

# Safe Machine Learning-Based Perception via Closed-Loop Analysis

Sydney Katz

Thesis Defense

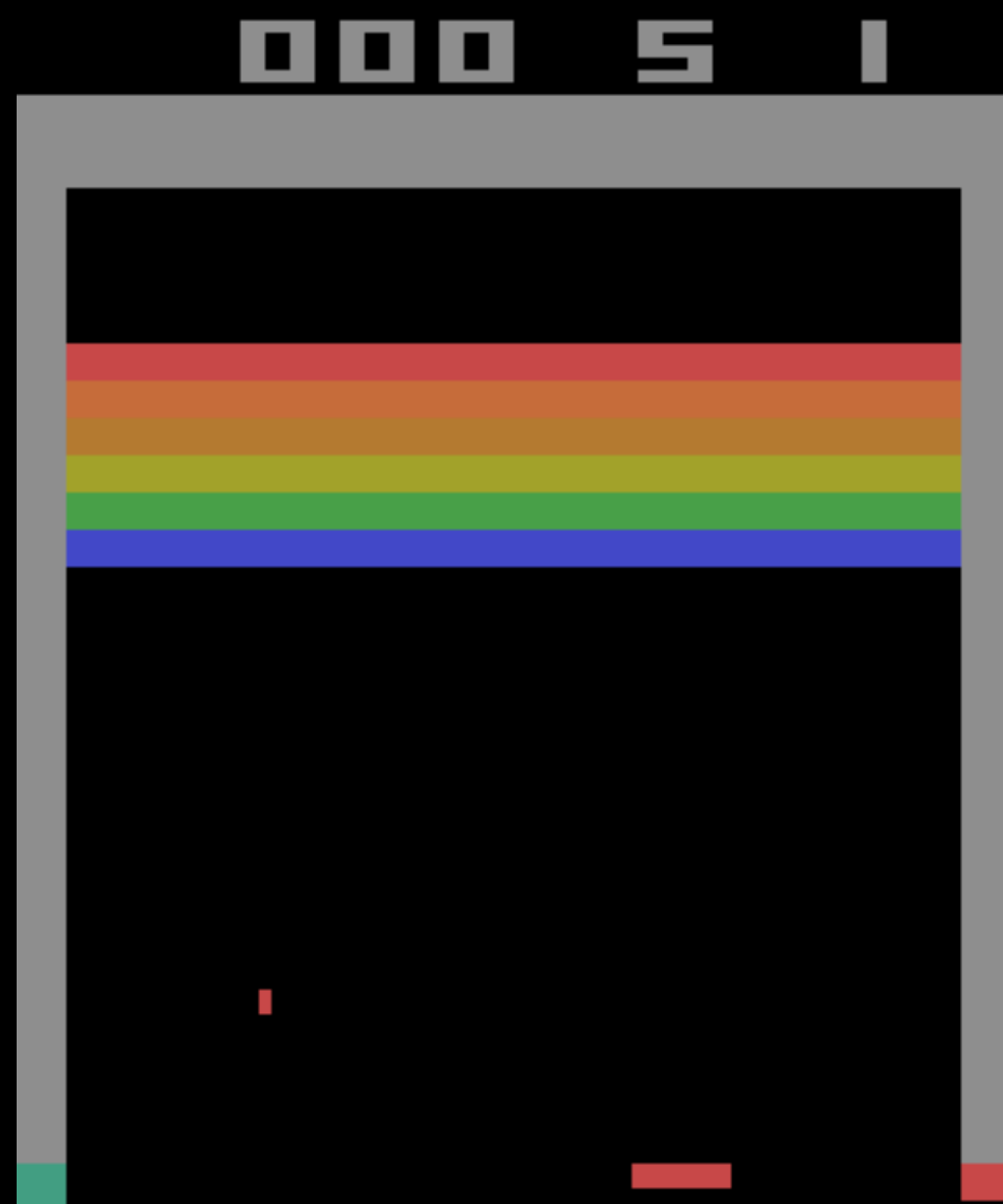
04.26.23

# Safe Machine Learning-Based Perception via Closed-Loop Analysis

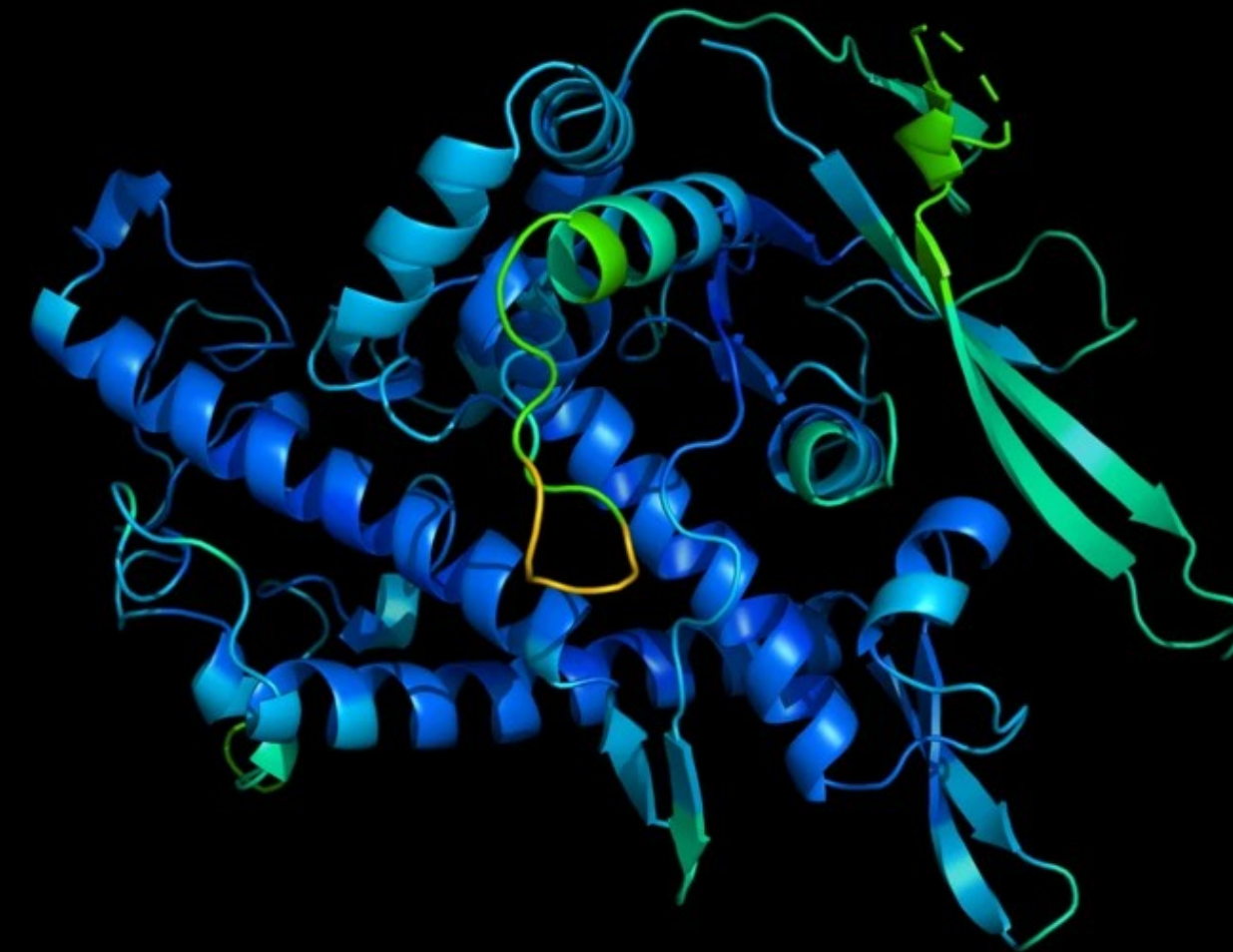
# Safe Machine Learning-Based Perception via Closed-Loop Analysis

# Safe Machine Learning-Based Perception via Closed-Loop Analysis

There is lots of opportunity for **Machine Learning!**



Credit: Deepmind



Credit: Deepmind



## Chat GPT

S Will I pass my PhD thesis defense?

As an AI language model, I do not have access to your personal information or your academic performance, and I cannot predict the outcome of your PhD thesis defense with certainty. However, with thorough preparation, practice, and a positive attitude, you will have the best chance of success. Good luck!

Credit: OpenAI

There is lots of opportunity for **Safe Machine Learning**!

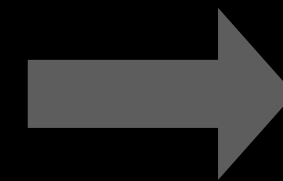
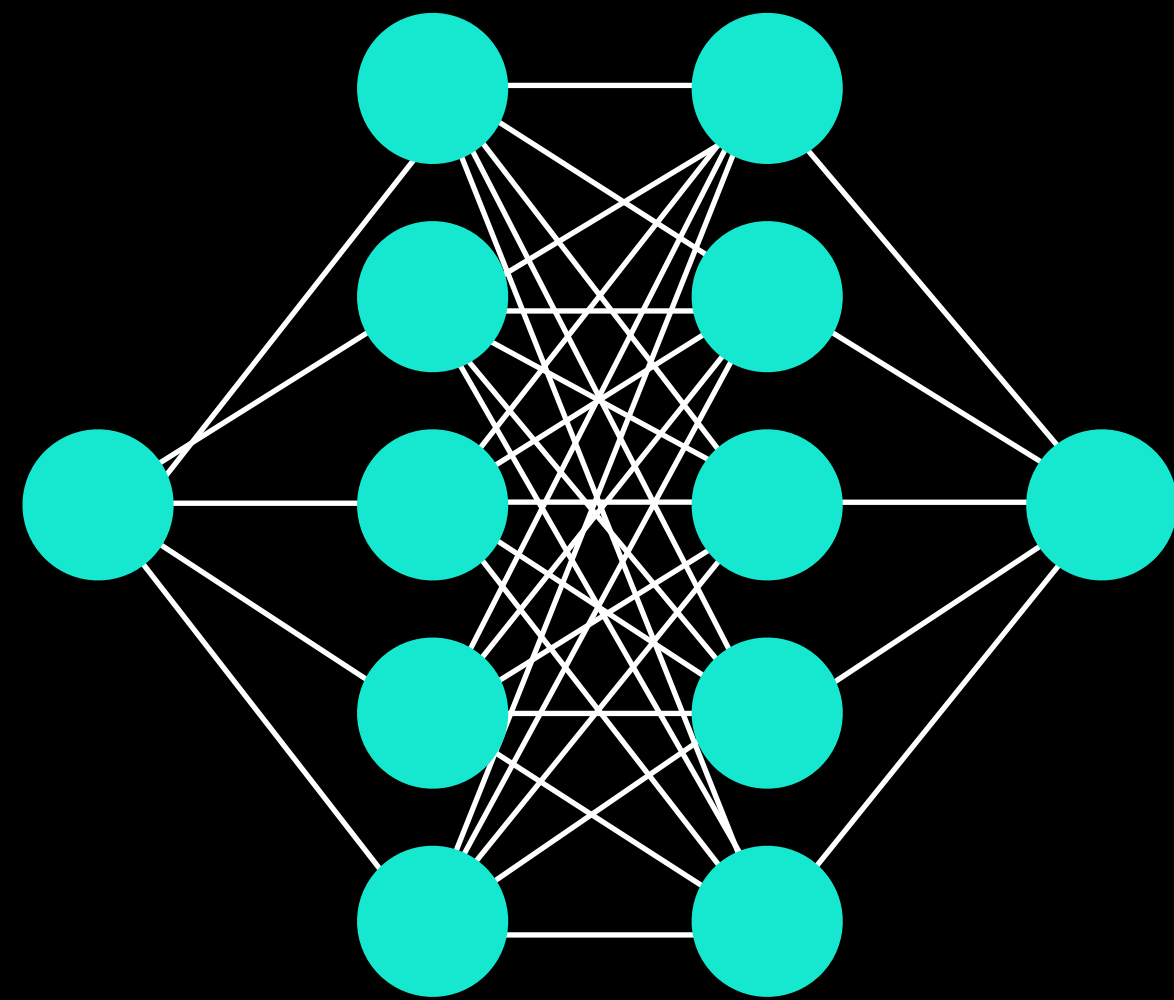


...but

automation of these systems tends to require **significant validation**

# Safe Machine Learning-Based Perception via Closed-Loop Analysis

automation of these systems tends to require **significant validation**  
...and machine learning adds **complexity**.



$$\begin{aligned} &0.05\max(0.2\max(-0.24x, 0) - 0.34\max(1.2x - 1.41, 0) - \\ &0.46\max(0.01x - 0.31, 0) - 0.21\max(1.11x - 3.16, 0) - \\ &0.42\max(0.45x - 5.99, 0), 0) + 1.64\max(0.14\max(-0.24x, 0) + \\ &0.33\max(1.2x - 1.41, 0) + 0.55\max(0.01x - 0.31, 0) - \\ &0.2\max(1.11x - 3.16, 0) - 2.66\max(0.45x - 5.99, 0) + 0.85, 0) - \\ &4.85\max(-0.08\max(-0.24x, 0) + 0.42\max(1.2x - 1.41, 0) - \\ &0.68\max(0.01x - 0.31, 0) - 0.5\max(1.11x - 3.16, 0) - \\ &1.26\max(0.45x - 5.99, 0) + 3.76, 0) - 0.63\max(0.33\max(-0.24x, 0) + \\ &0.17\max(1.2x - 1.41, 0) + 0.22\max(0.01x - 0.31, 0) - \\ &0.33\max(1.11x - 3.16, 0) - 0.13\max(0.45x - 5.99, 0) - 0.12, 0) + \\ &1.77\max(-0.41\max(-0.24x, 0) + 1.44\max(1.2x - 1.41, 0) - \\ &0.18\max(0.01x - 0.31, 0) + 1.16\max(1.11x - 3.16, 0) + \\ &1.09\max(0.45x - 5.99, 0) - 1.11, 0) - 0.46 \end{aligned}$$



We need to develop **new tools** to ensure the safe operation of machine learning-based systems.

# Safe Machine Learning-Based Perception via Closed-Loop Analysis

# Safe Machine Learning-Based **Perception** via Closed-Loop Analysis

# Safe Machine Learning-Based **Perception** via Closed-Loop Analysis

**Perception** is a key application area for machine learning.

Vision-Based Detect and Avoid



Vision-Based Taxi Navigation

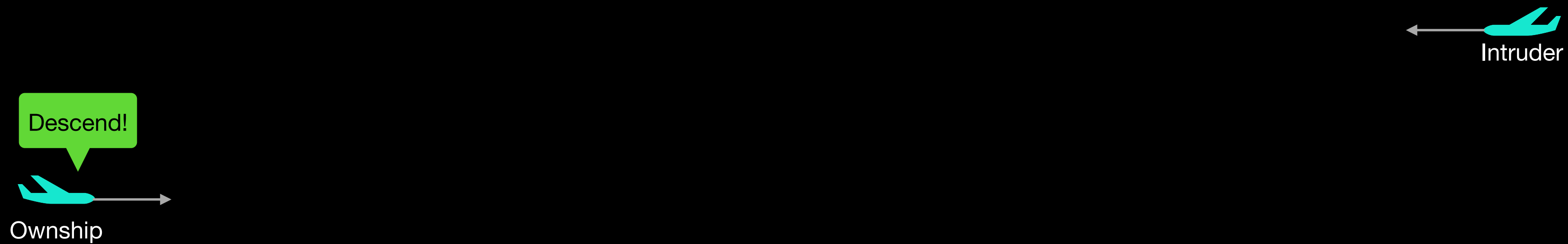


# Safe Machine Learning-Based Perception via Closed-Loop Analysis

# Safe Machine Learning-Based Perception via **Closed-Loop Analysis**

# Safe Machine Learning-Based Perception via **Closed-Loop Analysis**

**Closed-Loop Analysis** allows us to test **high-level** safety properties.

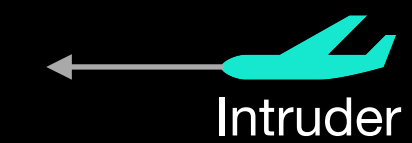
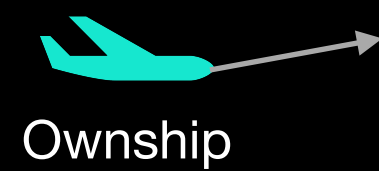


## **Open-Loop Property:**

“If the intruder is above the ownship and they are both in level flight, the ownship should always descend.”

# Safe Machine Learning-Based Perception via **Closed-Loop Analysis**

**Closed-Loop Analysis** allows us to test **high-level** safety properties.

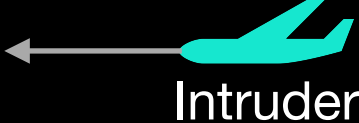
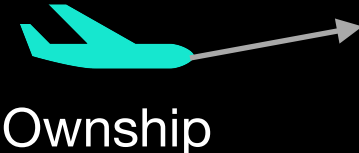


## **Open-Loop Property:**

“If the intruder is above the ownship and they are both in level flight, the ownship should always descend.”

# Safe Machine Learning-Based Perception via **Closed-Loop Analysis**

**Closed-Loop Analysis** allows us to test **high-level** safety properties.

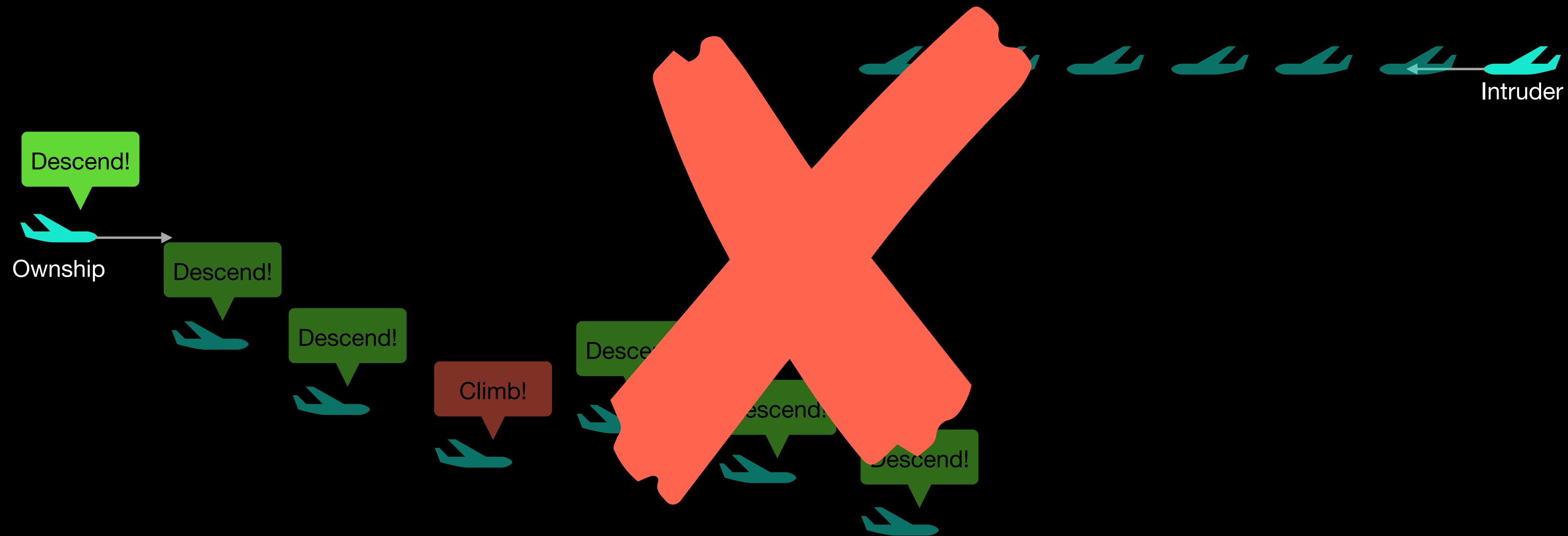


**Open-Loop Property:**

?

