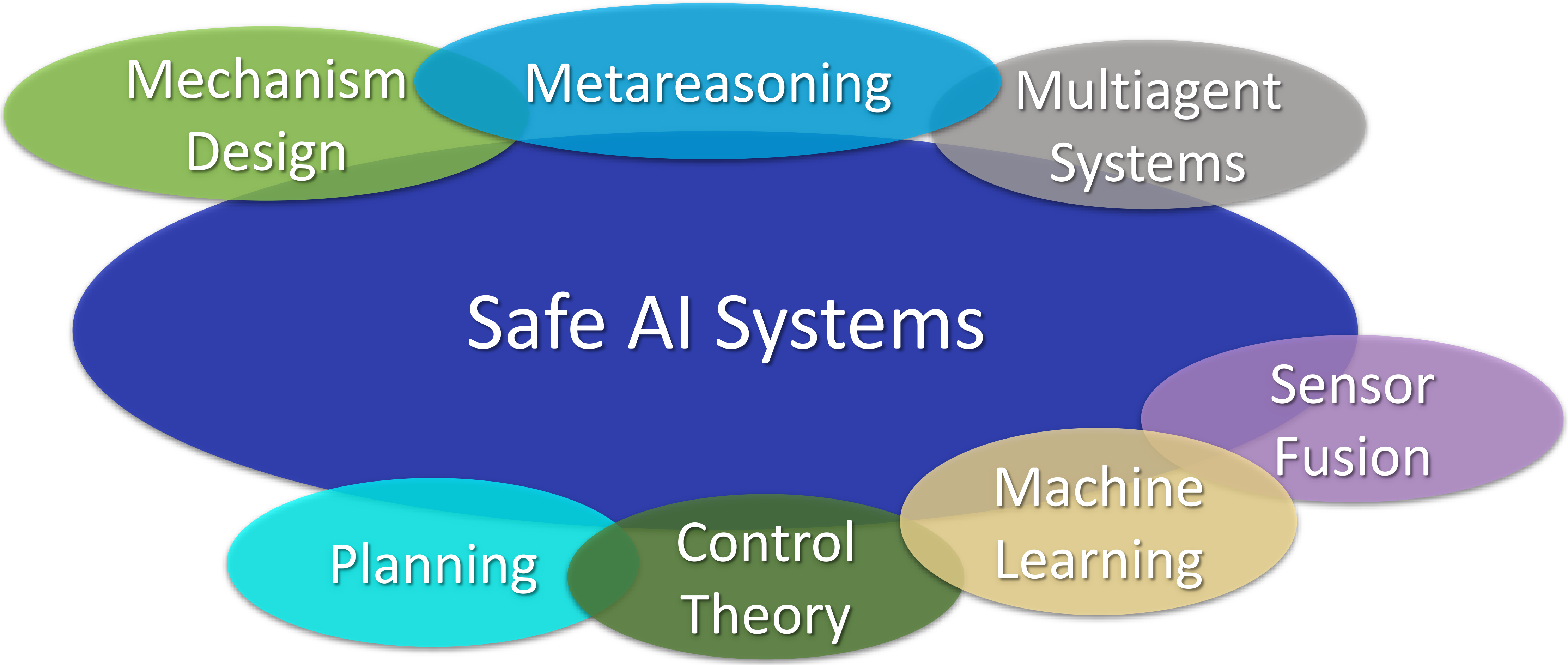
Relevance of Multiple Subdisciplines



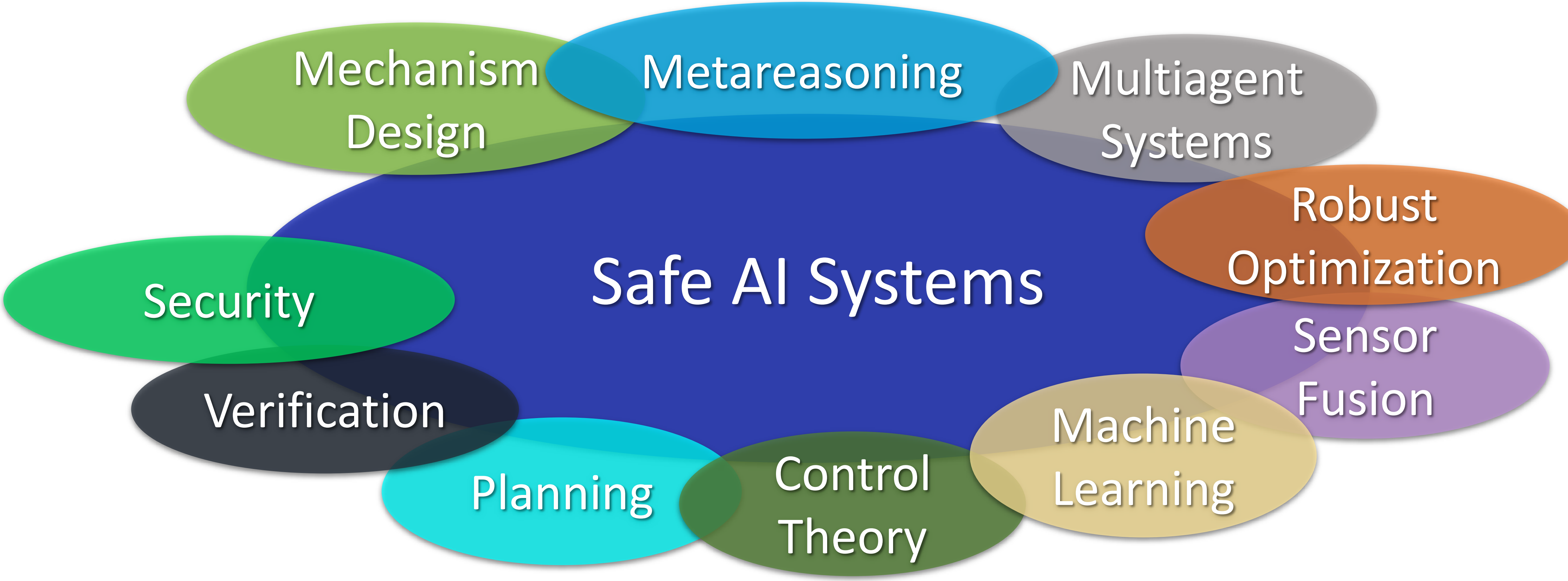
Relevance of Multiple Subdisciplines



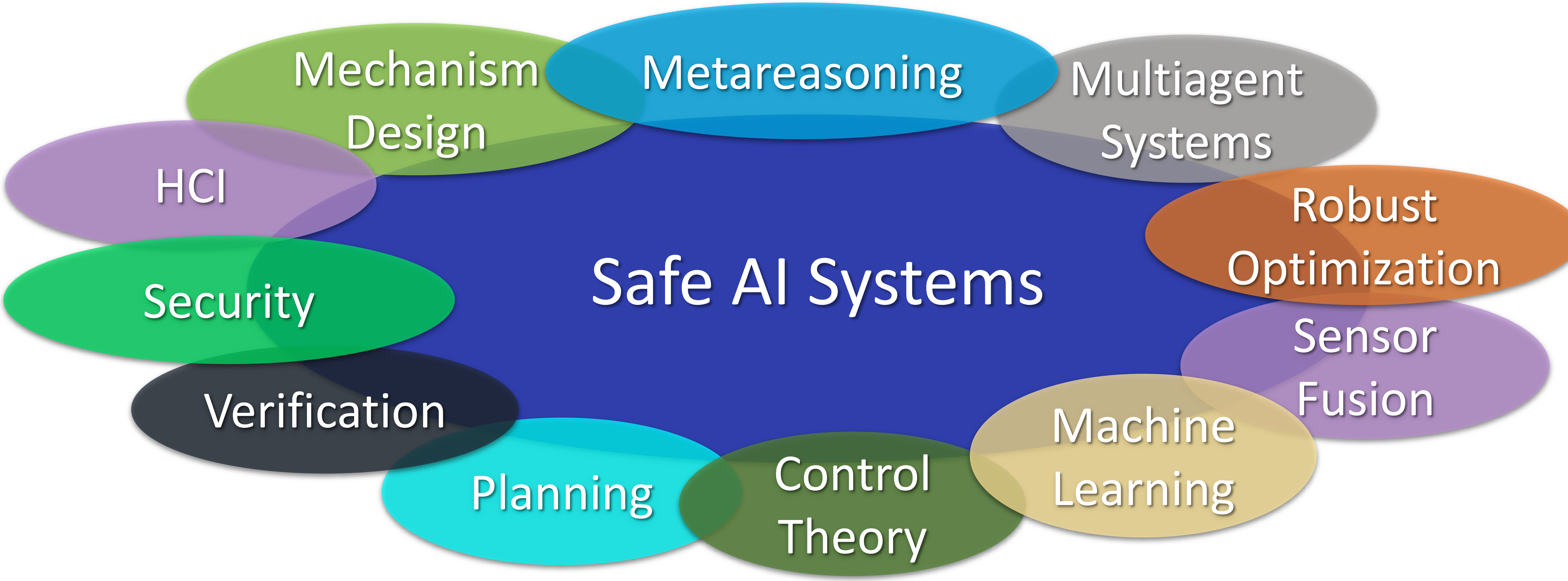
Relevance of Multiple Subdisciplines



Relevance of Multiple Subdisciplines



Relevance of Multiple Subdisciplines



Safe AI Systems

Mechanism Design

Metareasoning

Multiagent Systems

Robust Optimization

Sensor Fusion

Machine Learning

Control Theory

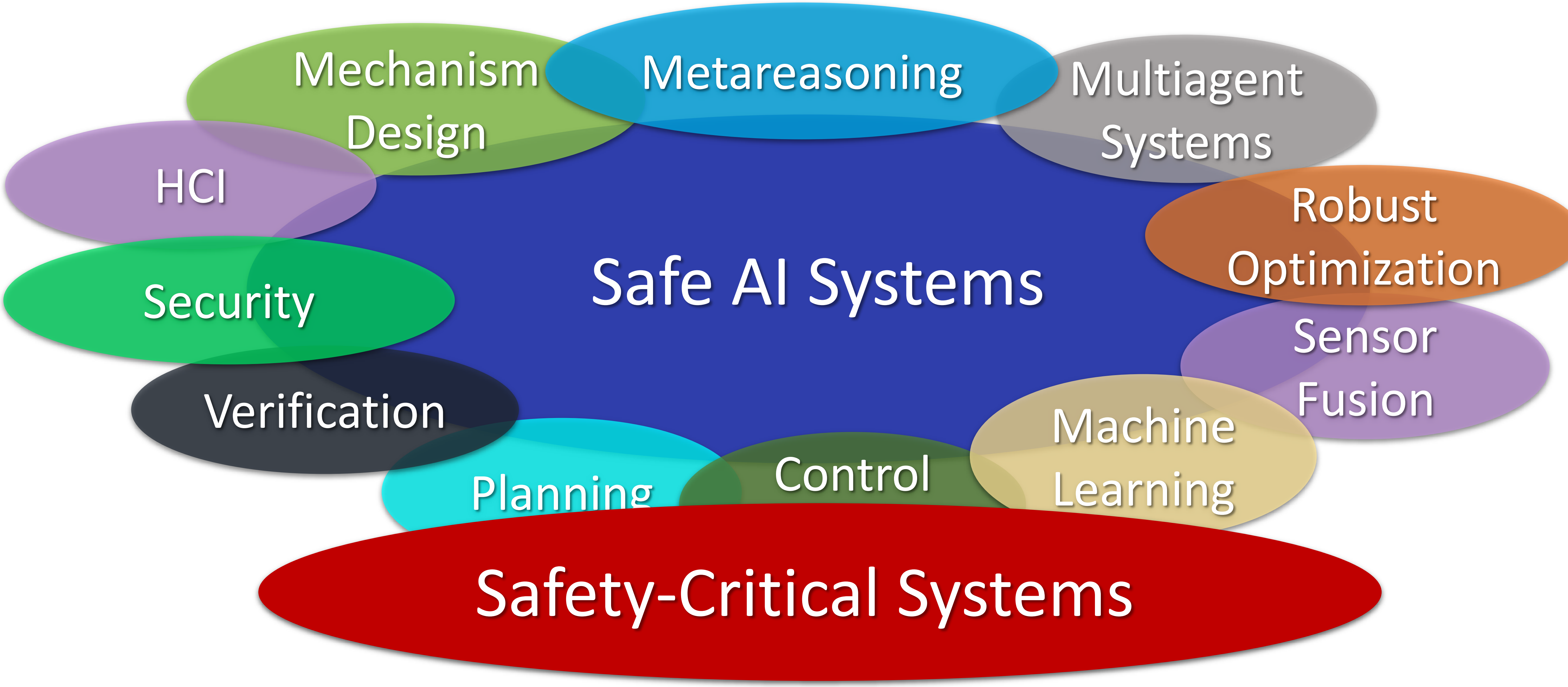
Planning

Verification

Security

HCI

Relevance of Multiple Subdisciplines



safety

ˈsāftē/

noun

1. the condition of being protected from or unlikely to cause danger, risk, or injury

safety-critical ˈsāftēˌkridək(ə)l/

adjective

1. systems whose failure could result in loss of life, significant property damage, or damage to the environment.

2. designed or needing to be fail-safe for safety purposes.

fail-safe \ˈfāl-ˌsāf\

noun

device or practice that, in the event of a failure, responds or results in a way that will cause no harm, or at least minimizes harm.

adjective

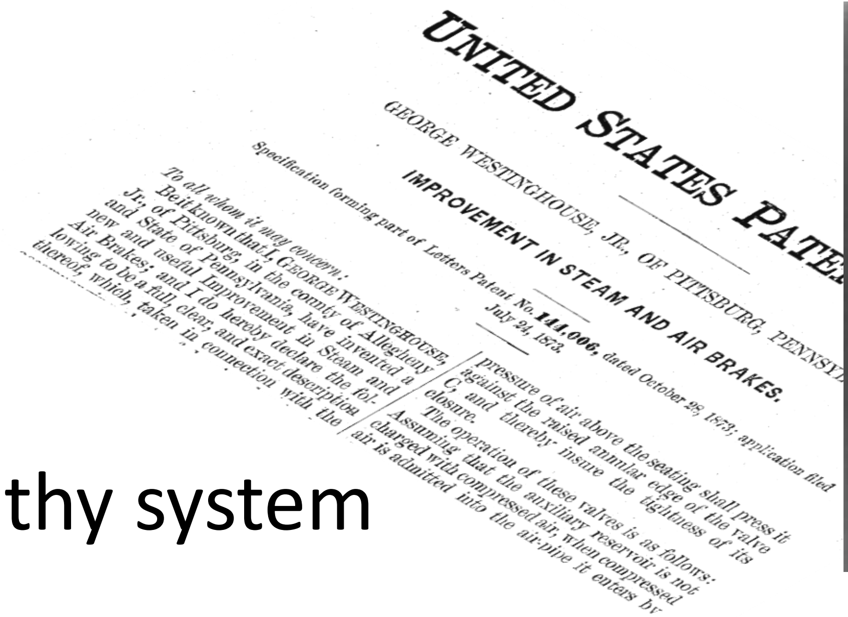
incorporating some feature for automatically counteracting the effect of an anticipated possible source of failure

Fail-safe

George Westinghouse, 1869
Train braking system

Brakes held "off" actively by healthy system

Brakes naturally resort to “on” if any failure of braking system



UNITED STATES PATE

GEORGE WESTINGHOUSE, JR., OF PITTSBURG, PENNSY

IMPROVEMENT IN STEAM AND AIR BRAKES.

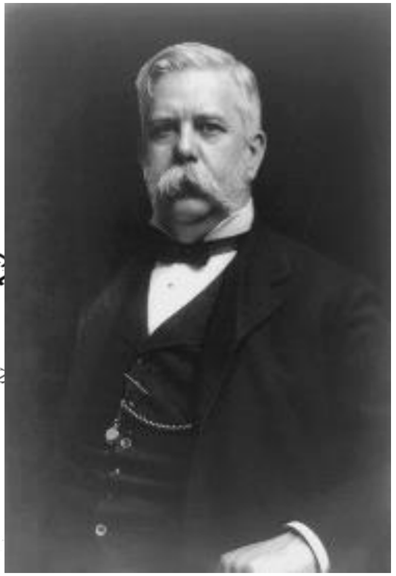
Specification forming part of Letters Patent No. **144,006**, dated October 28, 1873; application filed July 24, 1873.

To all whom it may concern:

Be it known that I, GEORGE WESTINGHOUSE, Jr., of Pittsburg, in the county of Allegheny and State of Pennsylvania, have invented a new and useful Improvement in Steam and Air Brakes; and I do hereby declare the following to be a full, clear, and exact description thereof, which, taken in connection with the

pressure of air above the seating shall press it against the raised annular edge of the valve C, and thereby insure the tightness of its closure.

The operation of these valves is as follows: Assuming that the auxiliary reservoir is not charged with compressed air, when compressed air is admitted into the air-pipe it enters by

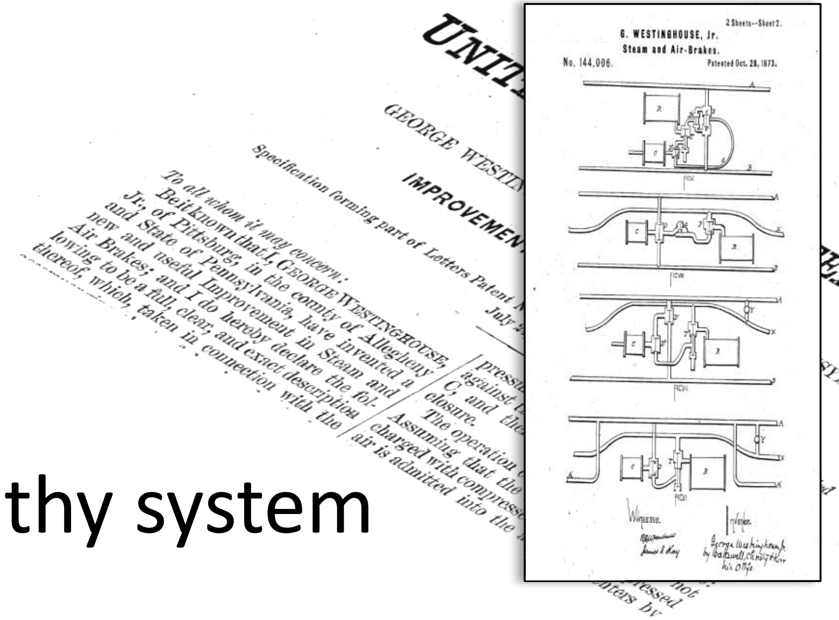


Fail-safe

George Westinghouse, 1869
Train braking system

Brakes held "off" actively by healthy system

Brakes naturally resort to “on” if any failure of braking system



UNIT

GEORGE WESTIN

IMPROVEMENT

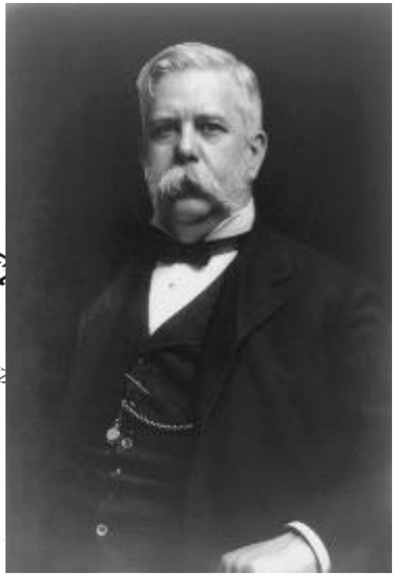
Specification forming part of Letters Patent N

July 2

To all whom it may concern:

Be it known that I, GEORGE WESTINGHOUSE, Jr., of Pittsburg, in the county of Allegheny and State of Pennsylvania, have invented a new and useful Improvement in Steam and Air Brakes; and I do hereby declare the following to be a full, clear, and exact description thereof, which, taken in connection with the

pressu against C, and the closure. The operation Assuming that the charged with compress air is admitted into the



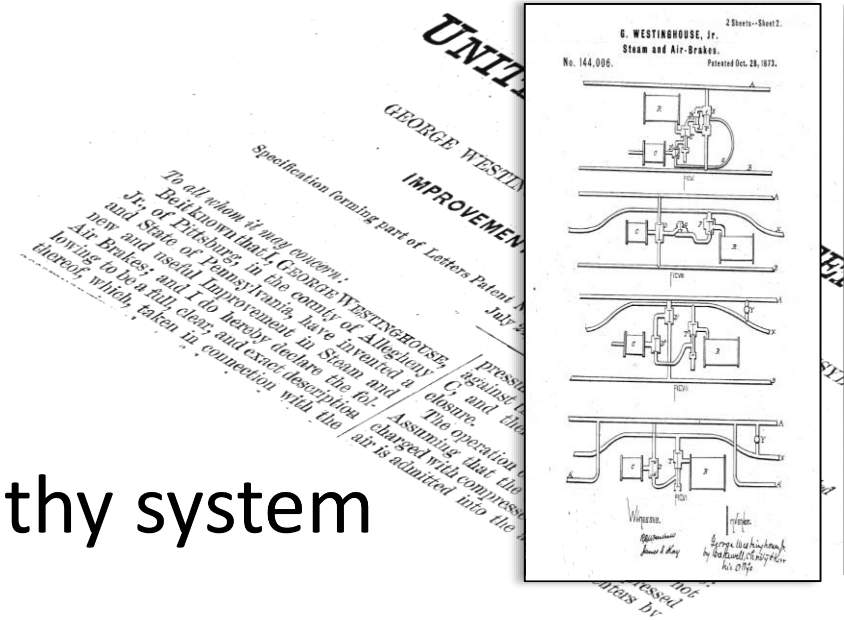
Fail-safe design
Air brakes

Fail-safe

George Westinghouse, 1869
Train braking system

Brakes held "off" actively by healthy system

Brakes naturally resort to “on” if any failure of braking system



UNIT

GEORGE WESTIN

IMPROVEMEN

Specification forming part of Letters Patent N

July 2

To all whom it may concern:

Be it known that I, GEORGE WESTINGHOUSE, Jr., of Pittsburg, in the county of Allegheny and State of Pennsylvania, have invented a new and useful Improvement in Steam and Air Brakes; and I do hereby declare the following to be a full, clear, and exact description thereof, which, taken in connection with the

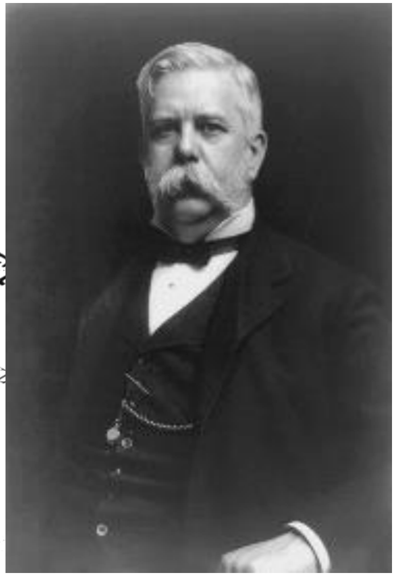
pressu against C, and the closure. The operation Assuming that the charged with compresse air is admitted into the

not ressed nters by

2 Sheets--Sheet 2.

G. WESTINGHOUSE, Jr.
Steam and Air-Brakes.

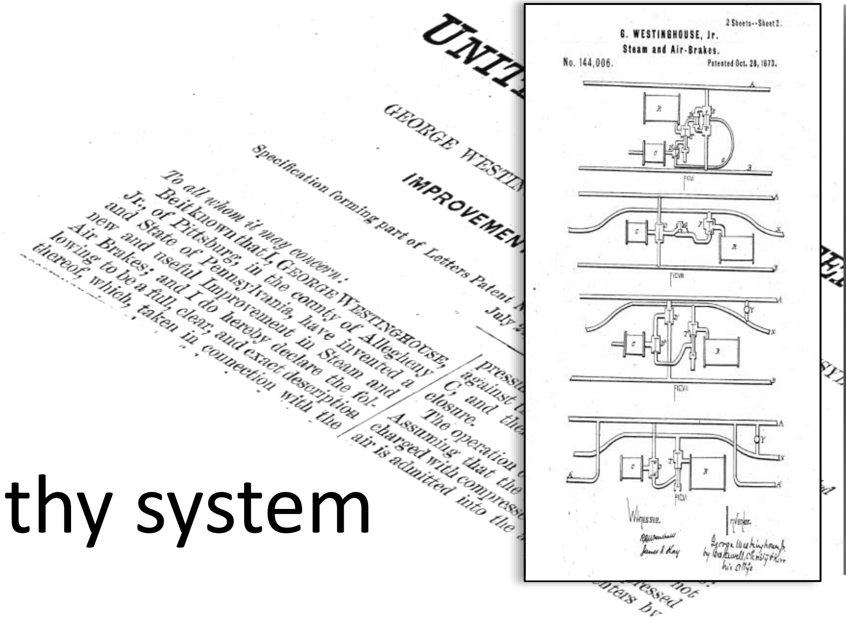
No. 144,006. Patented Oct. 28, 1873.



June 10, 1869
Union Station, Pittsburgh to Steubenville

Fail-safe

George Westinghouse, 1869
Train braking system



UNIT

GEORGE WESTIN

IMPROVEMEN

Specification forming part of Letters Patent N

July 2

To all whom it may concern:

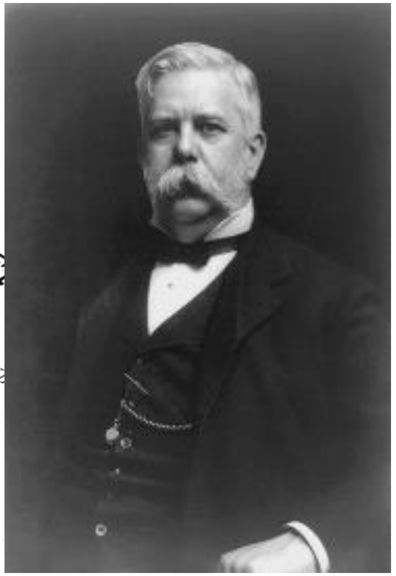
Be it known that I, GEORGE WESTINGHOUSE, Jr., of Pittsburg, in the county of Allegheny and State of Pennsylvania, have invented a new and useful Improvement in Steam and Air Brakes; and I do hereby declare the following to be a full, clear, and exact description thereof, which, taken in connection with the

pressu against C, and the closure. The operation Assuming that the charged with compresse air is admitted into the

2 Sheets--Sheet 2.

G. WESTINGHOUSE, Jr.
Steam and Air-Brakes.

No. 144,006. Patented Oct. 28, 1873.



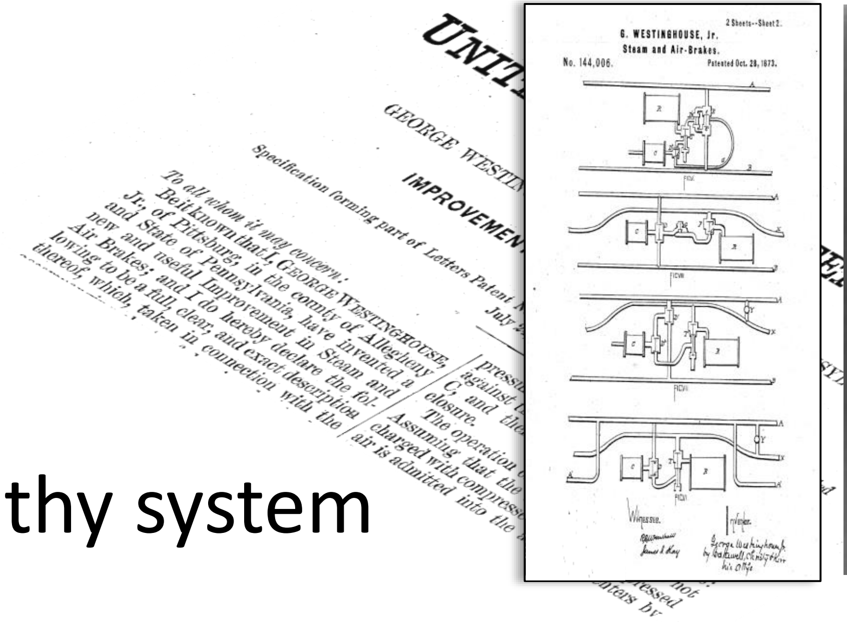
Brakes held "off" actively by healthy system

Brakes naturally resort to “on” if any failure of braking system

Fail-safe practice
Full-power throttle on arrested landing

Fail-safe

George Westinghouse, 1869
Train braking system



UNIT

GEORGE WESTIN

IMPROVEMEN

Specification forming part of Letters Patent N

July 2

To all whom it may concern:

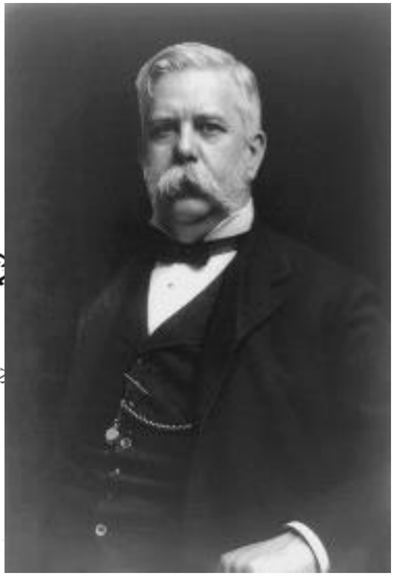
Be it known that I, GEORGE WESTINGHOUSE, Jr., of Pittsburg, in the county of Allegheny and State of Pennsylvania, have invented a new and useful Improvement in Steam and Air Brakes; and I do hereby declare the following to be a full, clear, and exact description thereof, which, taken in connection with the

2 Sheets--Sheet 2.

G. WESTINGHOUSE, Jr.

Steam and Air-Brakes.

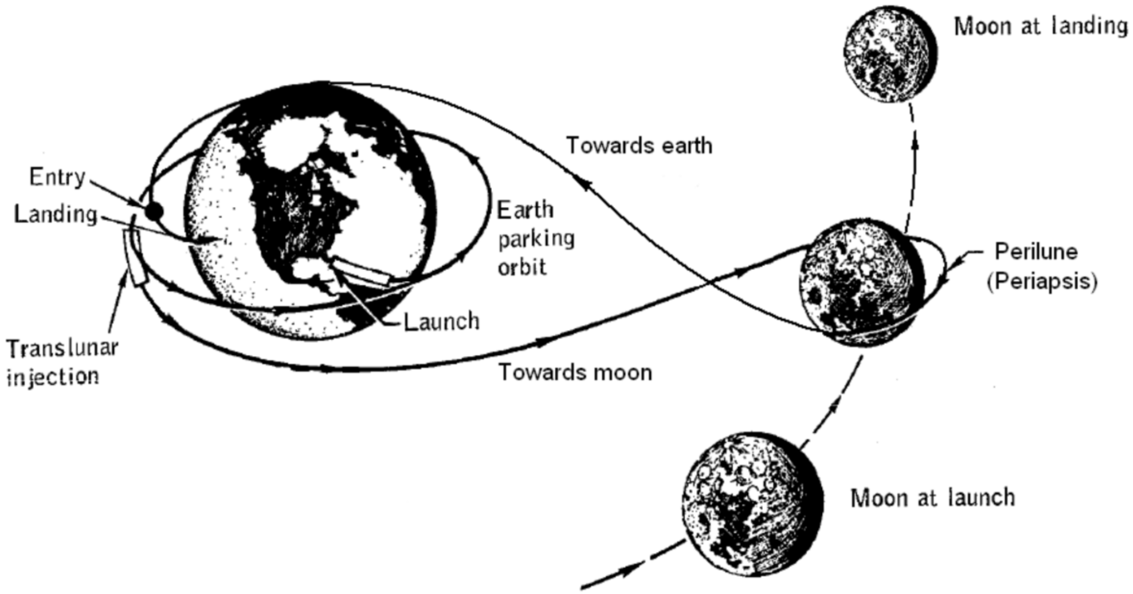
No. 144,006. Patented Oct. 28, 1873.



Brakes held "off" actively by healthy system

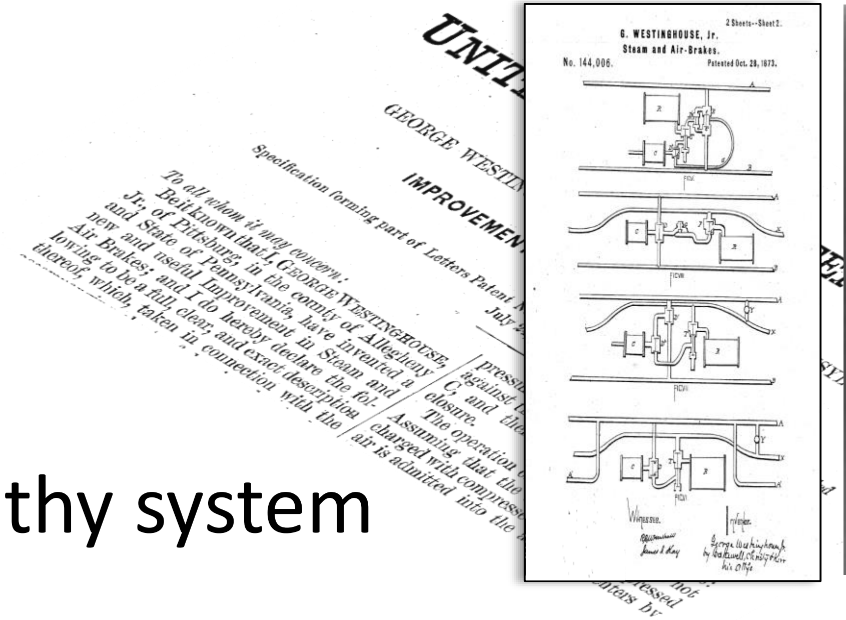
Brakes naturally resort to “on” if any failure of braking system

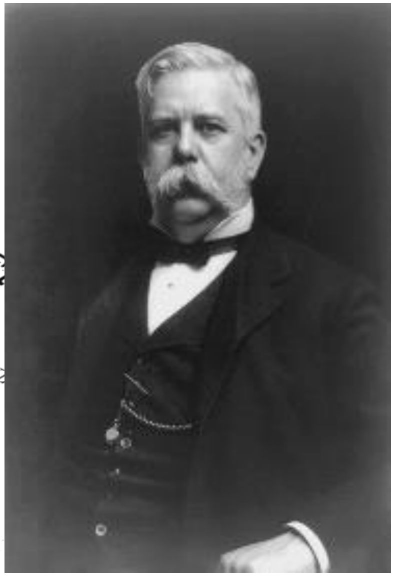
Fail-safe plan
Free return trajectory



Fail-safe

George Westinghouse, 1869
Train braking system


UNIT
GEORGE WESTIN
IMPROVEMENT
2 Sheets--Sheet 2.
G. WESTINGHOUSE, Jr.
Steam and Air-Brakes.
No. 144,006.
Patented Oct. 28, 1873.

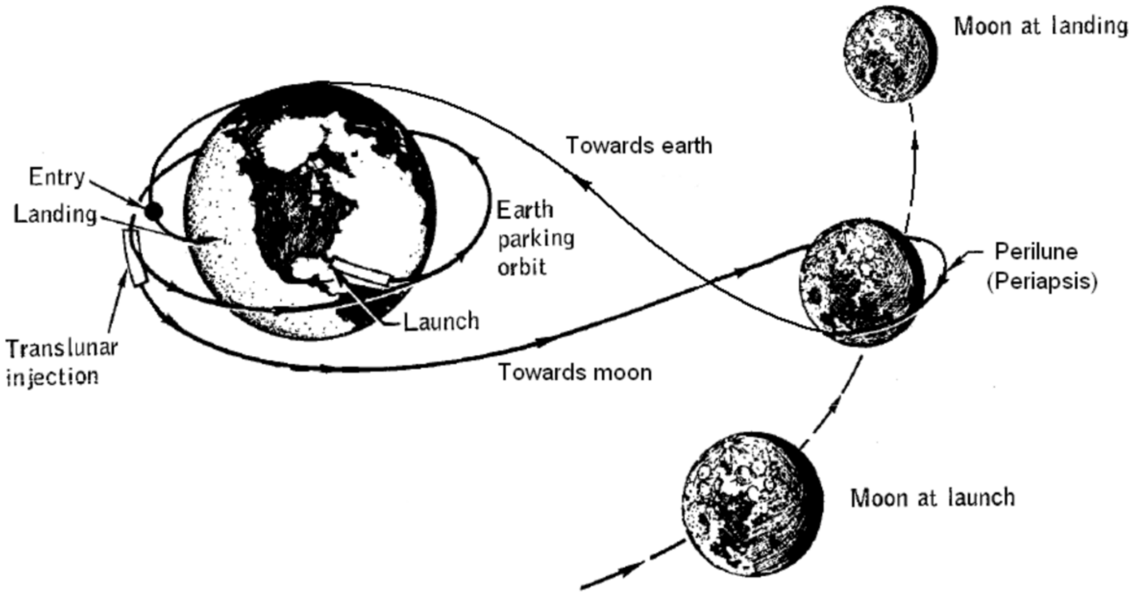


Brakes held "off" actively by healthy system

Brakes naturally resort to “on” if any failure of braking system

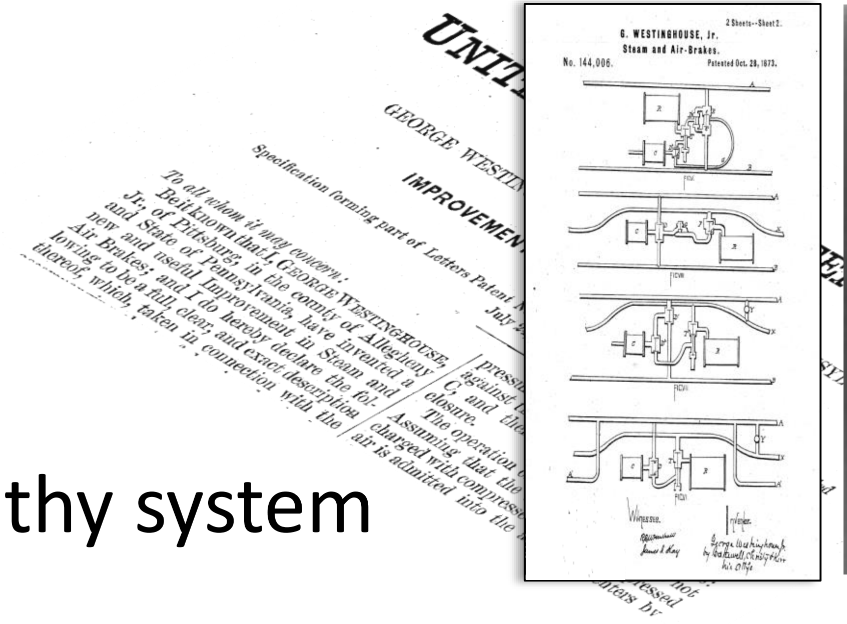
- ✓ Mechanism
- ✓ Practice
- ✓ Plan

Fail-safe plan
Free return trajectory



Fail-safe

George Westinghouse, 1869
Train braking system



UNIT

GEORGE WESTIN

IMPROVEMENT

Specification forming part of Letters Patent N

July 2

To all whom it may concern:

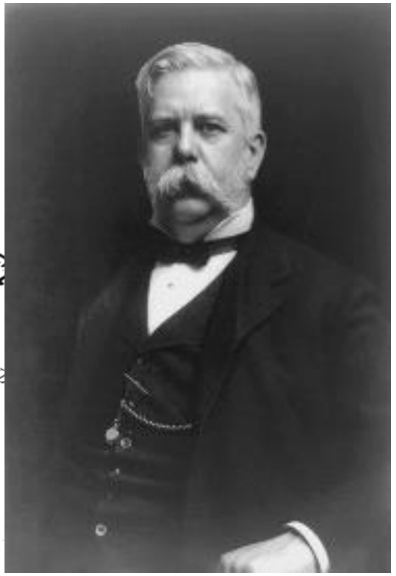
Be it known that I, GEORGE WESTINGHOUSE, Jr., of Pittsburg, in the county of Allegheny and State of Pennsylvania, have invented a new and useful Improvement in Steam and Air Brakes; and I do hereby declare the following to be a full, clear, and exact description thereof, which, taken in connection with the

pressu
against t
C, and the
closure.
The operation
Assuming that the
charged with compress
air is admitted into the

2 Sheets--Sheet 2.

G. WESTINGHOUSE, Jr.
Steam and Air-Brakes.

No. 144,006. Patented Oct. 28, 1873.



Brakes held "off" actively by healthy system

Brakes naturally resort to “on” if any failure of braking system

- ✓ Mechanism
- ✓ Practice
- ✓ Plan

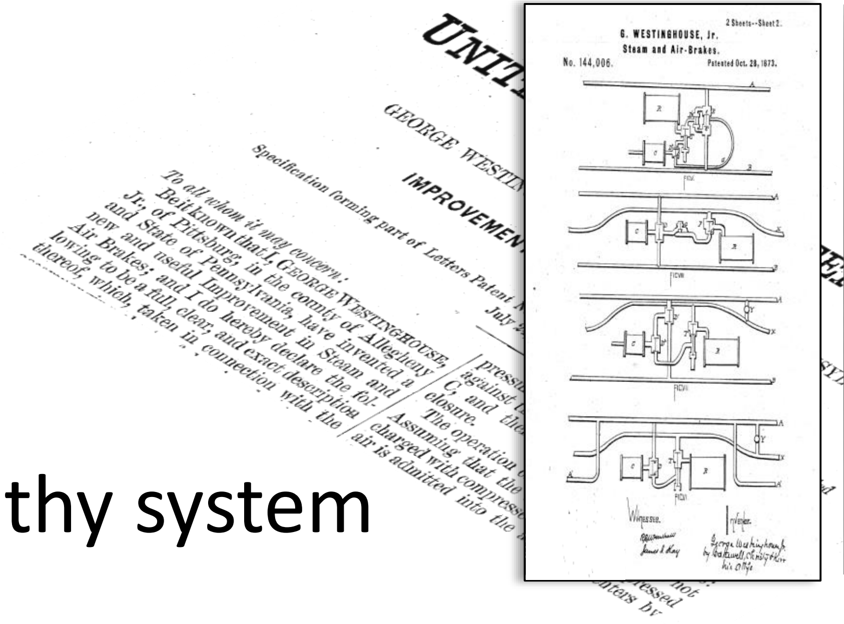
Fail-safe

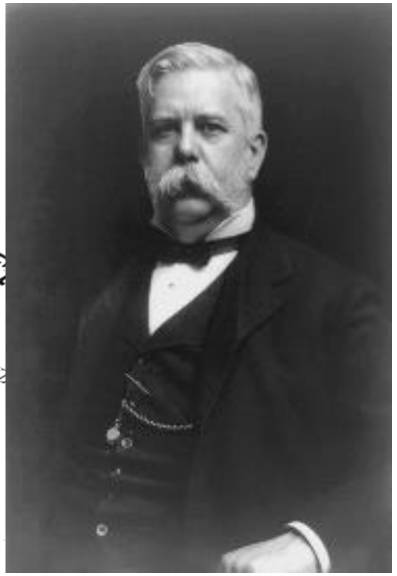
George Westinghouse, 1869
Train braking system

Brakes held "off" actively by healthy system

Brakes naturally resort to “on” if any failure of braking system

- ✓ Mechanism
- ✓ Practice
- ✓ Plan
- ✓ Monitoring


2 Sheets--Sheet 2.
G. WESTINGHOUSE, Jr.
Steam and Air-Brakes.
No. 144,006.
Patented Oct. 28, 1873.



AI in the Open World

Growing interest in issues & directions with AI in real-world settings

Grappling with uncertainty and more general incompleteness

AAAI President’s address (2008), “Artificial Intelligence in the Open World.”

AAAI President’s address (2016), “Steps Toward Robust Artificial Intelligence.”

E. Horvitz. Artificial Intelligence in the Open World, *AAAI President’s Address*, Chicago, IL, July 2008.

T. Dietterich, Steps Toward Robust Artificial Intelligences, *AAAI President's Address*, Phoenix, AX. February, 2016.

Special Considerations with AI

Open-world complexity → incomplete understanding

Uncertainties & poor-characterization of performance

Poor operating regimes, unfamiliar situations

Special Considerations with AI

Open-world complexity → incomplete understanding

Uncertainties & poor-characterization of performance

Poor operating regimes, unfamiliar situations

Rich ontology of failures

Numerous failure modalities

New attack surfaces (e.g., *machine learning attack*)

Self-modification & gaming (e.g., *modify reward fcn*)

Unmodeled influences

Special Considerations with AI

Open-world complexity → incomplete understanding

Uncertainties & poor-characterization of performance
Poor operating regimes, unfamiliar situations

Rich ontology of failures

Numerous failure modalities
New attack surfaces (e.g., *machine learning attack*)
Self-modification & gaming (e.g., *modify reward fcn*)
Unmodeled influences

Challenges of transfer across time & space

Challenge of coordinating human-machine collaborations

Operational opacity

AI & Open-World Complexity

Frame problem

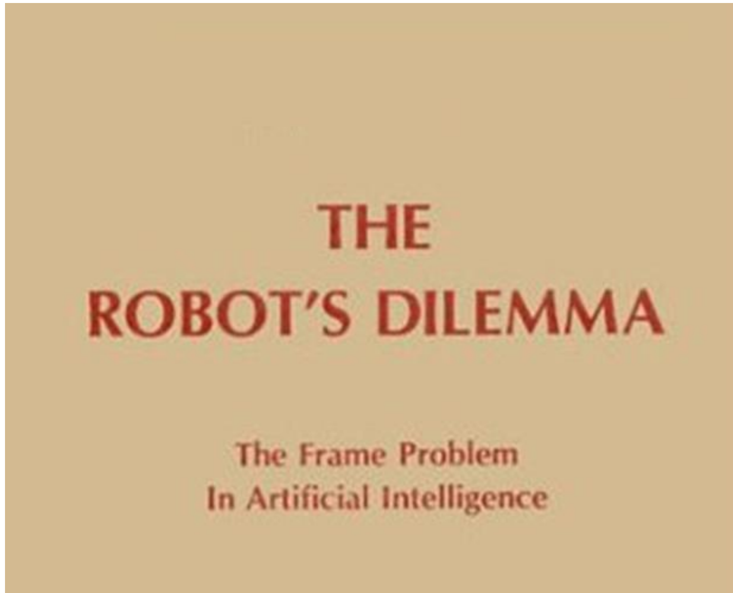
How to tractably derive consequences of an action?

Qualification problem

Understanding preconditions required for actions to have intended effects

Ramification problem

Understanding all important effects of action


THE ROBOT'S DILEMMA

The Frame Problem In Artificial Intelligence

AI & Open-World Complexity

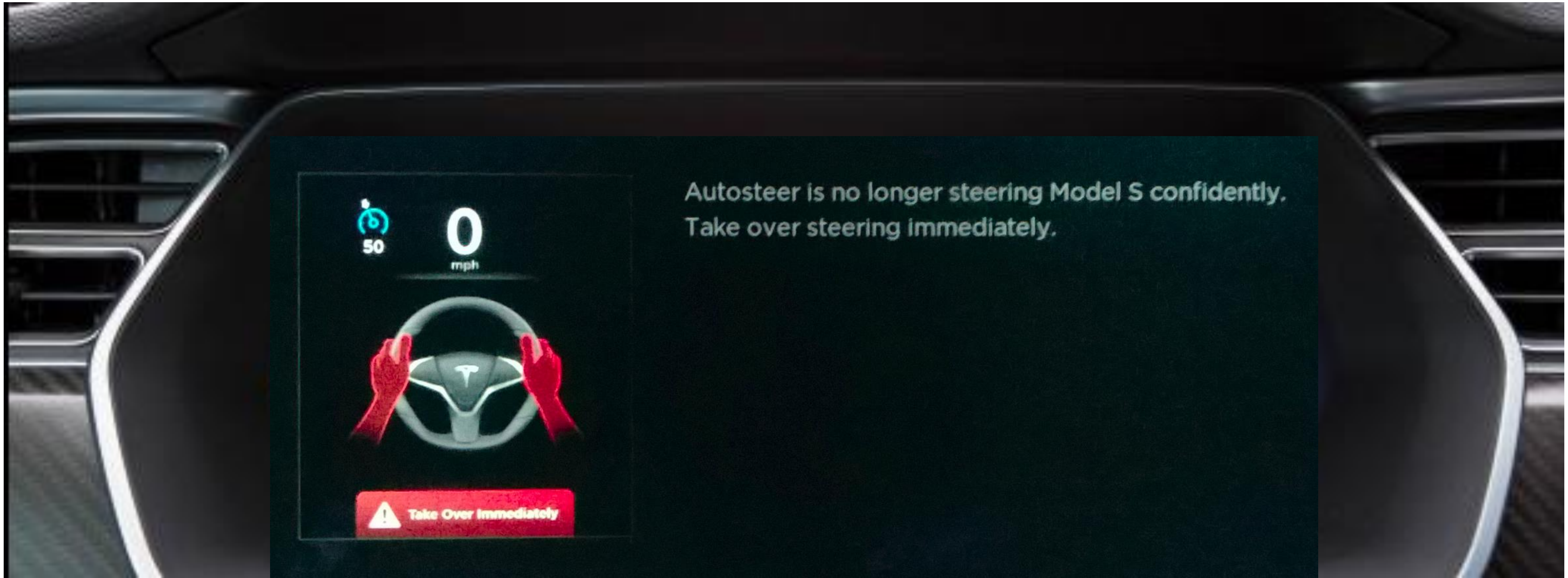
Rise of probabilistic methods: *known unknowns*

Recent attention to *unknown unknowns*

AI & Open-World Complexity

Rise of probabilistic methods: *known unknowns*

Recent attention to *unknown unknowns*

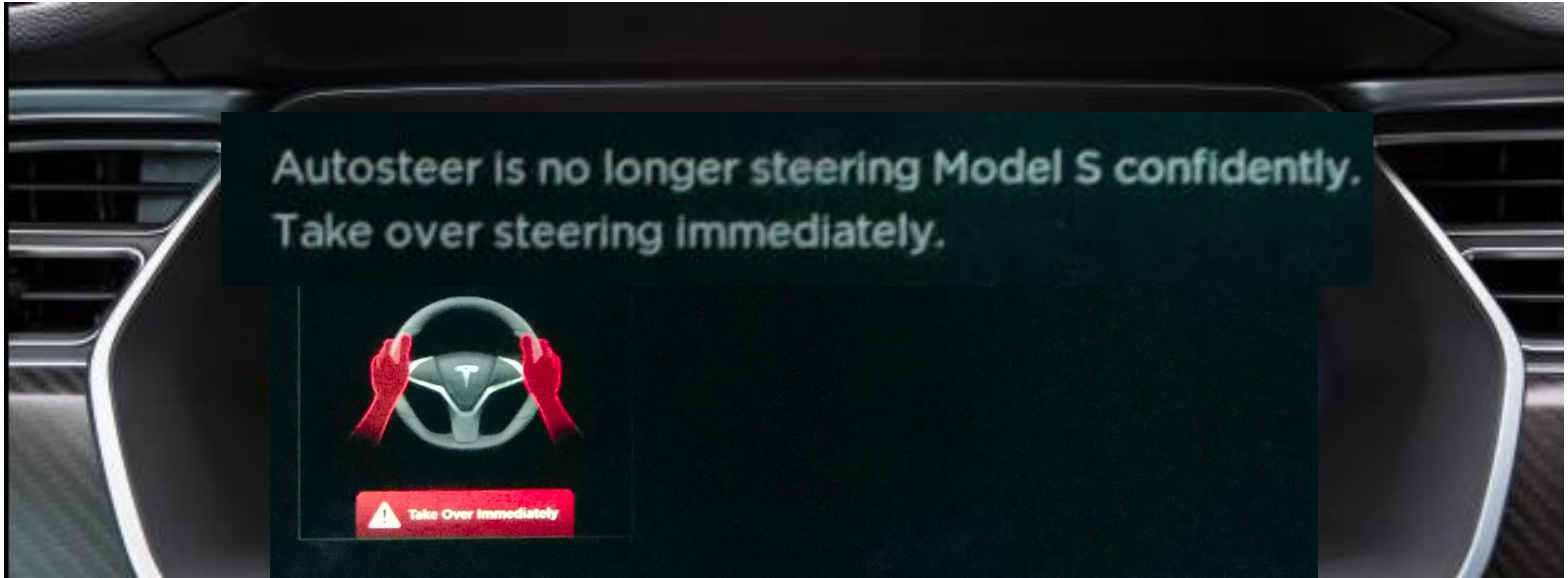


Autosteer is no longer steering Model S confidently. Take over steering immediately.

AI & Open-World Complexity

Rise of probabilistic methods: *known unknowns*

Recent attention to *unknown unknowns*



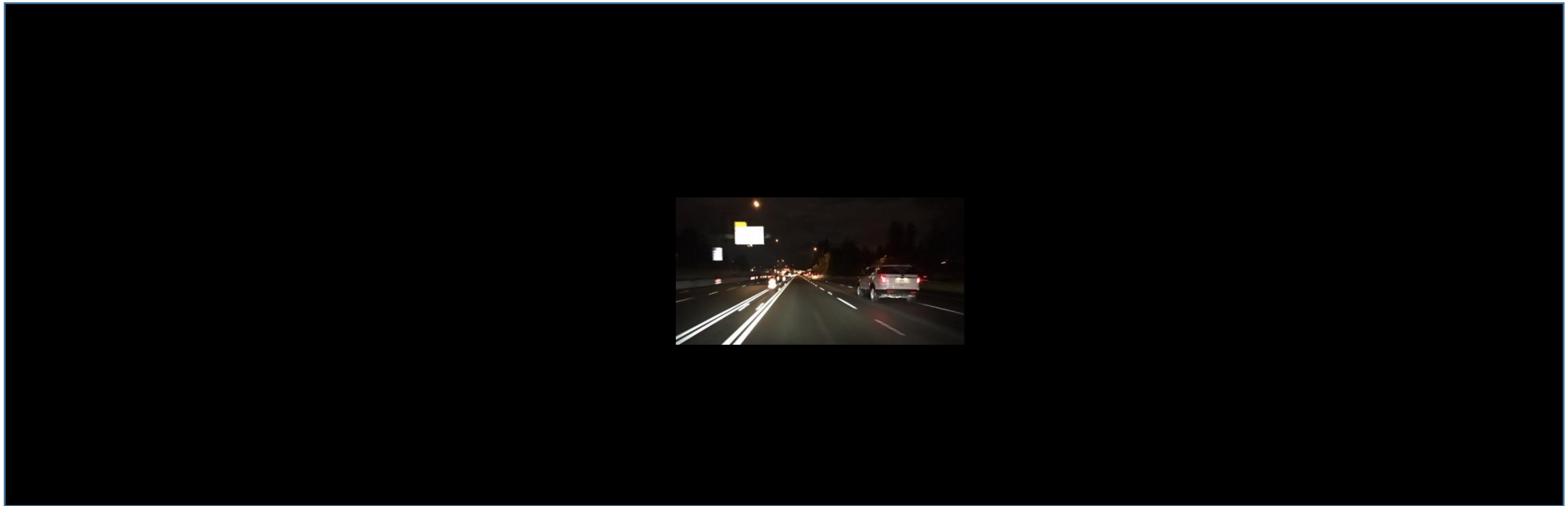
Autosteer is no longer steering Model S confidently.
Take over steering immediately.

AI & Open-World Complexity

Rise of probabilistic methods: *known unknowns*

Recent attention to *unknown unknowns*

Decision making under uncertainty & incompleteness

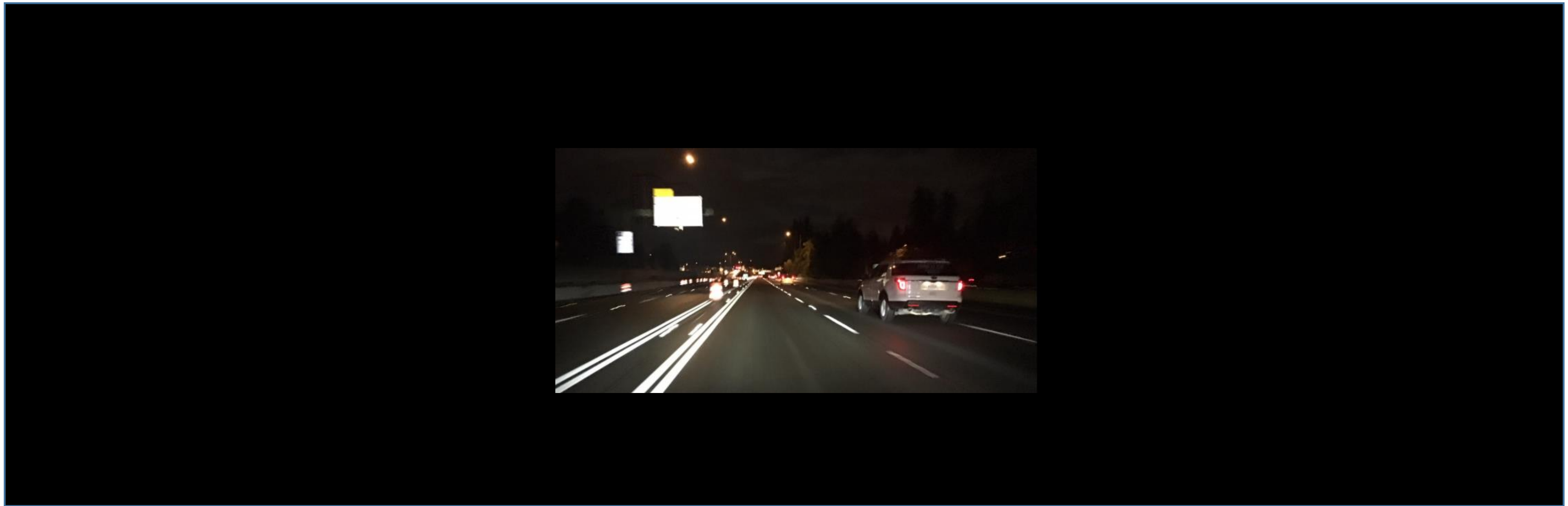


AI & Open-World Complexity

Rise of probabilistic methods: *known unknowns*

Recent attention to *unknown unknowns*

Decision making under uncertainty & incompleteness



AI & Open-World Complexity

Rise of probabilistic methods: *known unknowns*

Recent attention to *unknown unknowns*

Decision making under uncertainty & incompleteness

AI & Open-World Complexity

Rise of probabilistic methods: *known unknowns*

Recent attention to *unknown unknowns*

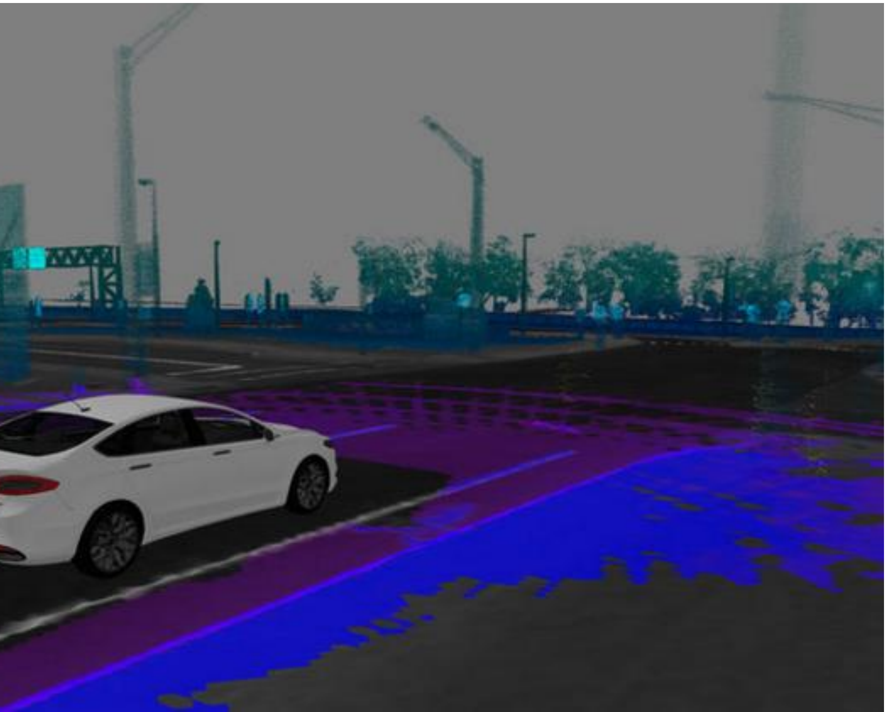
Decision making under uncertainty & incompleteness