# Safe Machine Learning-Based Perception via **Closed-Loop Analysis**

**Closed-Loop Analysis** allows us to test **high-level** safety properties.

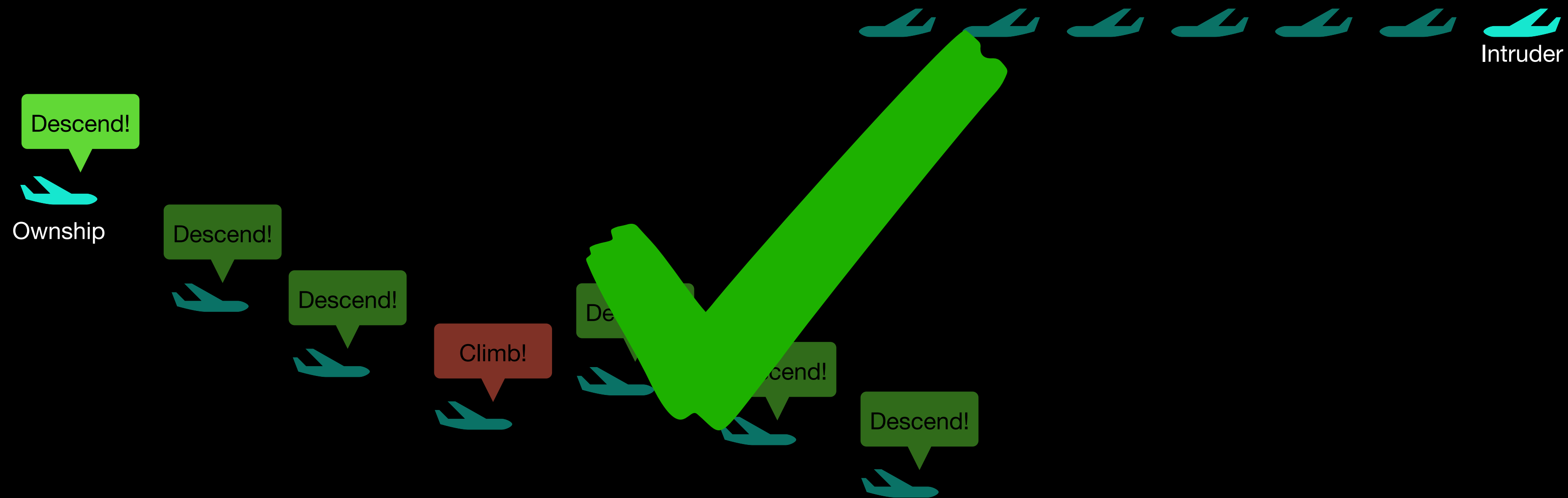


## **Open-Loop Property:**

“If the intruder is above the ownship and they are both in level flight, the ownship should always descend.”

# Safe Machine Learning-Based Perception via **Closed-Loop Analysis**

**Closed-Loop Analysis** allows us to test **high-level** safety properties.



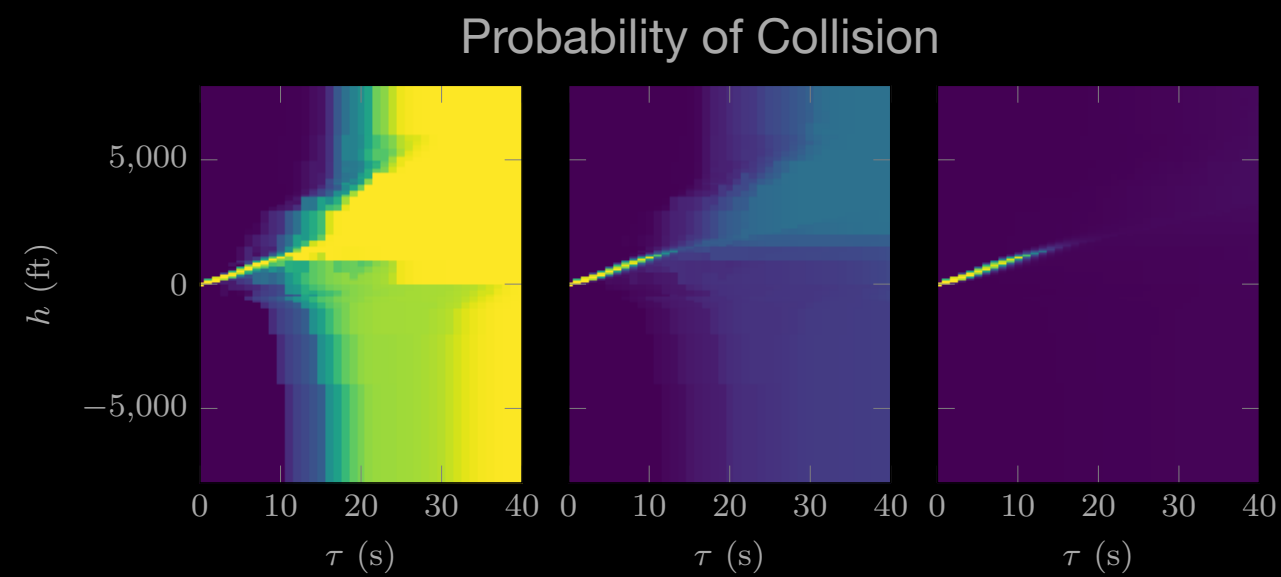
**Closed-Loop Property:**

“The aircraft should not collide.”

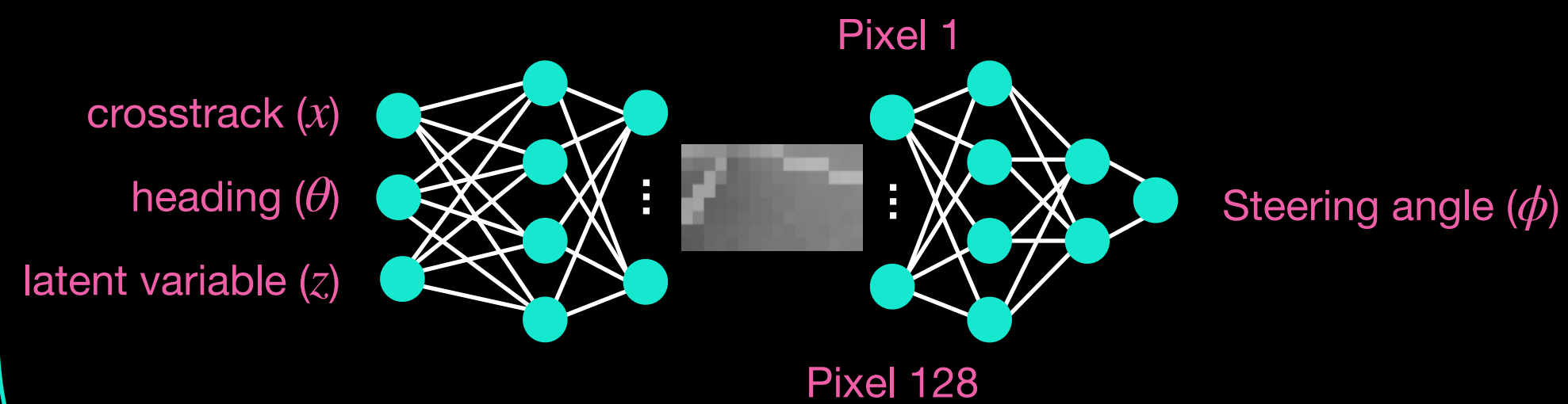
# Safety Verification



Generating Probabilistic Safety Guarantees for Neural Network Controllers



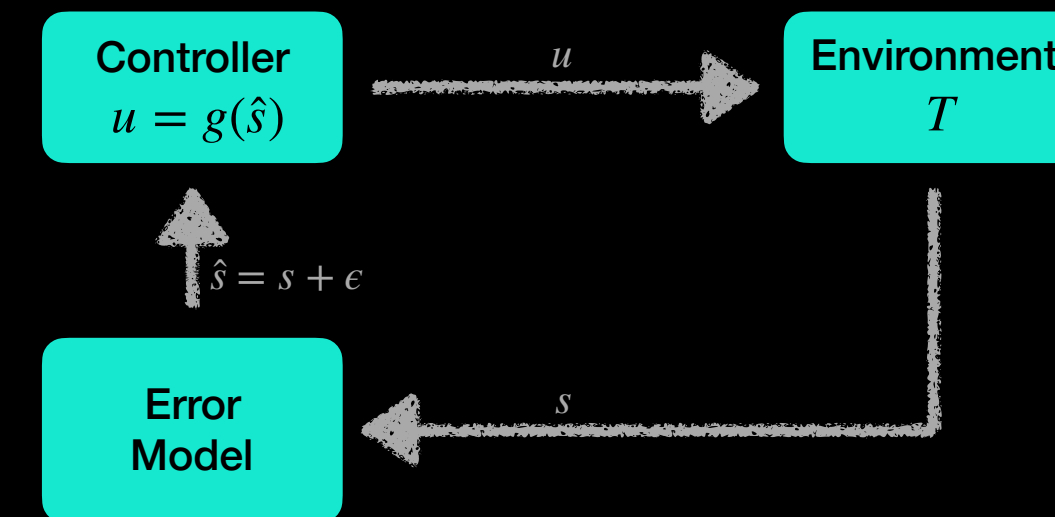
Verification of Image-Based Neural Network Controllers Using Generative Models



# Safe Design

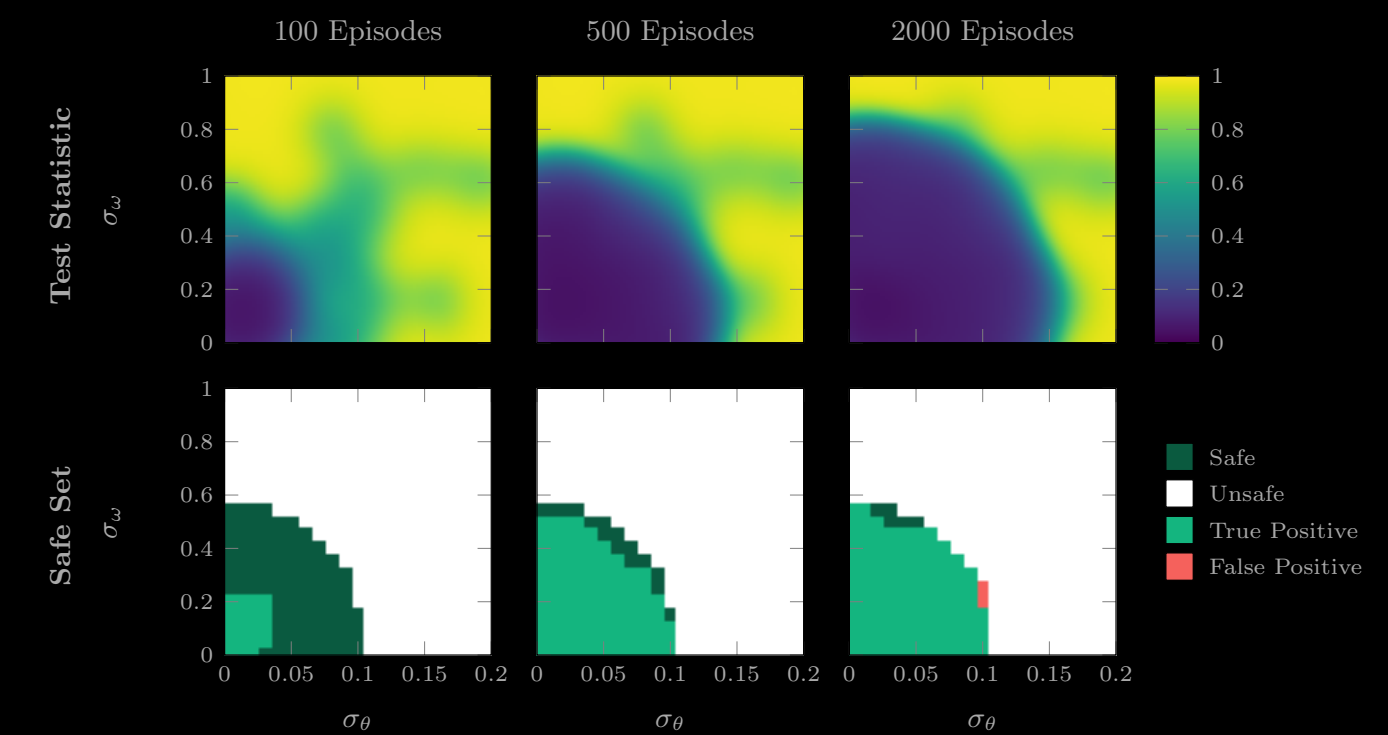


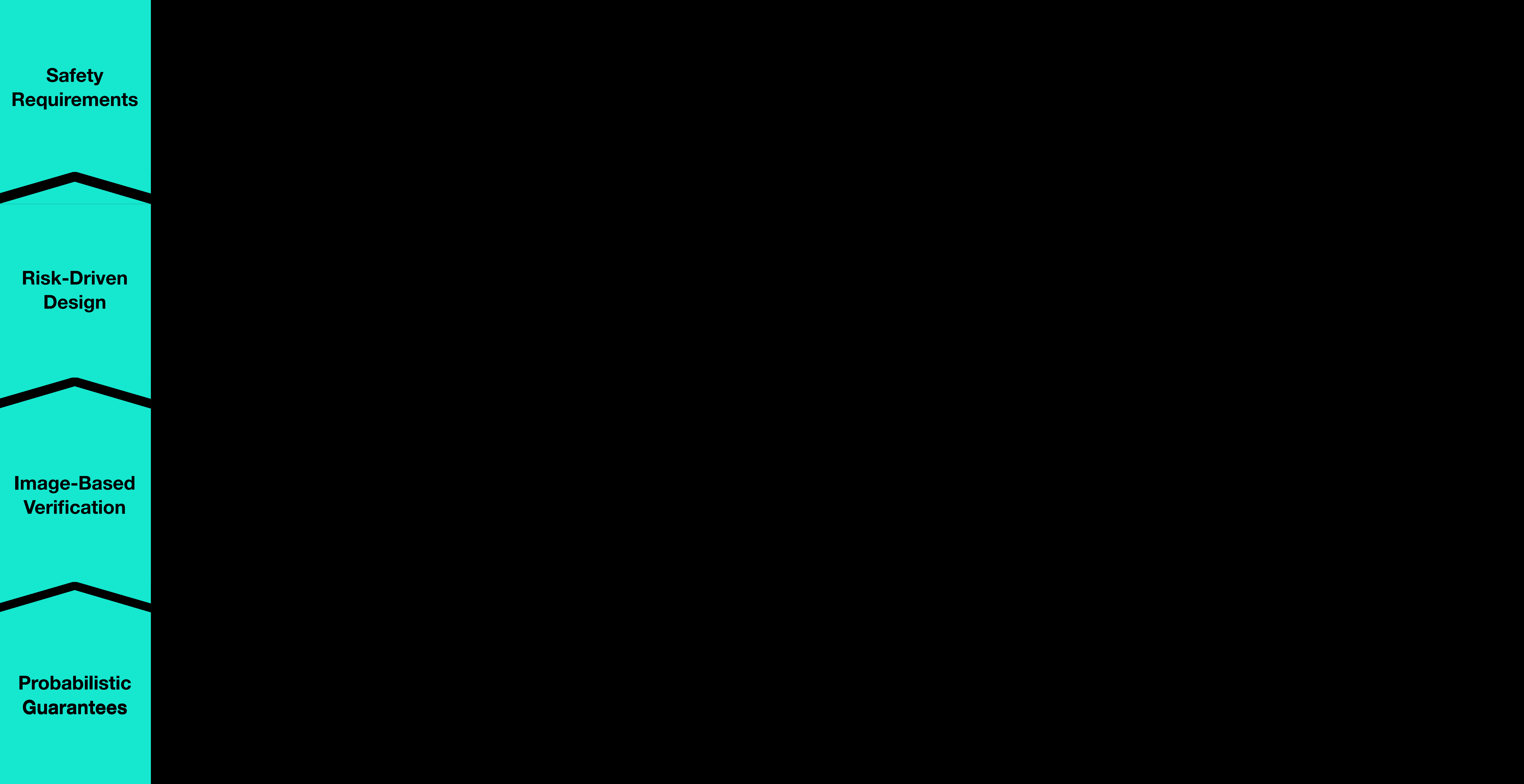
Risk-Driven Design of Perception Systems

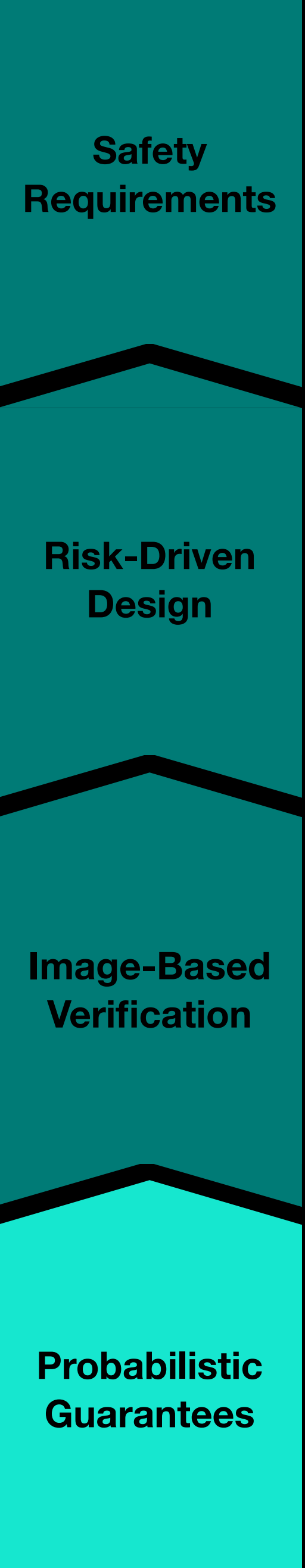


WIP

Efficient Determination of Safety Requirements for Perception Systems







# Generating Probabilistic Safety Guarantees for Neural Network Controllers

S. M. Katz, K. D. Julian, C. A. Strong, and M. J. Kochenderfer

Machine Learning Journal (2021)

Safety Requirements

Previous work on **reachability analysis** for neural network controllers did not account for **stochasticity** in the system **dynamics**.