AI & Open-World Complexity

Rise of probabilistic methods: *known unknowns*

Recent attention to *unknown unknowns*

Decision making under uncertainty & incompleteness

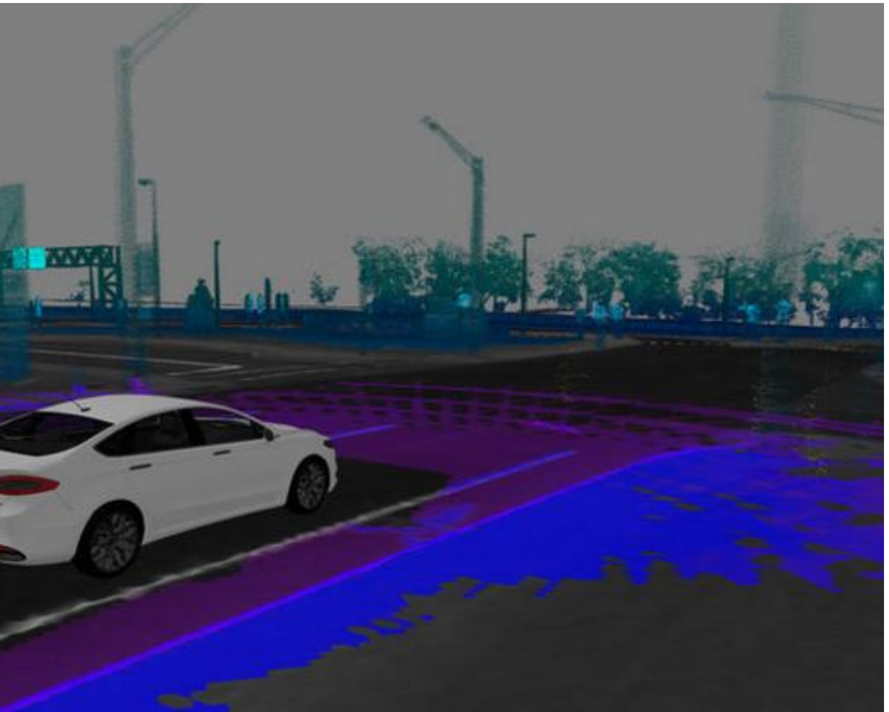


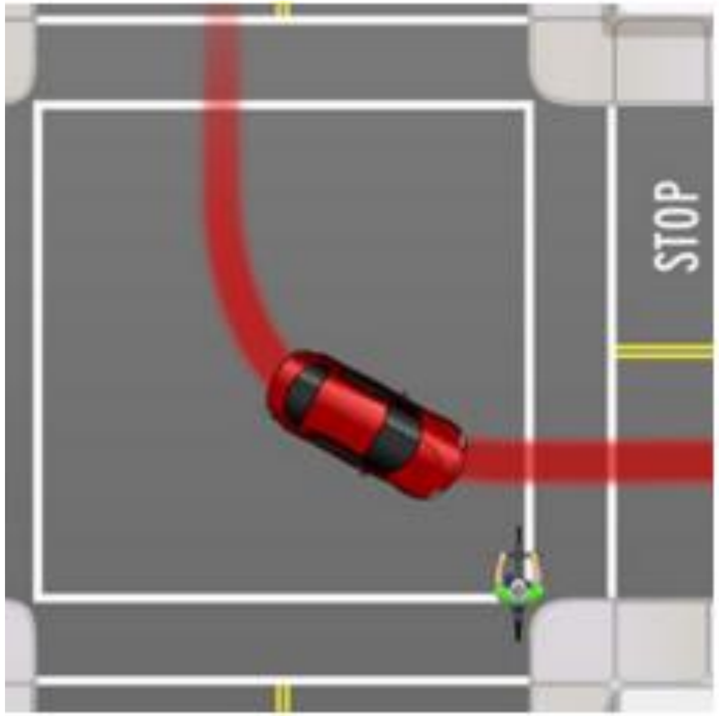
AI & Open-World Complexity

Rise of probabilistic methods: *known unknowns*

Recent attention to *unknown unknowns*

Decision making under uncertainty & incompleteness



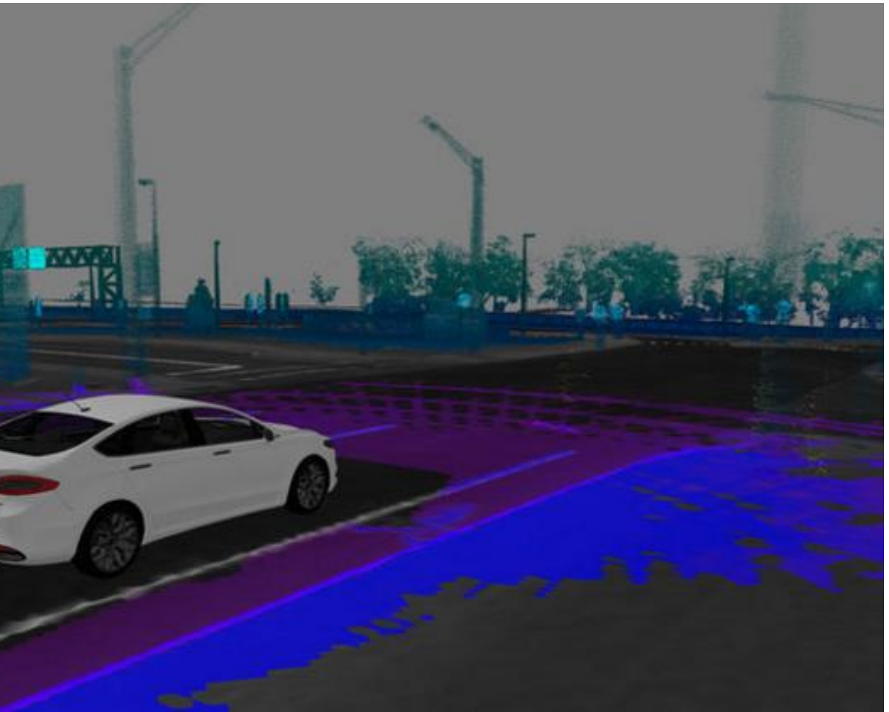


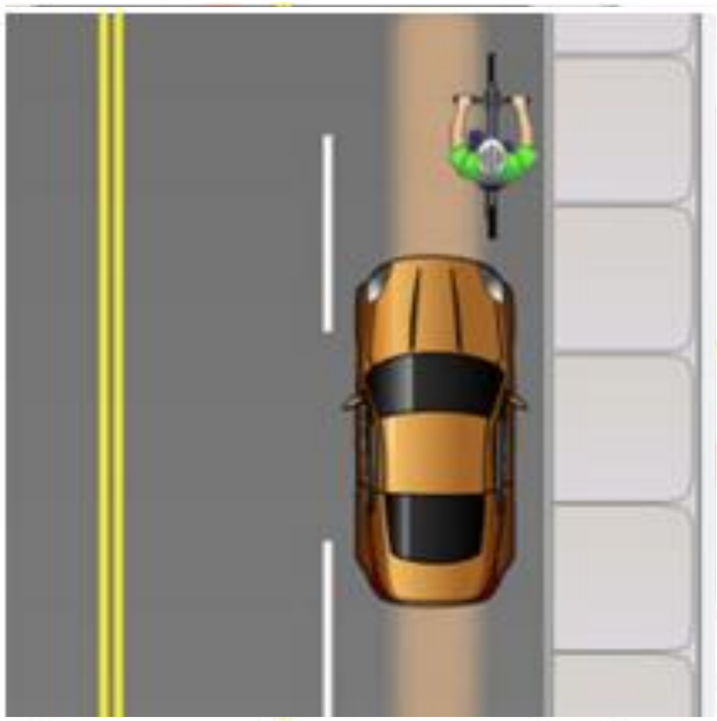
AI & Open-World Complexity

Rise of probabilistic methods: *known unknowns*

Recent attention to *unknown unknowns*

Decision making under uncertainty & incompleteness



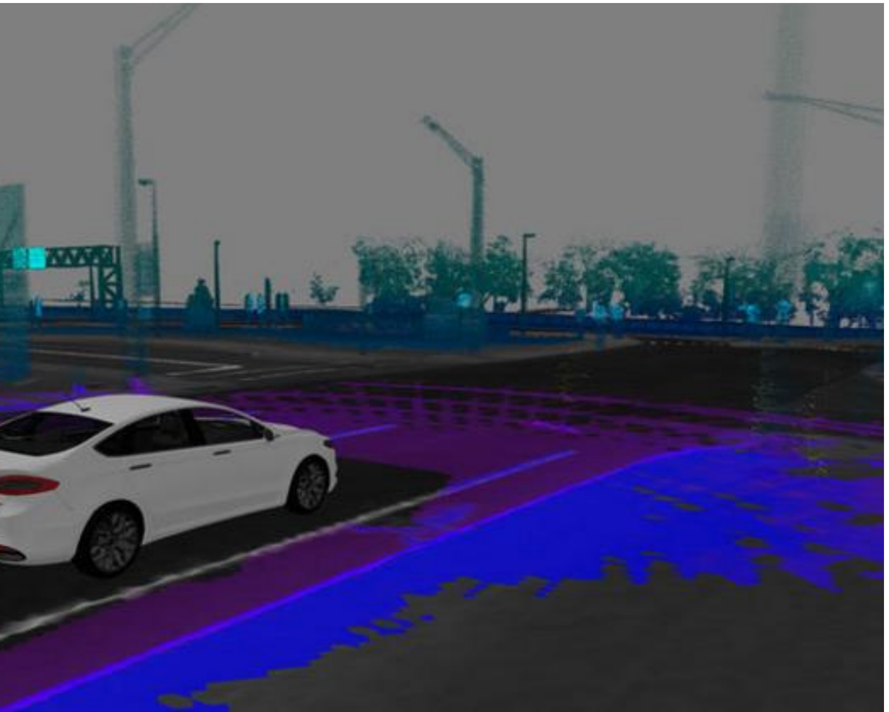


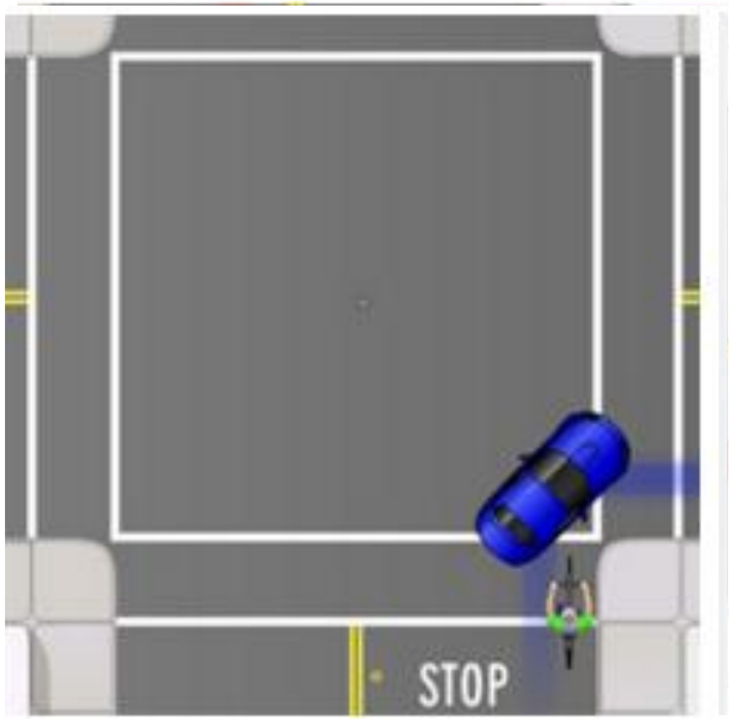
AI & Open-World Complexity

Rise of probabilistic methods: *known unknowns*

Recent attention to *unknown unknowns*

Decision making under uncertainty & incompleteness



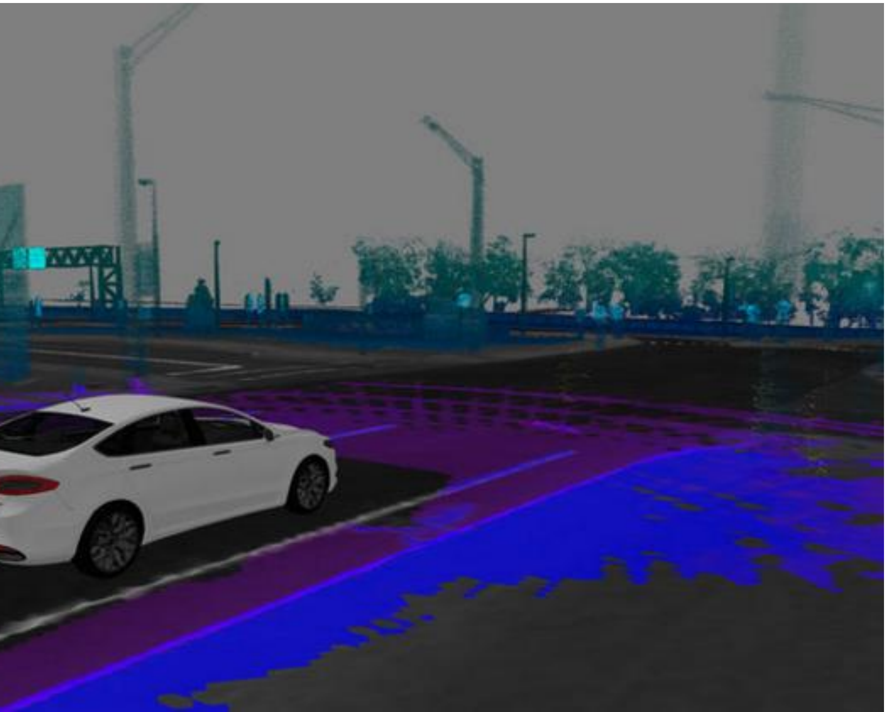


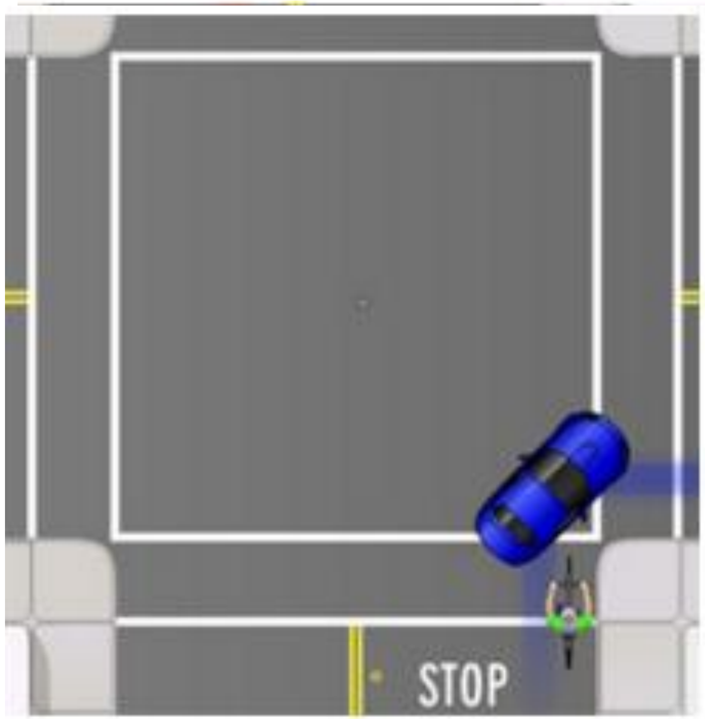
AI & Open-World Complexity

Rise of probabilistic methods: *known unknowns*

Recent attention to *unknown unknowns*

Decision making under uncertainty & incompleteness



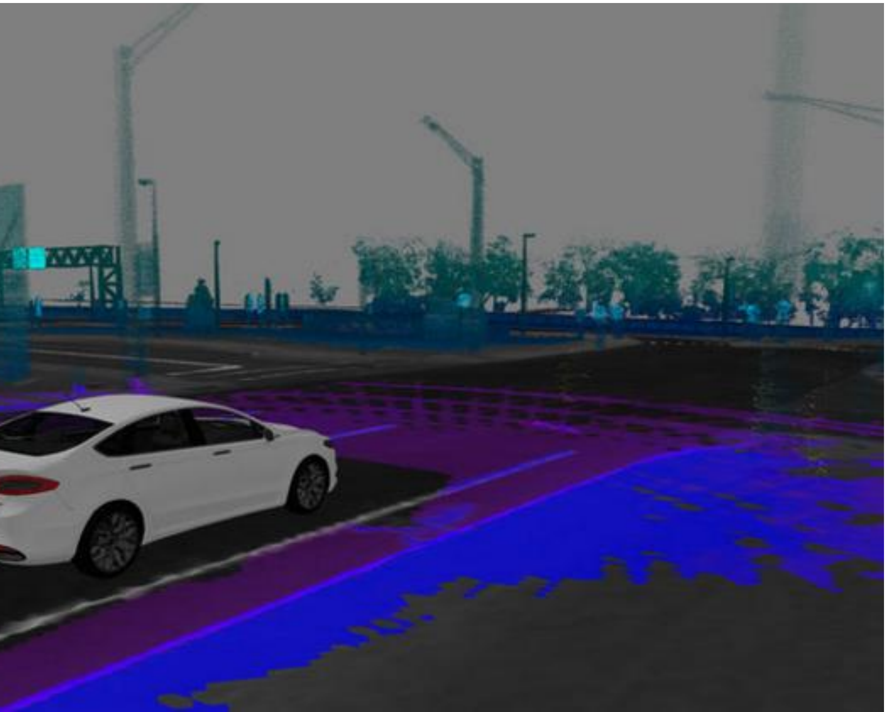


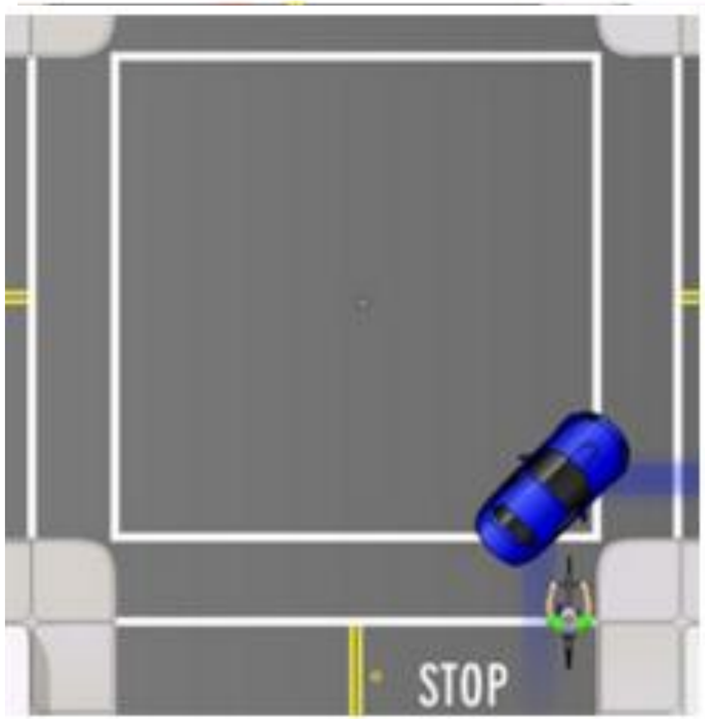
AI & Open-World Complexity

Rise of probabilistic methods: *known unknowns*

Recent attention to *unknown unknowns*

Decision making under uncertainty & incompleteness



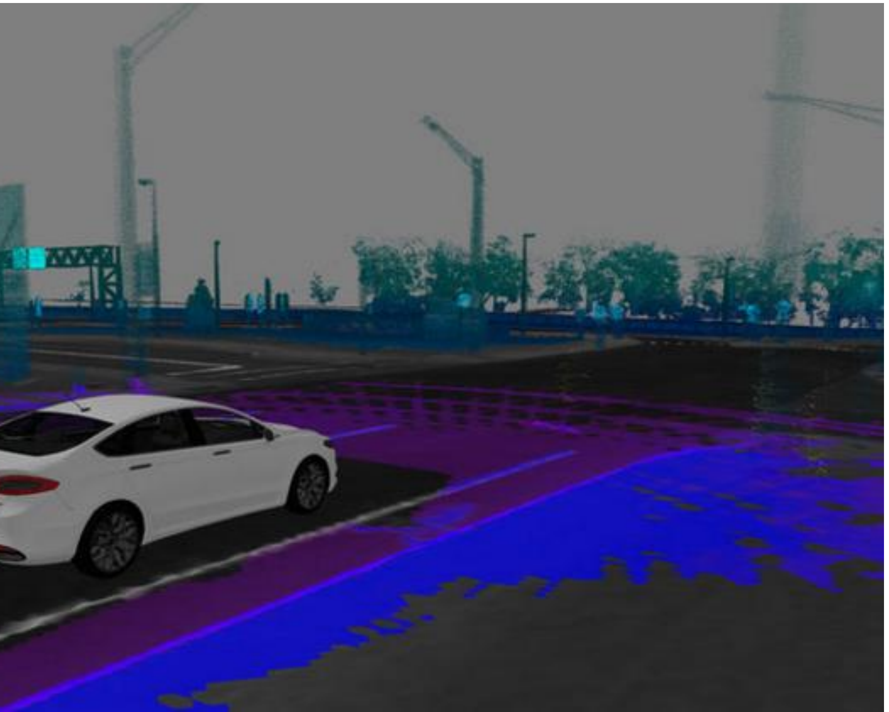


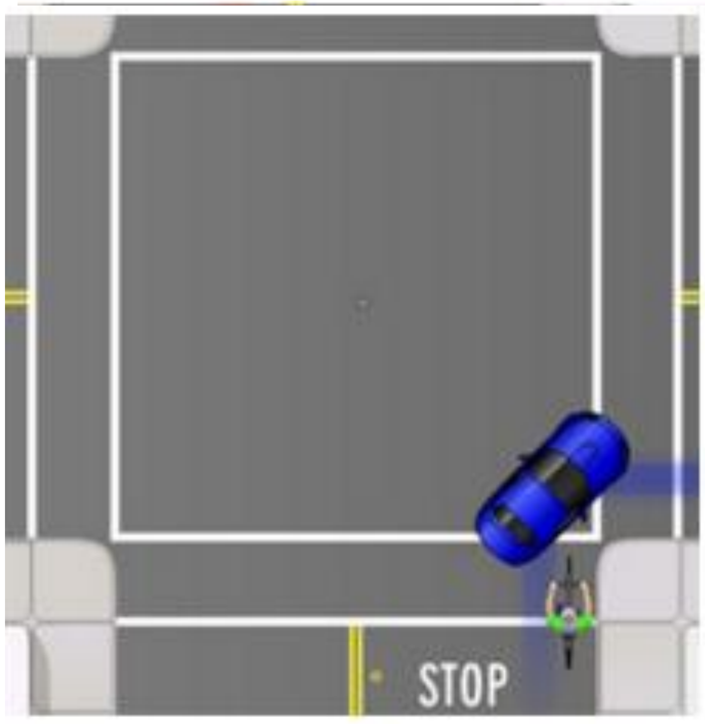
AI & Open-World Complexity

Rise of probabilistic methods: *known unknowns*

Recent attention to *unknown unknowns*

Decision making under uncertainty & incompleteness

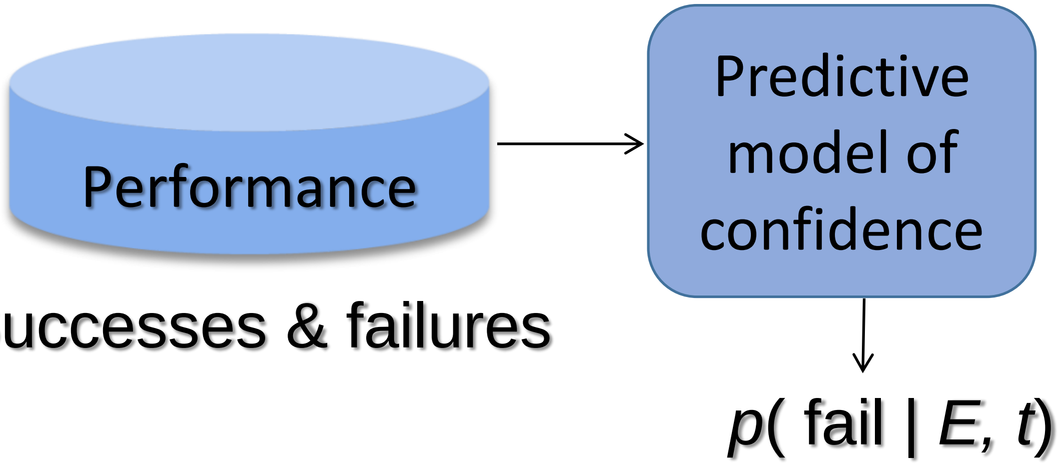


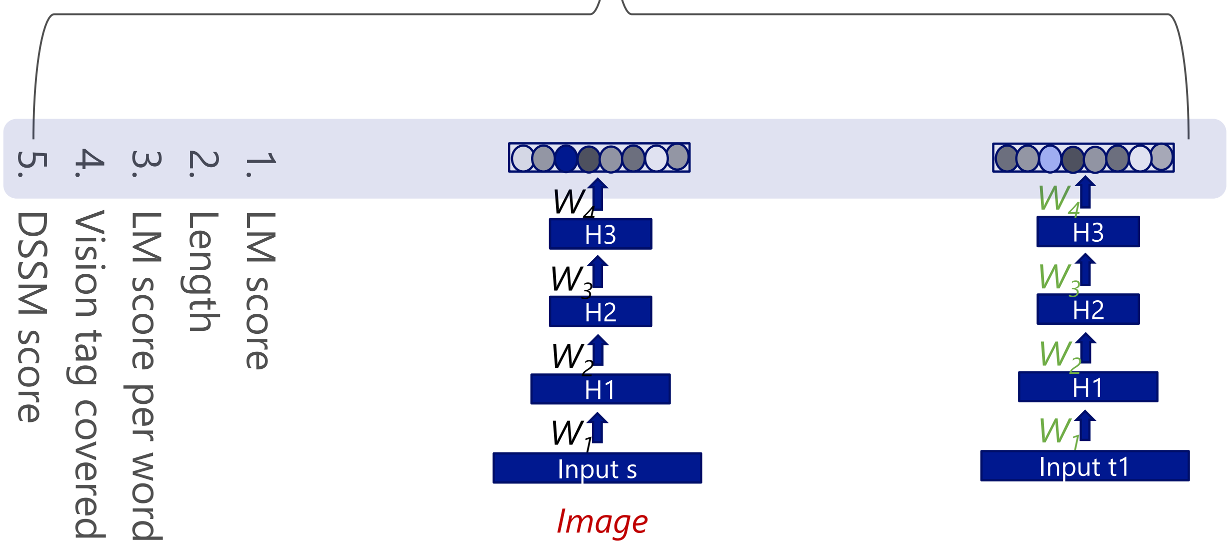


Direction: Learn about abilities & failures

Quality score [0,1] $s = \frac{e^{W\cdot f}}{1+e^{W\cdot f}}$

Deep learning about deep learning performance

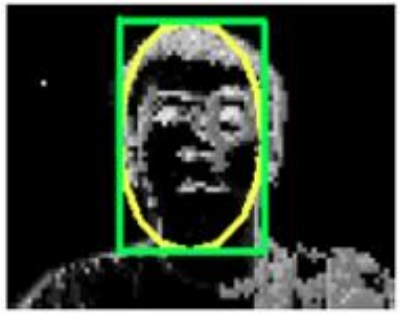


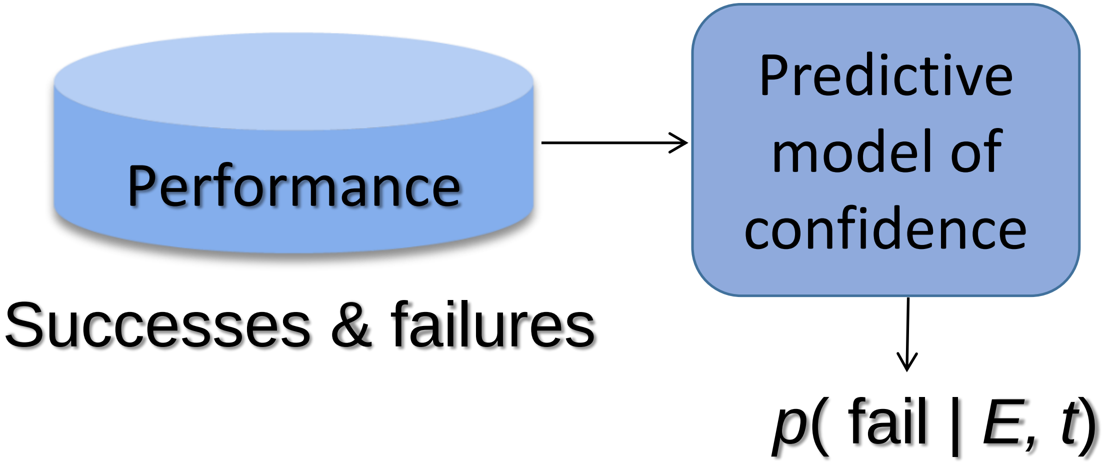


Caption:
a man holding a tennis racquet on a tennis court

Fang, et al., 2015

Direction: Learn about abilities & failures





Performance

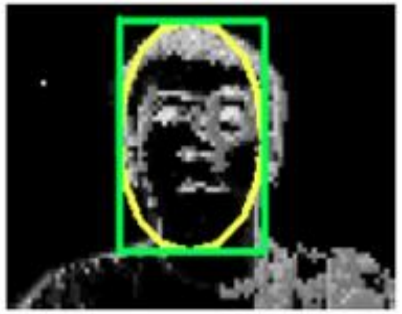
Successes & failures

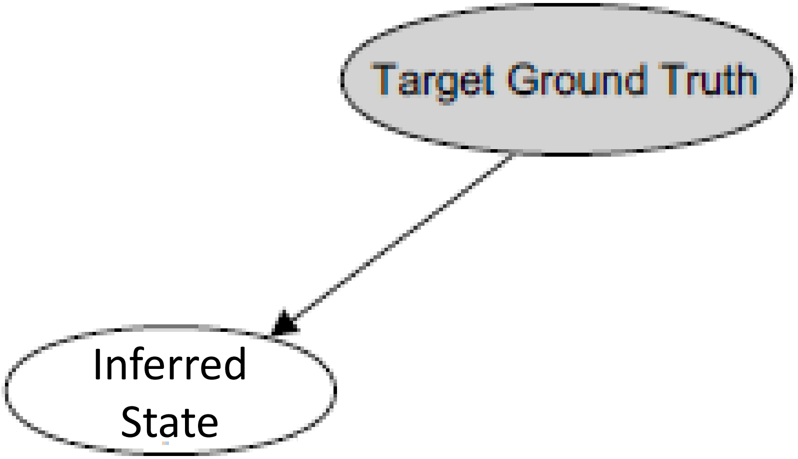
Predictive model of confidence

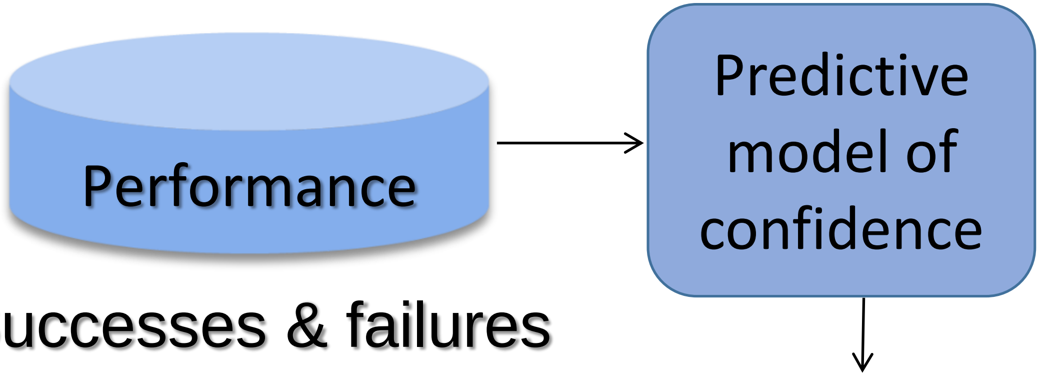
$p(\text{ fail } | E, t)$

Toyama & H. 2000

Direction: Learn about abilities & failures







Performance

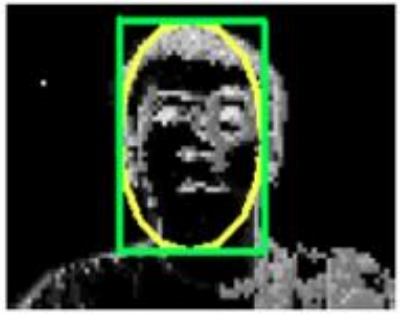
Predictive model of confidence

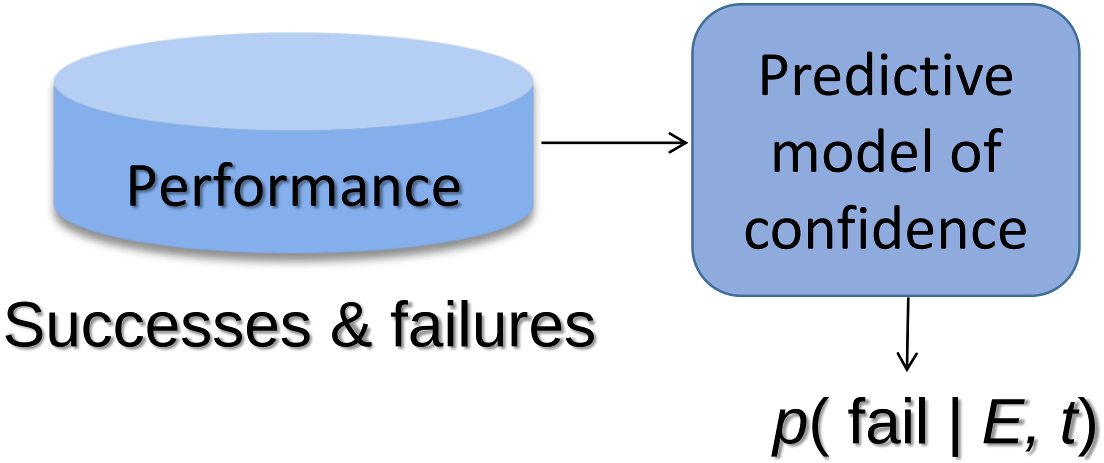
Successes & failures

$p(\text{ fail} \mid E, t)$

Toyama & H. 2000

Direction: Learn about abilities & failures



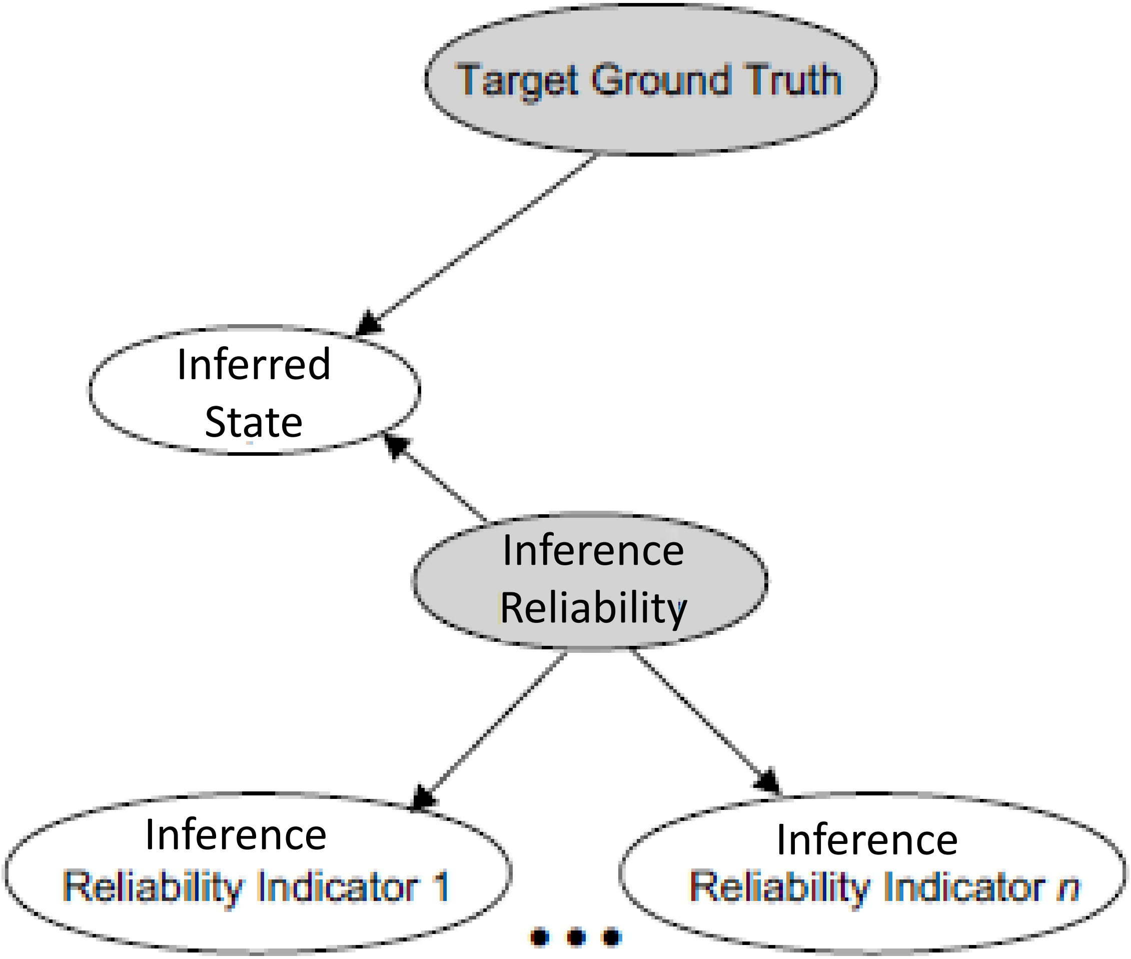


Performance

Successes & failures

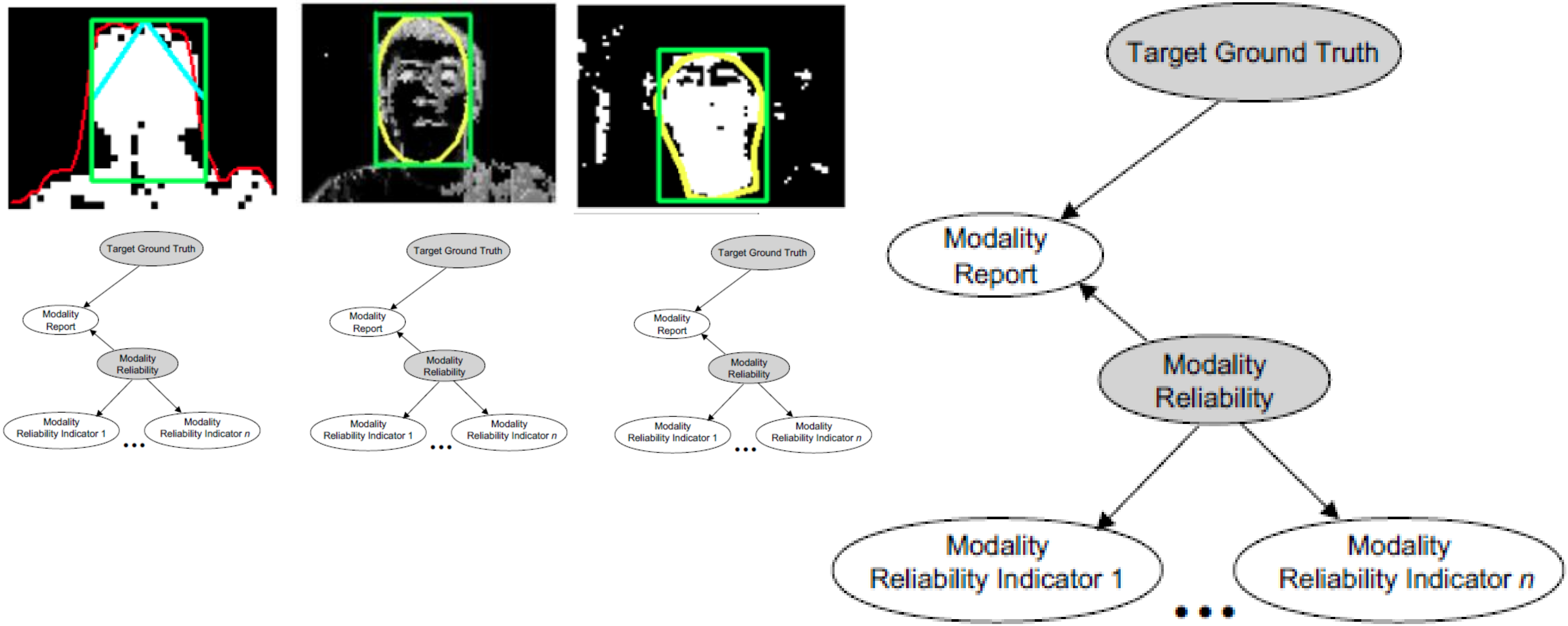
Predictive model of confidence

$p(\text{fail} \mid E, t)$


Target Ground Truth
Inferred State
Inference Reliability
Inference Reliability Indicator 1
...
Inference Reliability Indicator *n*

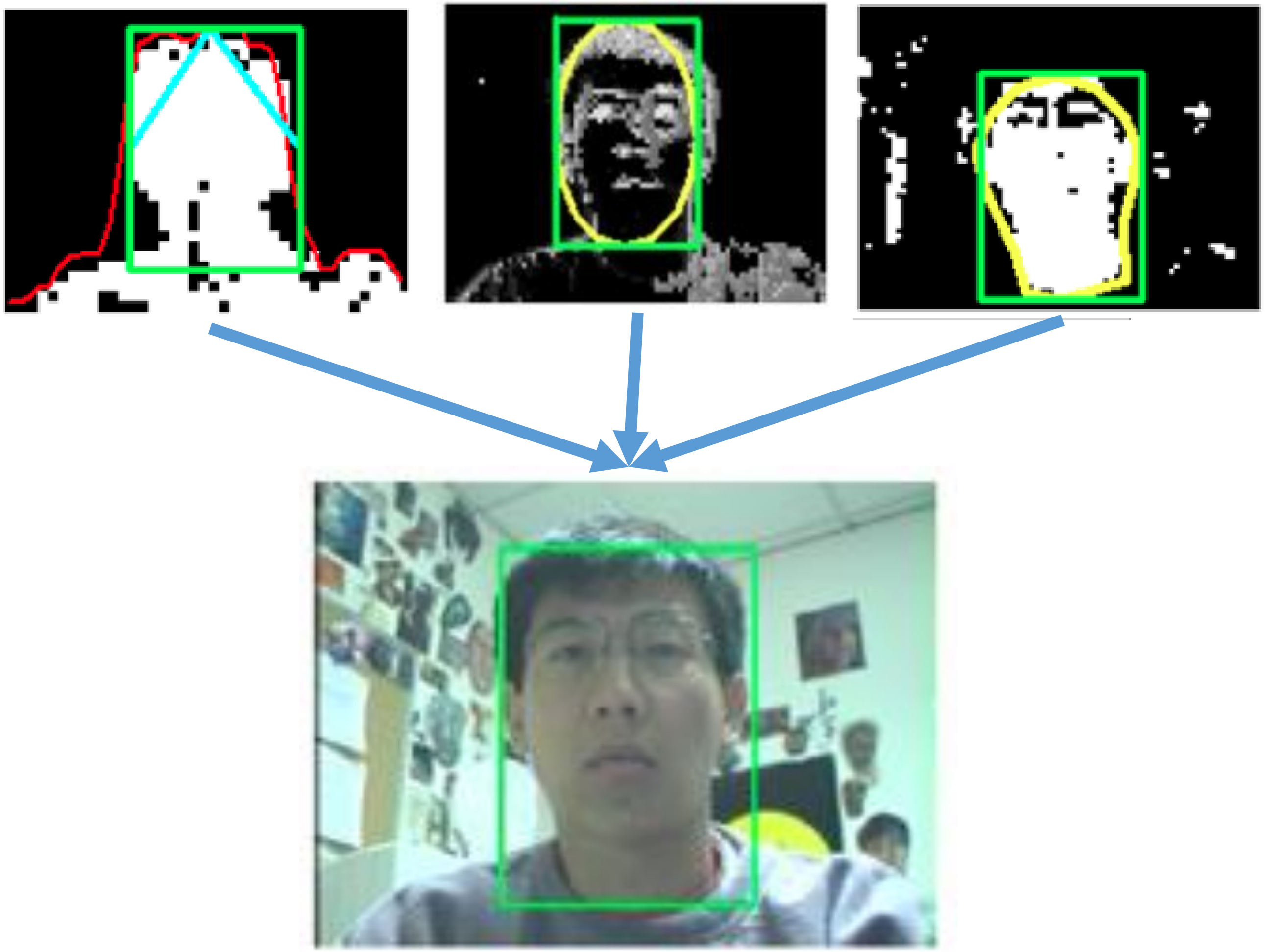
Toyama & H. 2000

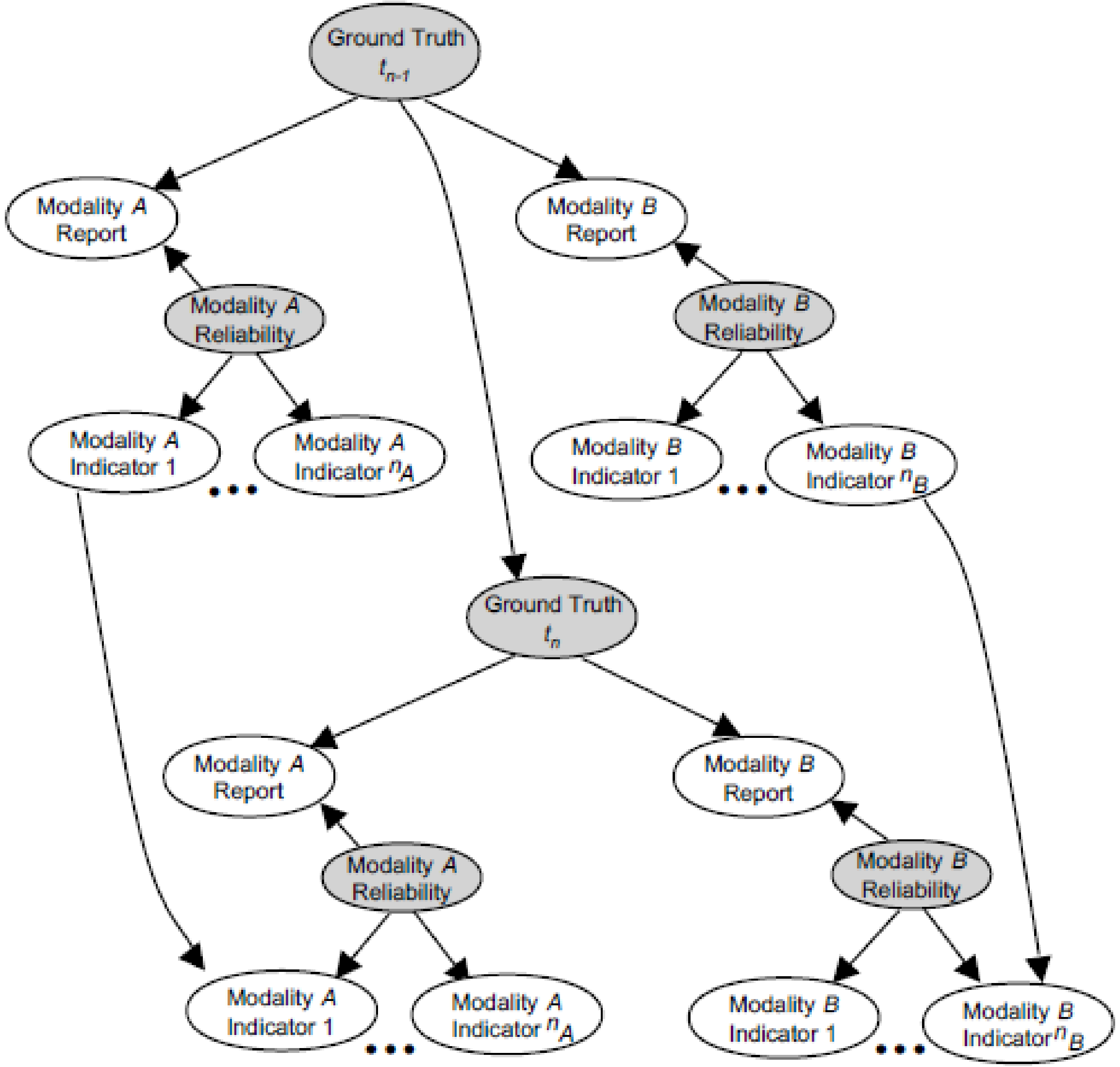
Direction: Robustness via analytical portfolios


Target Ground Truth
Modality Report
Modality Reliability
Modality Reliability Indicator 1
...
Modality Reliability Indicator *n*