Risk-Driven Design

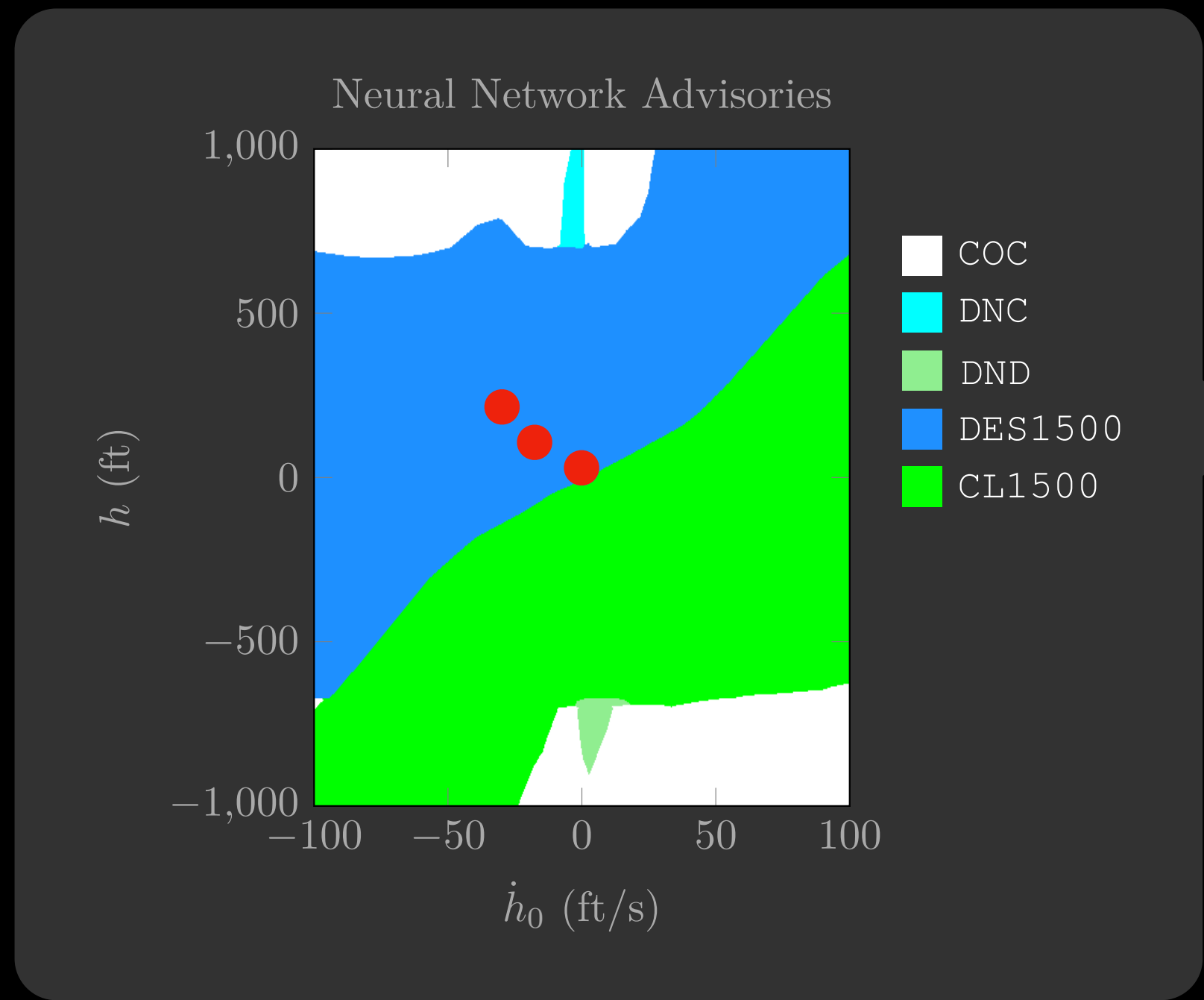
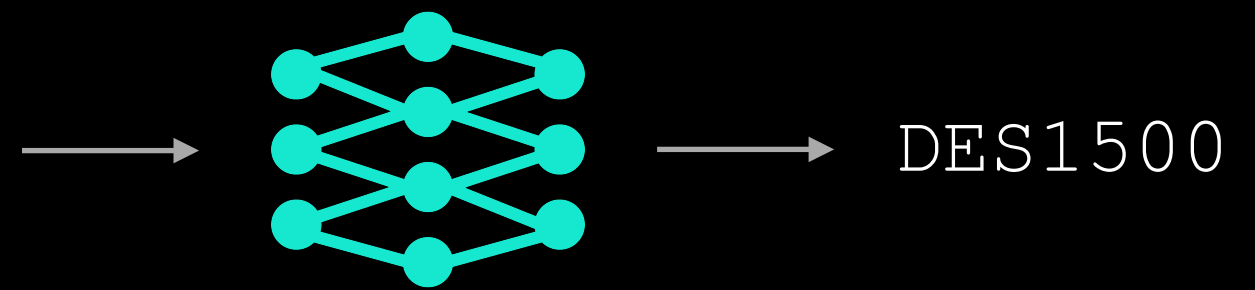


Image-Based Verification



Probabilistic Guarantees

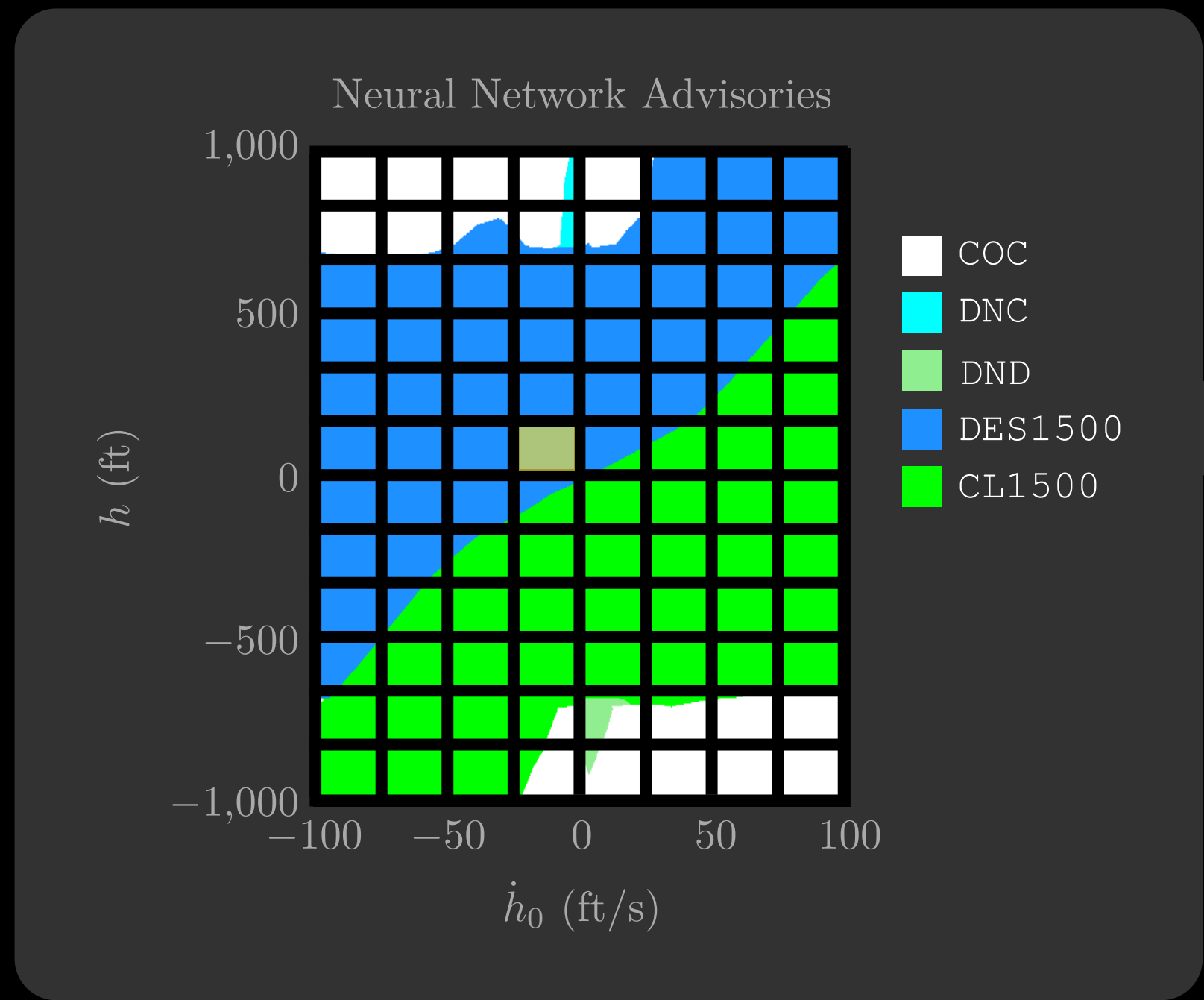


# Safe Machine Learning-Based Perception via Closed-Loop Analysis

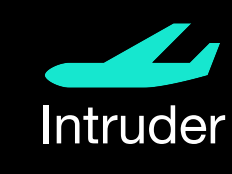
Safety Requirements

Previous work on **reachability analysis** for neural network controllers did not account for **stochasticity** in the system **dynamics**.

Risk-Driven Design



Ownship



Intruder

Image-Based Verification

Survey Paper



Intro Video



Neural Network Verification Tool

{DES1500}

Probabilistic Guarantees

Julian and Kochenderfer  
"Reachability analysis for neural network aircraft collision avoidance systems"  
(2021)

Safety Requirements

Previous work on **reachability analysis** for neural network controllers did not account for **stochasticity** in the system **dynamics**.

Risk-Driven Design

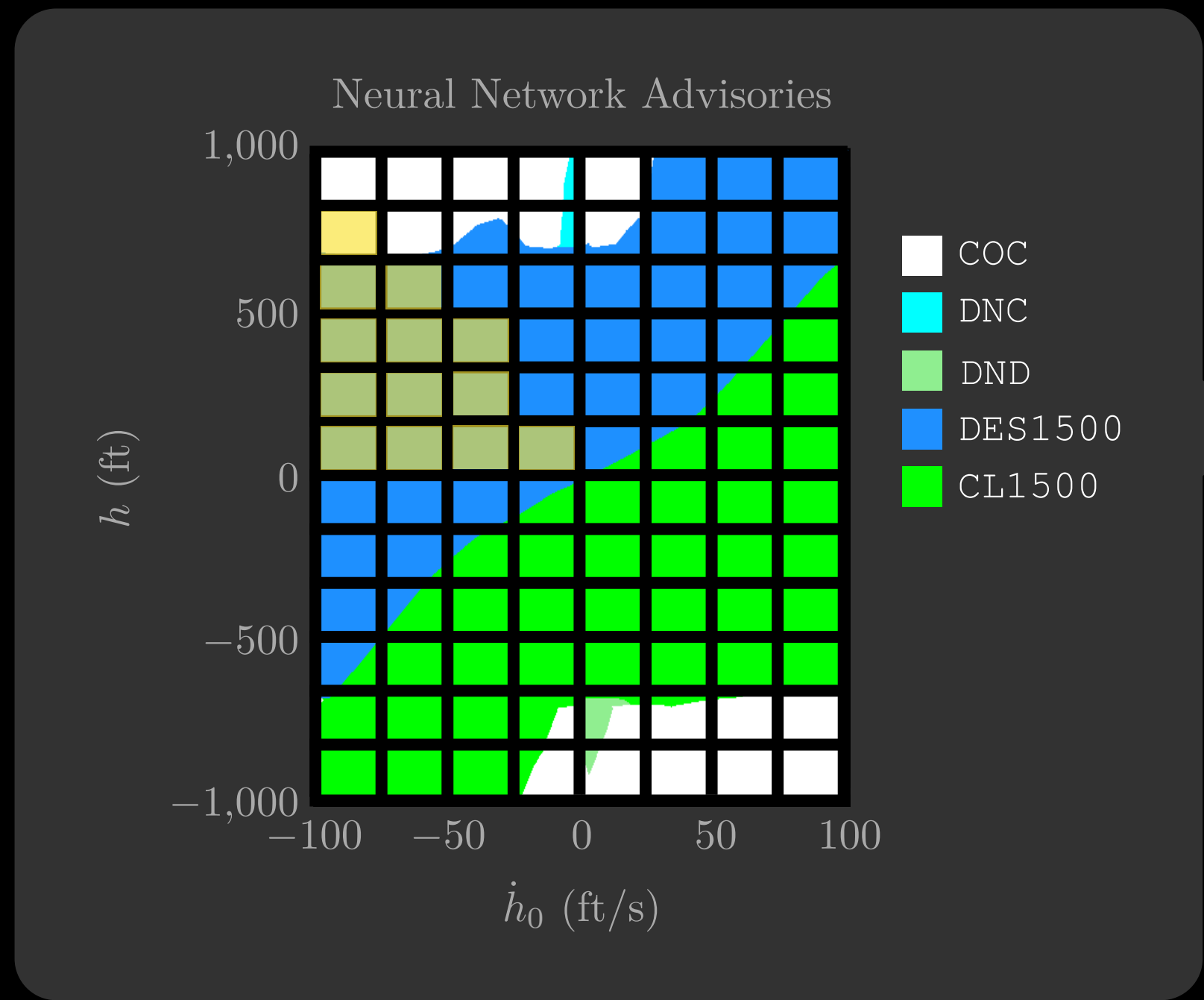
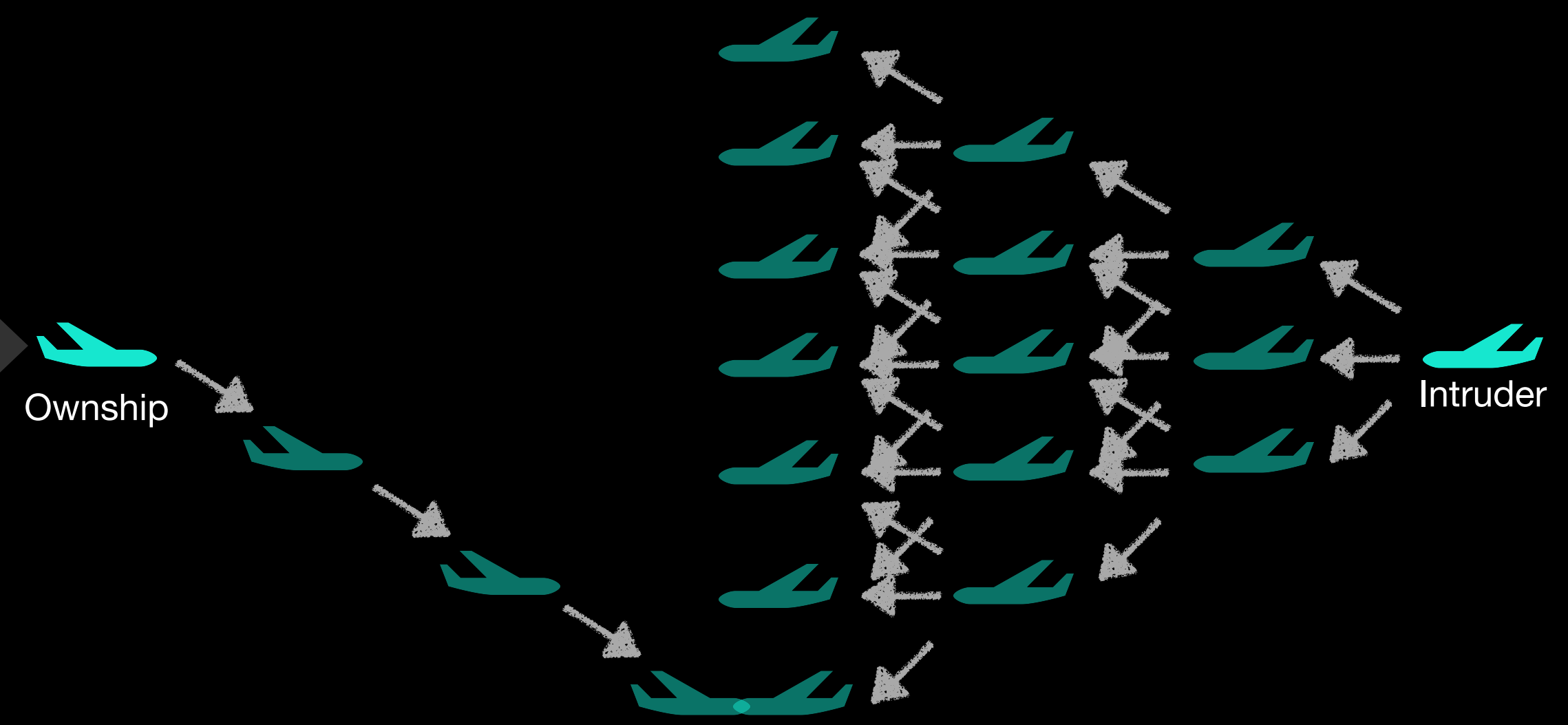


Image-Based Verification



Probabilistic Guarantees

Julian and Kochenderfer  
"Reachability analysis for  
neural network aircraft  
collision avoidance systems"  
(2021)

Safety Requirements

Previous work on **reachability analysis** for neural network controllers did not account for **stochasticity** in the system **dynamics**.