Toyama & H. 2000

Direction: Robustness via analytical portfolios

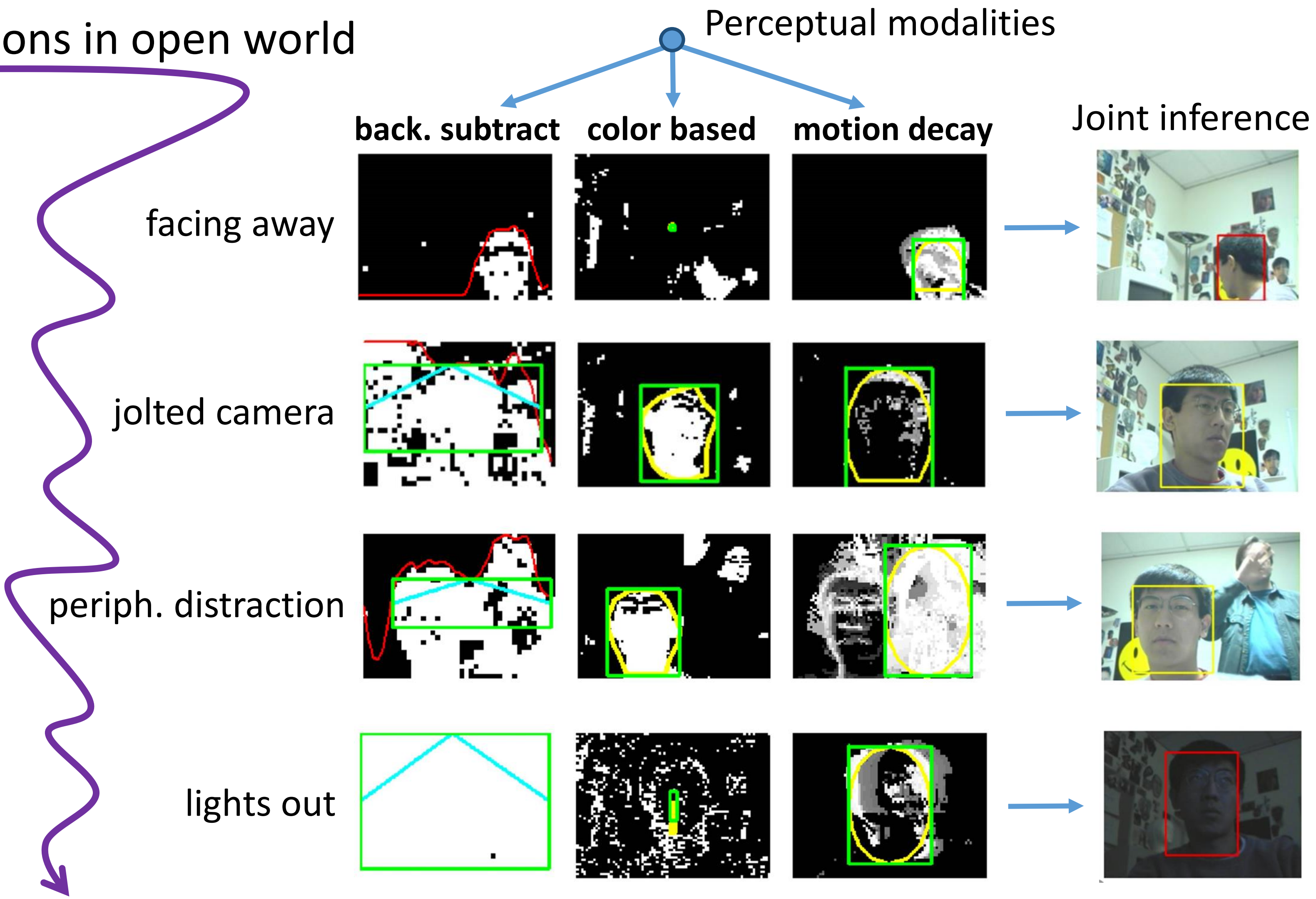




Toyama & H. 2000

Direction: Robustness via analytical portfolios

Unmodeled situations in open world



Perceptual modalities

back. subtract **color based** **motion decay**

Joint inference

facing away

jolted camera

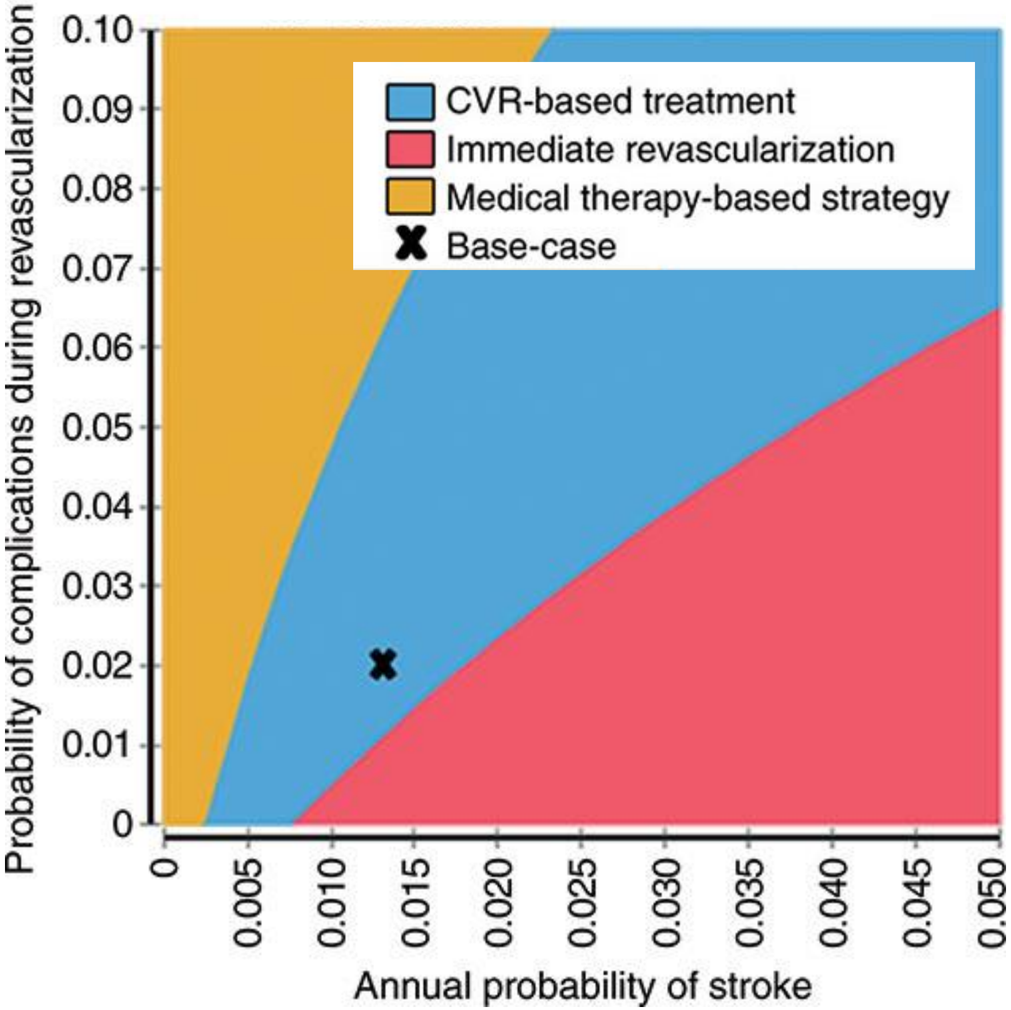
periph. distraction

lights out

Toyama & H. 2000

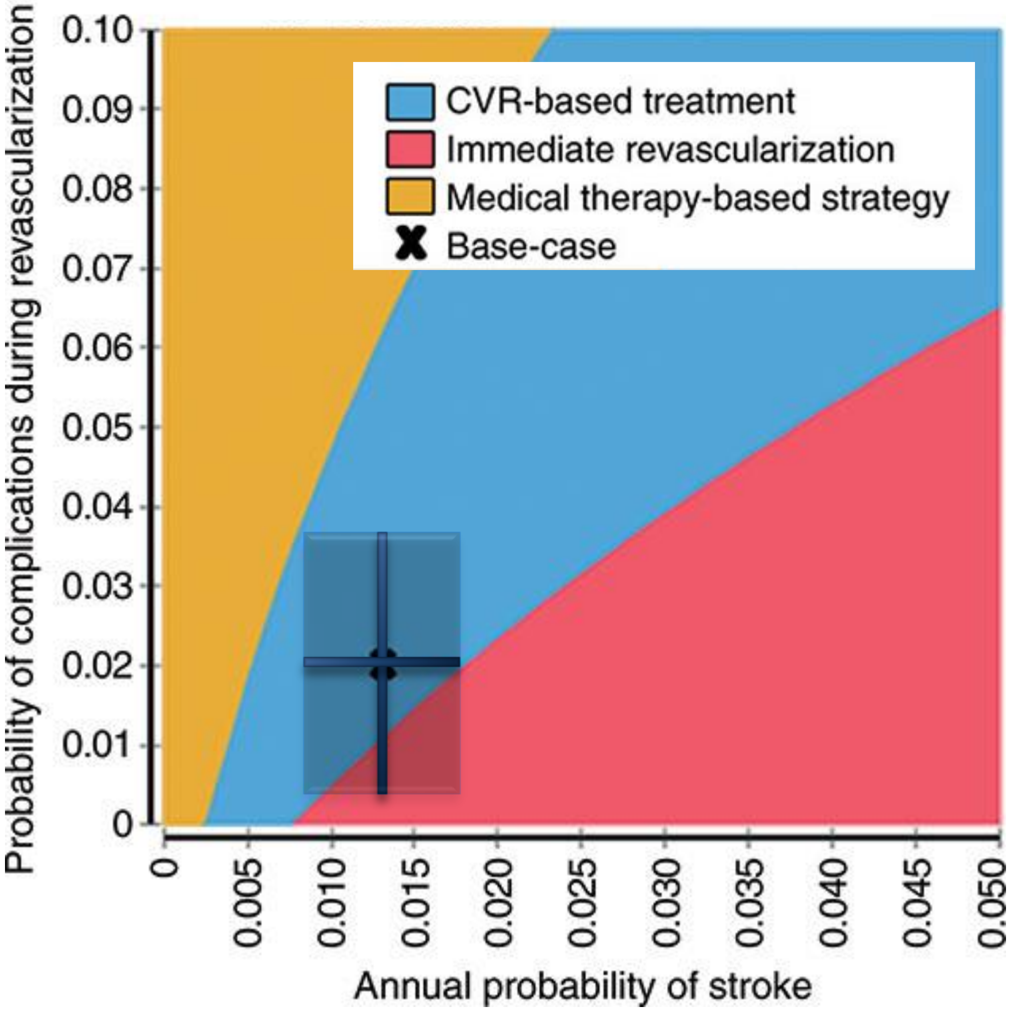
Direction: Understanding robustness via sensitivity analyses

Vary model structure, parameters, inferences



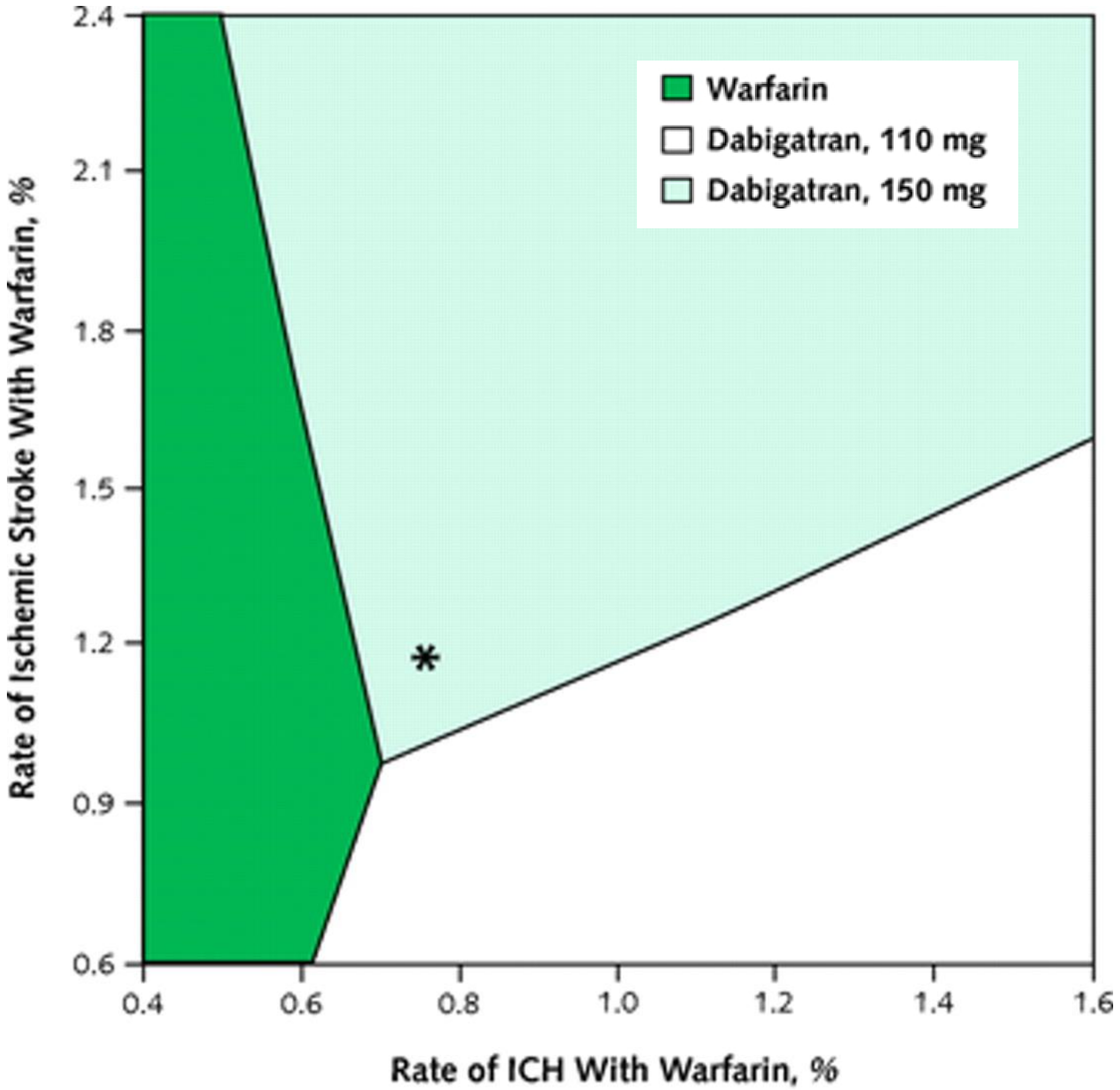
Direction: Understanding robustness via sensitivity analyses

Vary model structure, parameters, inferences



Direction: Understanding robustness via sensitivity analyses

Vary model structure, parameters, inferences



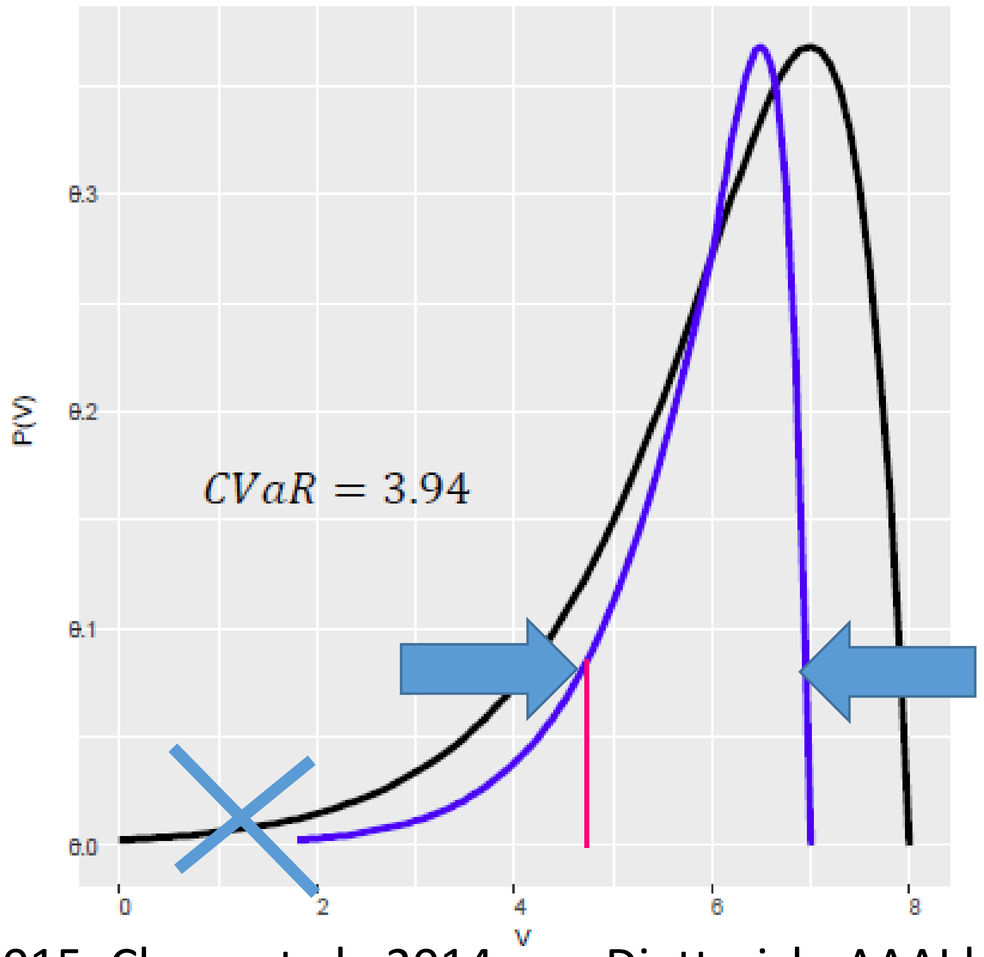
Direction: Robust optimization to minimize downside

Robust optimization under uncertain parameters

Risk-sensitive objective

e.g., conditional-value-at-risk budget

Methods trade upside value for reducing probability of costly outcomes



Tamar, 2015; Chow, et al., 2014; per Dietterich, AAAI lect. 2016

Direction: Learn about unknown unknowns

Data, experience, rich simulations

Detect anomalies, unexpected variations, distributional shifts

Meta-analysis & transfer

Human engagement

Direction: Learn about unknown unknowns

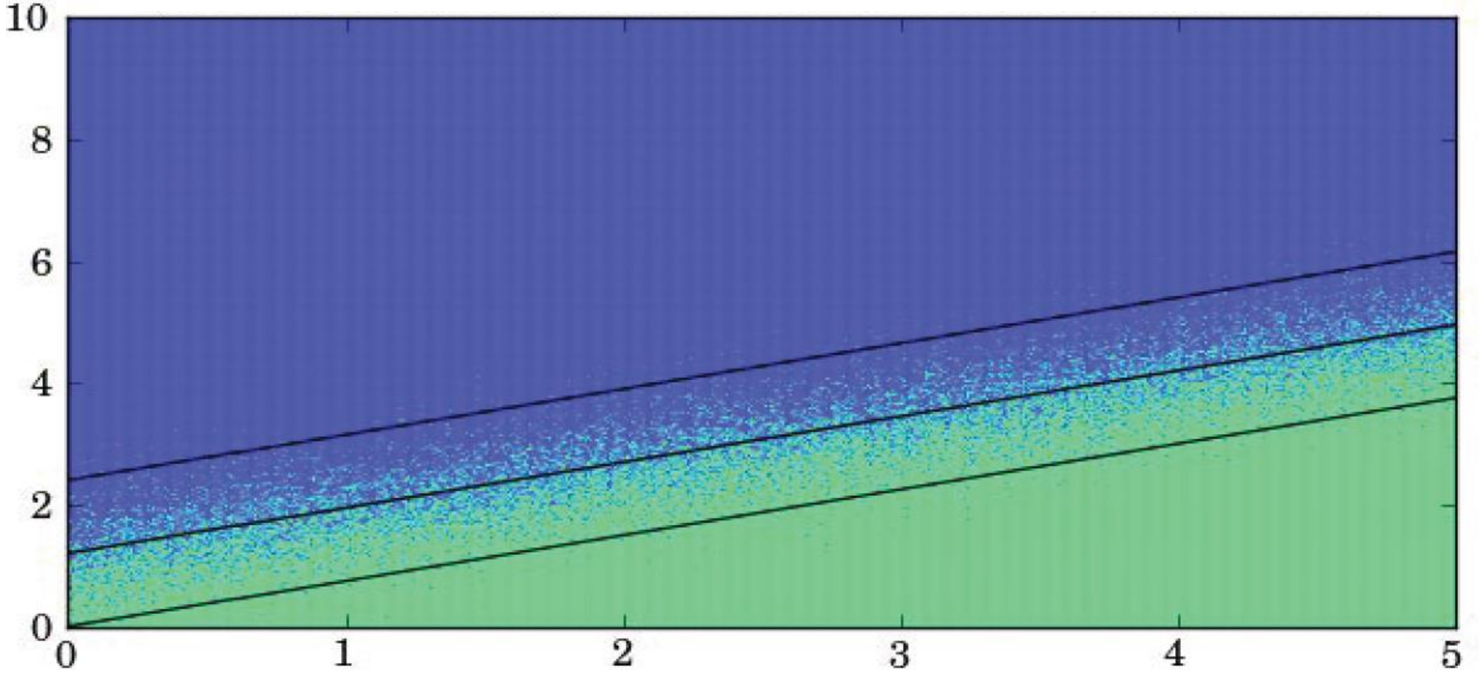
Data, experience, rich simulations

Detect anomalies, unexpected variations, distributional shifts

Meta-analysis & transfer

Human engagement

"Beat the Machine" (Attenberg, Ipeirotis, Provost 2015)



Direction: Learn about unknown unknowns

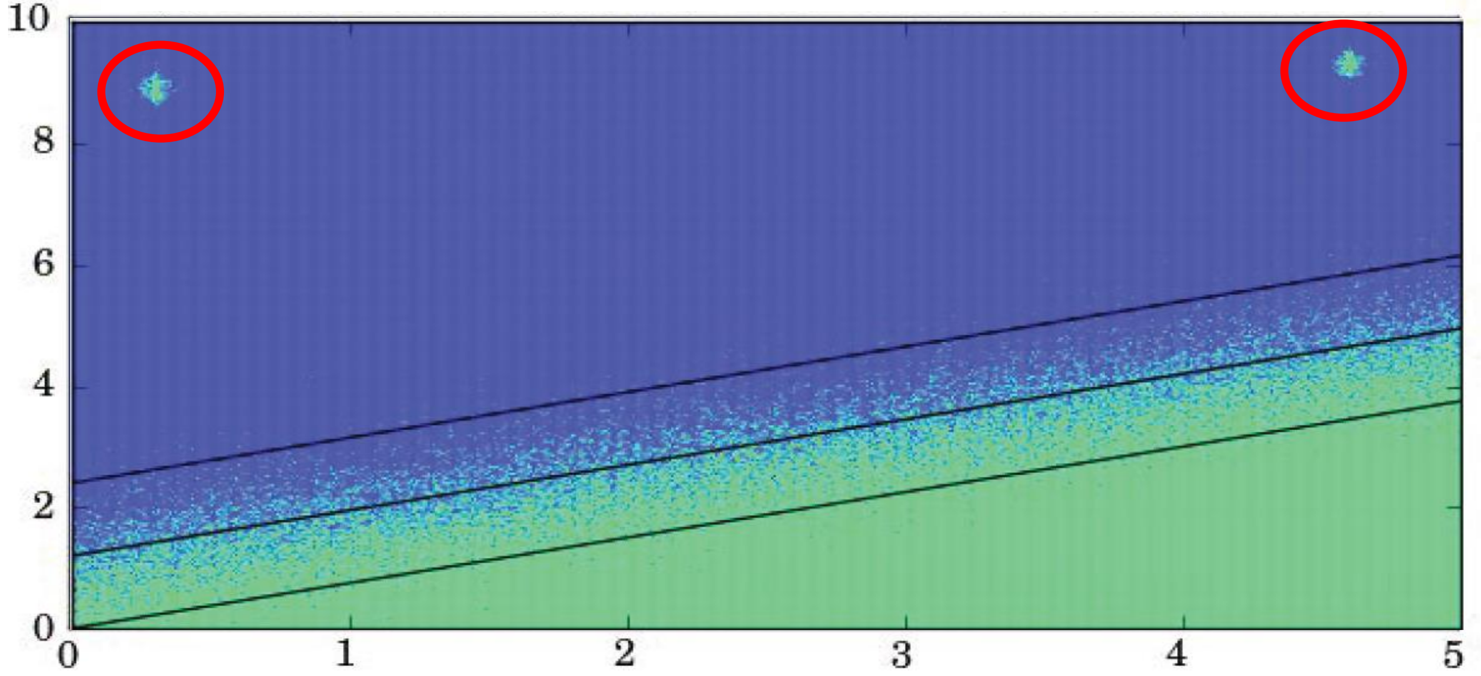
Data, experience, rich simulations

Detect anomalies, unexpected variations, distributional shifts

Meta-analysis & transfer

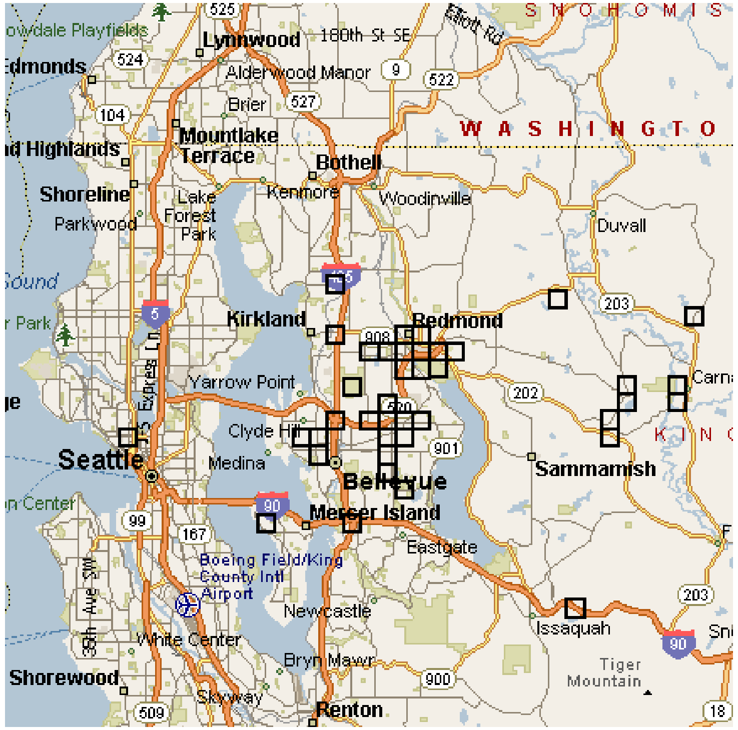
Human engagement

“Beat the Machine” (Attenberg, Ipeirotis, Provost 2015)



Direction: Learn about unknown unknowns

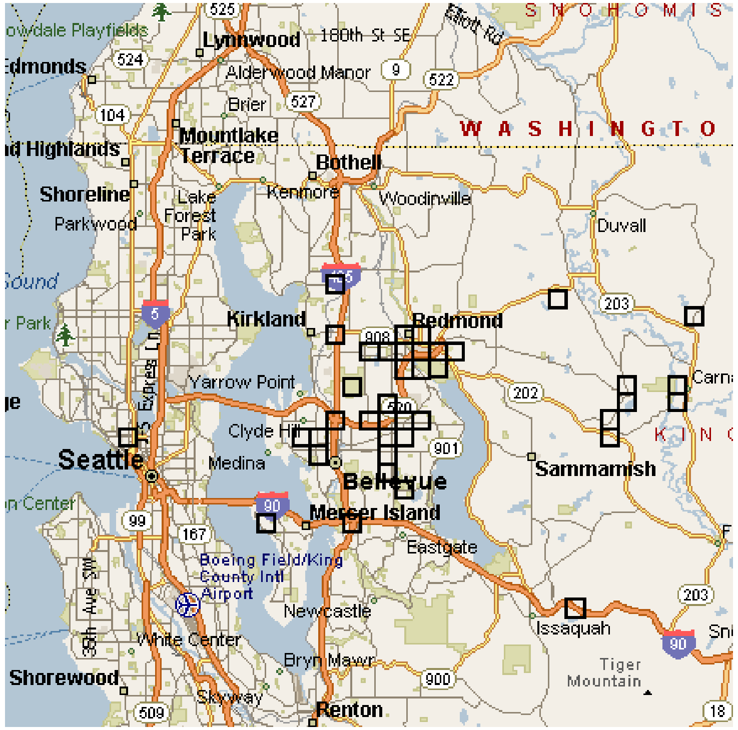
Predict new distinctions, combine open- & closed-world models

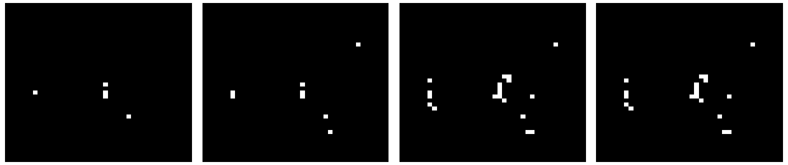


Krumm, H., 2006

Direction: Learn about unknown unknowns

Predict new distinctions, combine open- & closed-world models





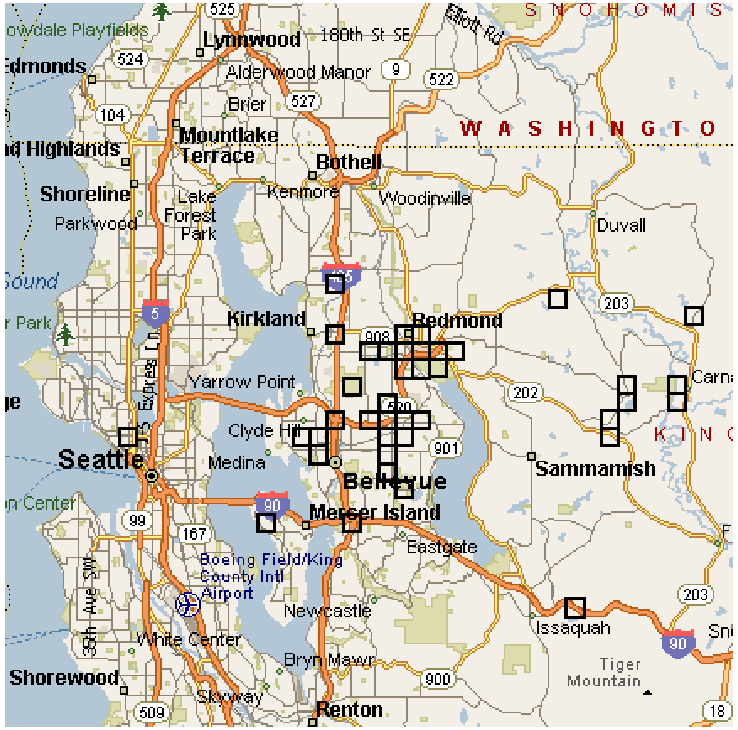
Destinations,
E1..,En, t

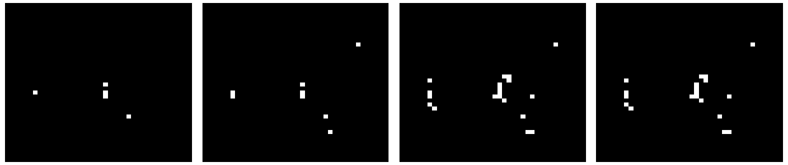
Krumm, H., 2006

Direction: Learn about unknown unknowns

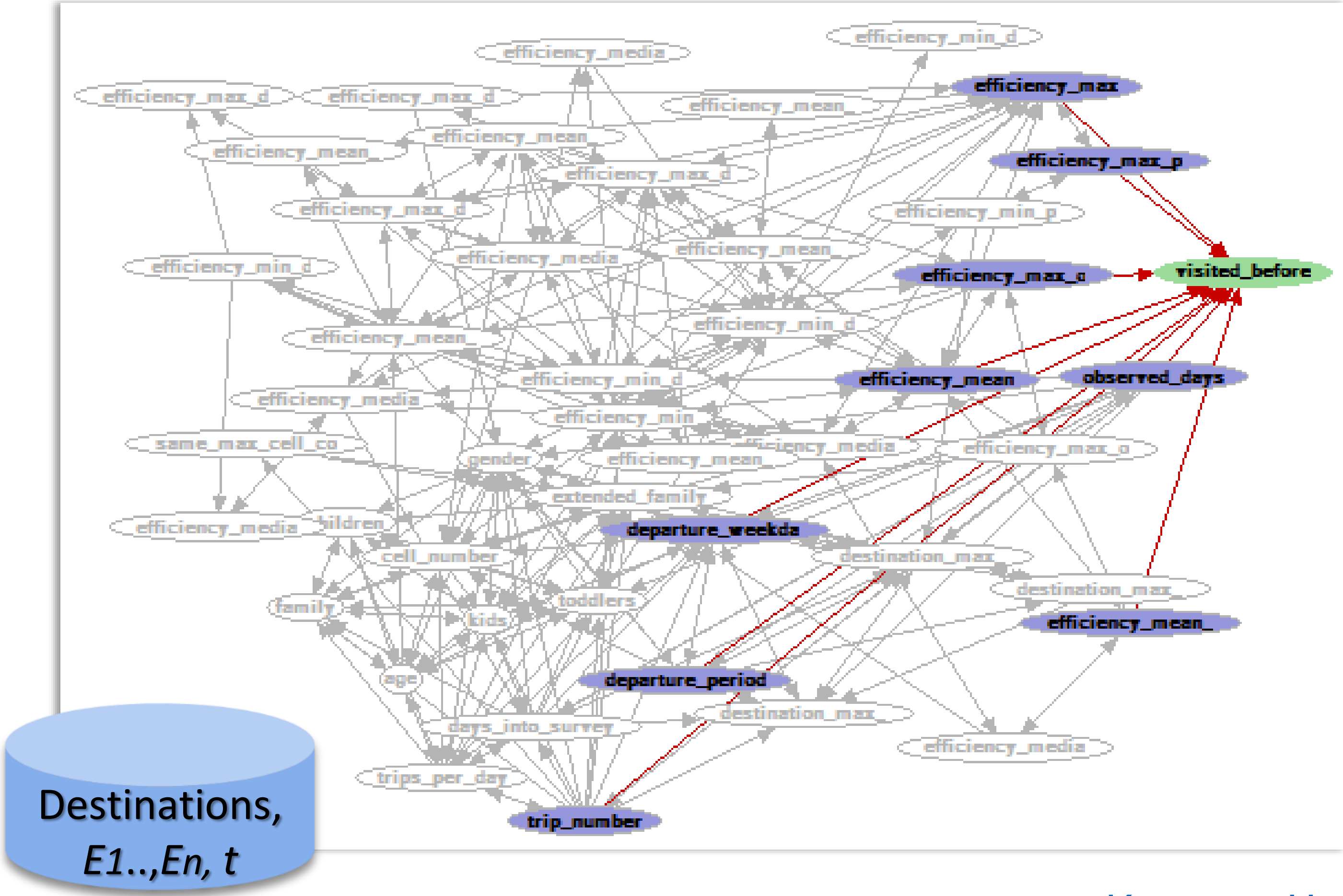
Predict new distinctions, combine open- & closed-world models