Risk-Driven Design

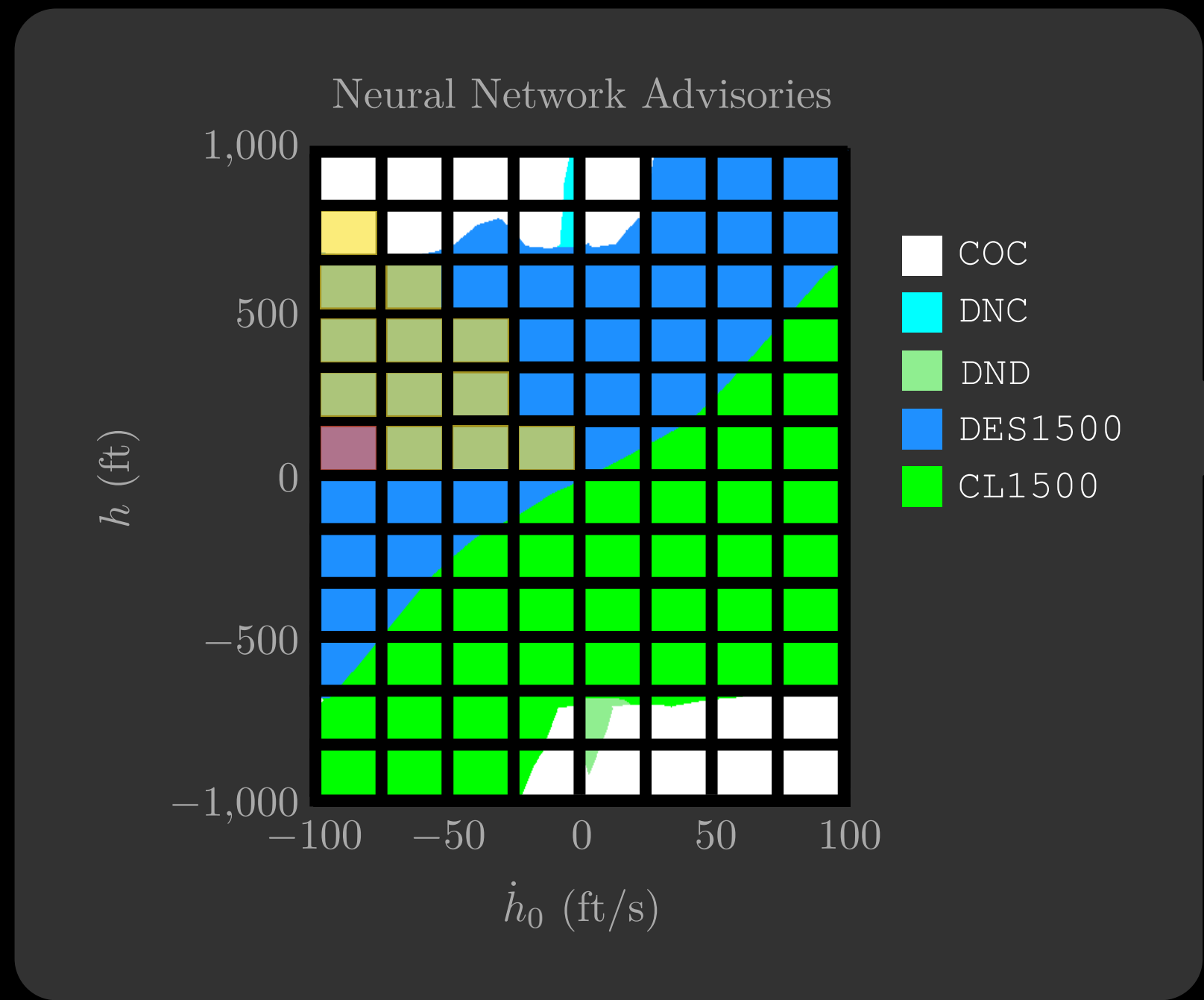
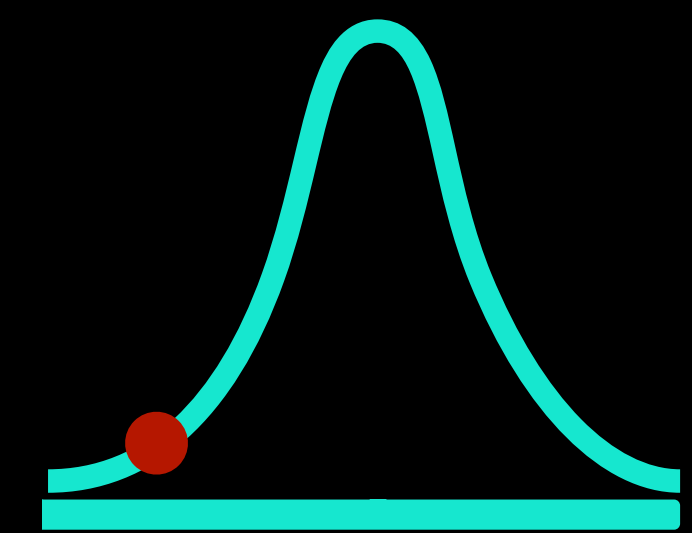


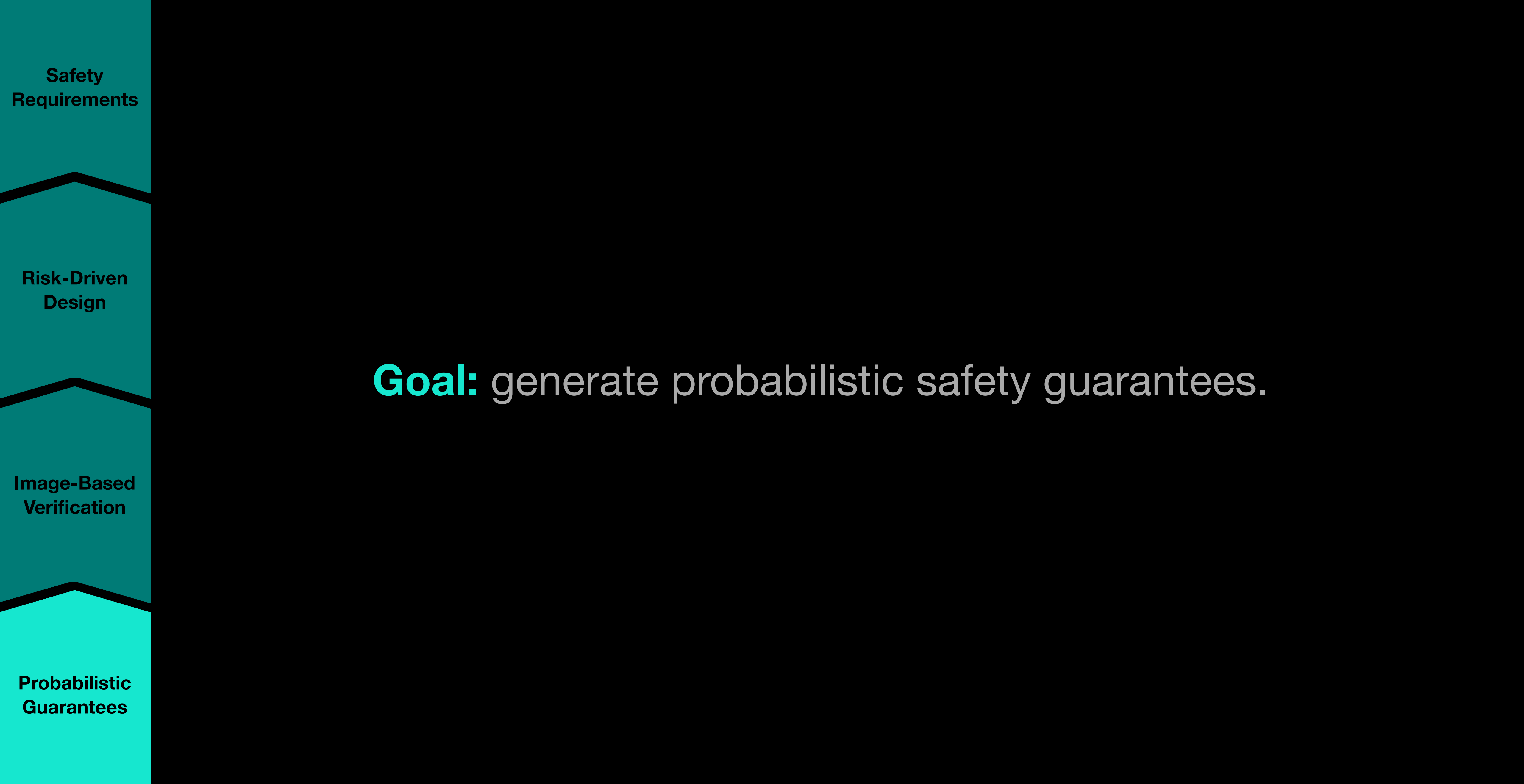
Image-Based Verification



Probabilistic Guarantees

Julian and Kochenderfer  
"Reachability analysis for neural network aircraft collision avoidance systems"  
(2021)





**Goal:** generate probabilistic safety guarantees.

Safety Requirements

Risk-Driven Design

Image-Based Verification

Probabilistic Guarantees

1

Modify traditional **MDP model checking** formulation to work with neural network controllers.

2

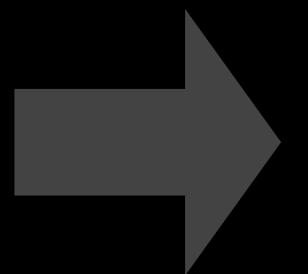
Minimize **overapproximation error** along the way.

1

Modify traditional **MDP model checking** formulation to work with neural network controllers.

**MDP model checking**

dynamic programming



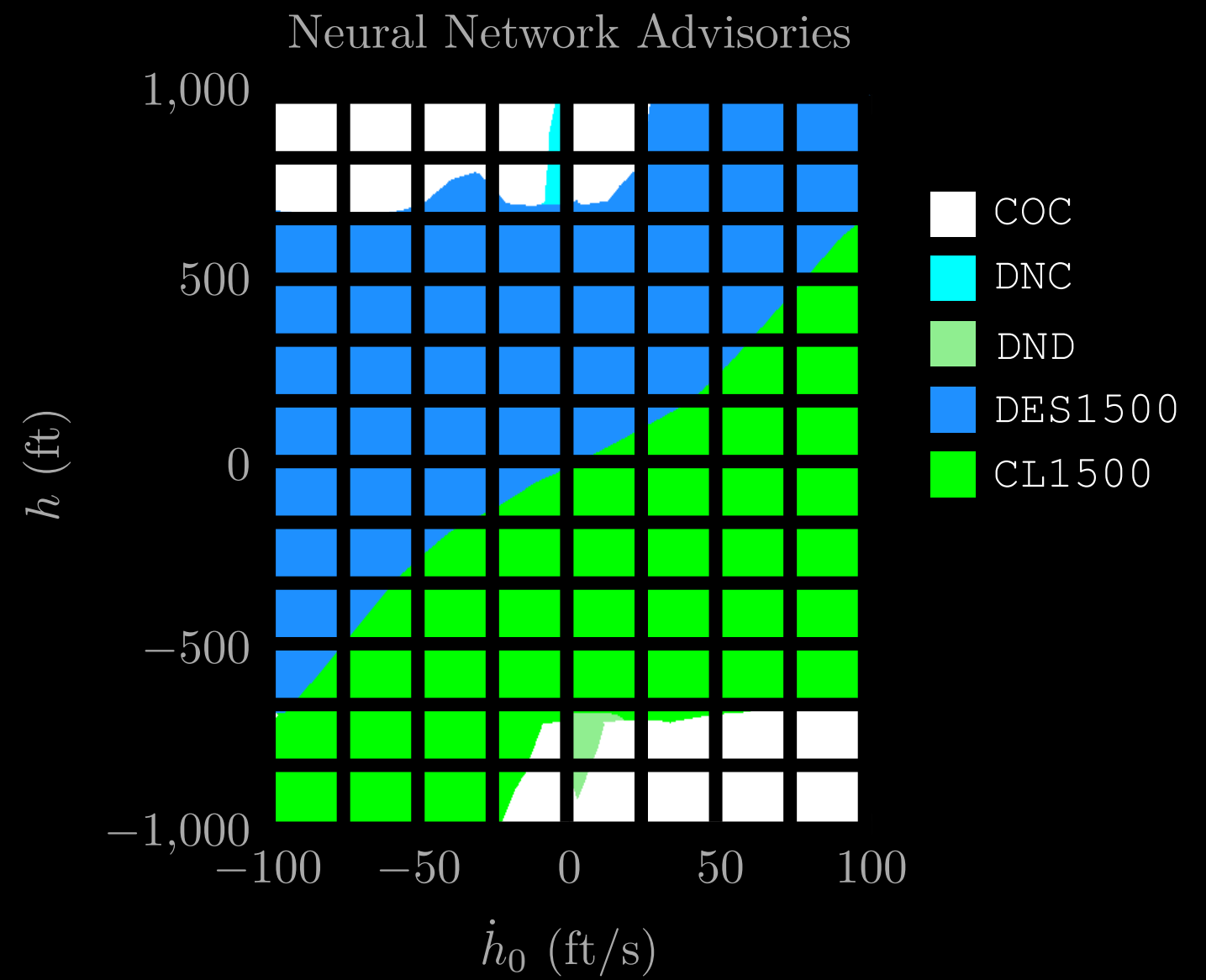
$\Pr^\pi(s)$  : probability of NMAC when starting from state  $s$ .

Safety Requirements

1

Modify traditional **MDP model checking** formulation to work with neural network controllers.

Risk-Driven Design



$\Pr^{\pi}(s)$  : probability of NMAC when starting from state  $s$ .

+ overapproximation

$\Pr^{\tilde{\pi}}(c)$  : overapproximated probability of NMAC when starting from cell  $c$ .

Image-Based Verification

Probabilistic Guarantees

Safety Requirements

2

Minimize **overapproximation error** along the way.

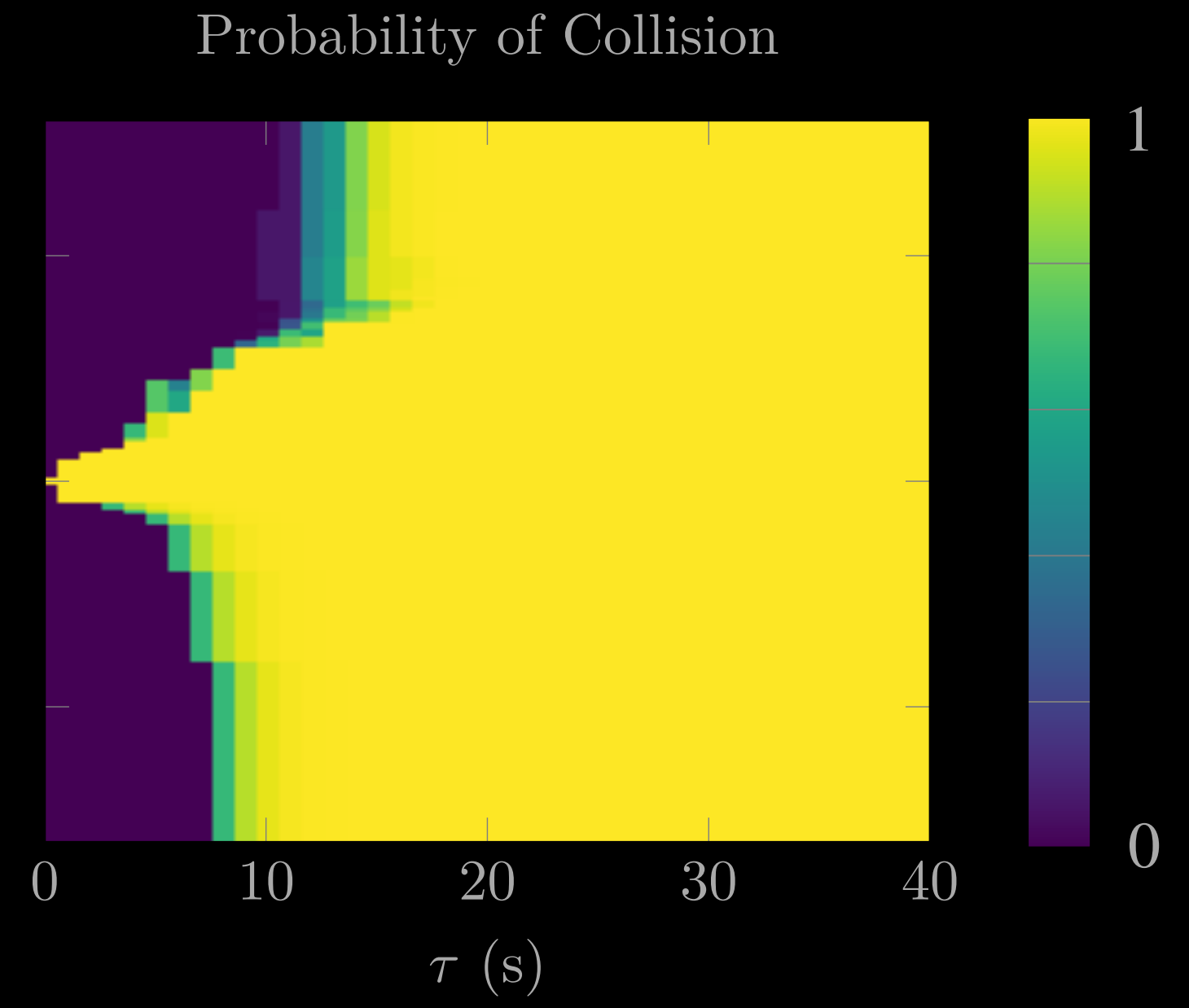
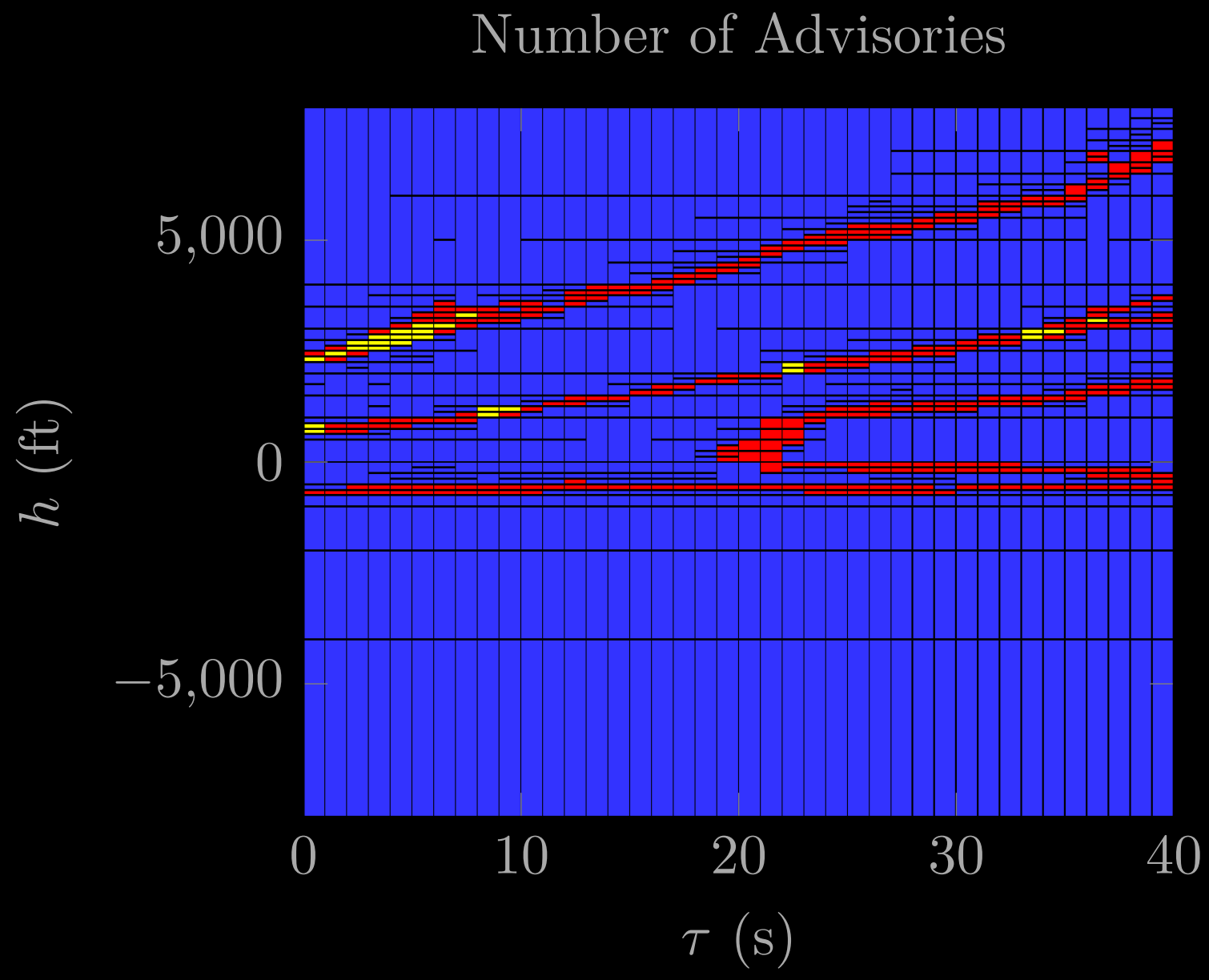
$\Pr^{\tilde{\pi}}(c)$  : overapproximated probability of NMAC when starting from cell  $c$ .

Risk-Driven Design

- 1 action
- 2 actions
- 3+ actions

Image-Based Verification

Probabilistic Guarantees



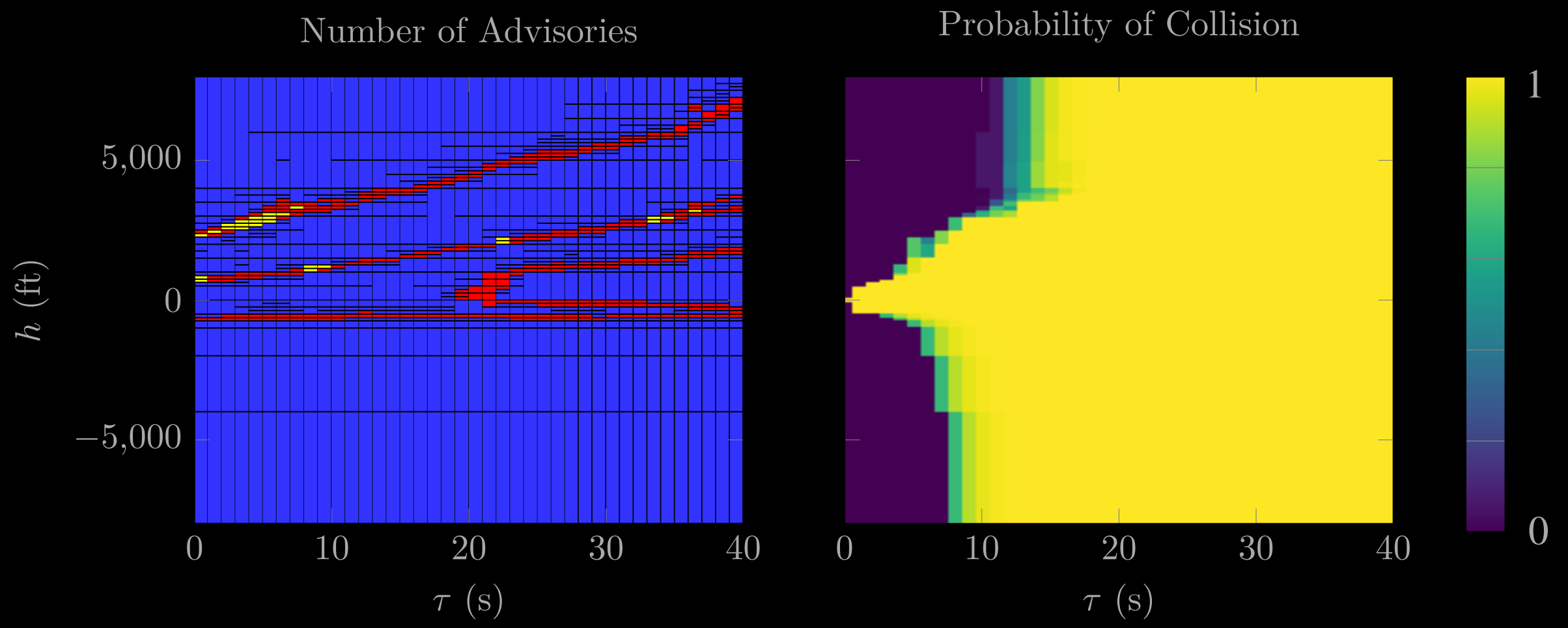
2

Minimize **overapproximation error** along the way.

$\Pr^{\tilde{\pi}}(c)$  : overapproximated probability of NMAC when starting from cell  $c$ .

We draw on ideas from **state abstraction** to develop on **online splitting heuristic** to further reduce overapproximation error **the solving process**.

- 1 action
- 2 actions
- 3+ actions



Safety Requirements

2

Minimize **overapproximation error** along the way.

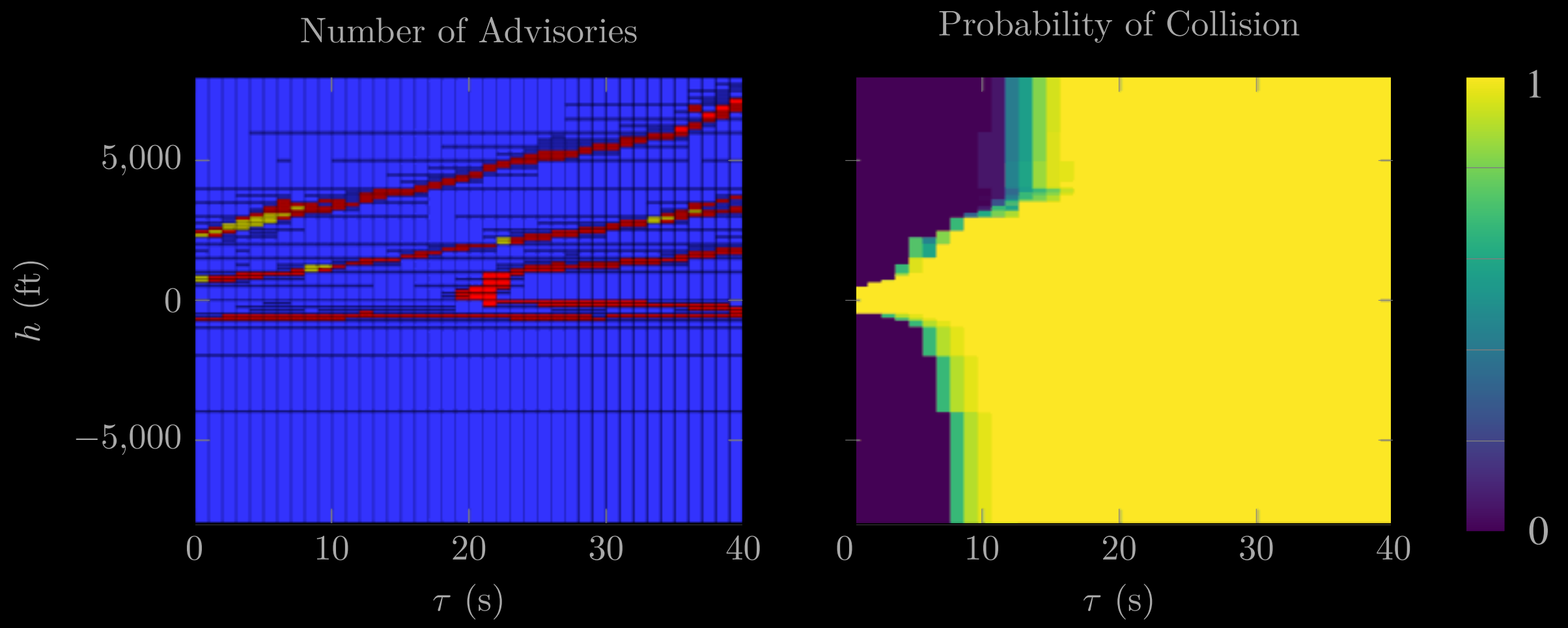
$\Pr^{\tilde{\pi}}(c)$  : overapproximated probability of NMAC when starting from cell  $c$ .

Risk-Driven Design

We draw on ideas from **state abstraction** to develop on **online splitting heuristic** to further reduce overapproximation error **the solving process**.

Image-Based Verification

- 1 action
- 2 actions
- 3+ actions



Probabilistic Guarantees

# Generating Probabilistic Safety Guarantees for Neural Network Controllers

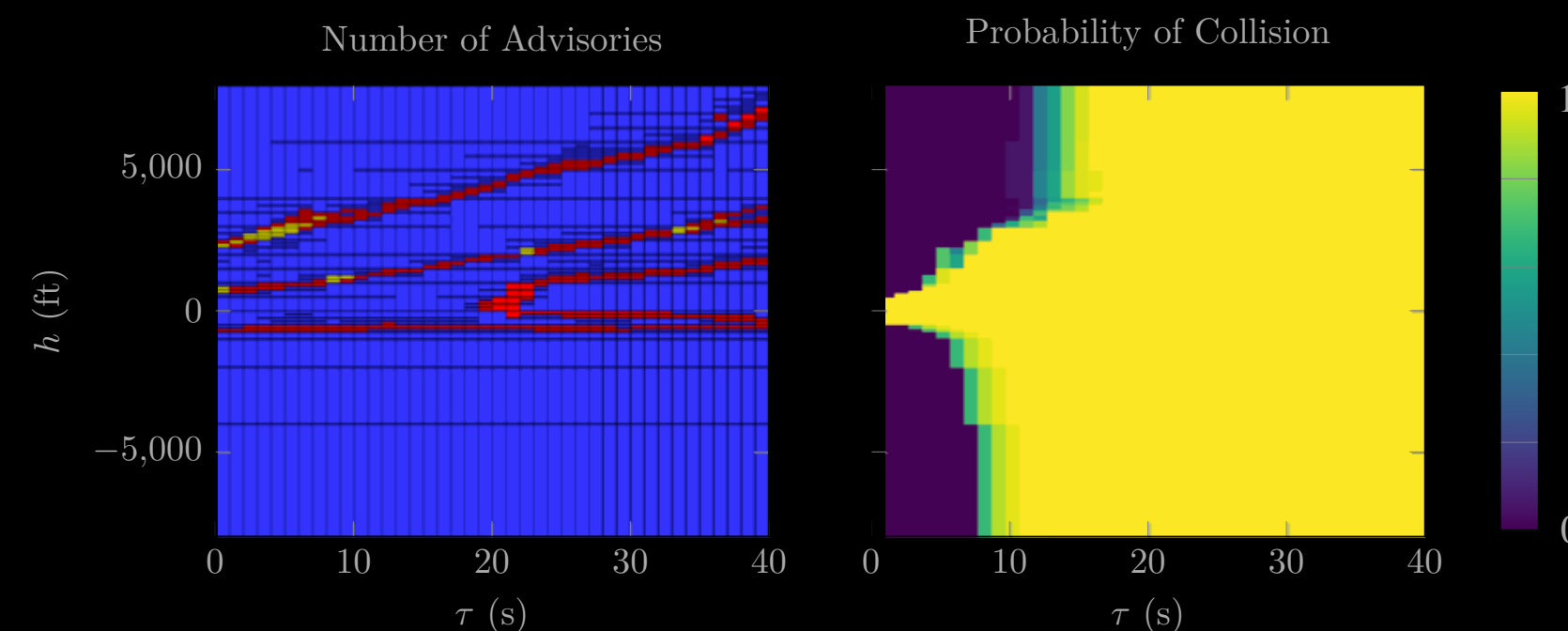
1

Modify traditional **MDP model checking** formulation to work with neural network controllers.

$\Pr^{\tilde{\pi}}(c)$  : overapproximated probability of NMAC when starting from cell  $c$ .

2

Minimize **overapproximation error** along the way.