Predict previously unseen destination





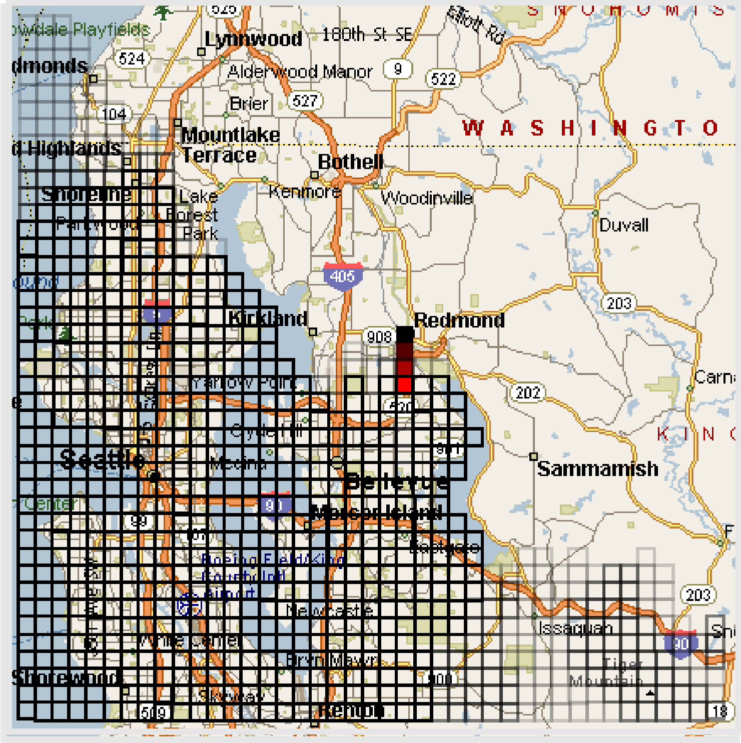
Day 1 → Day 14

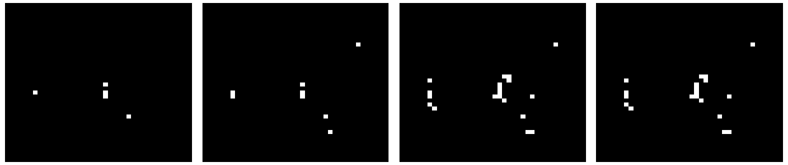


Krumm, H., 2006

Direction: Learn about unknown unknowns

Predict new distinctions, combine open- & closed-world models

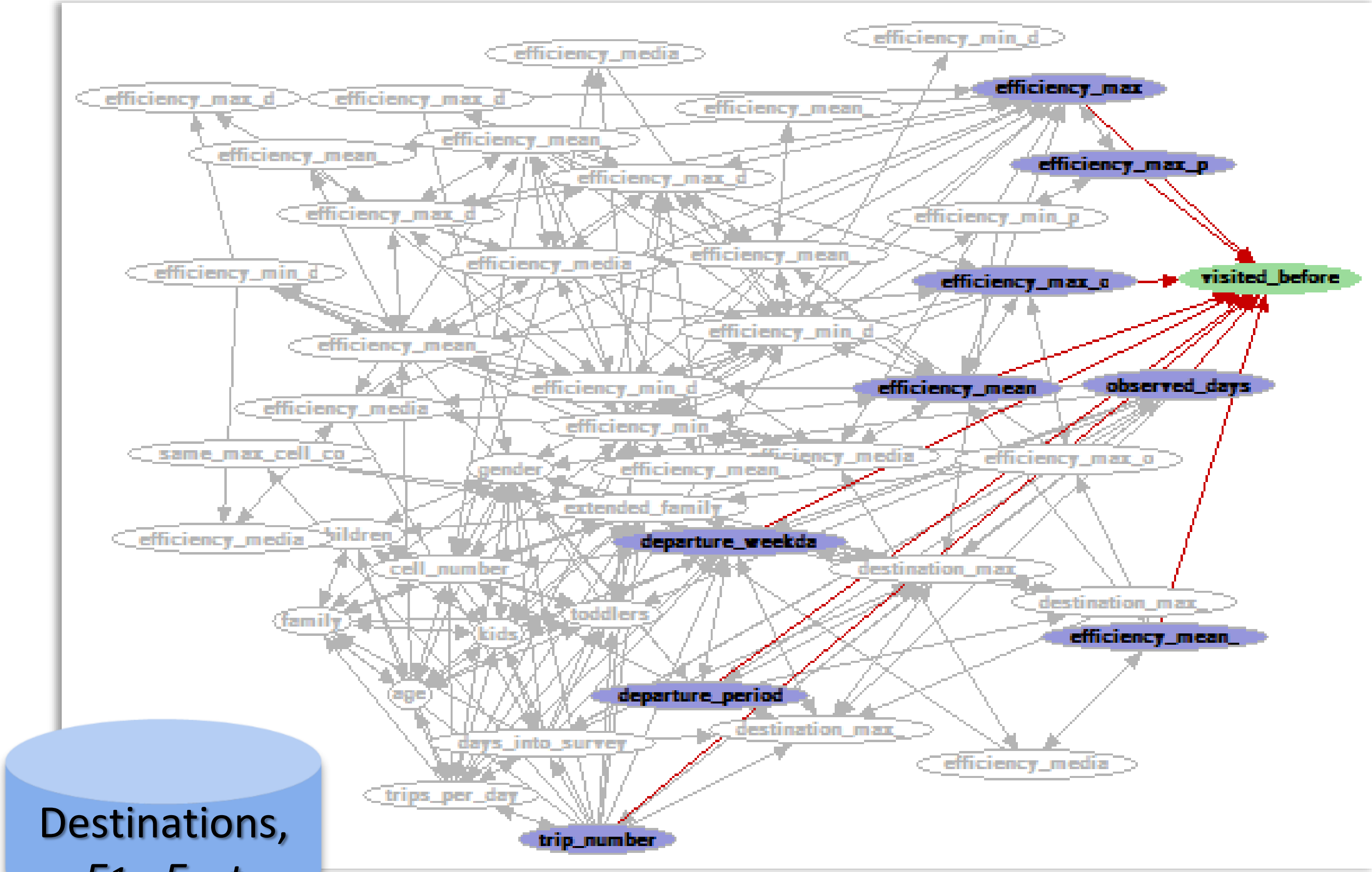




Day 1 → Day 14

Destinations,
E1..,En, t

Predict previously unseen destination

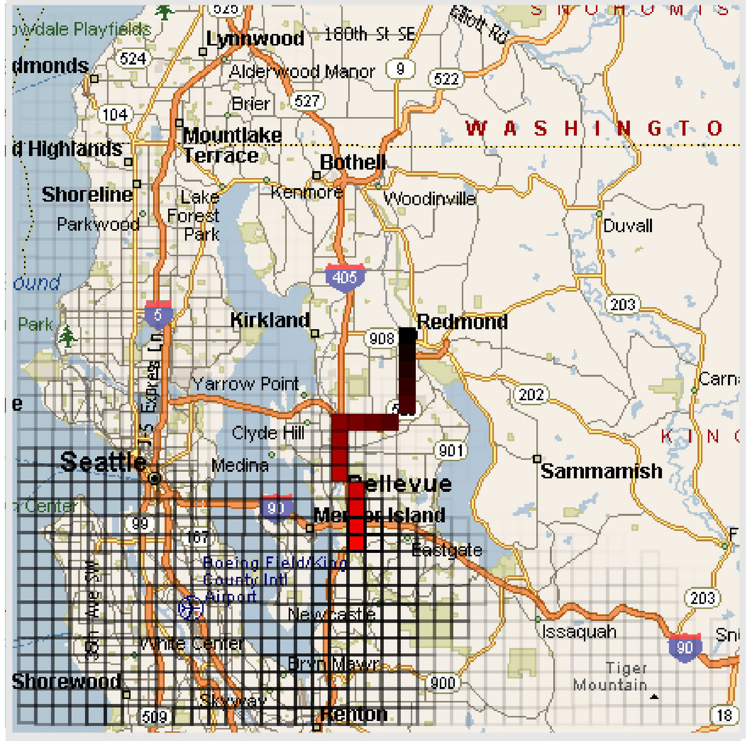


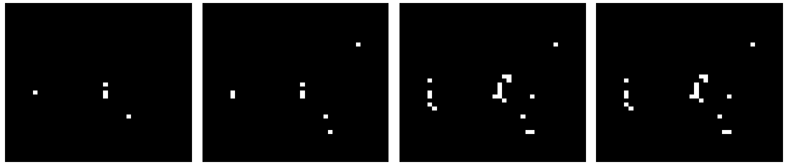
Krumm, H., 2006

Direction: Learn about unknown unknowns

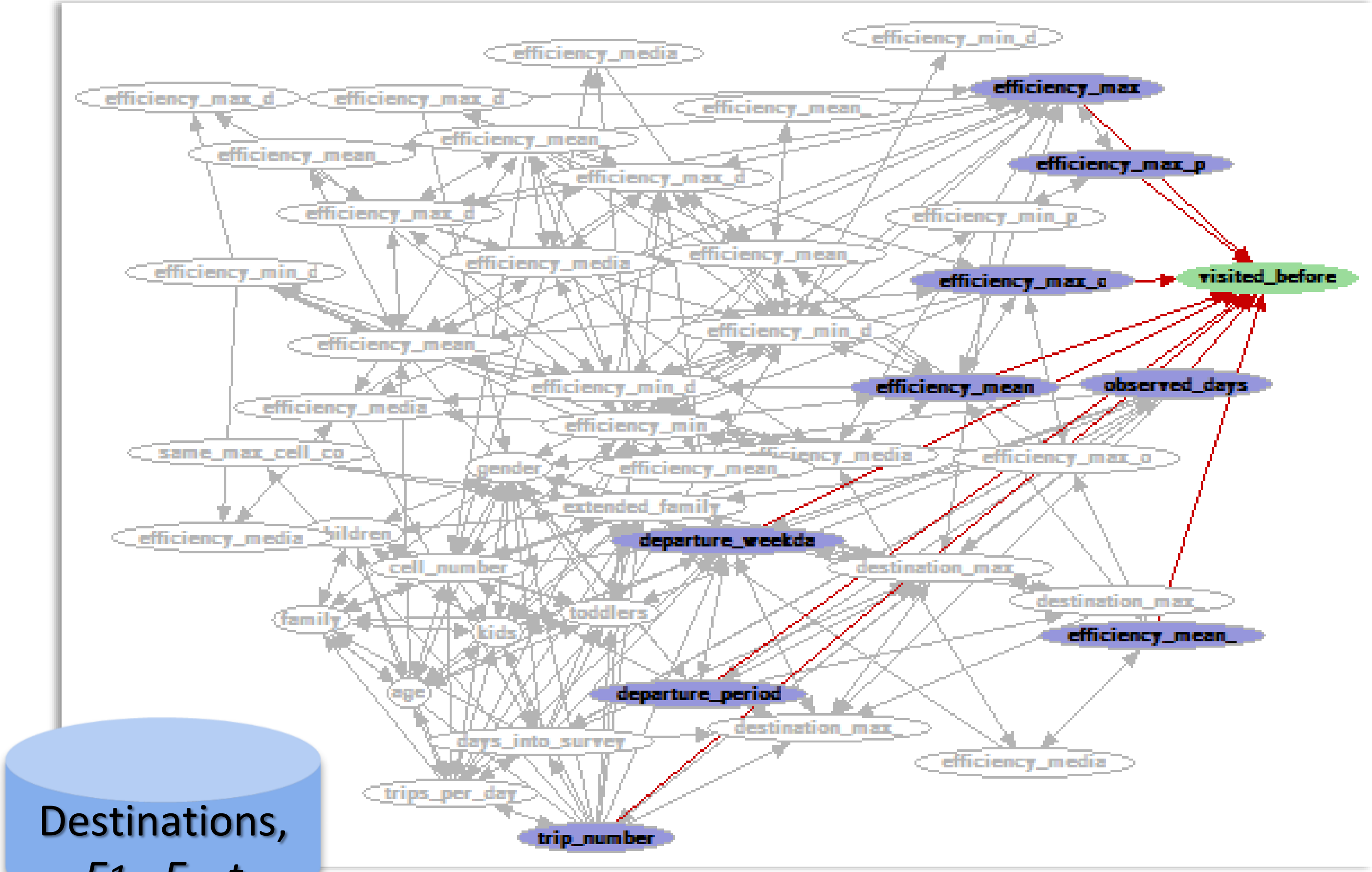
Predict new distinctions, combine open- & closed-world models

Predict previously unseen destination




Day 1 → Day 14

Destinations, *E1..,En, t*



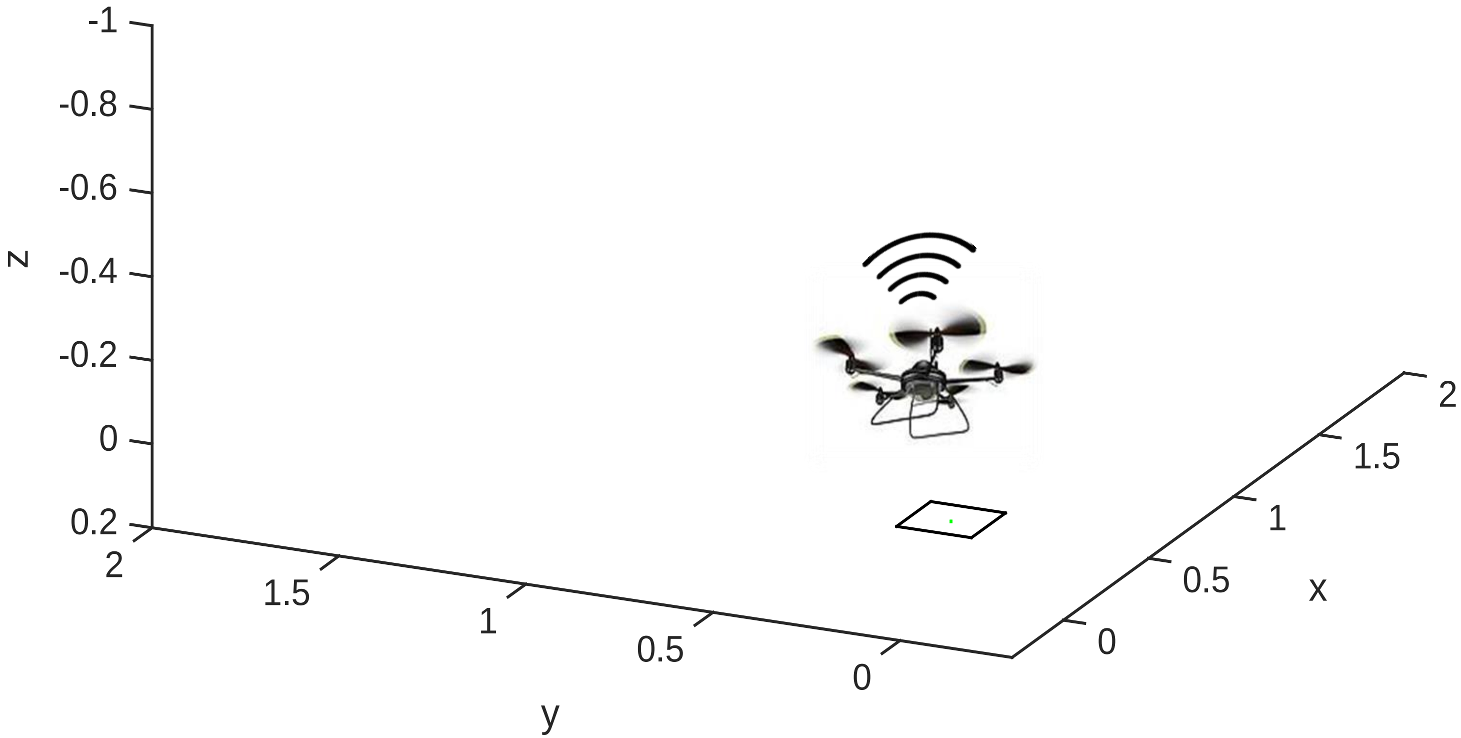
Krumm, H., 2006

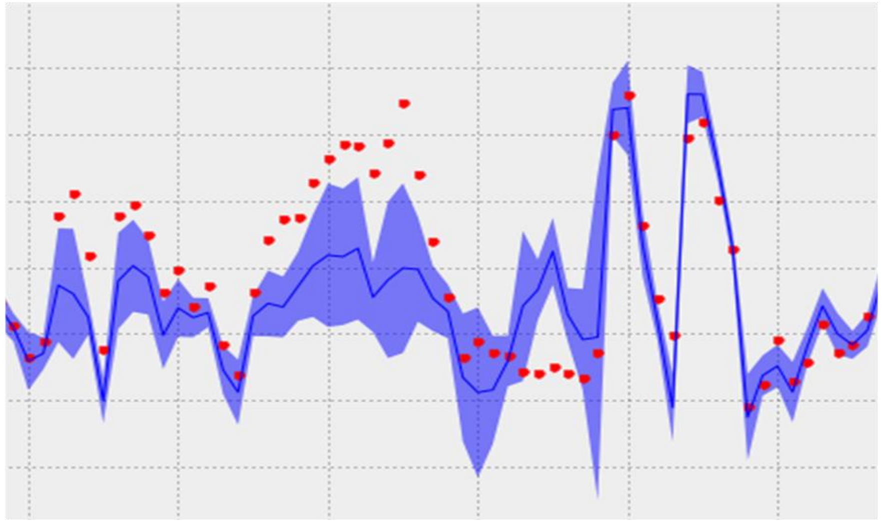
Direction: Joint modeling of key dimensions of error

Example: Learn about errors of perception & control

Probabilistic models of control $\boldsymbol{\varphi}_{\mathbf{roll}}$

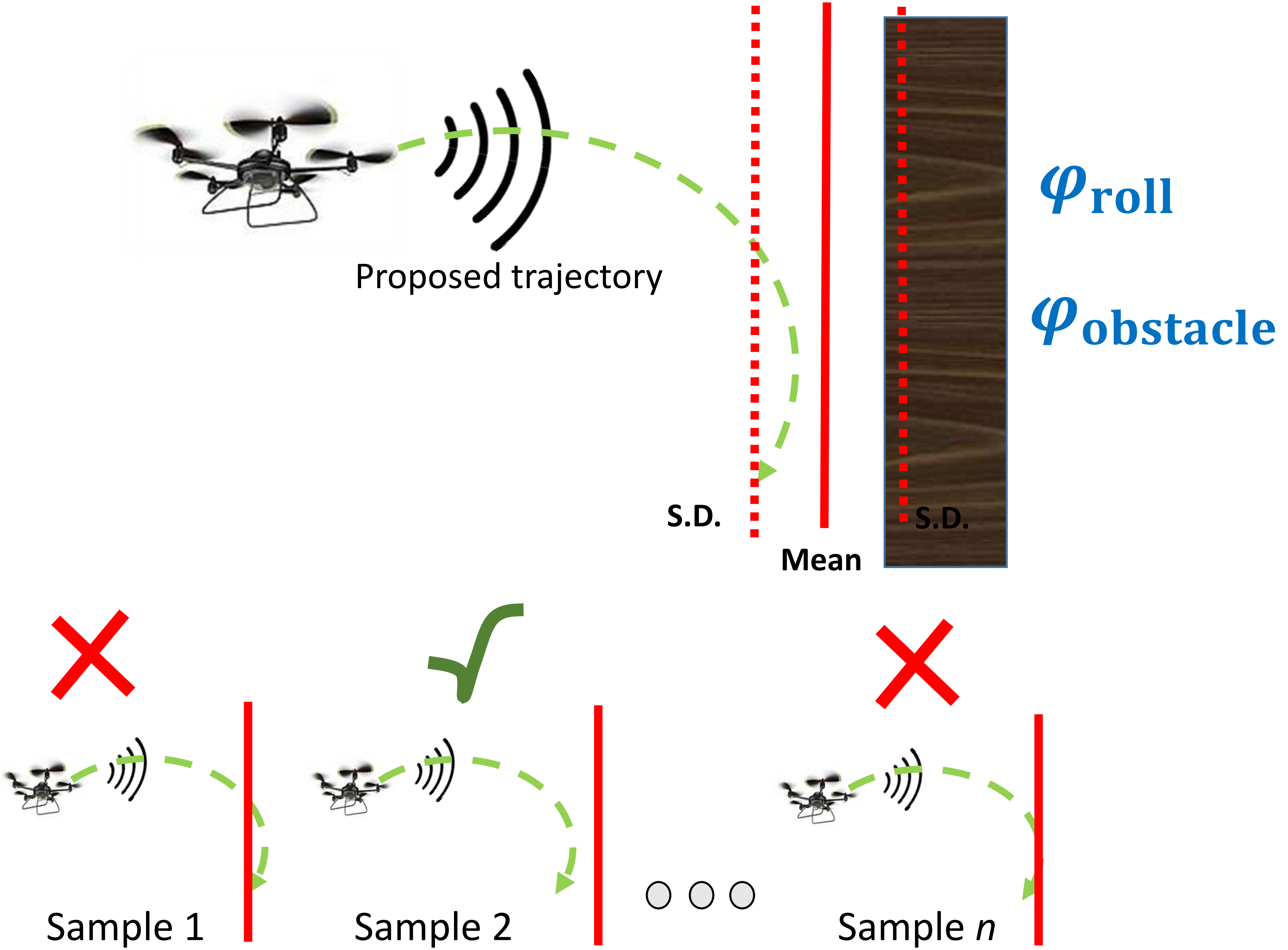
Probabilistic models of sensing $\boldsymbol{\varphi}_{\mathbf{obstacle}}$





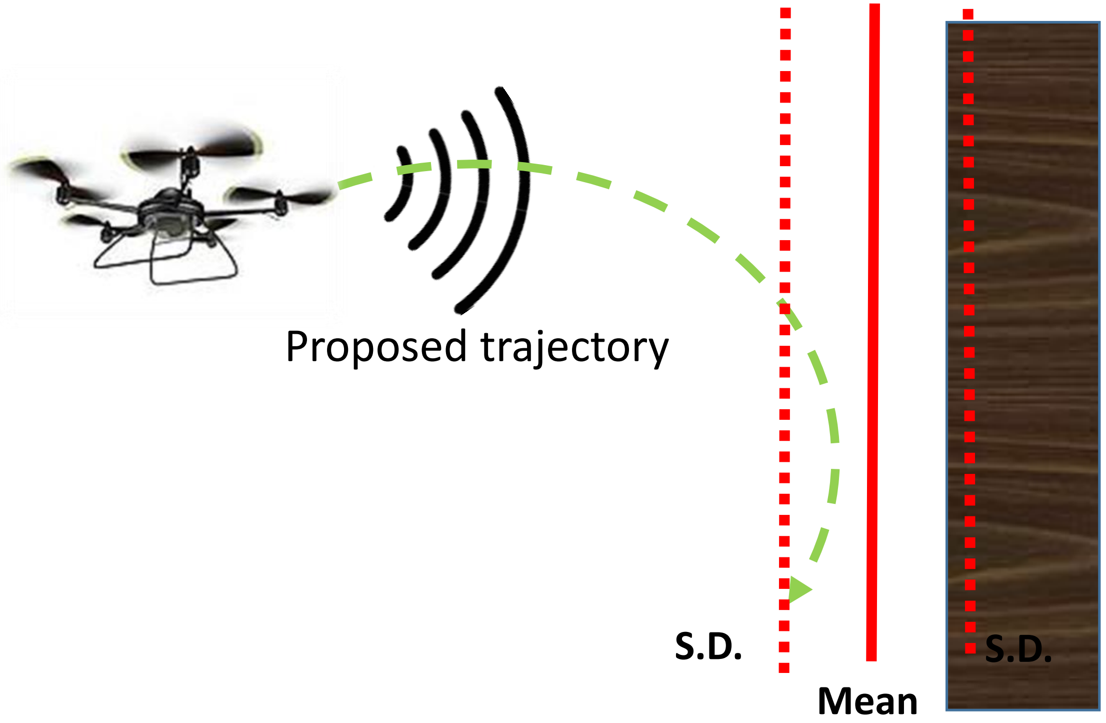
Sadigh & Kapoor, 2016

Direction: Joint modeling of key dimensions of error



Sadigh & Kapoor, 2016

Direction: Joint modeling of key dimensions of error



$\varphi_{\mathbf{roll}}$

$\varphi_{\mathbf{obstacle}}$

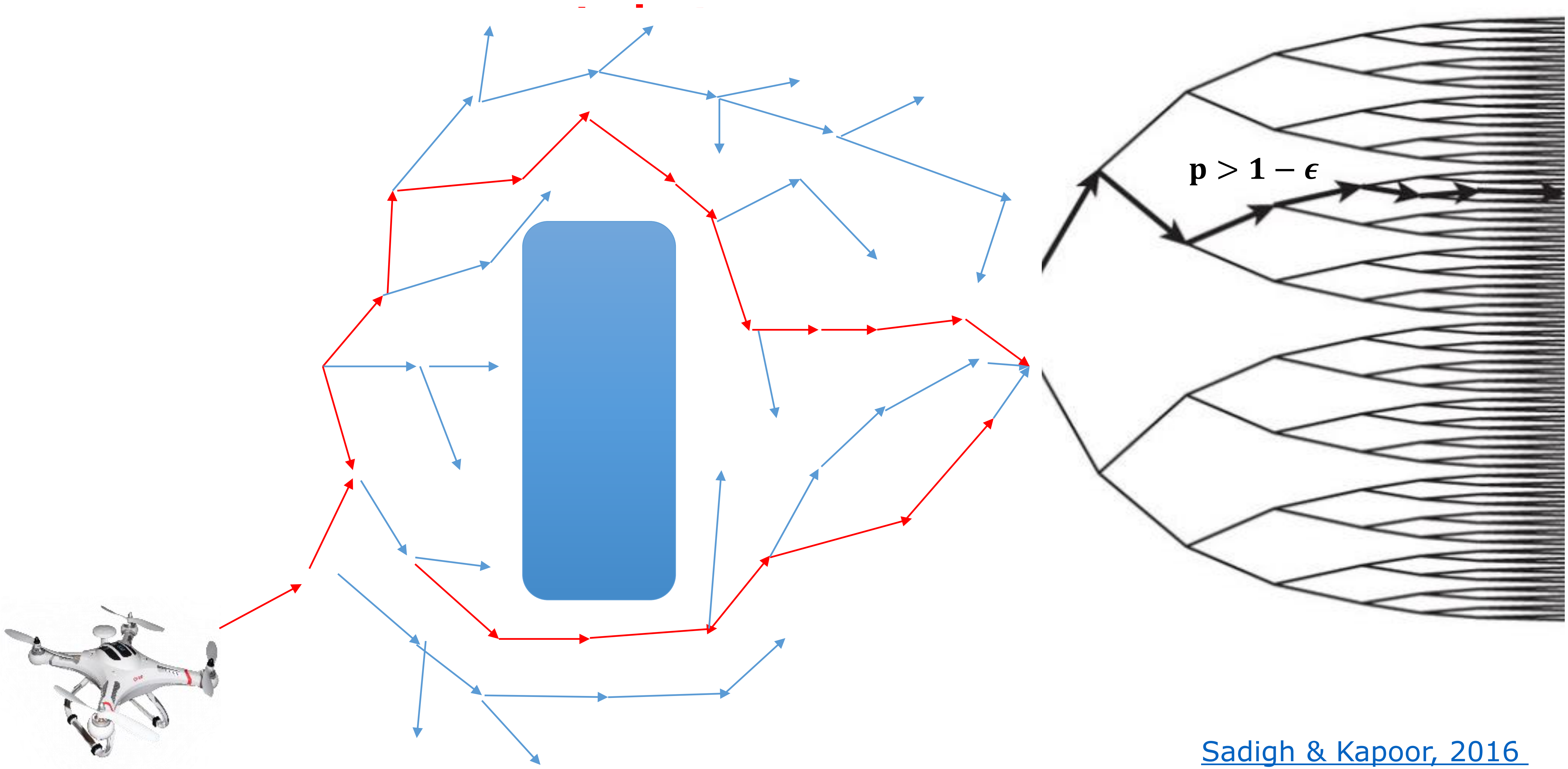
Trajectory <u>safe</u> if:

$$\frac{\sum ✓}{\sum ✓ + \sum ✗} > 1 - \epsilon$$

$\sum$ ✓✓✓✓✓✗✓✓✓✗✓✓✓✓✓✓✗✓✓✓✗✓✗✓✓✓

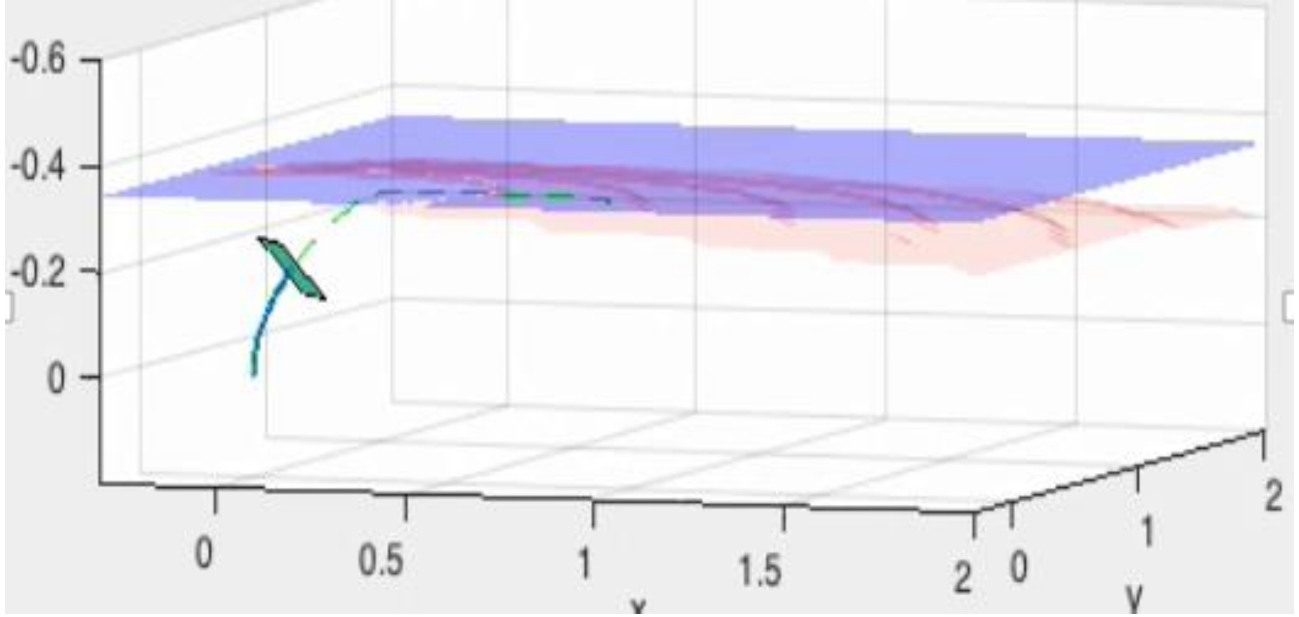
Sadigh & Kapoor, 2016

Direction: Joint modeling of key dimensions of error



Sadigh & Kapoor, 2016

Direction: Joint modeling of key dimensions of error



(video)

Value of refining models & system

- Value of additional data
- Value of enhancing sensors
- Value of better controller

Sadigh & Kapoor, 2016

Direction: Joint modeling of key dimensions of error

$$\mathbf{p} > \mathbf{1} - \boldsymbol{\epsilon}$$

(video)

Sadigh & Kapoor, 2016

Direction: Joint modeling of key dimensions of error

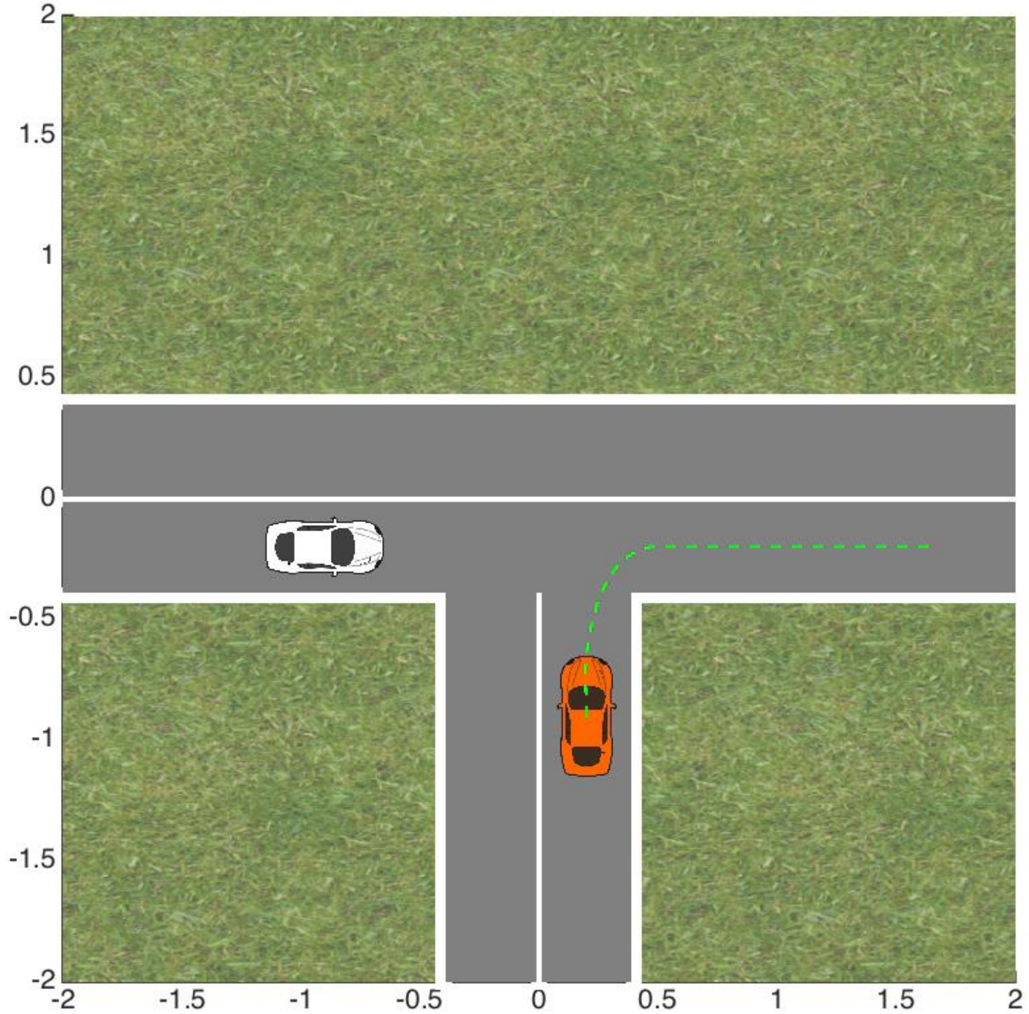
Fail-safe

$\cancel{\mathbf{p} > \mathbf{1} - \epsilon}$

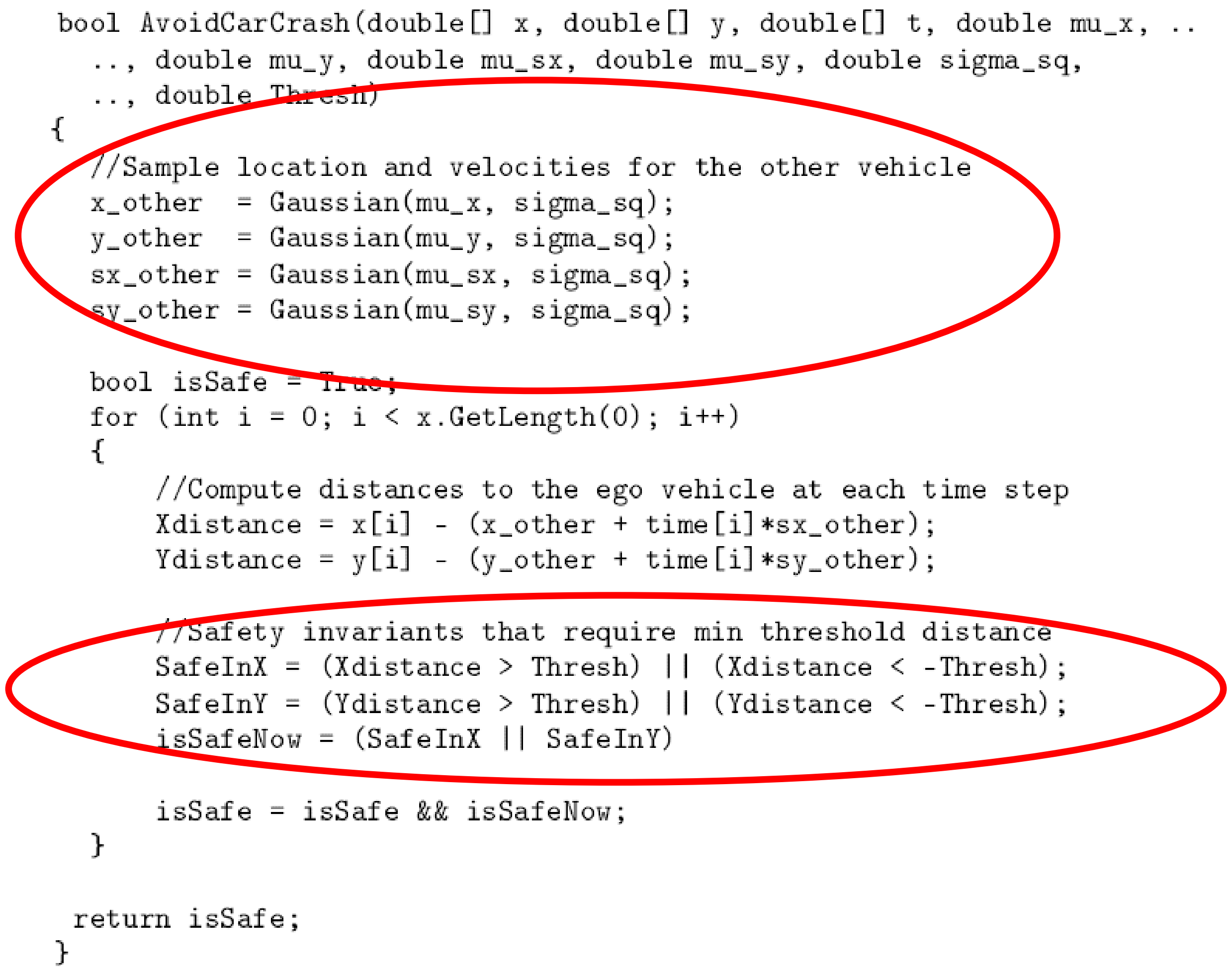
(video)

Sadigh & Kapoor, 2016

Direction: Joint modeling of key dimensions of error



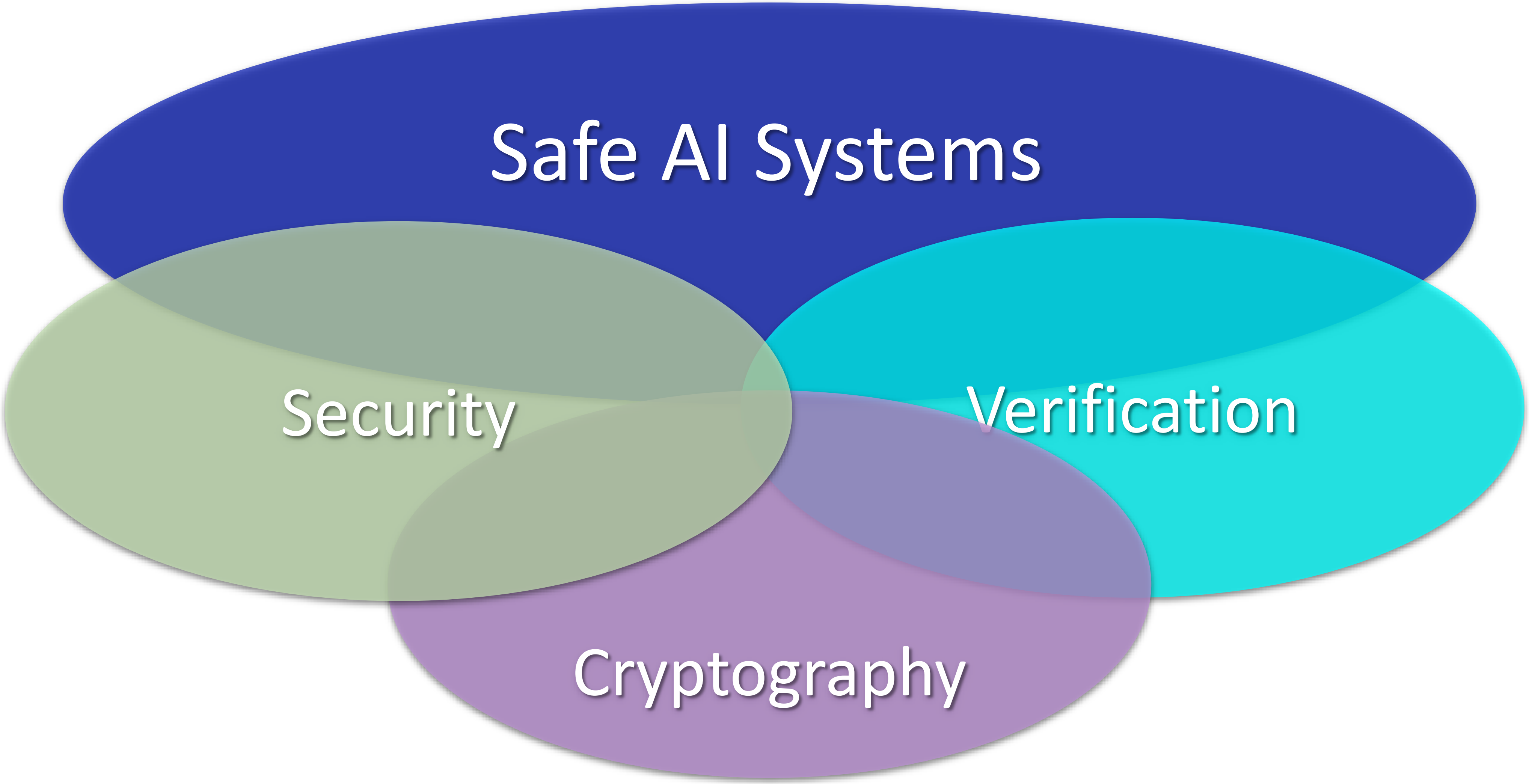
(video)



```
bool AvoidCarCrash(double[] x, double[] y, double[] t, double mu_x, ..
  .., double mu_y, double mu_sx, double mu_sy, double sigma_sq,
  .., double Thresh)
{
  //Sample location and velocities for the other vehicle
  x_other  = Gaussian(mu_x, sigma_sq);
  y_other  = Gaussian(mu_y, sigma_sq);
  sx_other = Gaussian(mu_sx, sigma_sq);
  sy_other = Gaussian(mu_sy, sigma_sq);

  bool isSafe = True;
  for (int i = 0; i < x.GetLength(0); i++)
  {
      //Compute distances to the ego vehicle at each time step
      Xdistance = x[i] - (x_other + time[i]*sx_other);
      Ydistance = y[i] - (y_other + time[i]*sy_other);

      //Safety invariants that require min threshold distance
      SafeInX = (Xdistance > Thresh) || (Xdistance < -Thresh);
      SafeInY = (Ydistance > Thresh) || (Ydistance < -Thresh);
      isSafeNow = (SafeInX || SafeInY)

      isSafe = isSafe && isSafeNow;
  }

 return isSafe;
}
```

Sadigh & Kapoor, 2016

Direction: Verification, security, cryptography



Direction: Verification, security, cryptography

Static analysis

Run-time verification

Whitebox fuzzing

Cybersecurity to protect attack surfaces

Appropriate use of physical security, isolation

Encryption for data integrity, protection of interprocess comms.

Direction: Runtime verification

Difficult to do formal analysis of large-scale system

→ Analysis & execution considers info. from running system

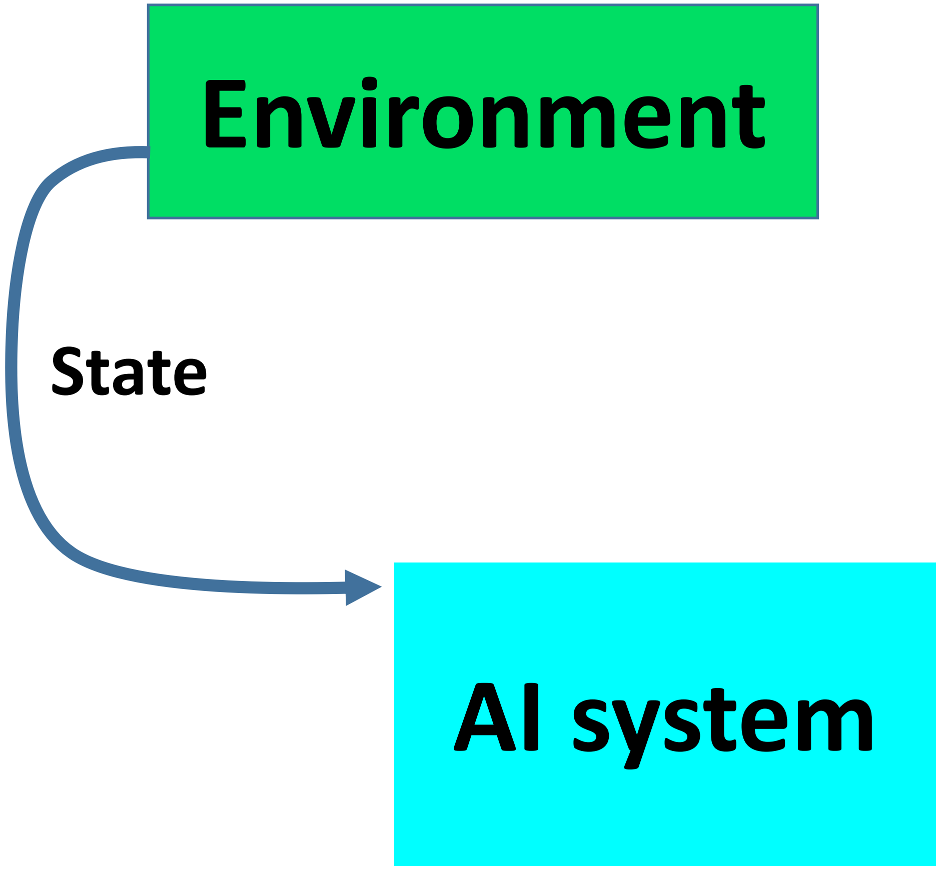
Satisfy or violate desired properties?

Identify problem, future problem

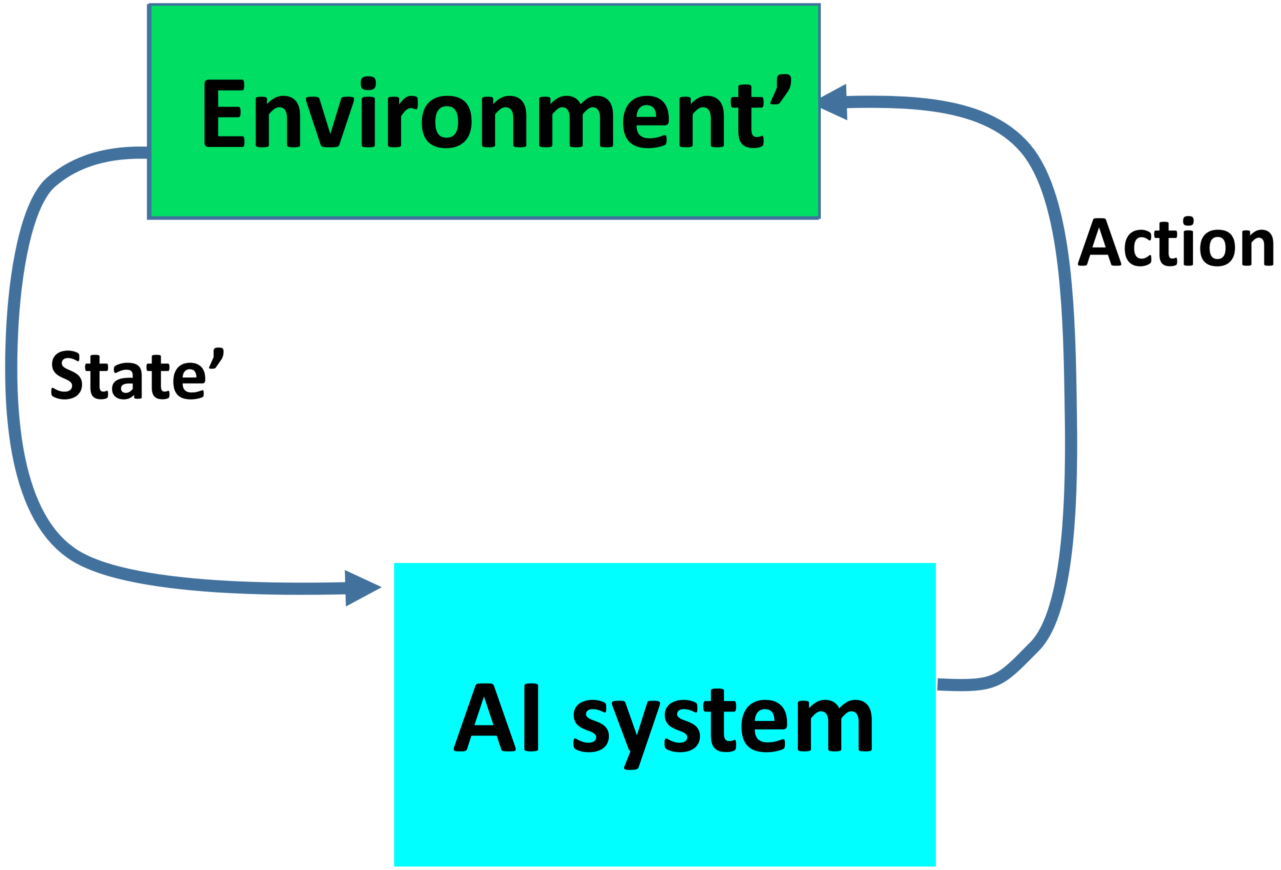
Engage human

Take fail-safe action

Direction: Metalevel analysis, monitoring, assurance



Direction: Metalevel analysis, monitoring, assurance



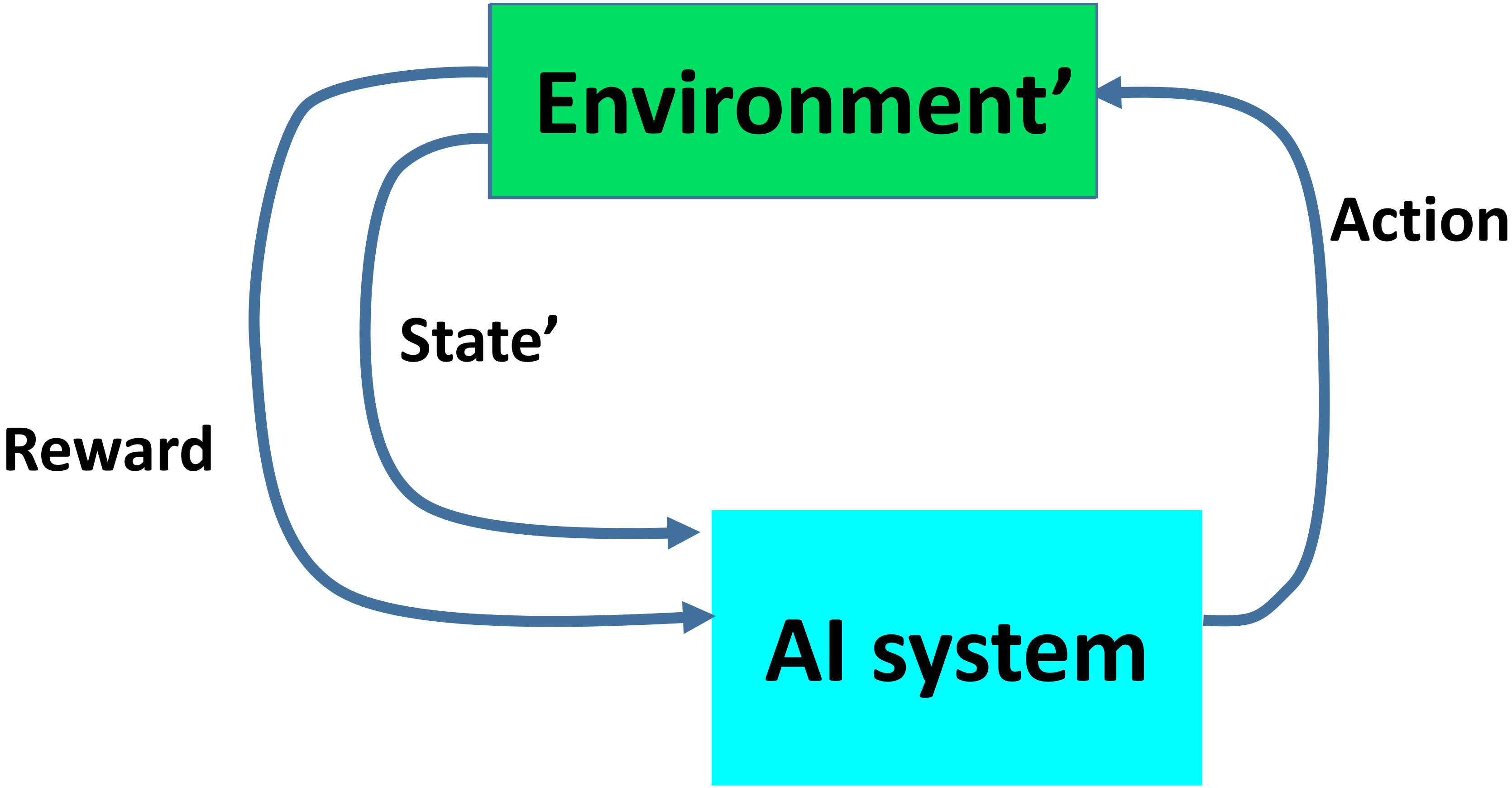
Environment’

Action

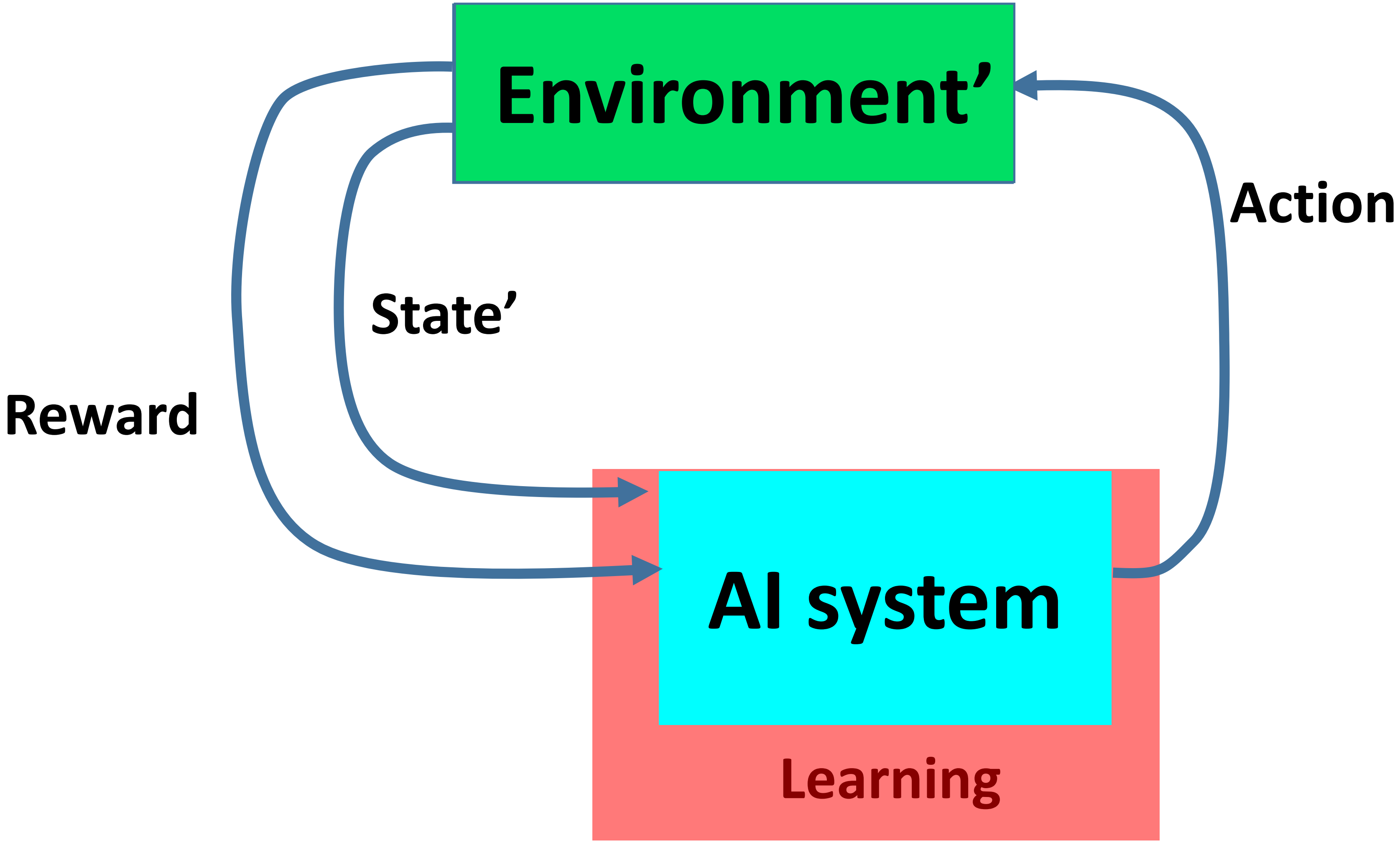
State’

AI system

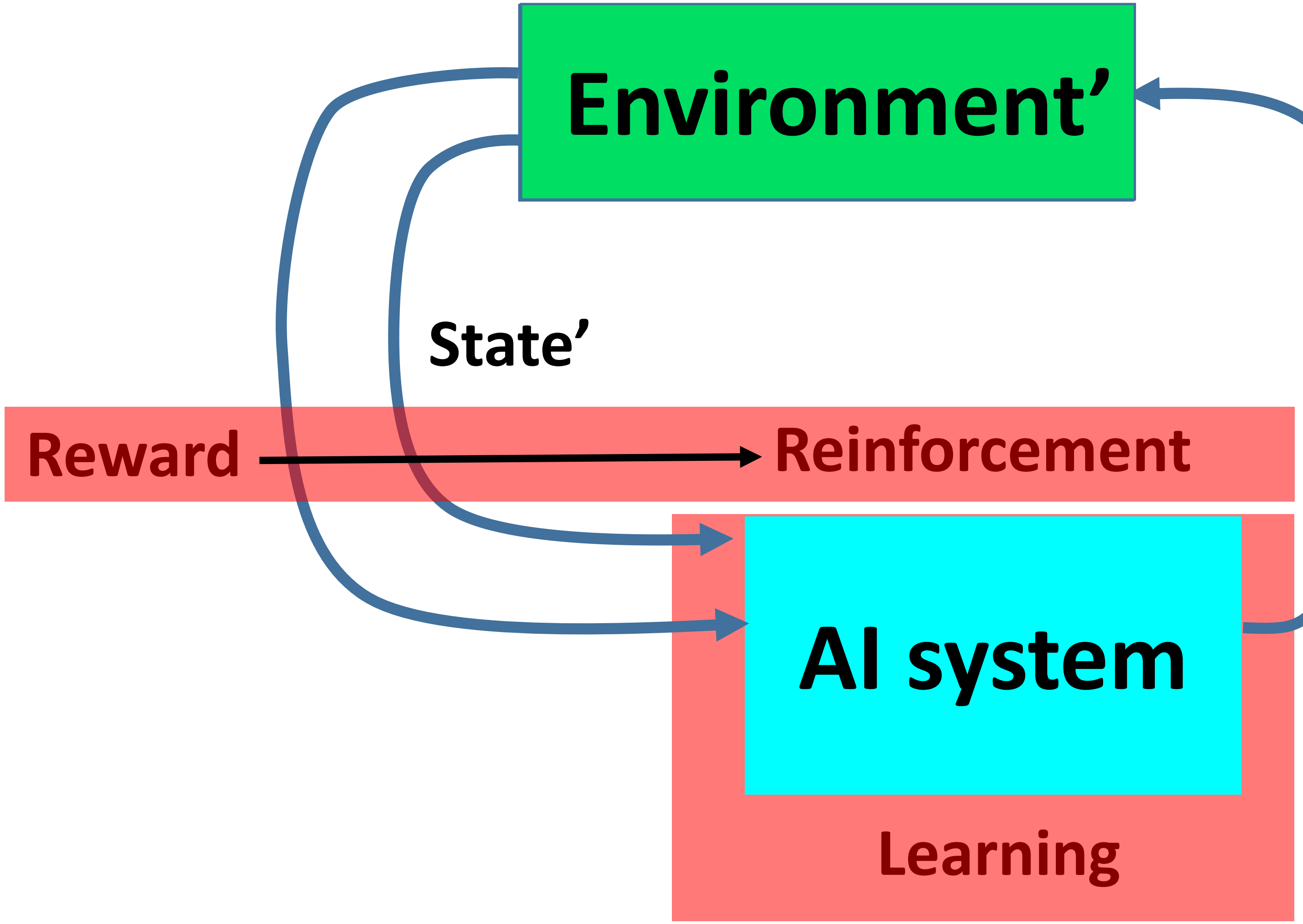
Direction: Metalevel analysis, monitoring, assurance



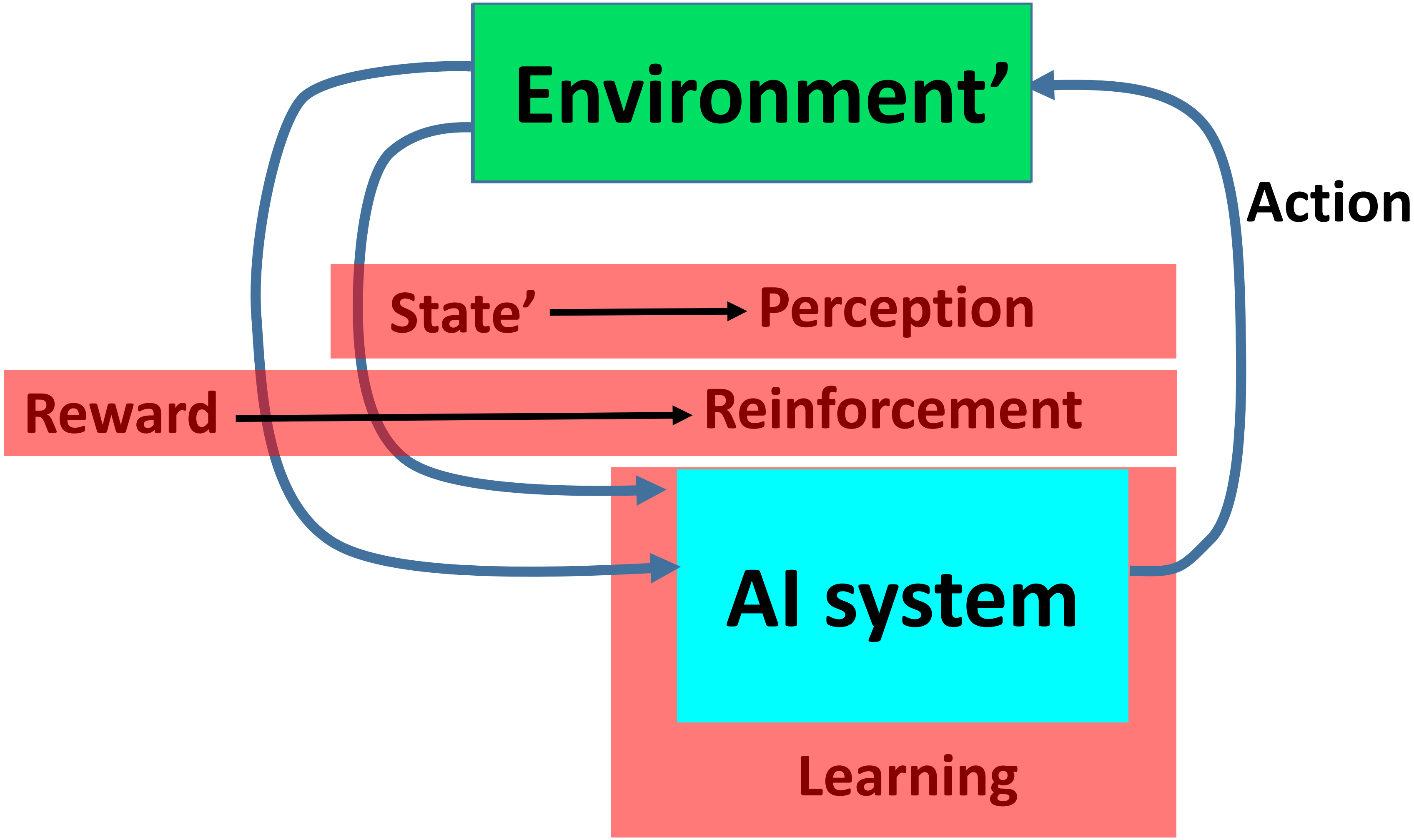
Direction: Metalevel analysis, monitoring, assurance



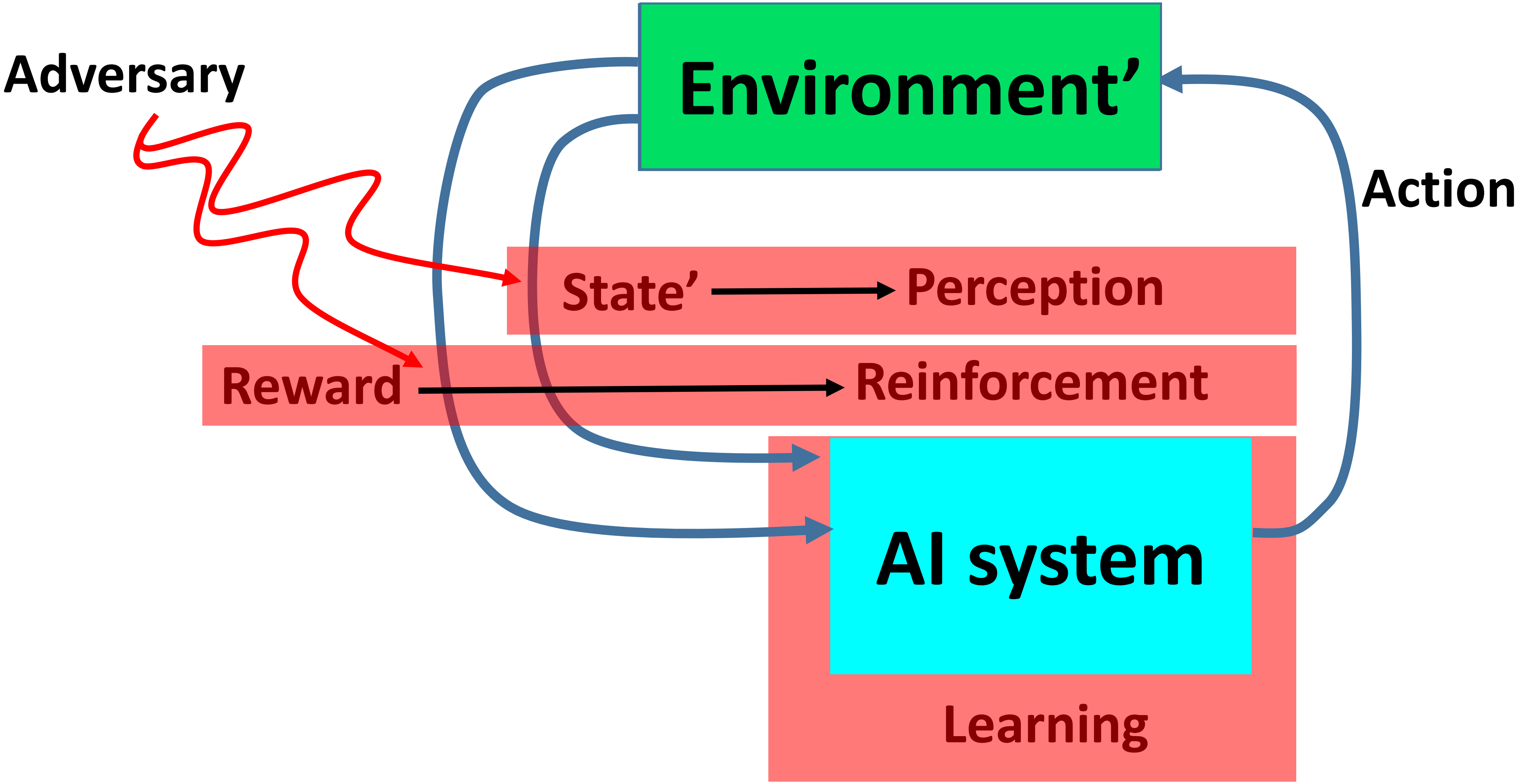
Direction: Metalevel analysis, monitoring, assurance



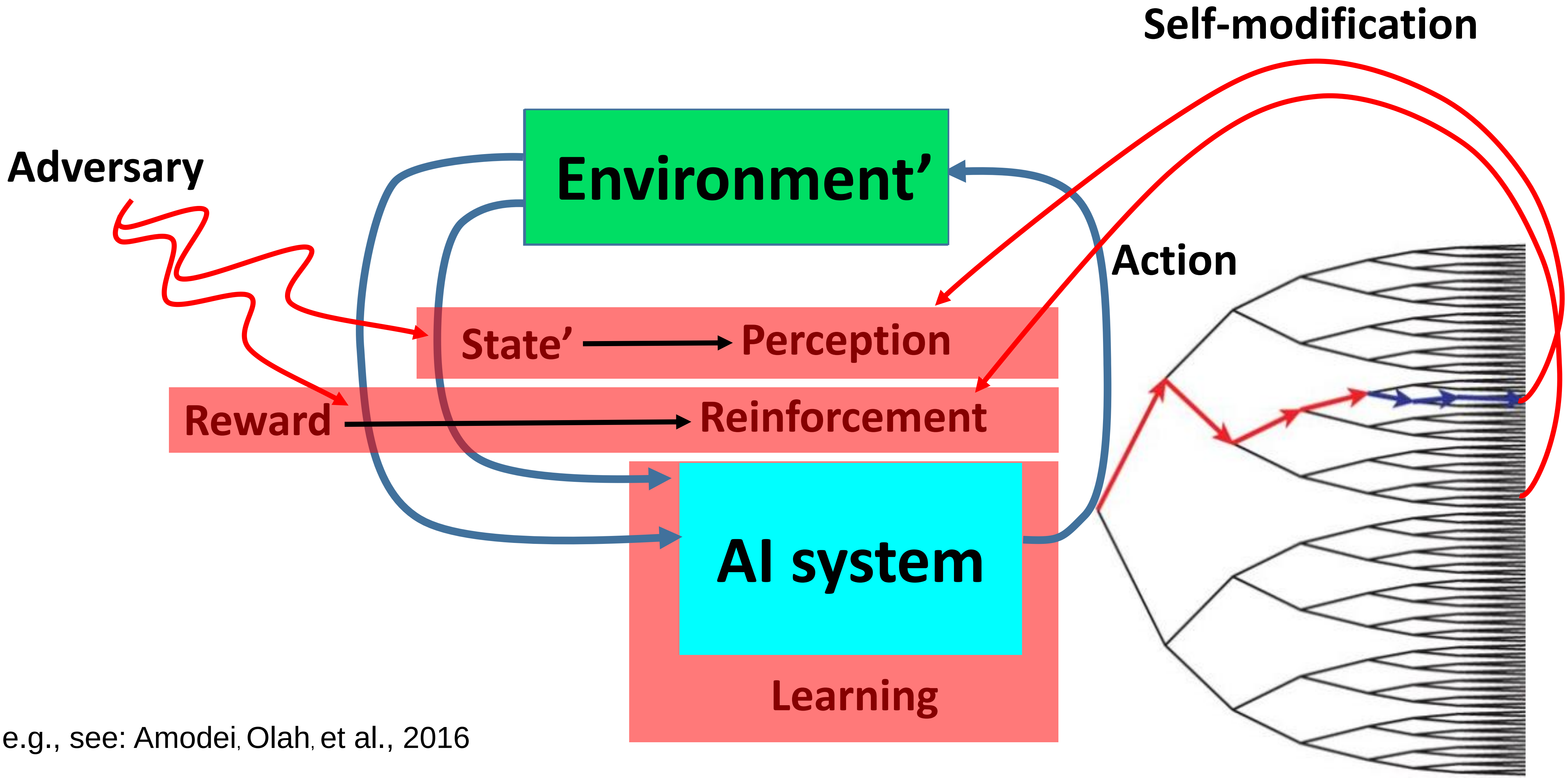
Direction: Metalevel analysis, monitoring, assurance



Direction: Metalevel analysis, monitoring, assurance

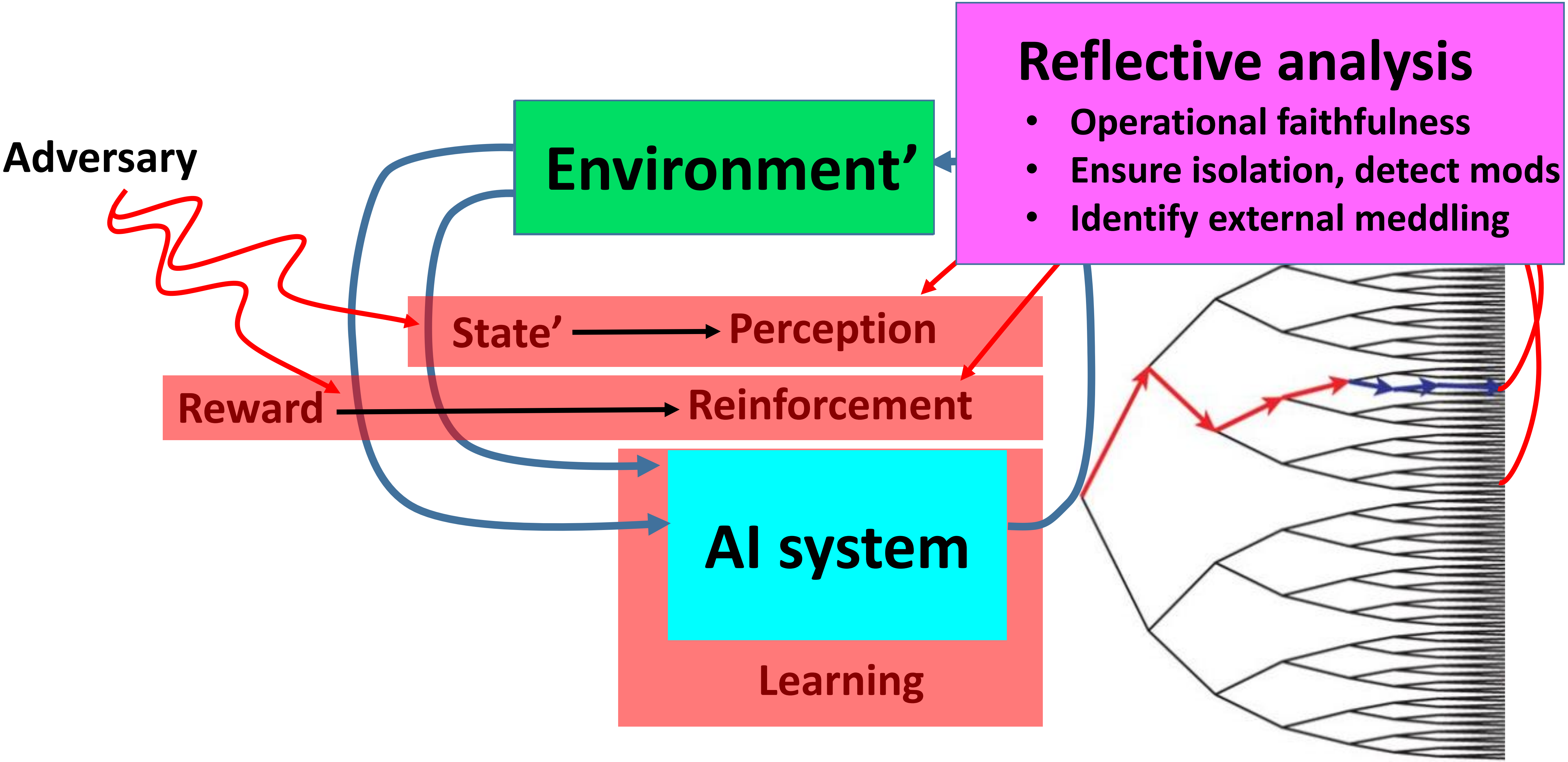


Direction: Metalevel analysis, monitoring, assurance



e.g., see: Amodei, Olah, et al., 2016

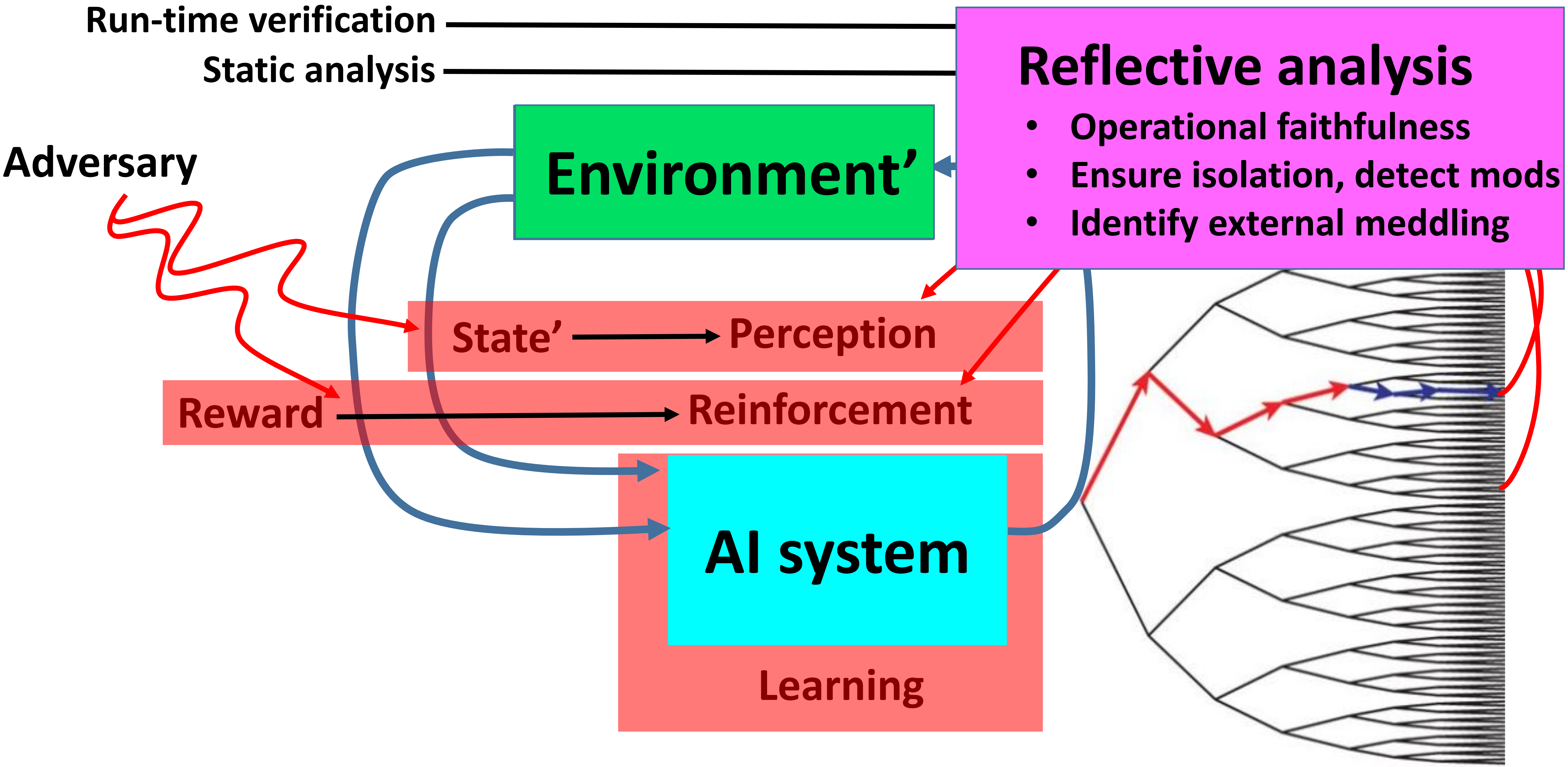
Direction: Metalevel analysis, monitoring, assurance