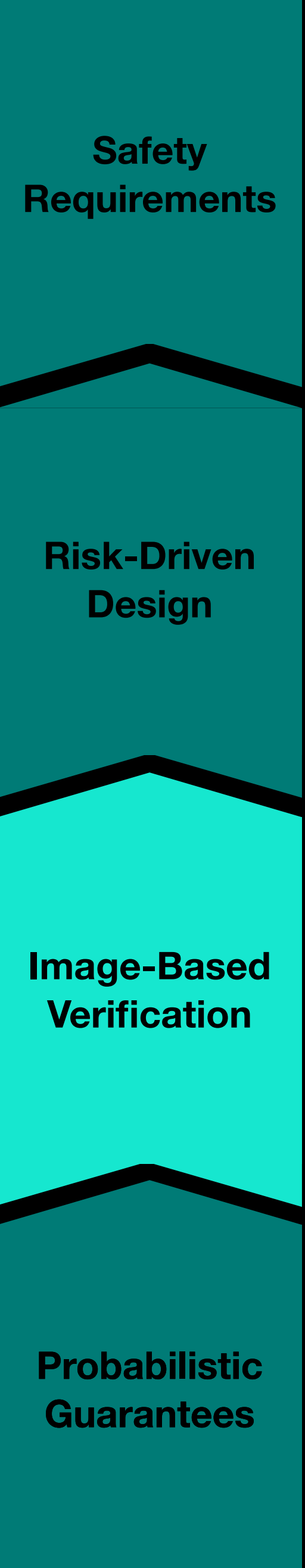


Safety Requirements

Risk-Driven Design

Image-Based Verification

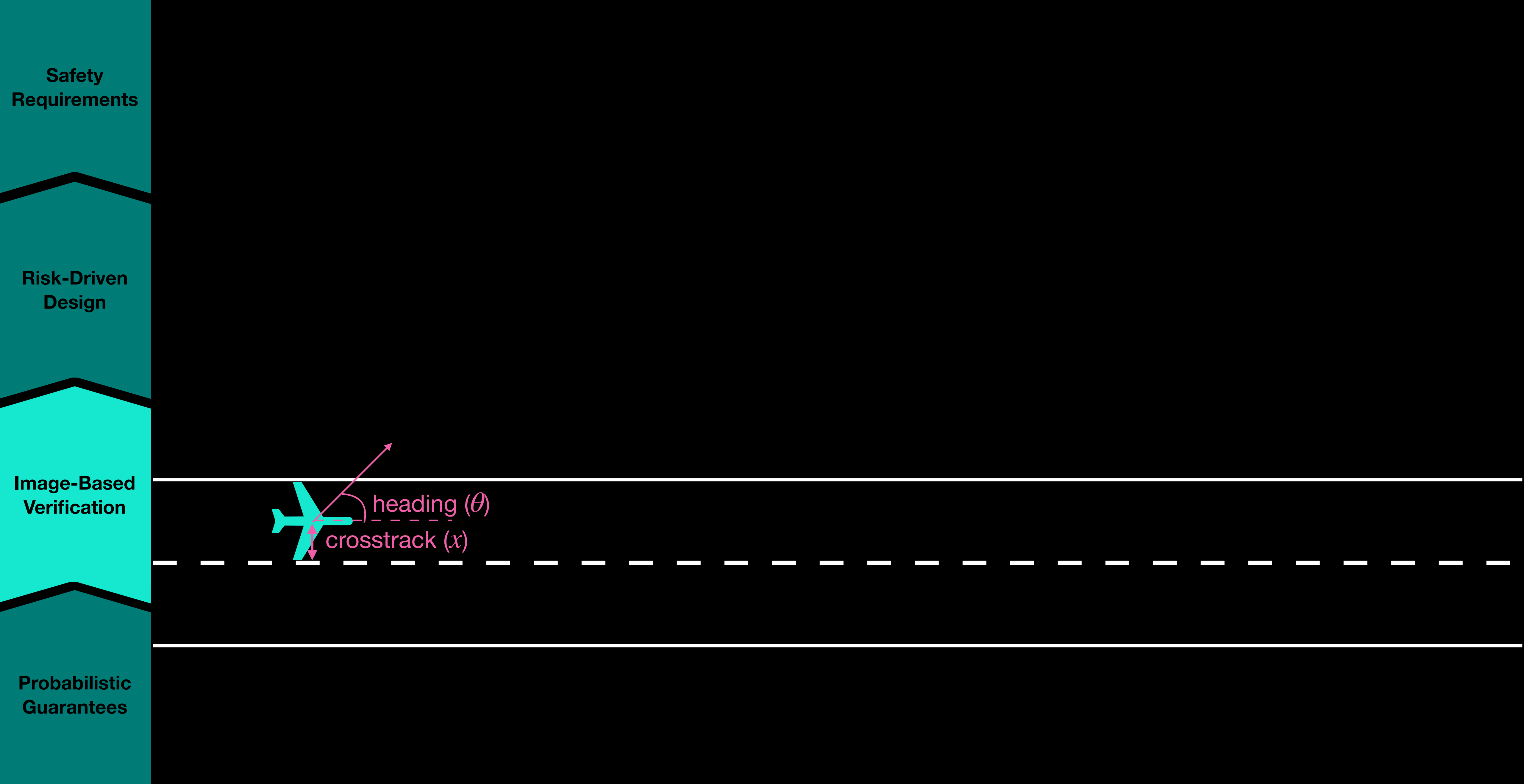
Probabilistic Guarantees

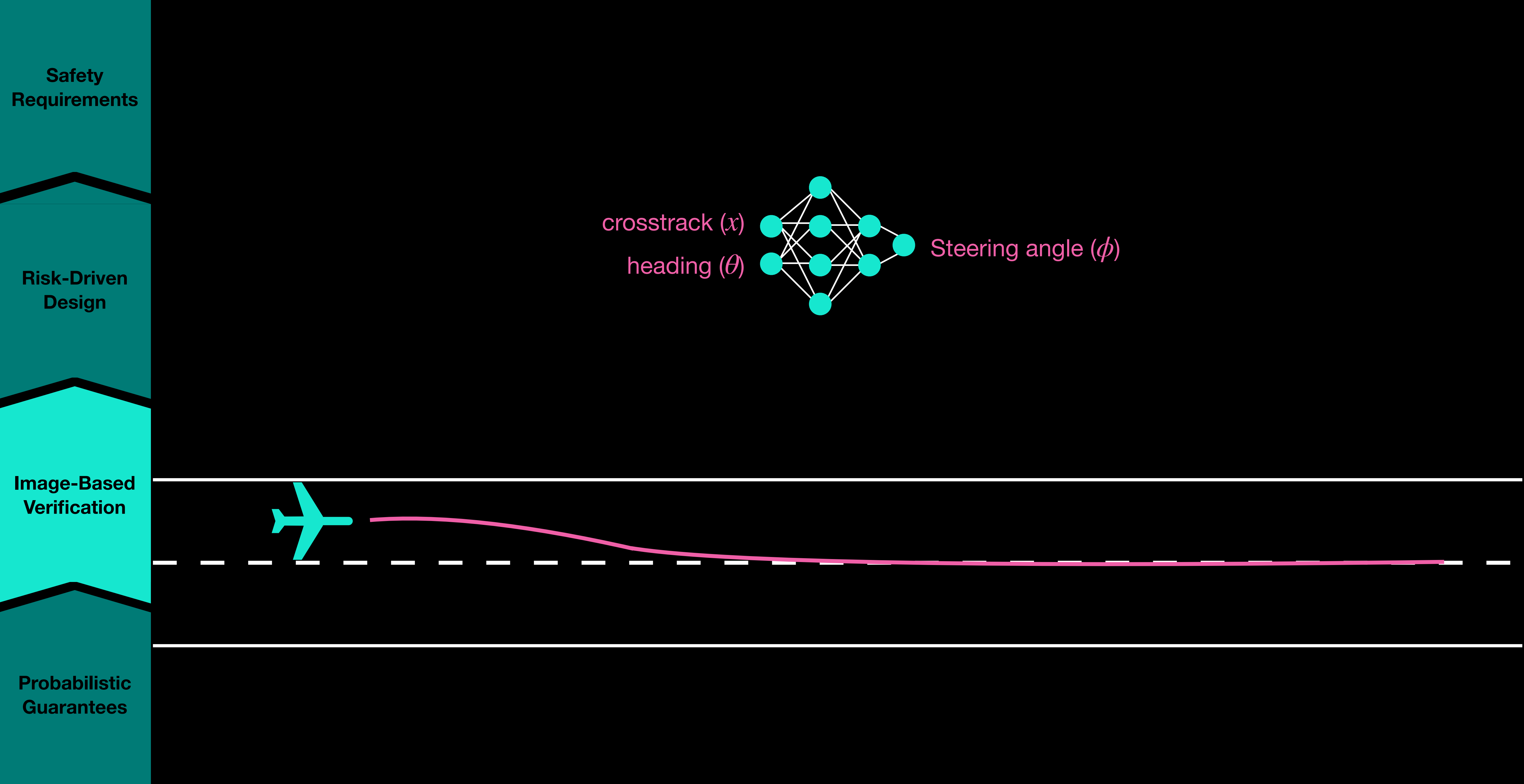


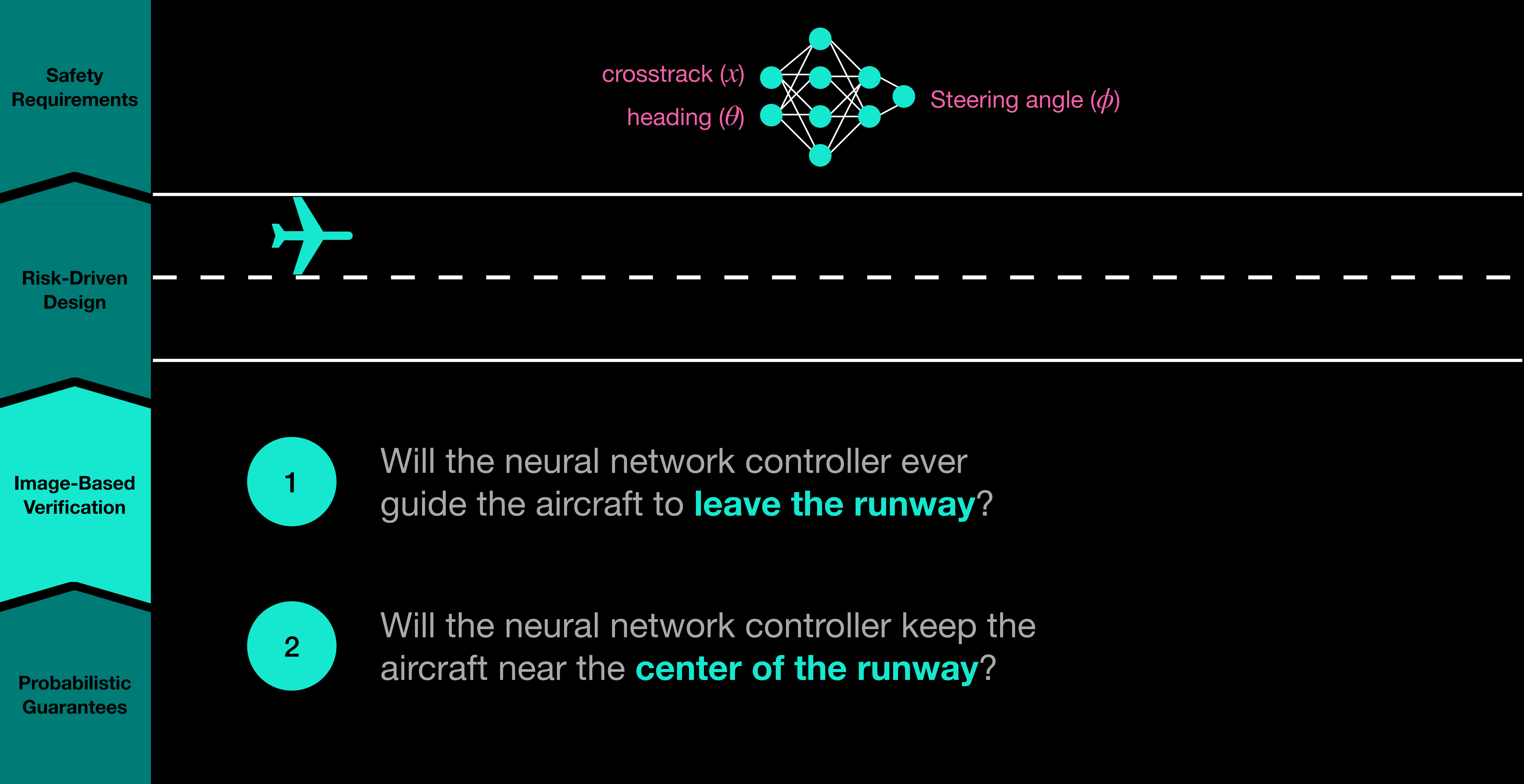
# Verification of Image-Based Neural Network Controllers Using Generative Models

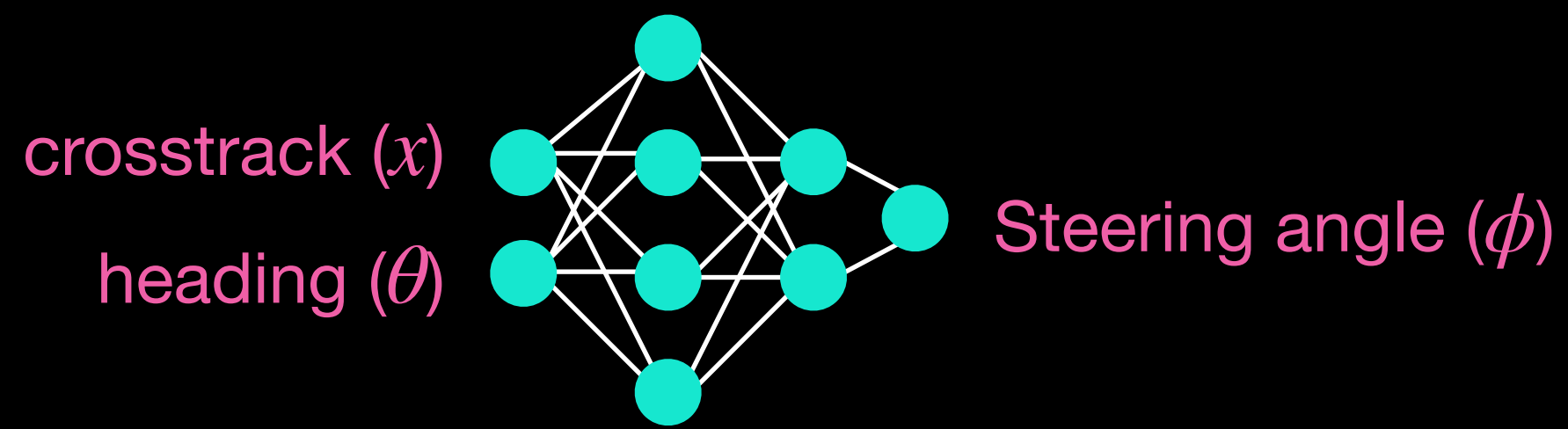
S. M. Katz\*, A. L. Corso\*, C. A. Strong\*, and M. J. Kochenderfer

Digital Avionics Systems Conference (2021), Journal of Aerospace Information Systems (2022)



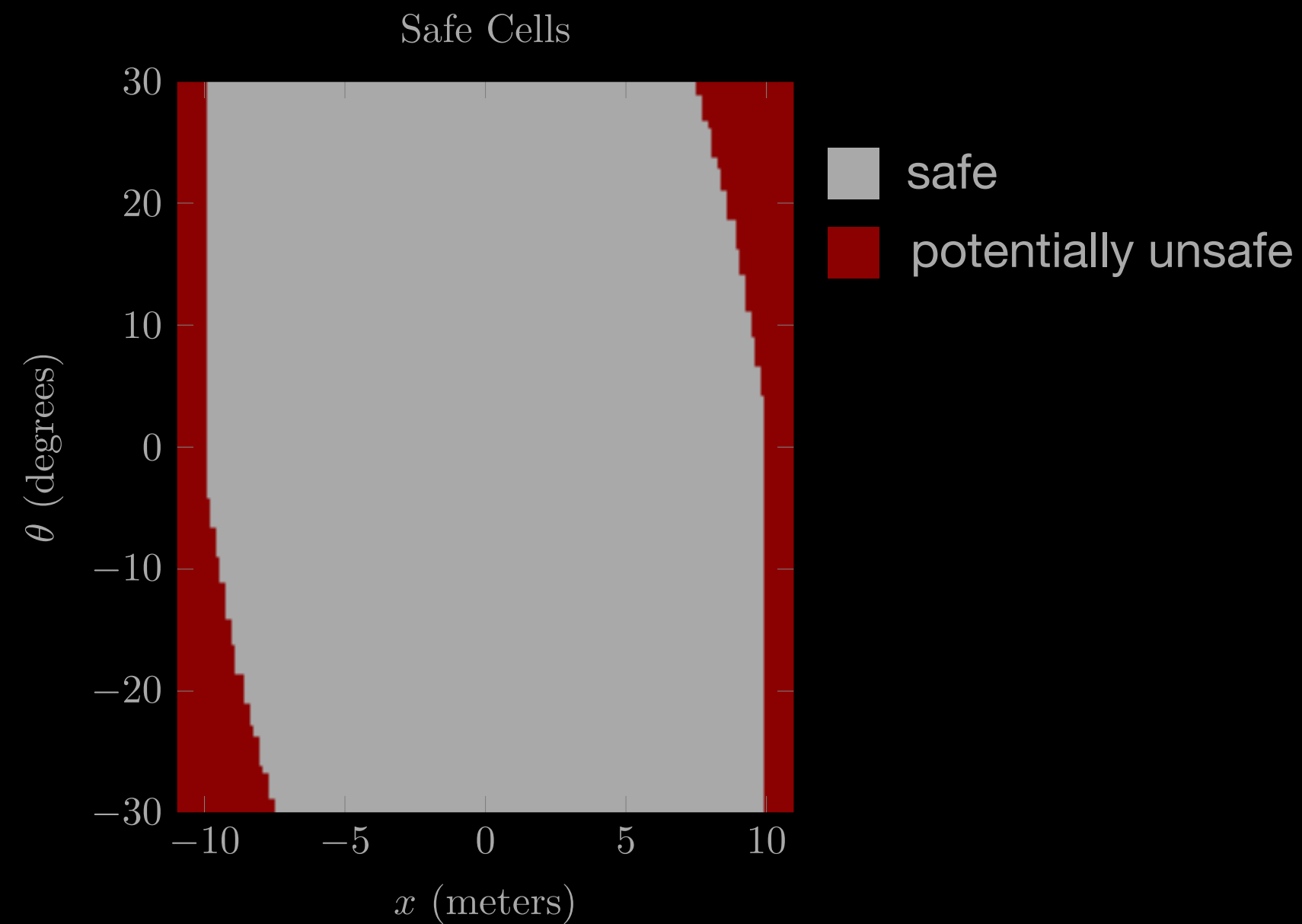




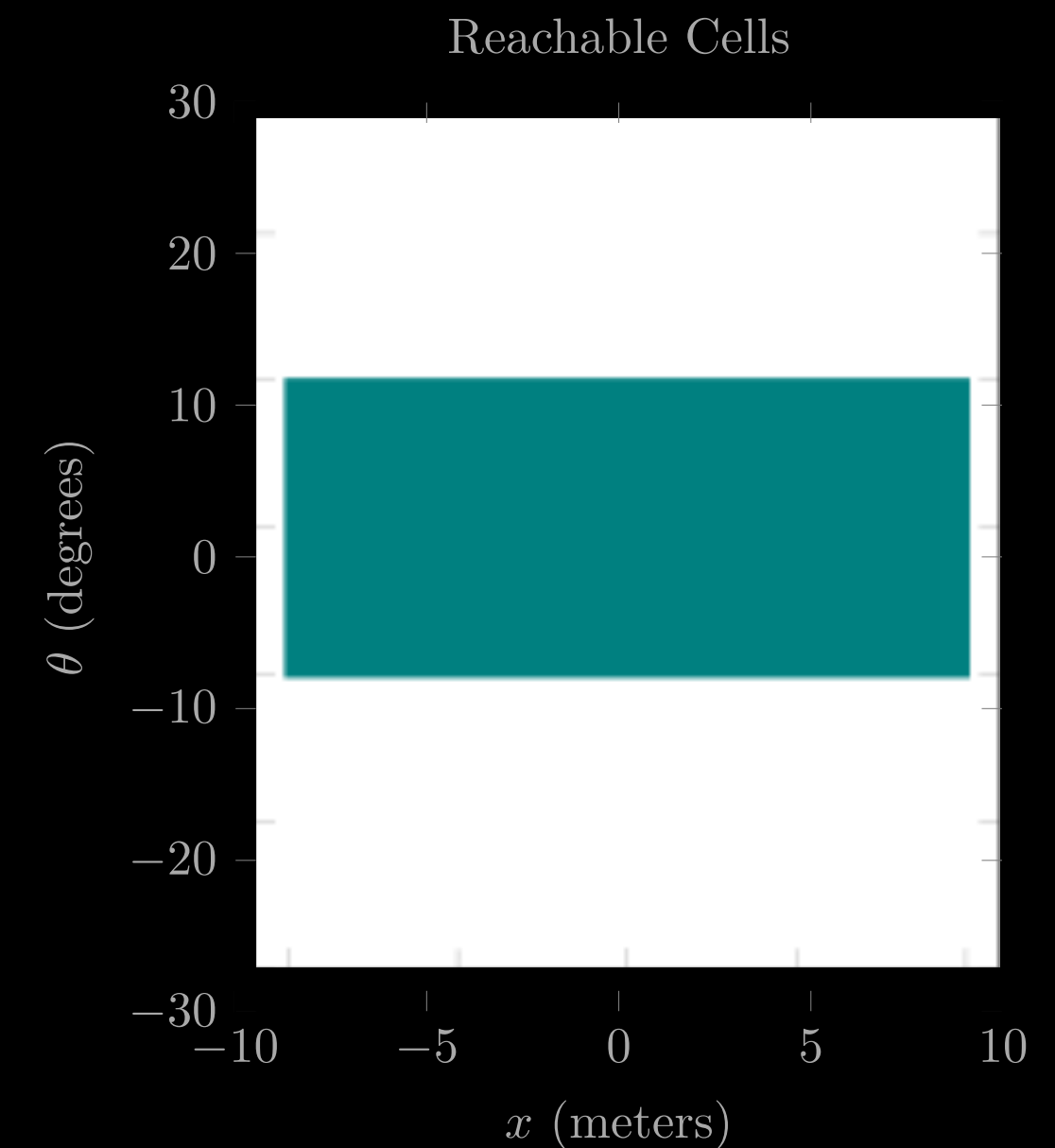


1 Will the neural network controller ever guide the aircraft to **leave the runway**?

2 Will the neural network controller keep the aircraft near the **center of the runway**?

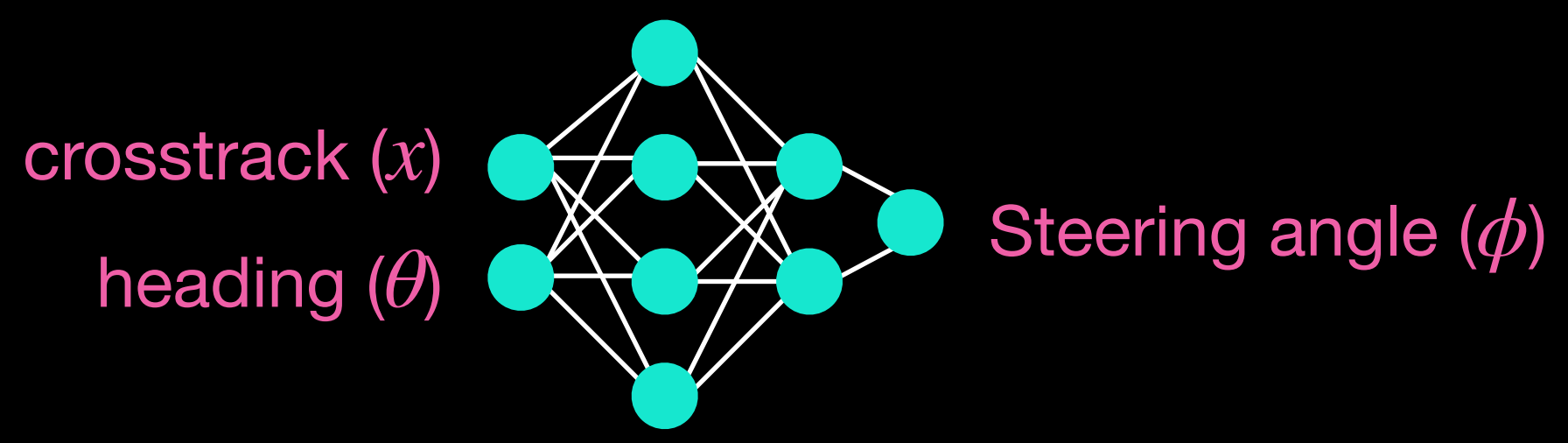


Katz et al. (2021)



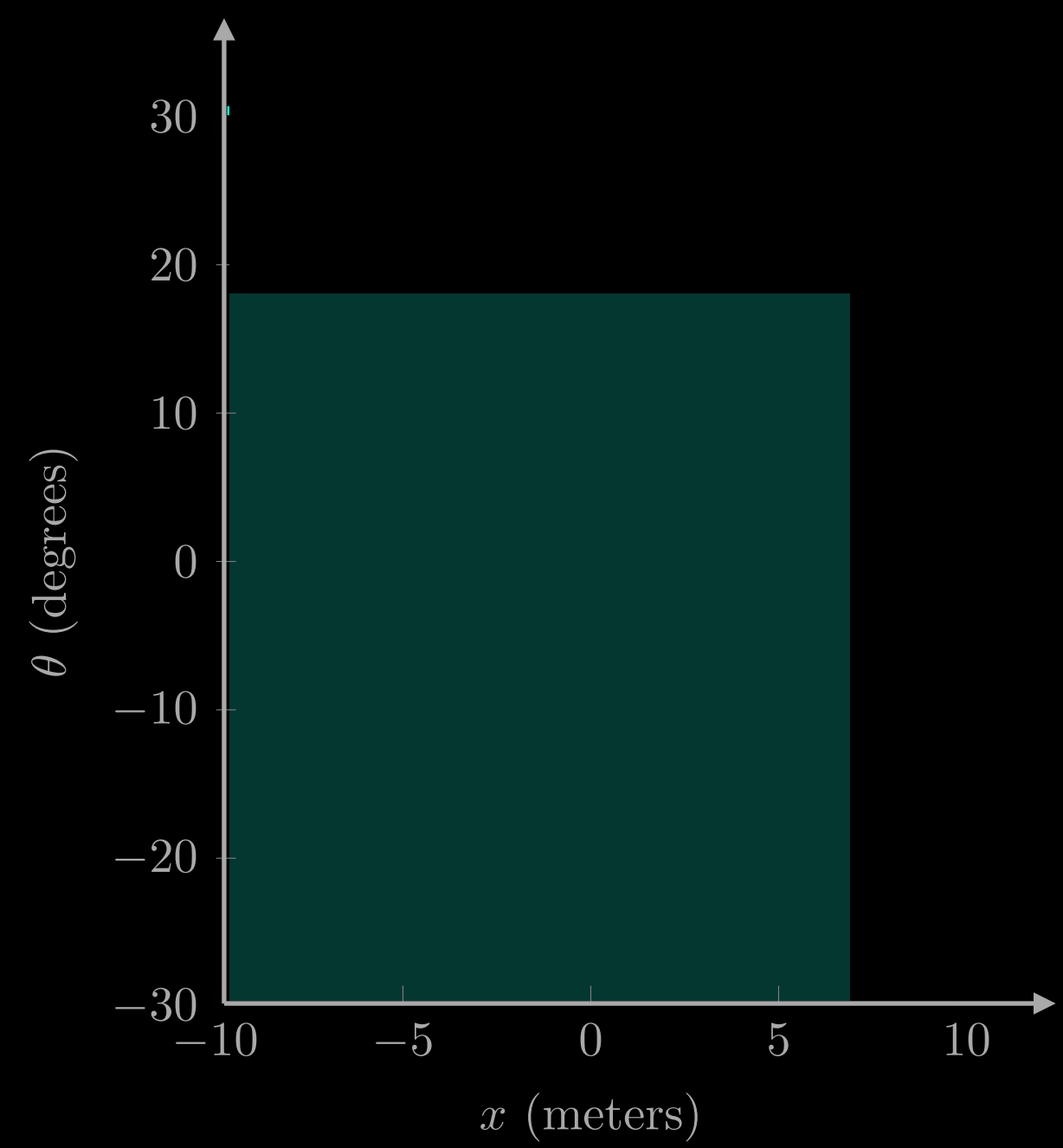
Julian and Kochenderfer (2021)