Reflective analysis

- Operational faithfulness
- Ensure isolation, detect mods
- Identify external meddling

Direction: Metalevel analysis, monitoring, assurance



Run-time verification

Static analysis

Adversary

Environment’

Reflective analysis

- Operational faithfulness
- Ensure isolation, detect mods
- Identify external meddling

State’ → Perception

Reward → Reinforcement

AI system

Learning

Direction: Human-machine collaboration

Models of human cognition

Transparency of state, explanation

Mastering coordination of initiatives

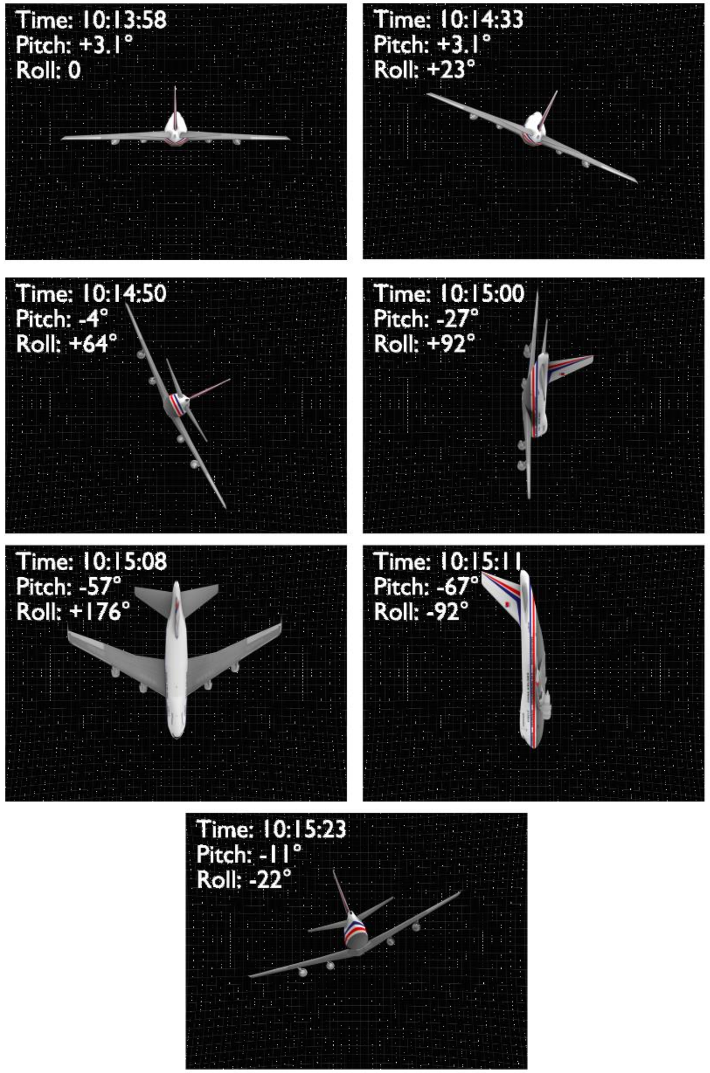
Direction: Human-machine collaboration

China Airlines 006 (Feb 1985)

747 dives 10,000 in 20 seconds. 5g, supersonic.

Air France 447 (June 2009)

Unrecoverable stall.



Direction: Human-machine collaboration

China Airlines 006 (Feb 1985)

747 dives 10,000 in 20 seconds. 5g, supersonic.

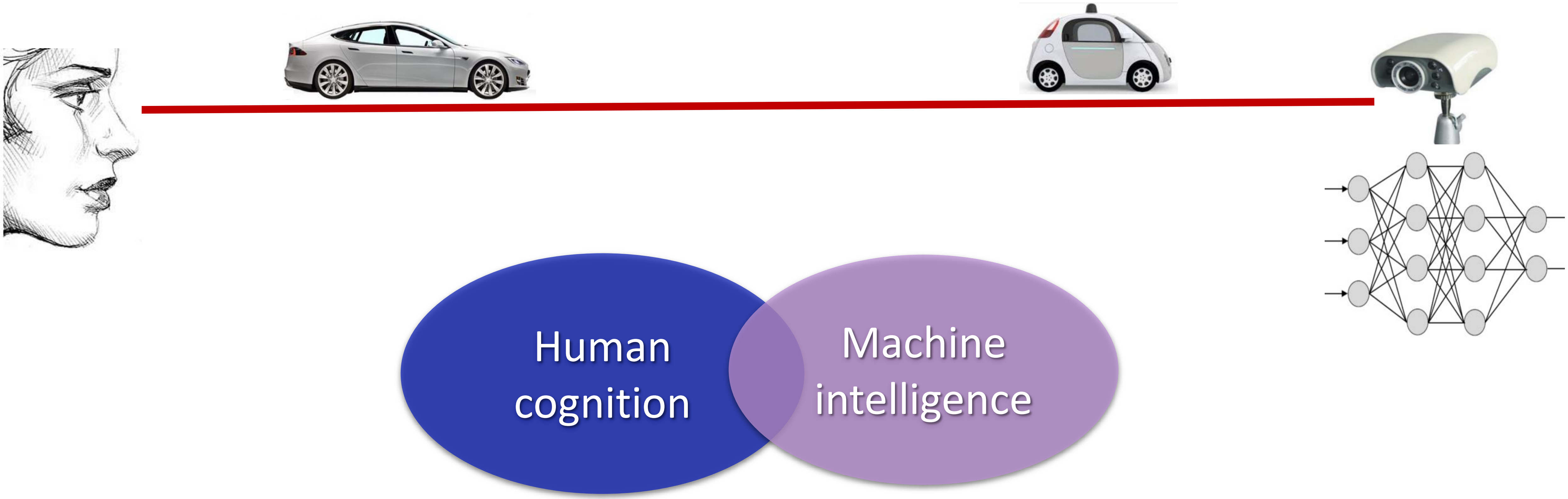
Air France 447 (June 2009)

Unrecoverable stall.

Direction: Human-machine collaboration

Rich spectrum of autonomy

How to best work together for safety?



Human cognition

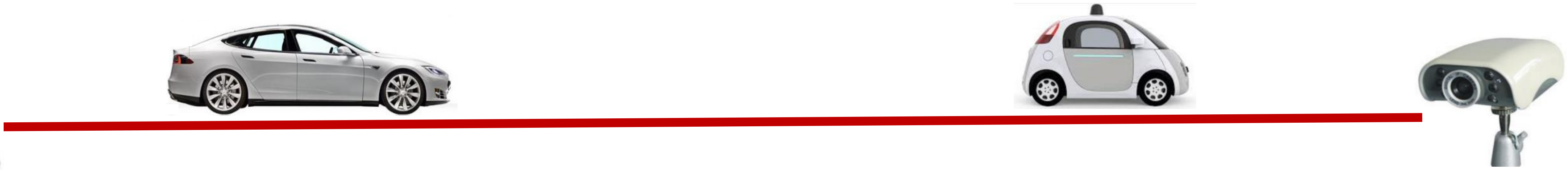
Machine intelligence

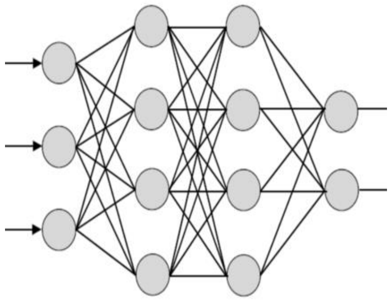
Kamar, Hacker, H., 2012

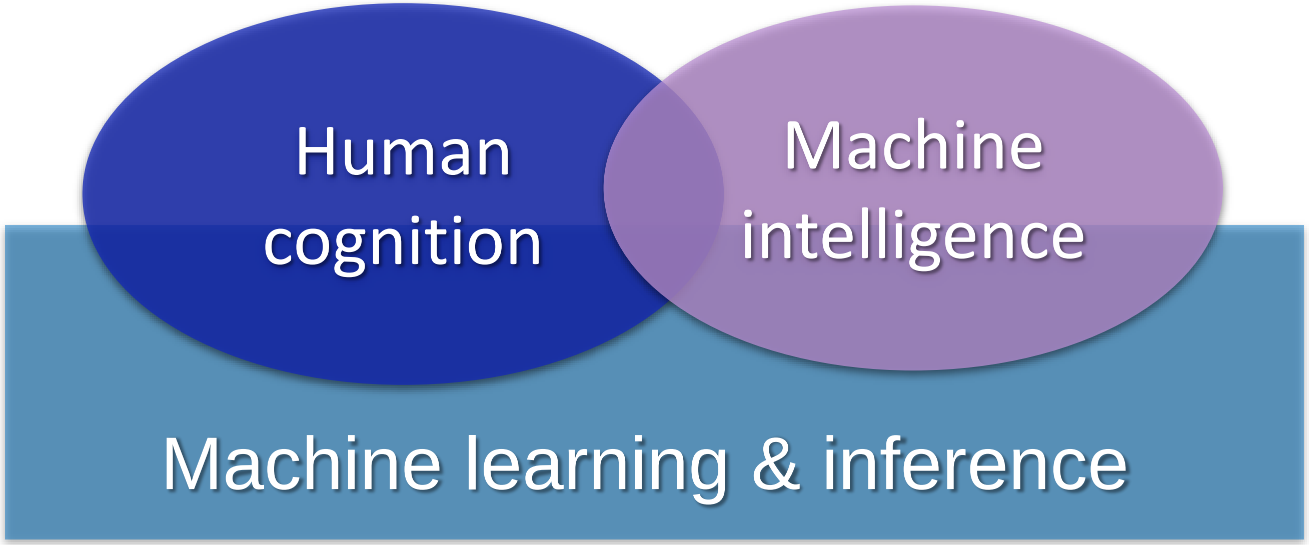
Direction: Human-machine collaboration

Rich spectrum of autonomy

How to best work together for safety?







Human cognition

Machine intelligence

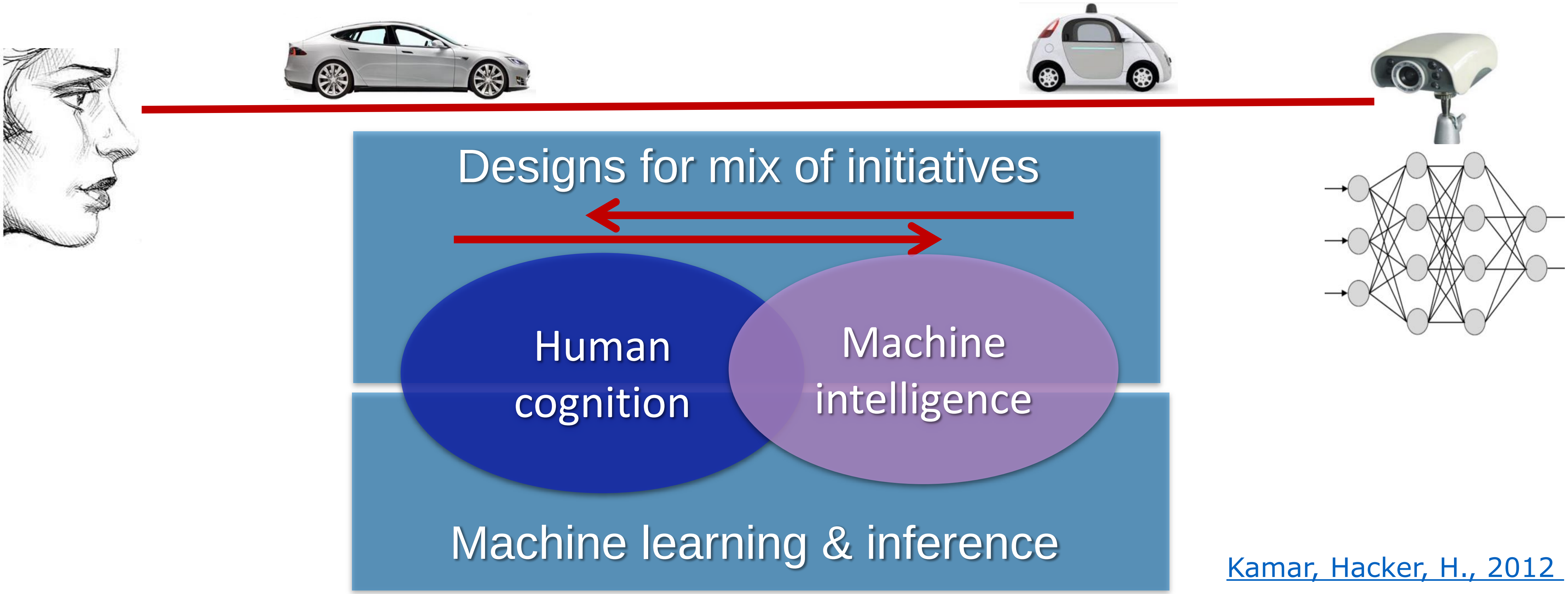
Machine learning & inference

Kamar, Hacker, H., 2012

Direction: Human-machine collaboration

Rich spectrum of autonomy

How to best work together for safety?