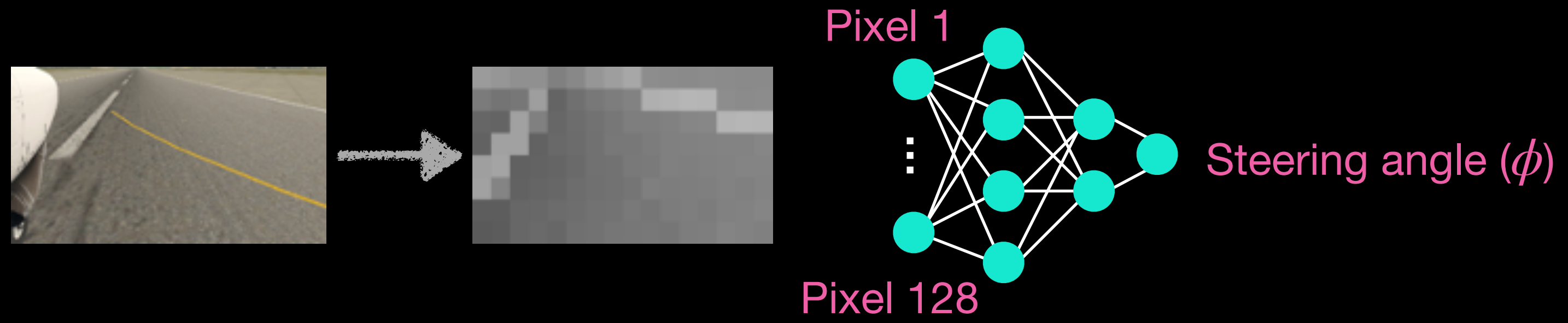


1 The inputs are **low-dimensional**.

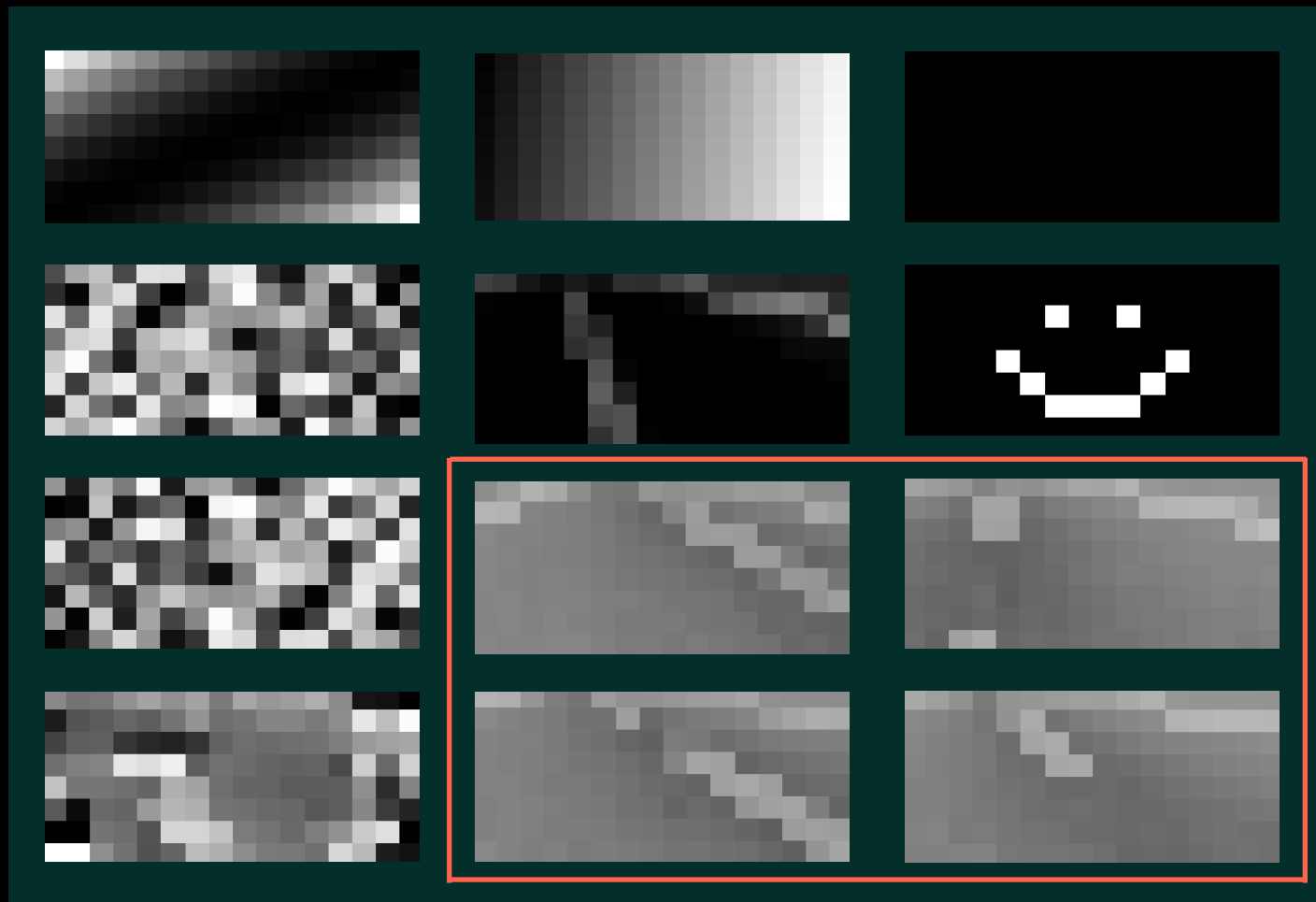
2 The set of plausible inputs is **well-defined**.



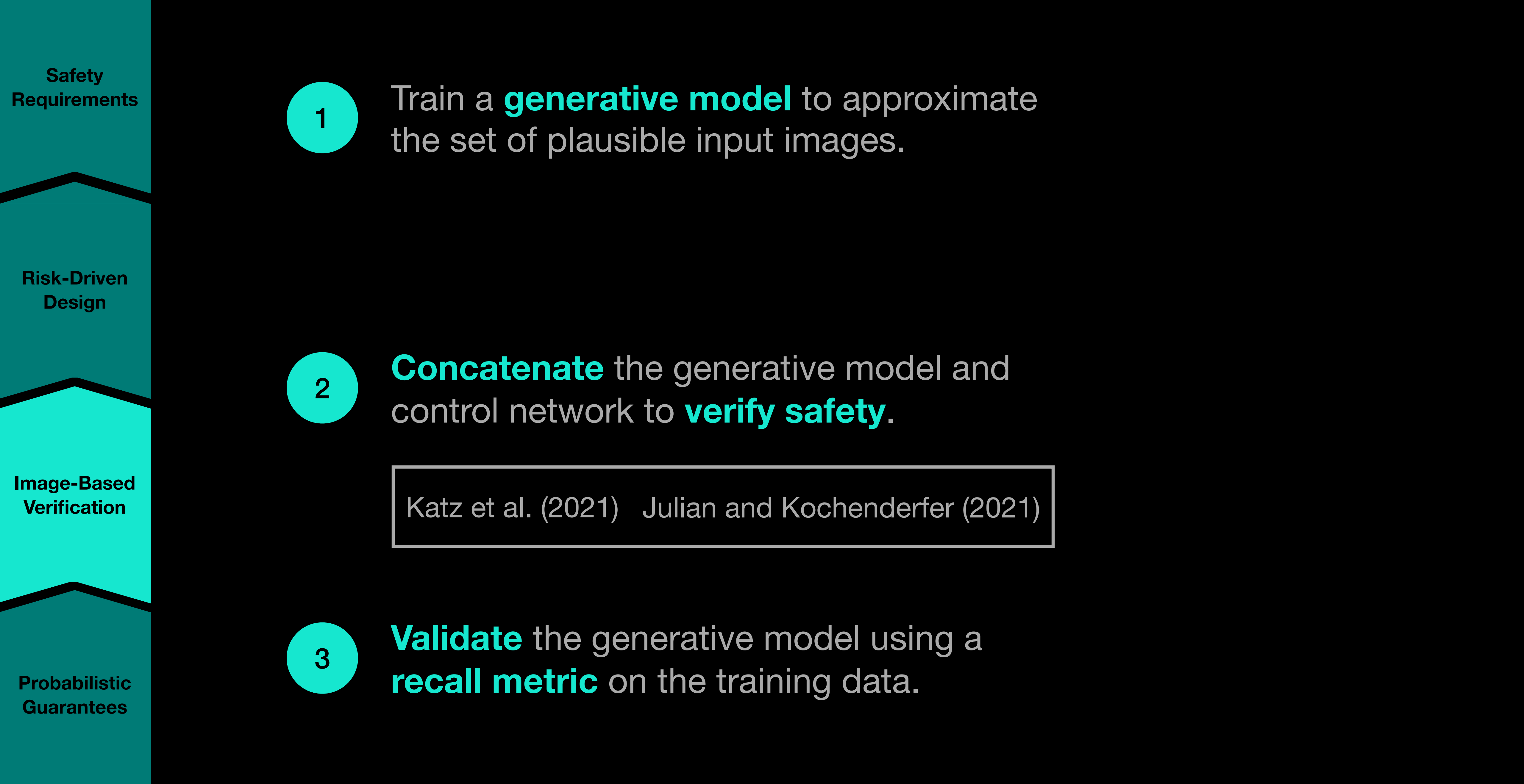
Katz et al. (2021) Julian and Kochenderfer (2021)



- 1 The inputs are ~~low-dimensional~~ **high**.
- 2 The set of plausible inputs is ~~well-defined~~ **not**.



Katz et al. (2021) Julian and Kochenderfer (2021)



Safety Requirements

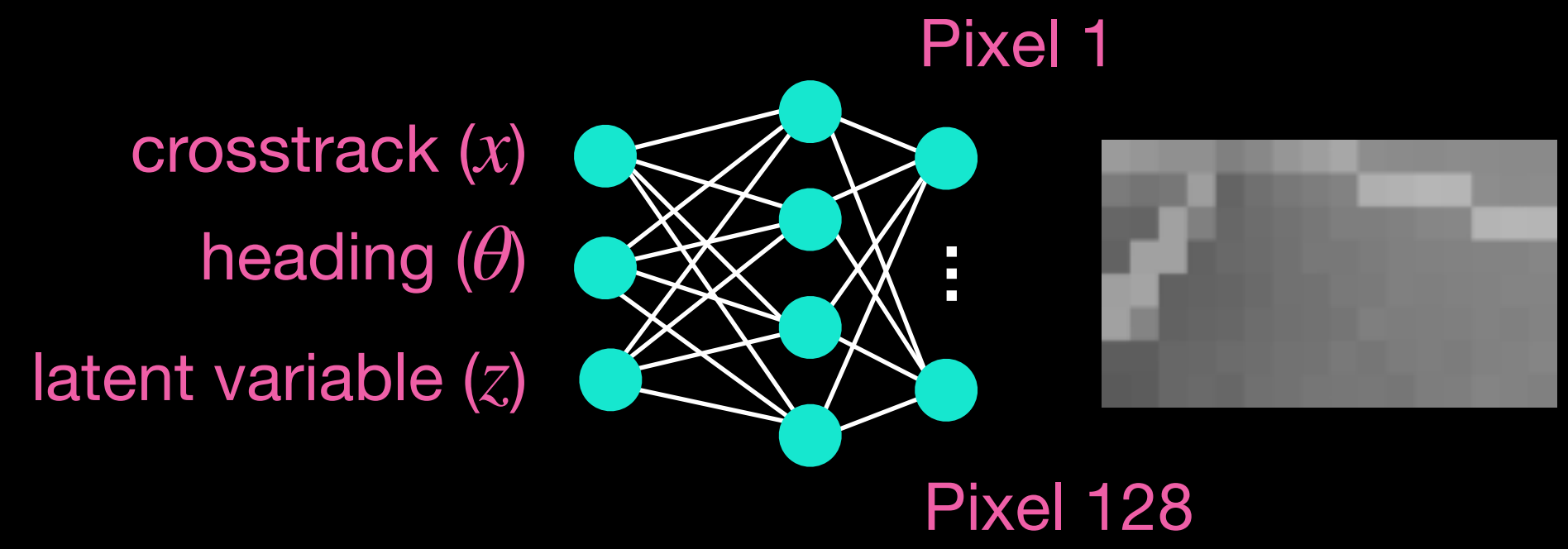
1

Train a **generative model** to approximate the set of plausible input images.

Risk-Driven Design

Generative Adversarial Network (GAN)

Image-Based Verification

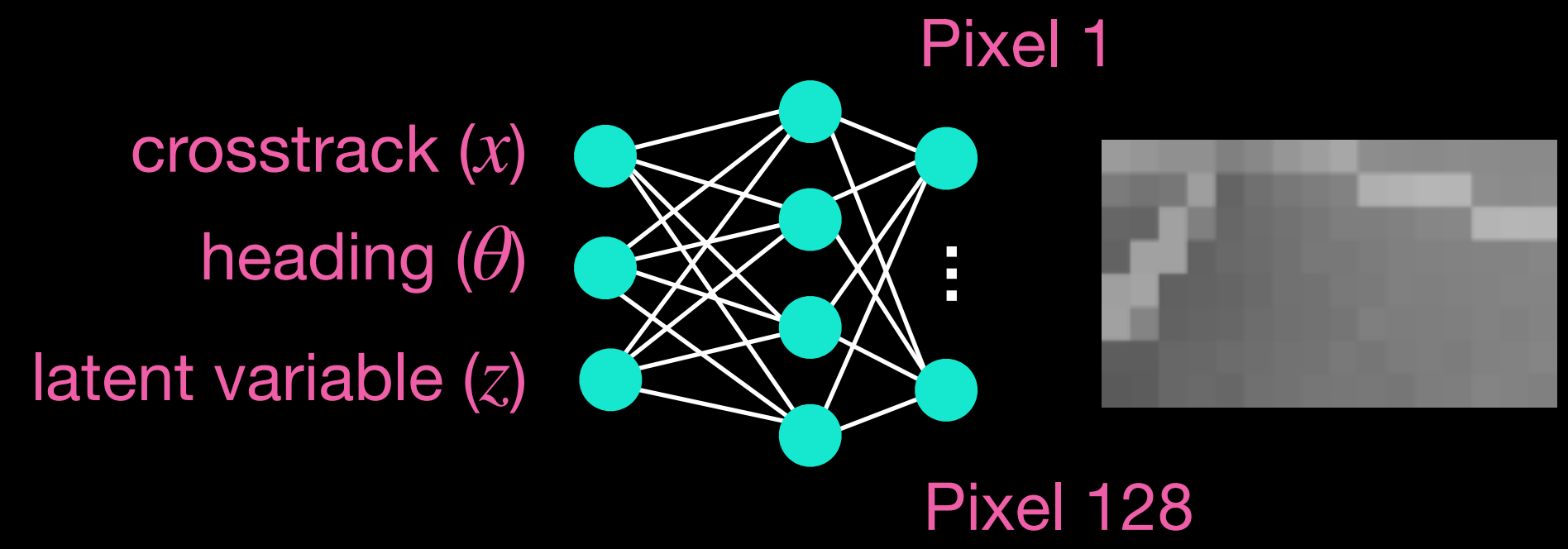


Probabilistic Guarantees

1

Train a **generative model** to approximate the set of plausible input images.