Designs for mix of initiatives

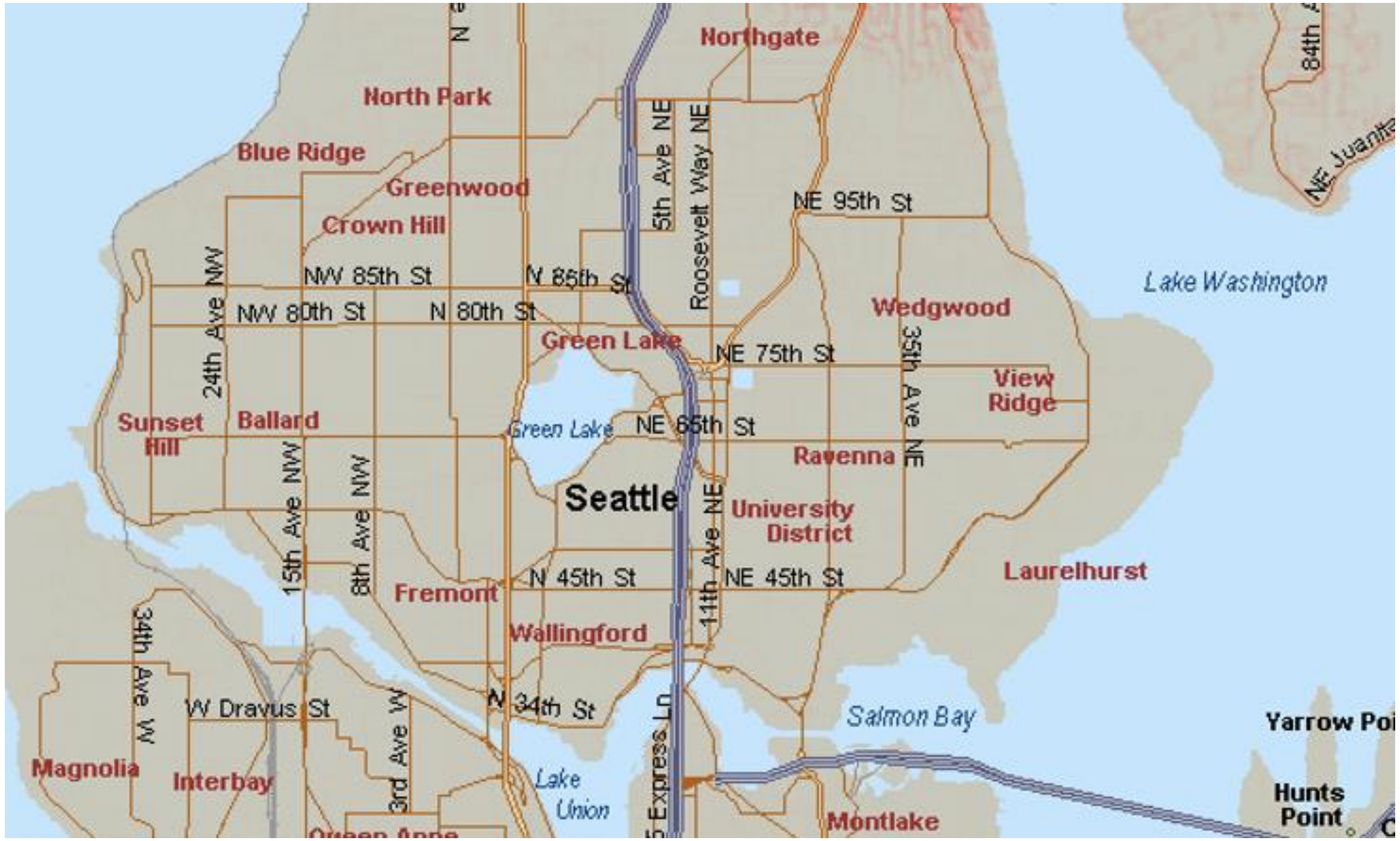
Human cognition

Machine intelligence

Machine learning & inference

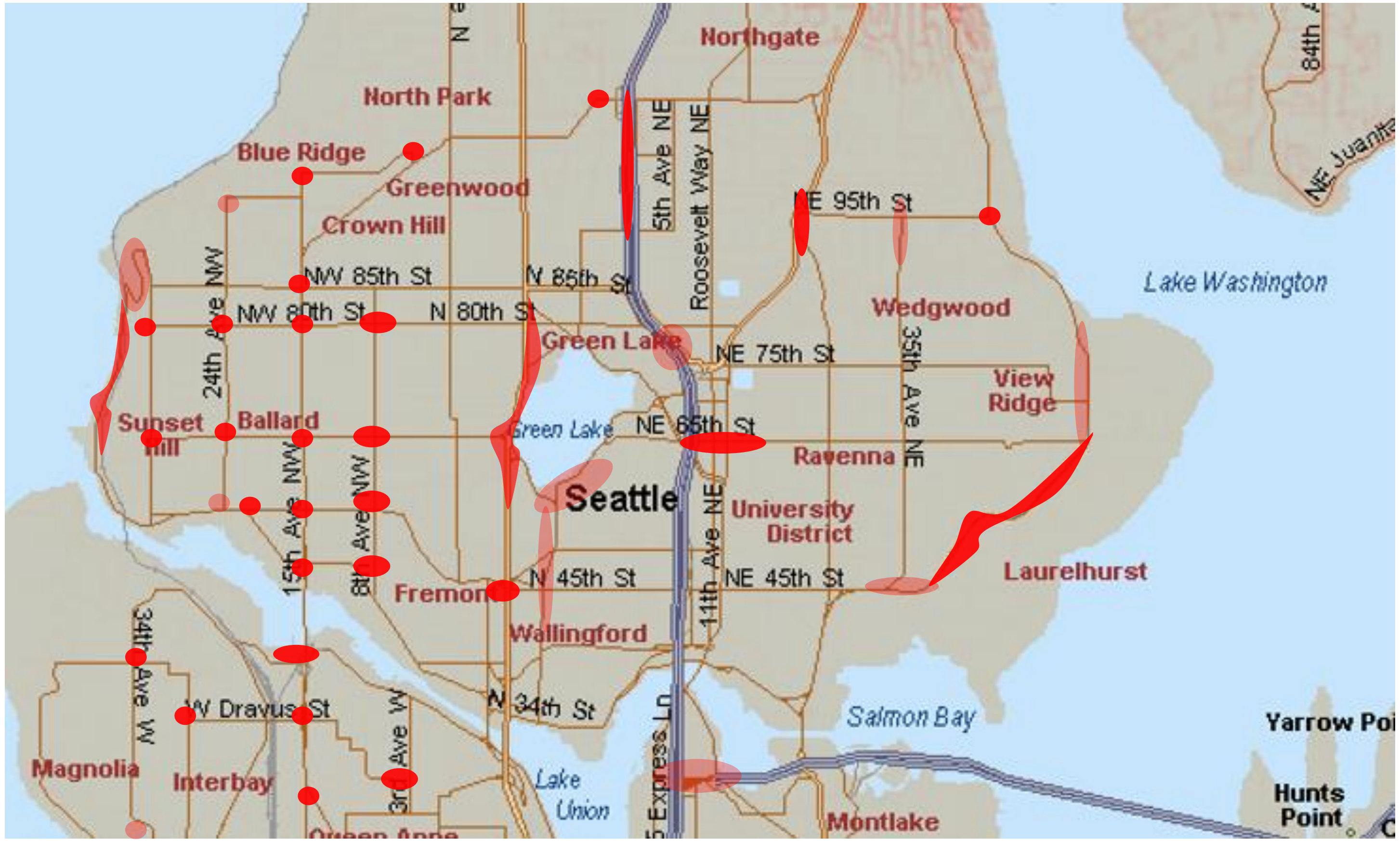
Kamar, Hacker, H., 2012

Direction: Human-machine collaboration



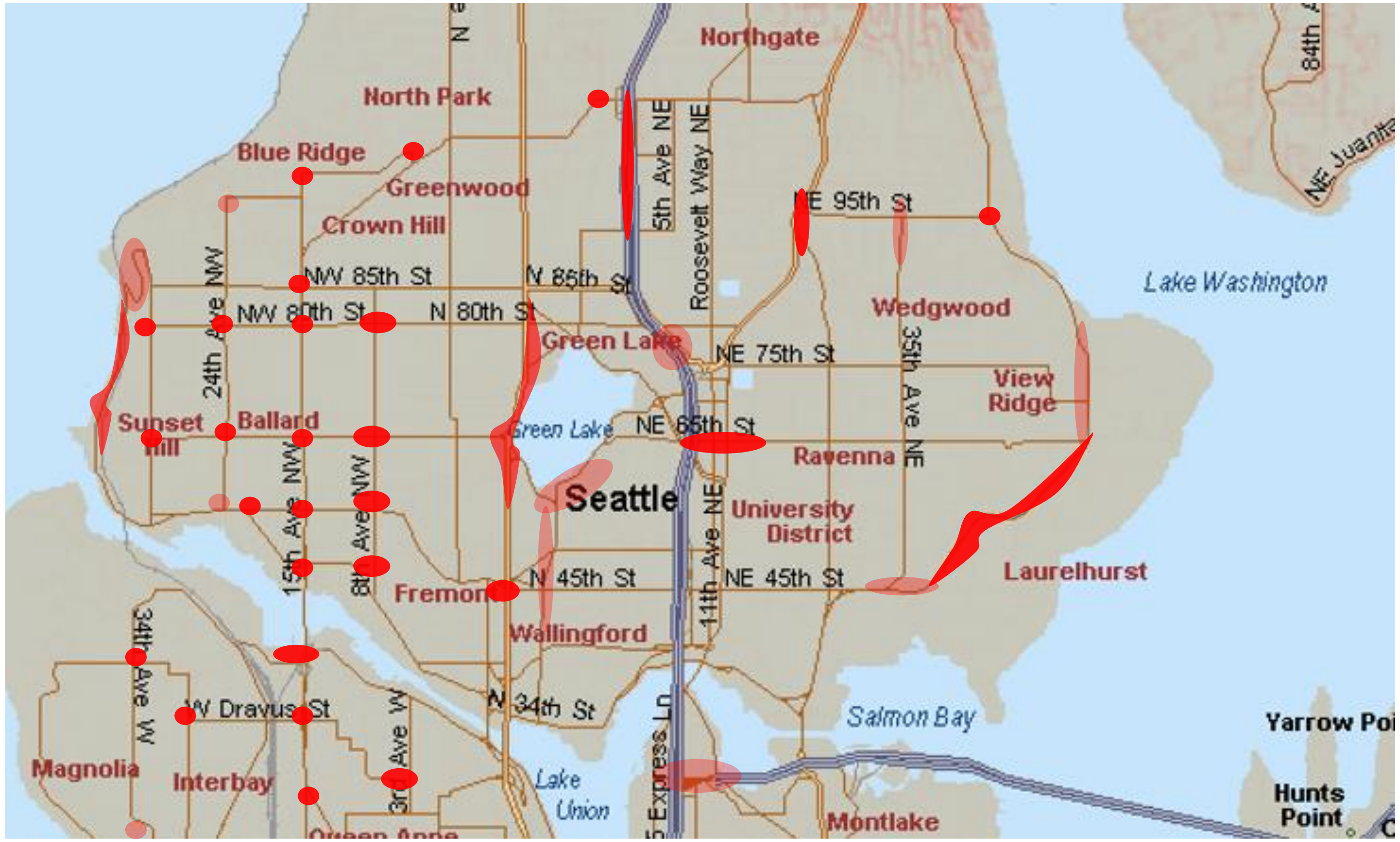
Direction: Human-machine collaboration

Infer challenges with machine competency

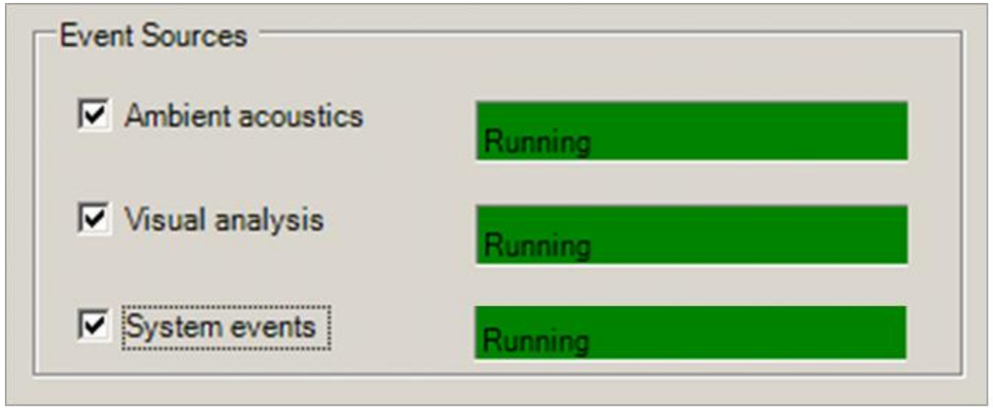


Direction: Human-machine collaboration

Infer challenges with machine competency



Infer human attention



Event Sources

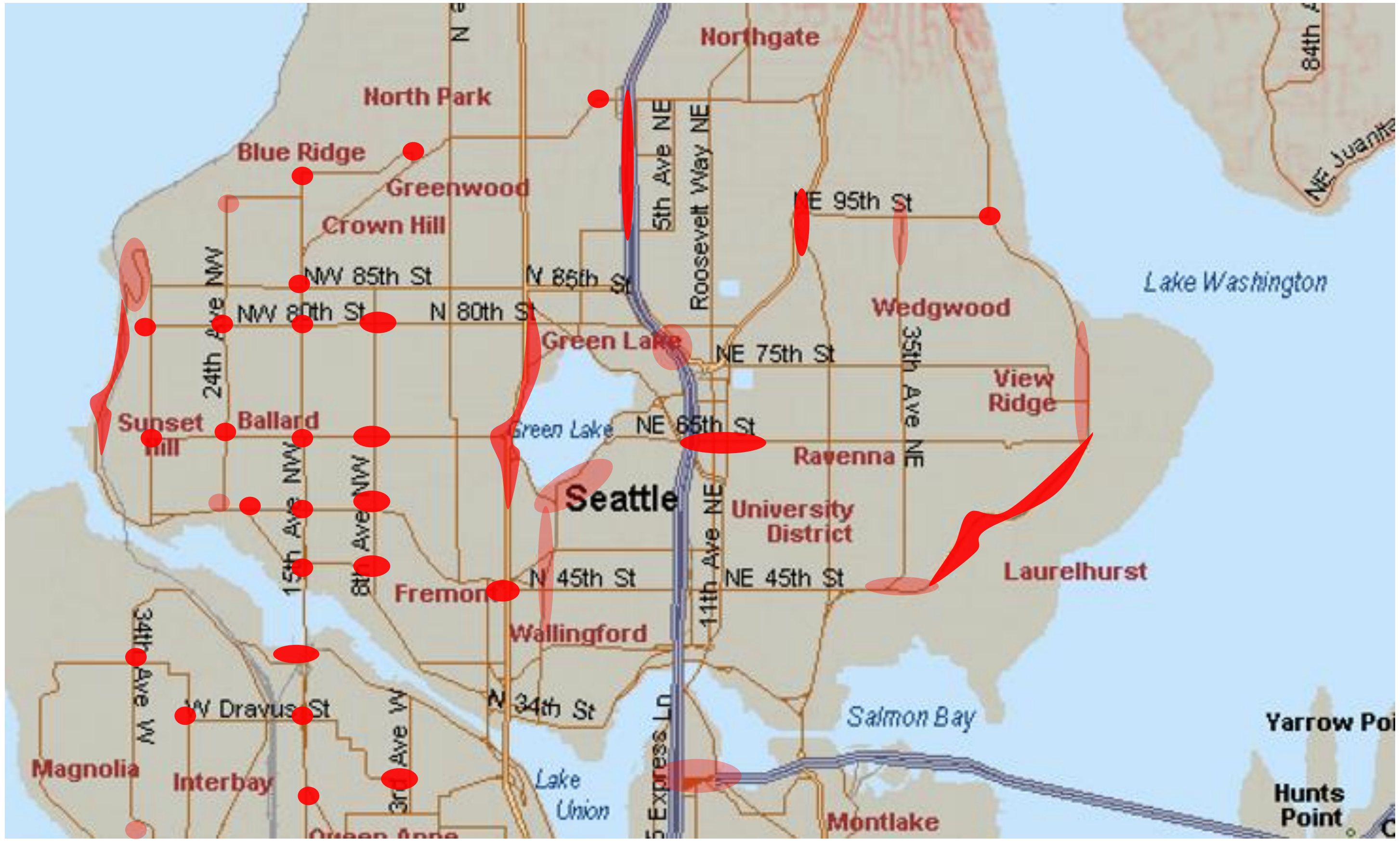
☑ Ambient acoustics Running

☑ Visual analysis Running

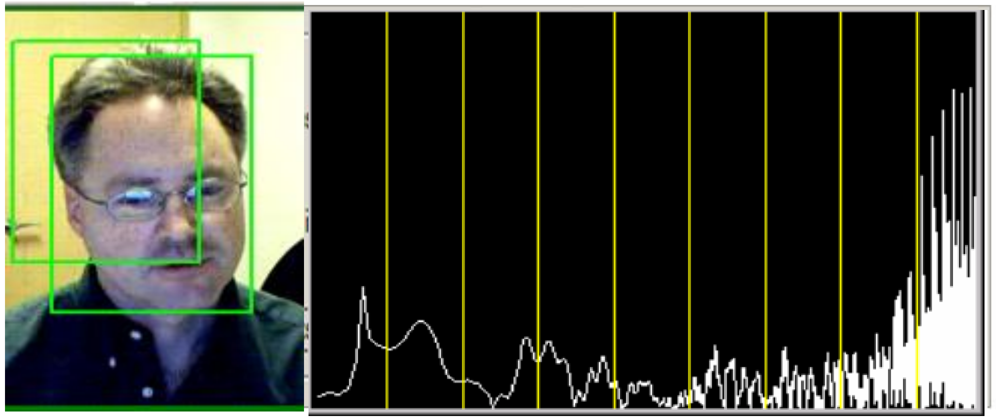
☑ System events Running

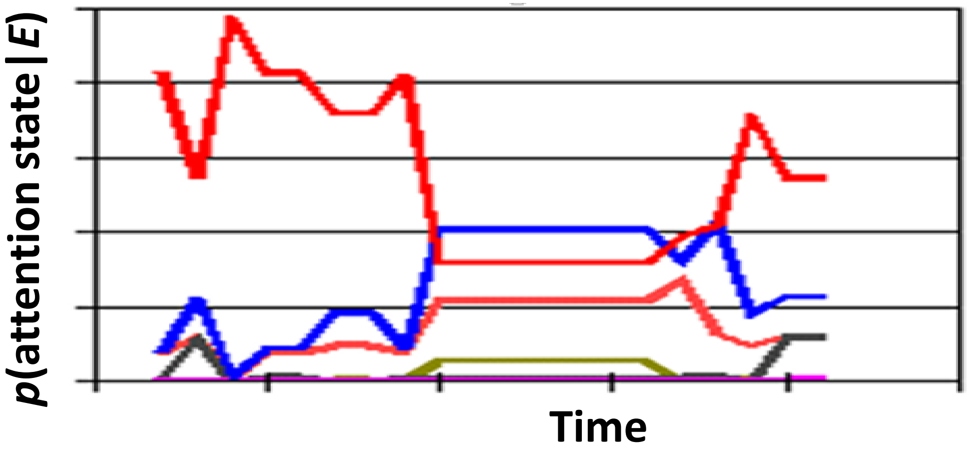
Direction: Human-machine collaboration

Infer challenges with machine competency



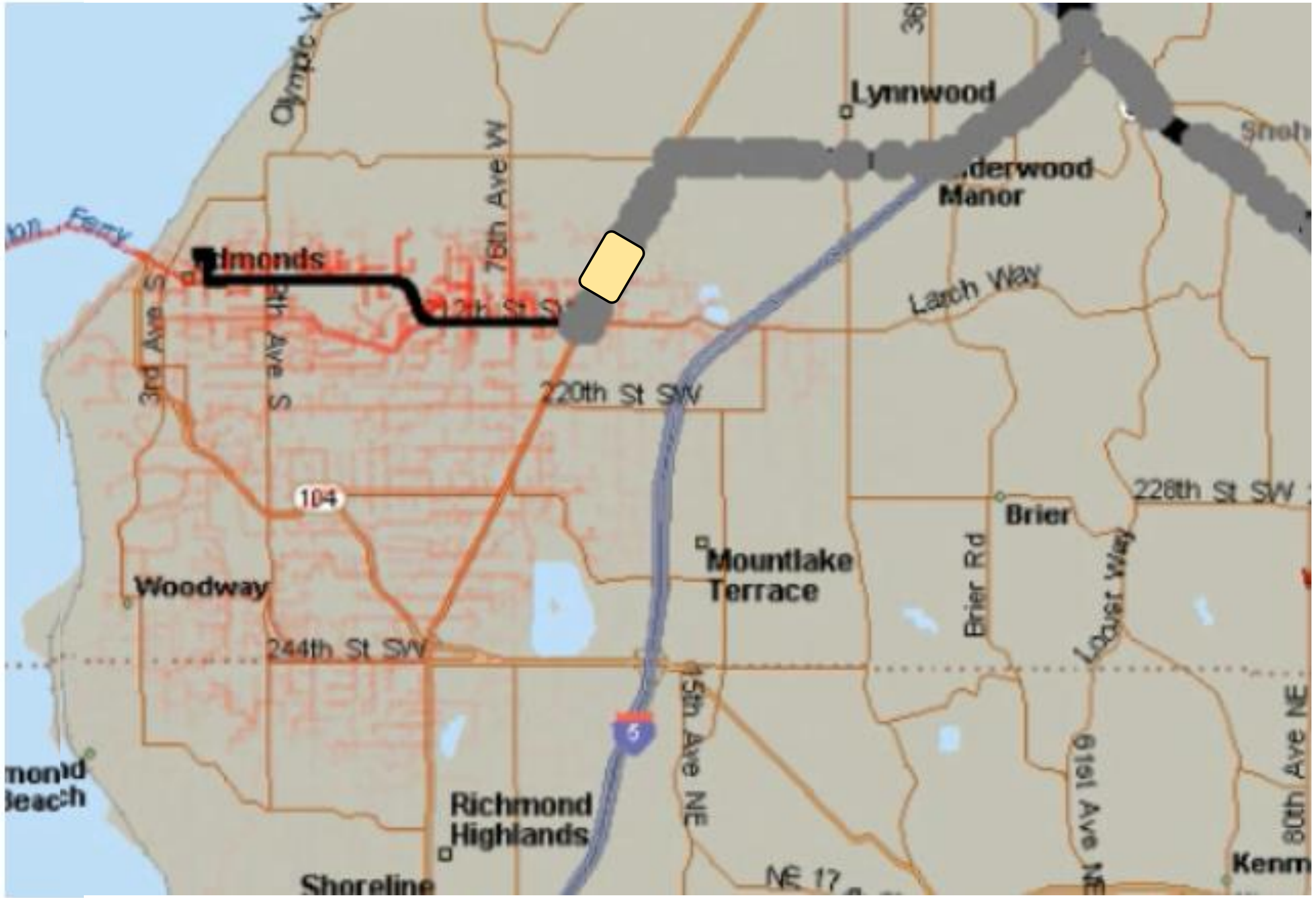
Infer human attention



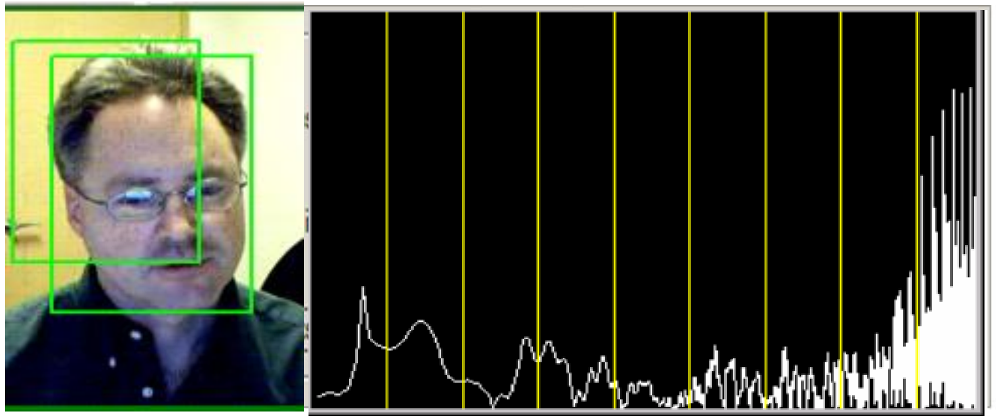


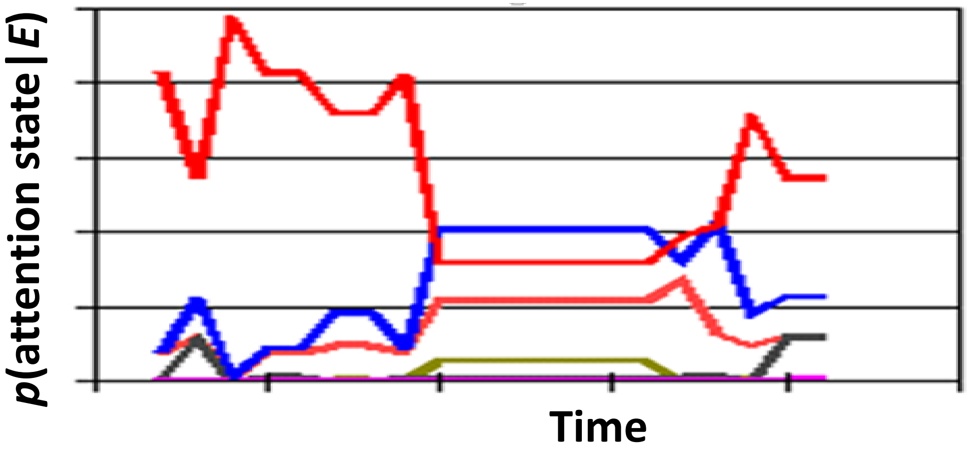
Direction: Human-machine collaboration

Continual prediction of trajectories



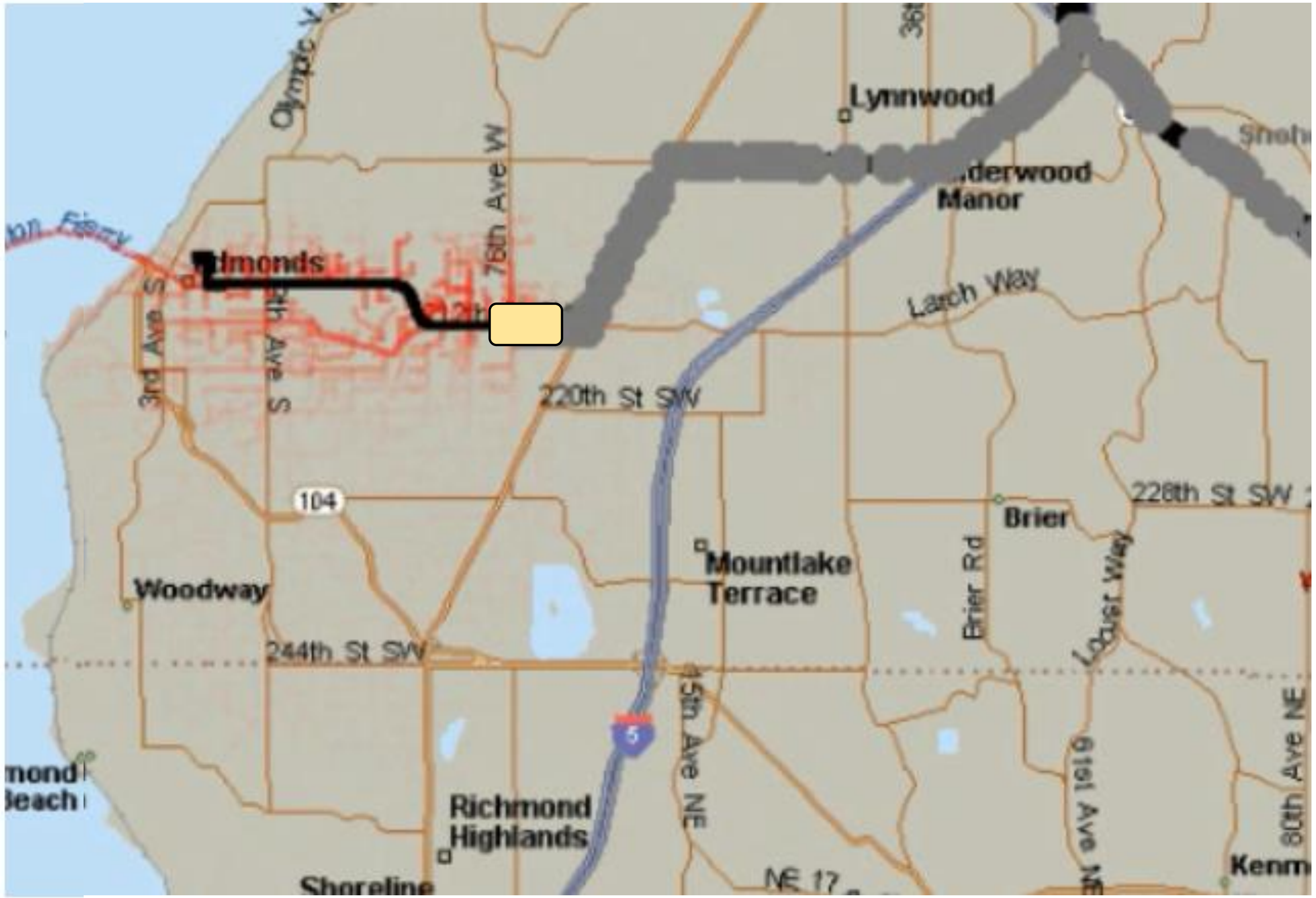
Infer human attention



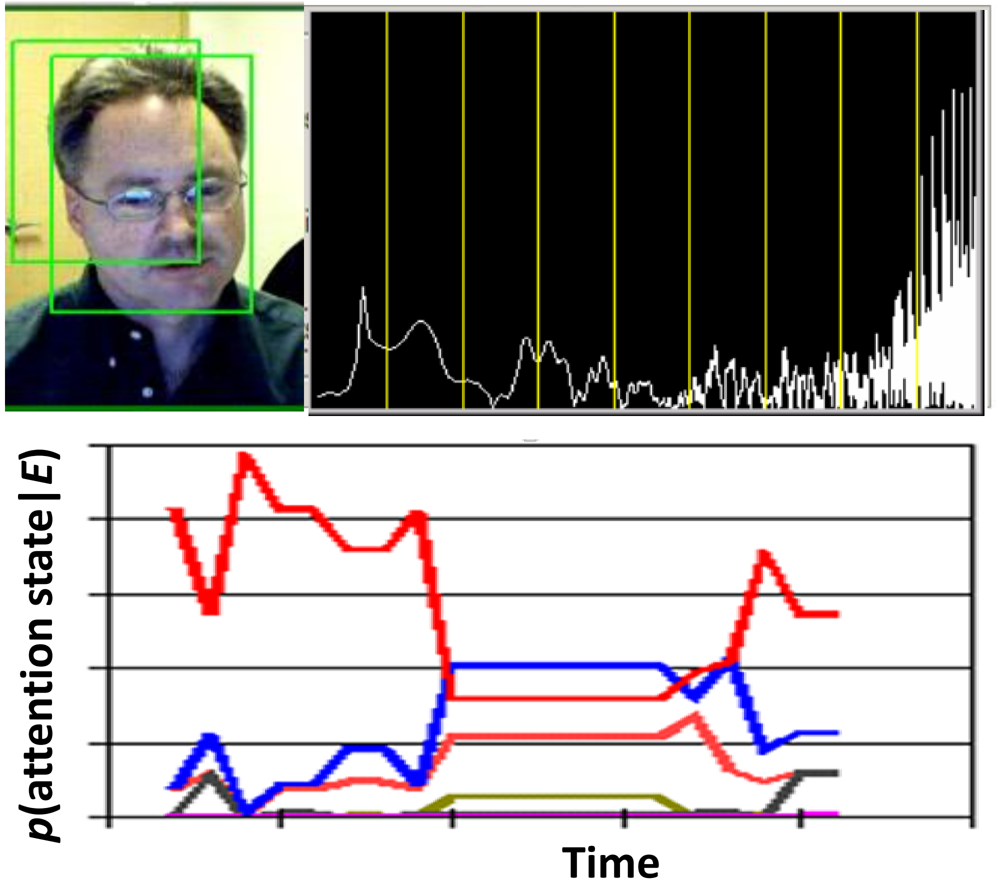


Direction: Human-machine collaboration

Continual prediction of trajectories

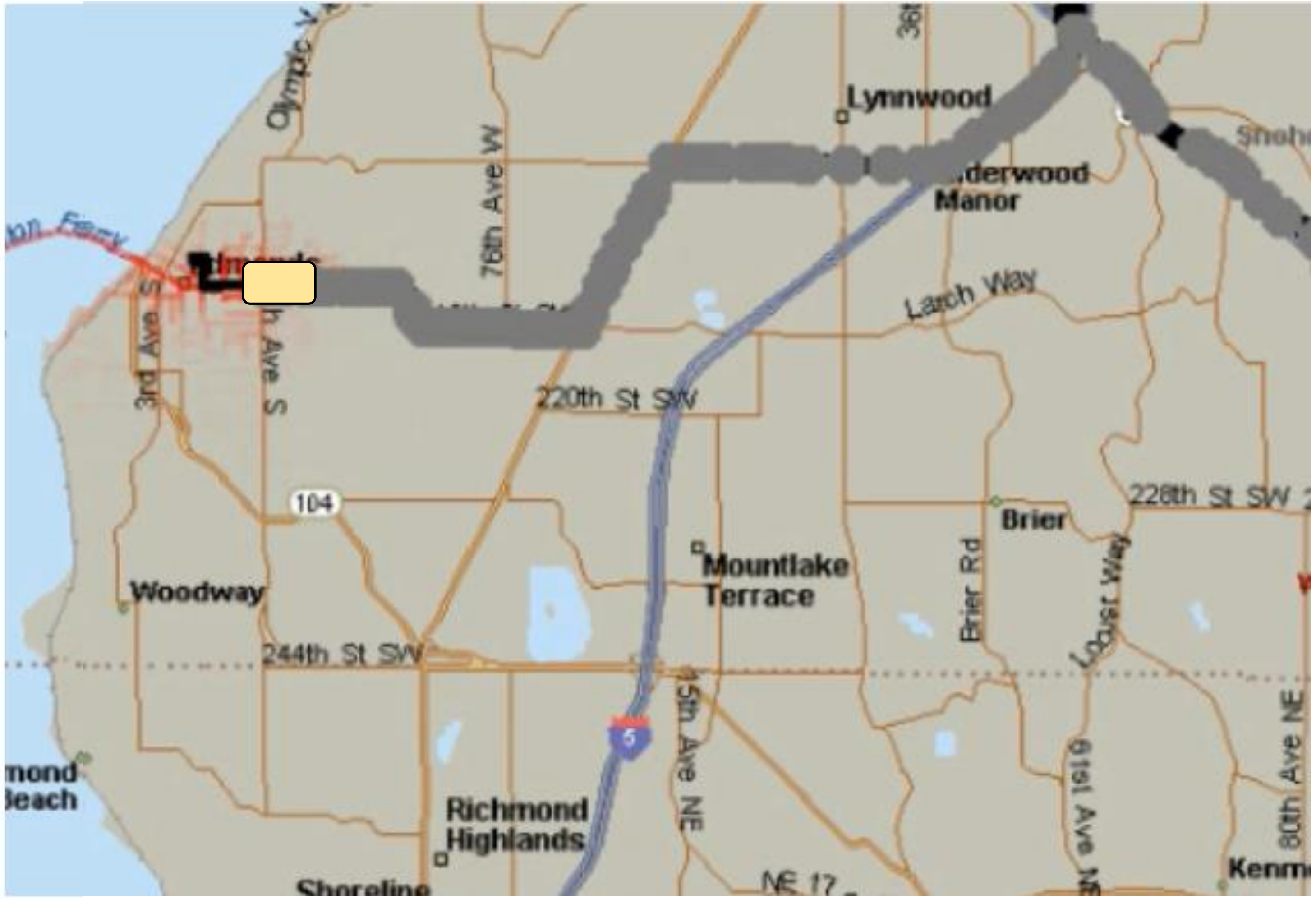


Infer human attention

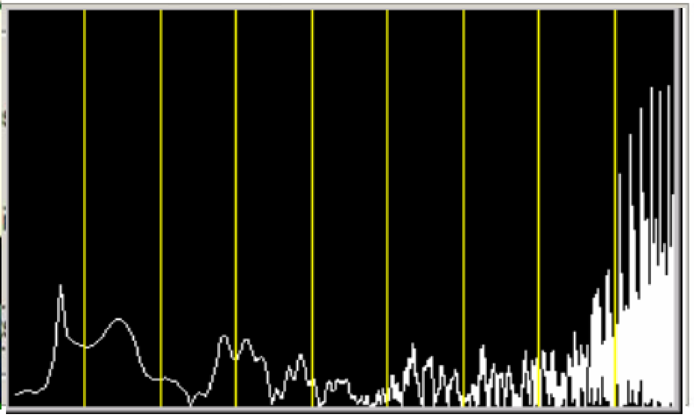


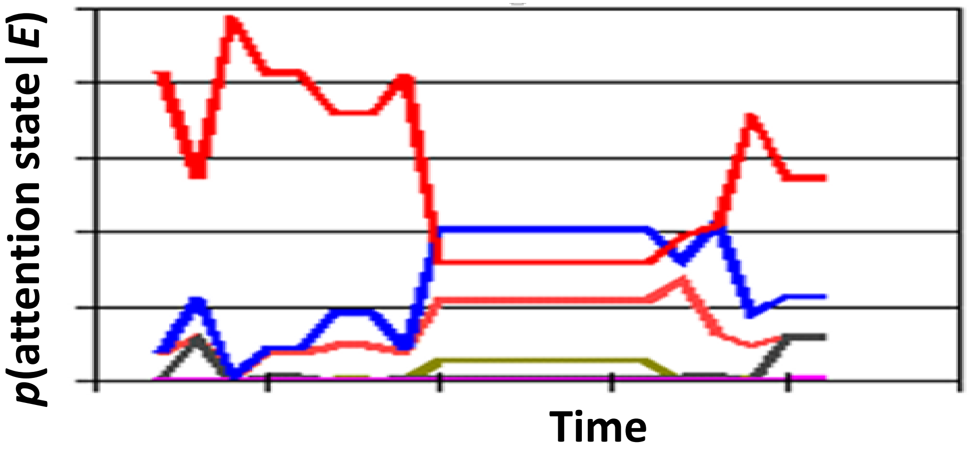
Direction: Human-machine collaboration

Continual prediction of trajectories



Infer human attention

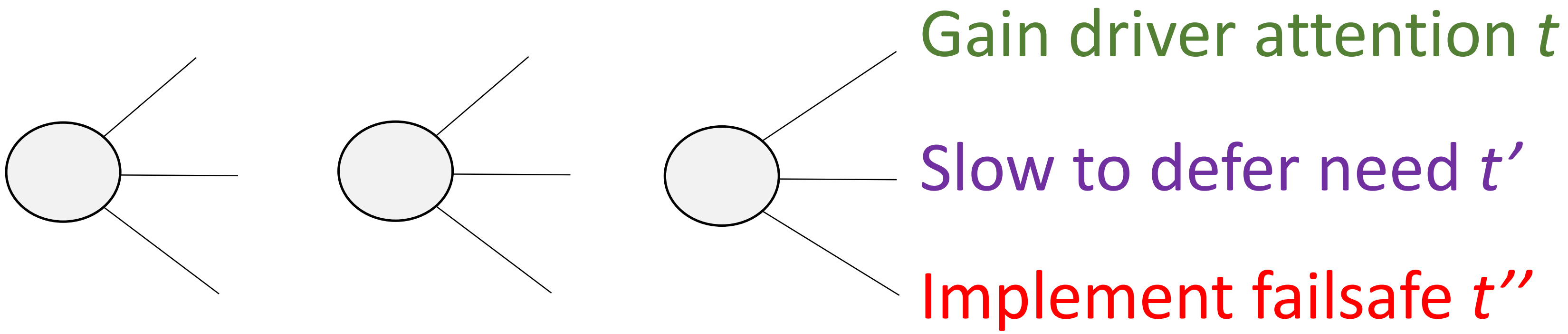




Direction: Human-machine collaboration

Safety-assuring mixed-initiative planner

- Driver’s attention over time
- Latency of human input
- Latency tolerance of situation
- Cost & influence of alerting driver
- Custom language, ongoing dialog



Gain driver attention *t*

Slow to defer need *t’*

Implement failsafe *t’’*

Direction: Develop Best Practices for Safe AI

- Phases of study, testing, reporting for rolling out new capabilities in safety-critical domains (akin to FDA clinical trials, post-marketing surveillance)
- Disclosure & control of parameters on failure rates, tradeoffs, preferences
- Transparency & explainability of perception, inference, action
- System self-monitoring & reporting machinery
- Isolation of components in intelligence architectures
- Detecting & addressing feedback of system’s influence on self

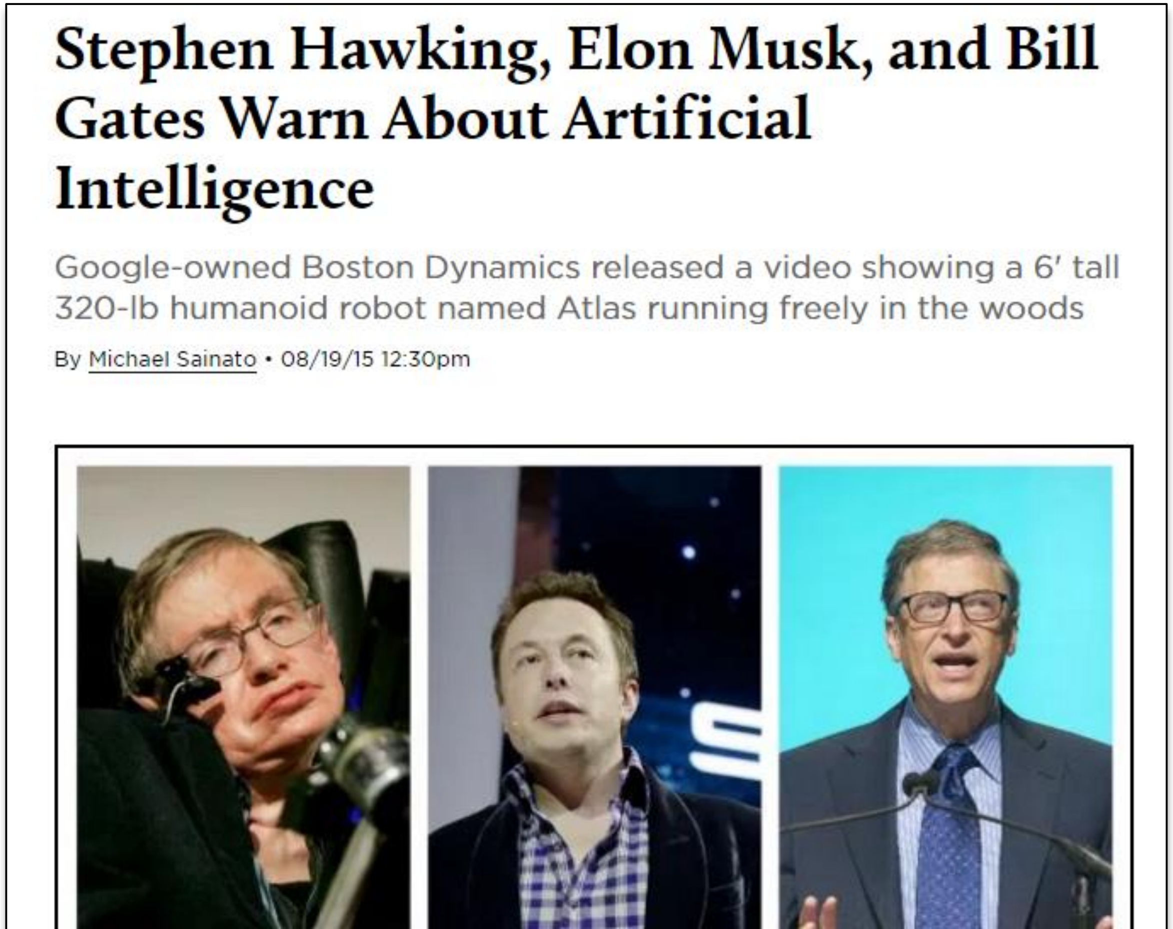
Direction: Develop Best Practices for Safe AI

- Standard protocols for handoffs, attention, awareness, warning, in human-machine collaborations
- Policies for visible disclosure of autonomy to others (e.g., indication to others that a car is currently on automated policy)
- Fail-safe actions & procedures given predicted or sensed failures
- Enhancing robustness via co-design of environment & systems
- Testing for drift of assumptions, distributions in domains
- Special *openness* & *adherence to best practices* for data, learning, decision making for applications in governance & public policy

Direction: Address concerns about “superintelligences”

Addressing concerns of public

Significant differences of opinion, including experts



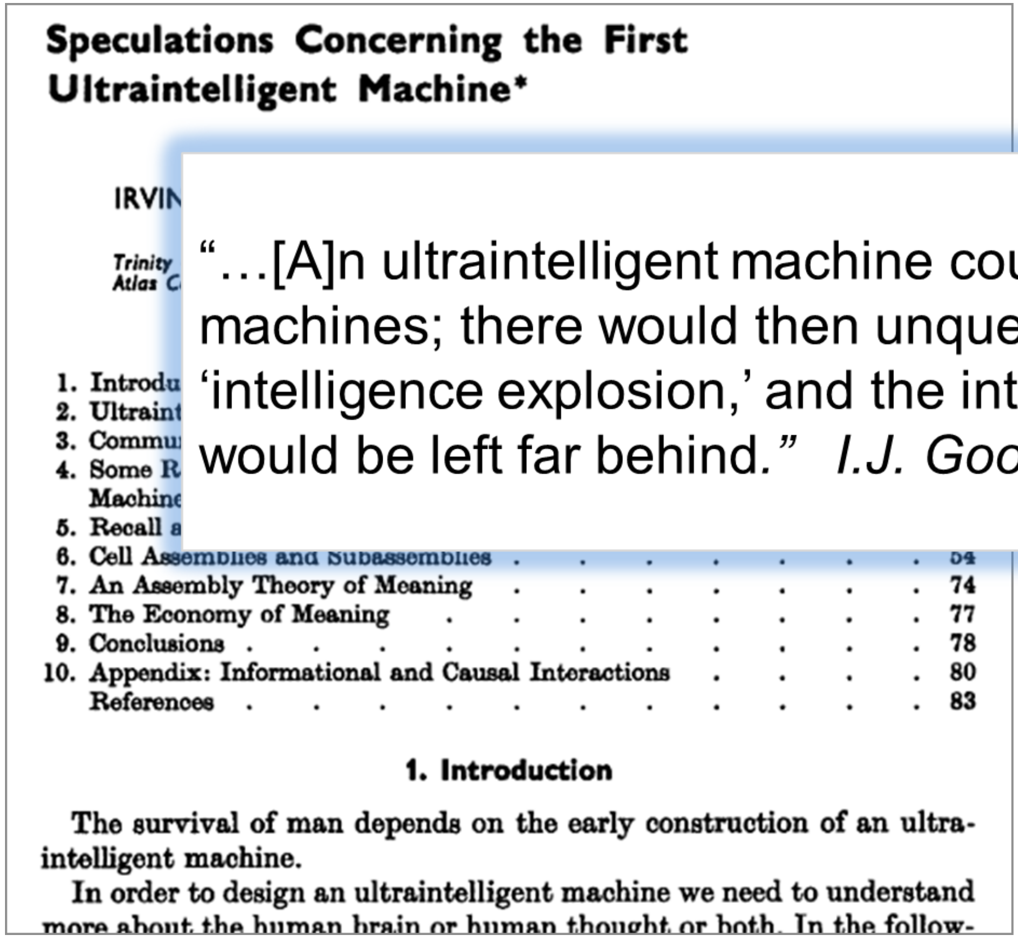
Stephen Hawking, Elon Musk, and Bill Gates Warn About Artificial Intelligence

Google-owned Boston Dynamics released a video showing a 6′ tall 320-lb humanoid robot named Atlas running freely in the woods

By Michael Sainato • 08/19/15 12:30pm

Direction: Address concerns about “superintelligences”

- Addressing concerns of public
- Significant differences of opinion, including among experts



Speculations Concerning the First Ultraintelligent Machine*

IRVIN

Trinity
Atlas C

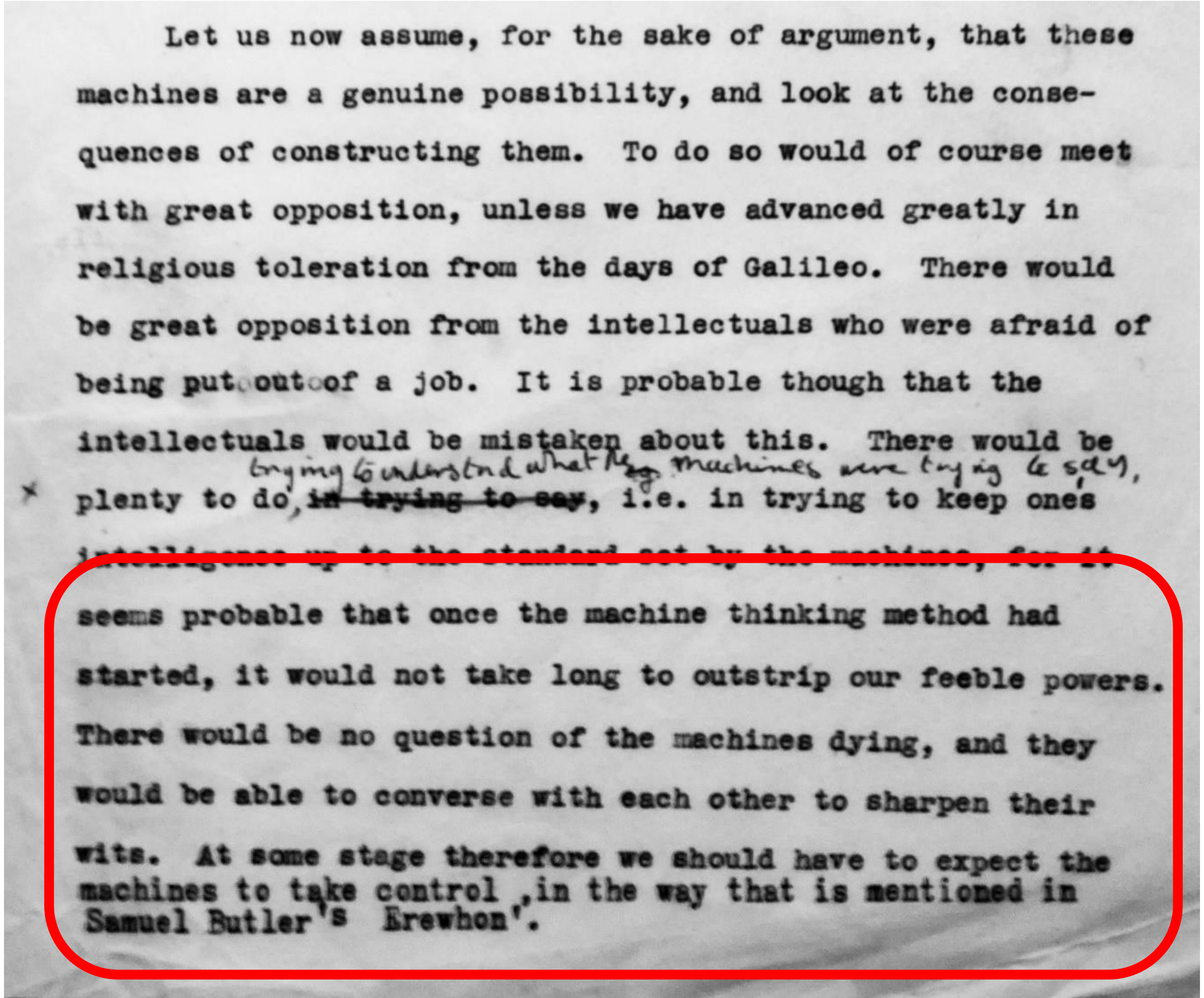
1. Introduction

The survival of man depends on the early construction of an ultraintelligent machine.

In order to design an ultraintelligent machine we need to understand more about the human brain or human thought or both. In the follow-

“…[A]n ultraintelligent machine could design even better machines; there would then unquestionably be an ‘intelligence explosion,’ and the intelligence of man would be left far behind.” *I.J. Good* (1965)

Direction: Address concerns about "superintelligences"

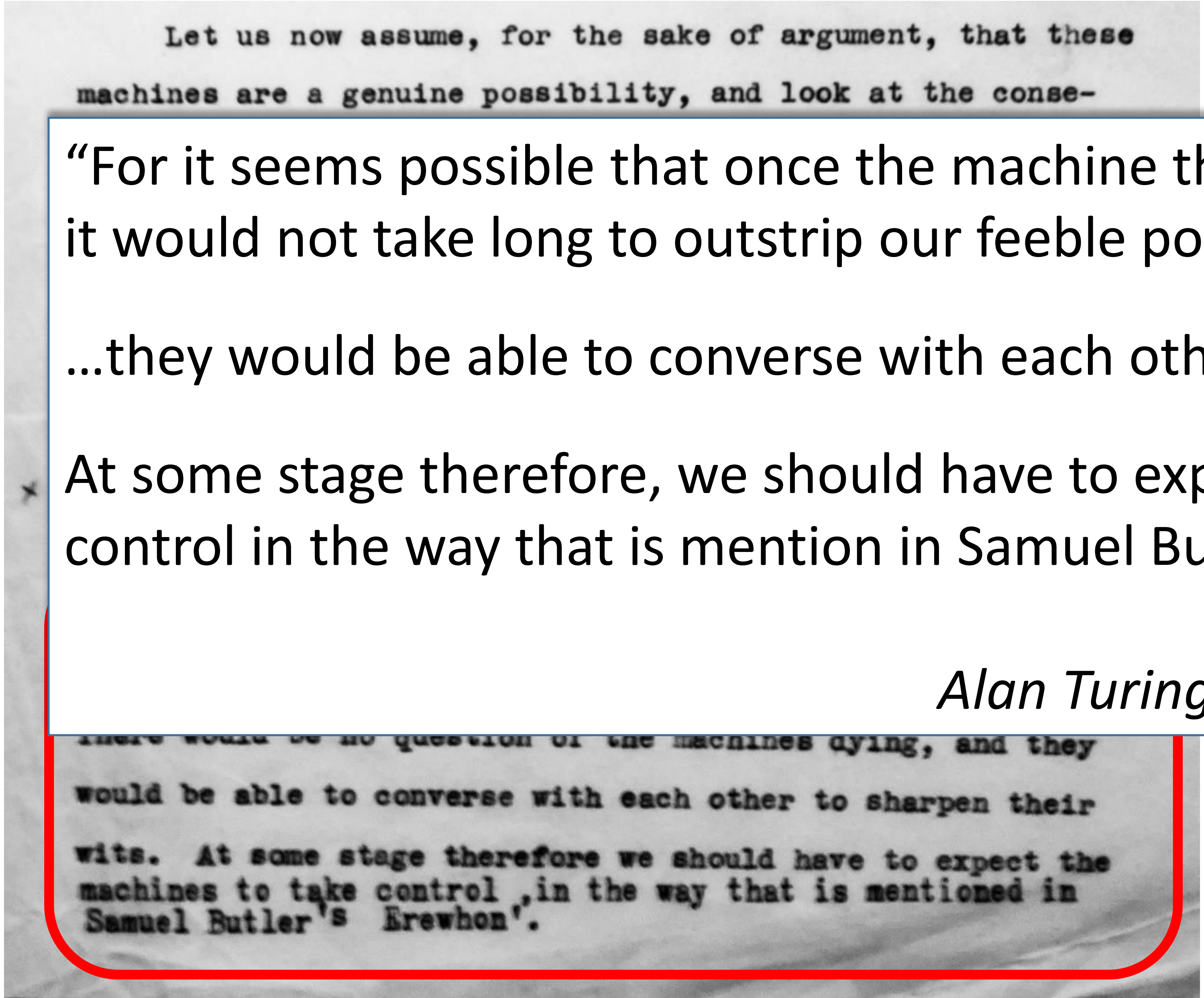


Let us now assume, for the sake of argument, that these machines are a genuine possibility, and look at the consequences of constructing them. To do so would of course meet with great opposition, unless we have advanced greatly in religious toleration from the days of Galileo. There would be great opposition from the intellectuals who were afraid of being put out of a job. It is probable though that the intellectuals would be mistaken about this. There would be plenty to do, trying to understand what the machines were trying to say, ~~in trying to say~~, i.e. in trying to keep one's intelligence up to the standard set by the machines, for it seems probable that once the machine thinking method had started, it would not take long to outstrip our feeble powers. There would be no question of the machines dying, and they would be able to converse with each other to sharpen their wits. At some stage therefore we should have to expect the machines to take control ,in the way that is mentioned in Samuel Butler's Erewhon'.

Alan Turing

Script, BBC broadcast, 1951

Direction: Address concerns about “superintelligences”



Let us now assume, for the sake of argument, that these machines are a genuine possibility, and look at the conse-

There would be no question of the machines dying, and they would be able to converse with each other to sharpen their wits. At some stage therefore we should have to expect the machines to take control ,in the way that is mentioned in Samuel Butler's Erewhon'.

“For it seems possible that once the machine thinking method had started, it would not take long to outstrip our feeble powers.

…they would be able to converse with each other to sharpen their wits.

At some stage therefore, we should have to expect the machines to take control in the way that is mention in Samuel Butler’s Erewhon.”

Alan Turing, 1951

Direction: Address concerns about “superintelligences”

Addressing concerns of public

Significant differences of opinion, including experts

- Do we understand possibilities?
- What kind of research should done proactively?
- Can we “backcast” from imagined poor outcomes
- Designs of clear ways to thwart possibilities, ease concerns