### Generative Adversarial Network (GAN)



True Image



Downsampled Image



Generated Image



Safety Requirements

2

**Concatenate** the generative model and control network to **verify safety**.

Risk-Driven Design

Generative Adversarial Network (GAN)

Control Network



Image-Based Verification

Probabilistic Guarantees

Safety Requirements

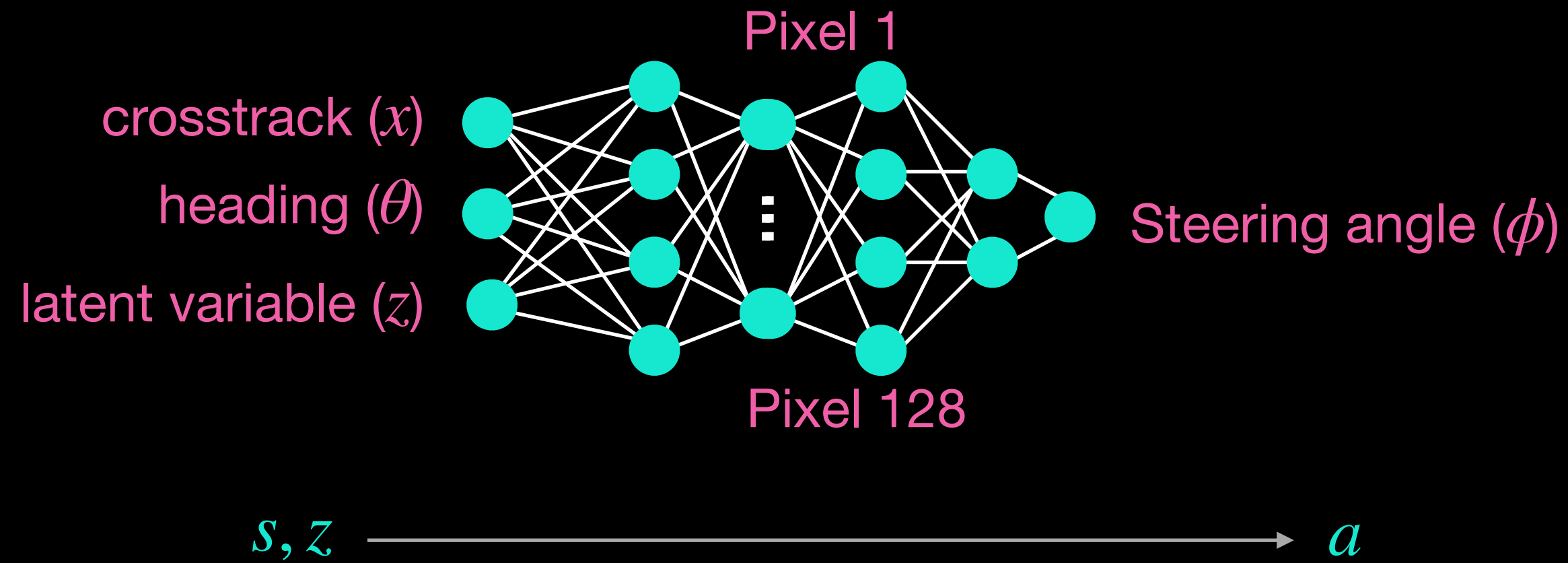
Risk-Driven Design

Image-Based Verification

Probabilistic Guarantees

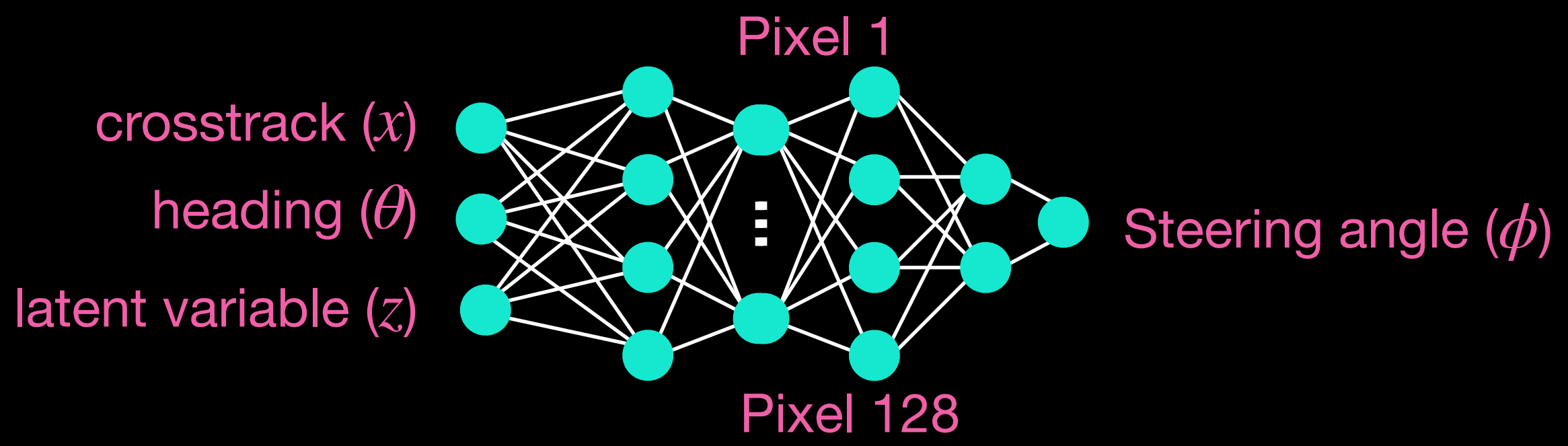
2

**Concatenate** the generative model and control network to **verify safety**.



2

**Concatenate** the generative model and control network to **verify safety**.

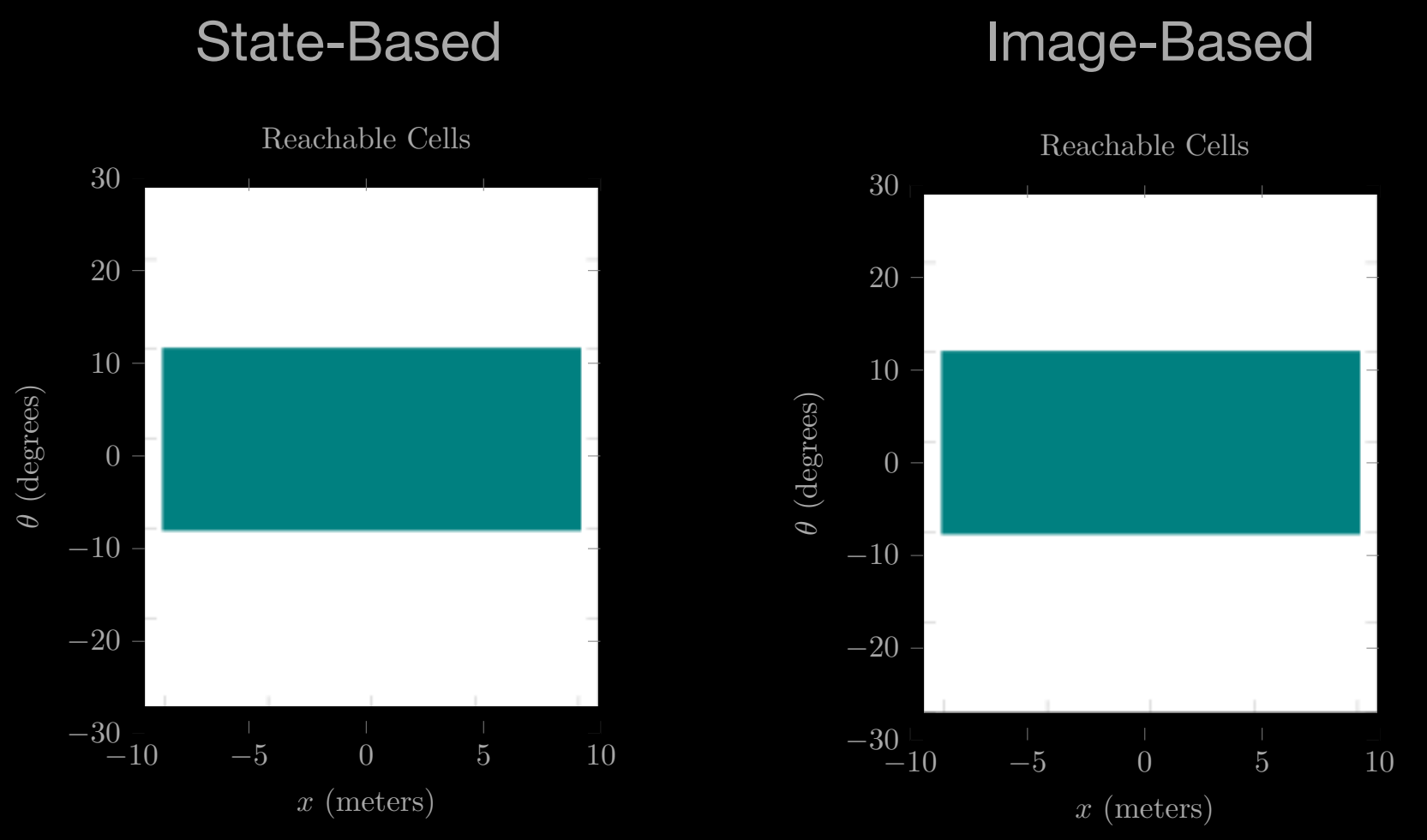
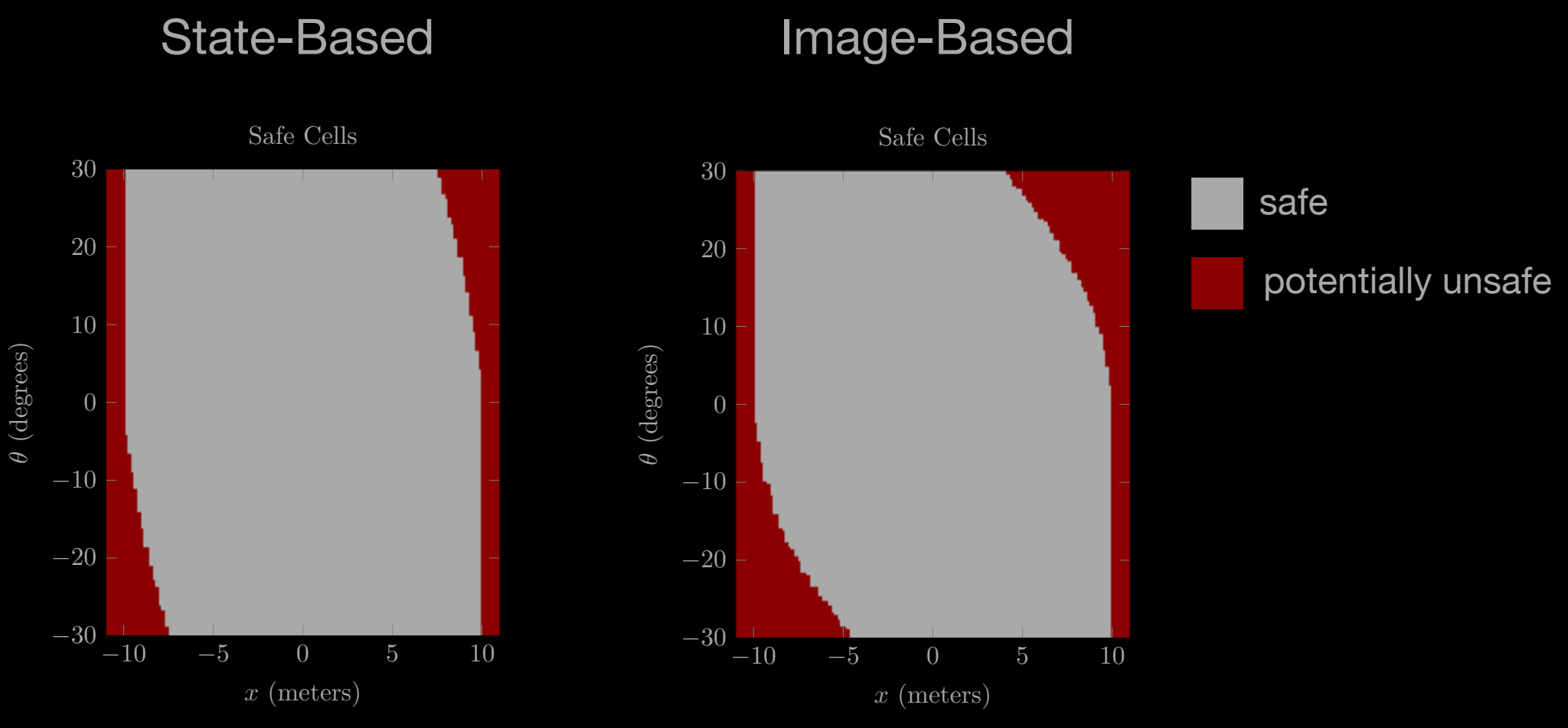


1

Will the neural network controller ever guide the aircraft to **leave the runway**?

2

Will the neural network controller keep the aircraft near the **center of the runway**?



Safety Requirements

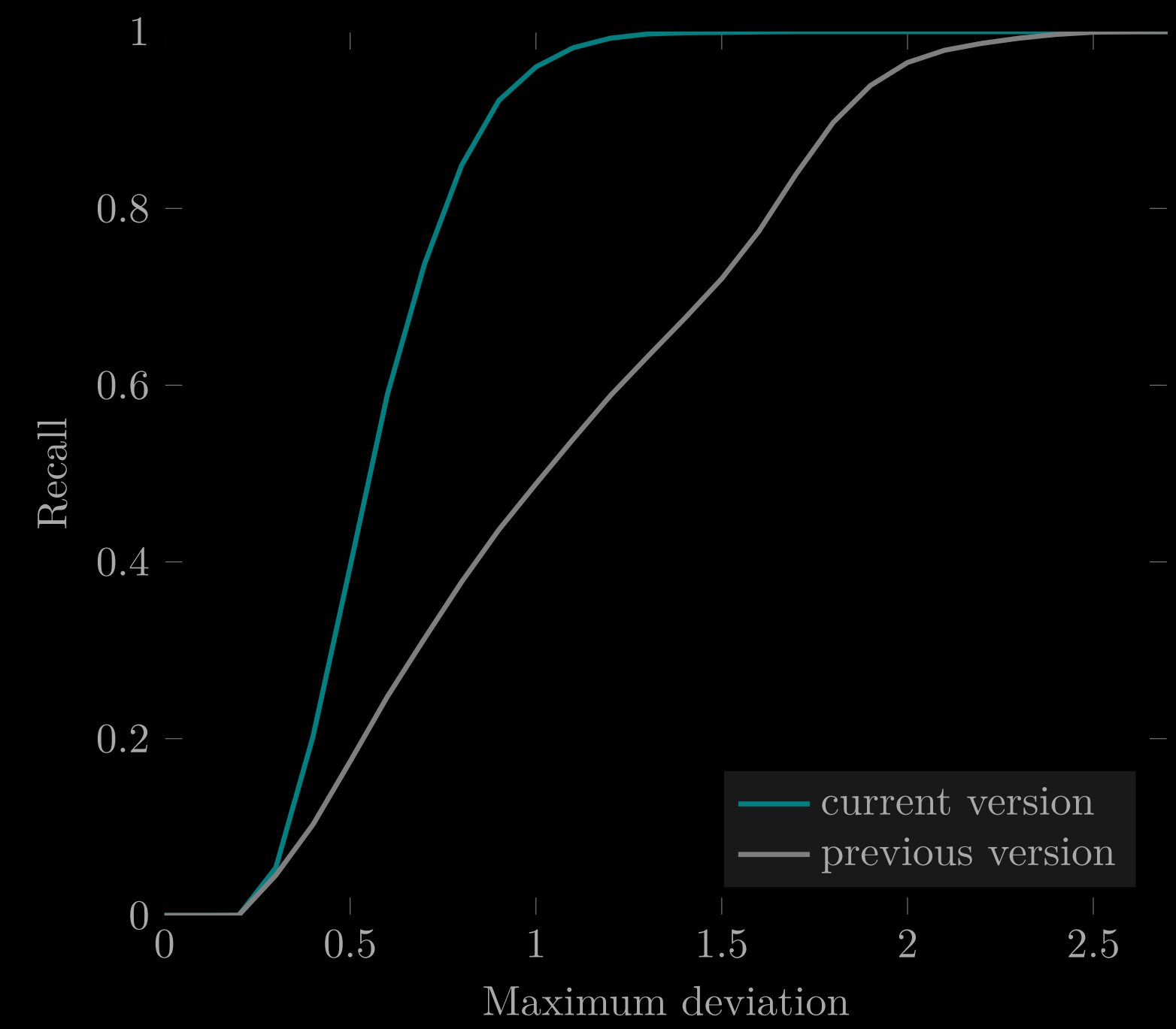
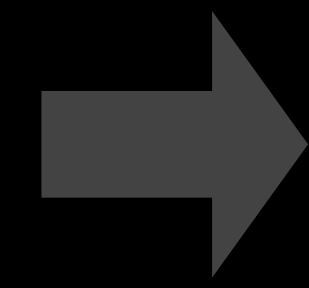
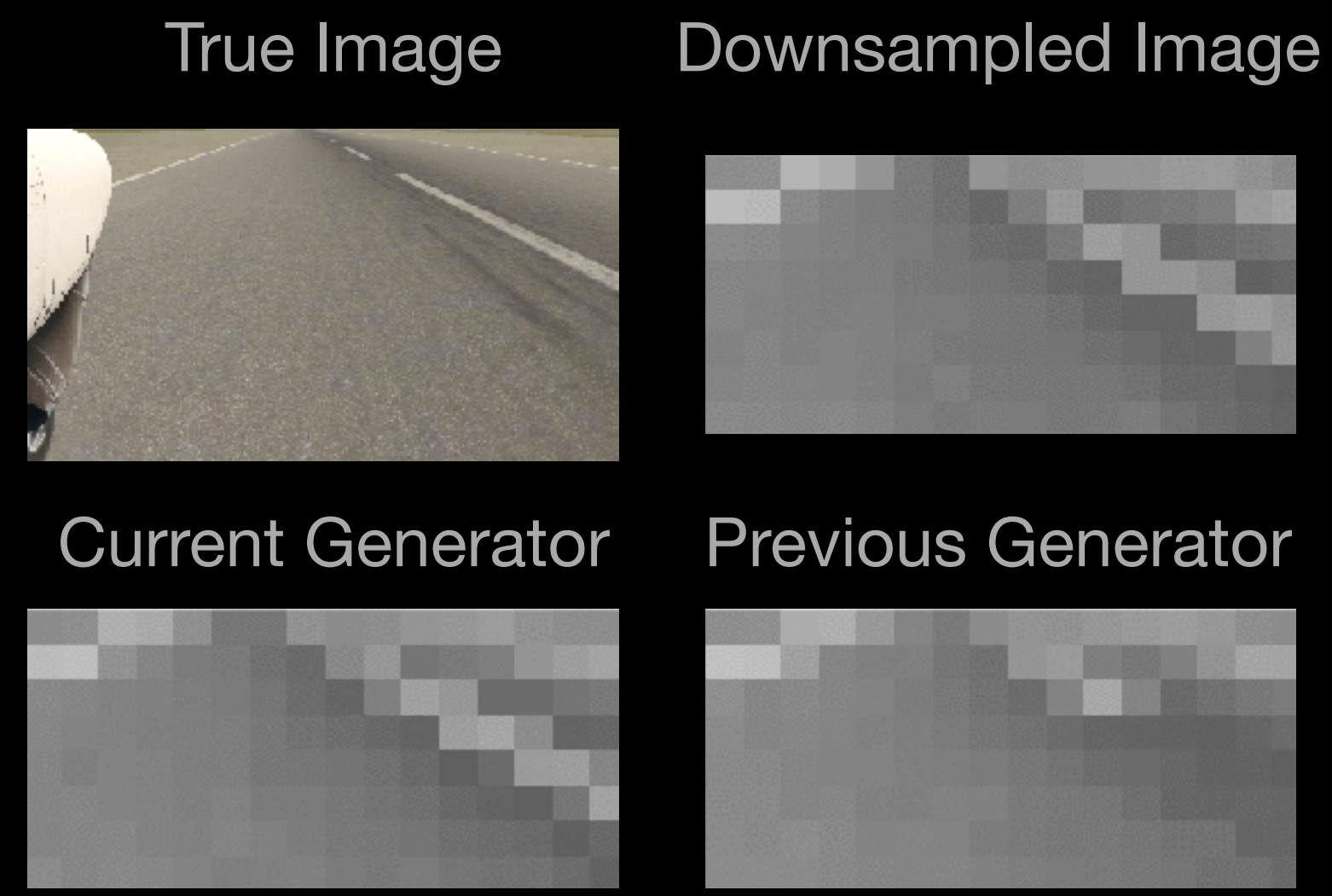
3

**Validate** the generative model using a **recall metric** on the training data.

Risk-Driven Design

How well can the generative model **produce the real images**?

Image-Based Verification



Probabilistic Guarantees

# Verification of Image-Based Neural Network Controllers Using Generative Models

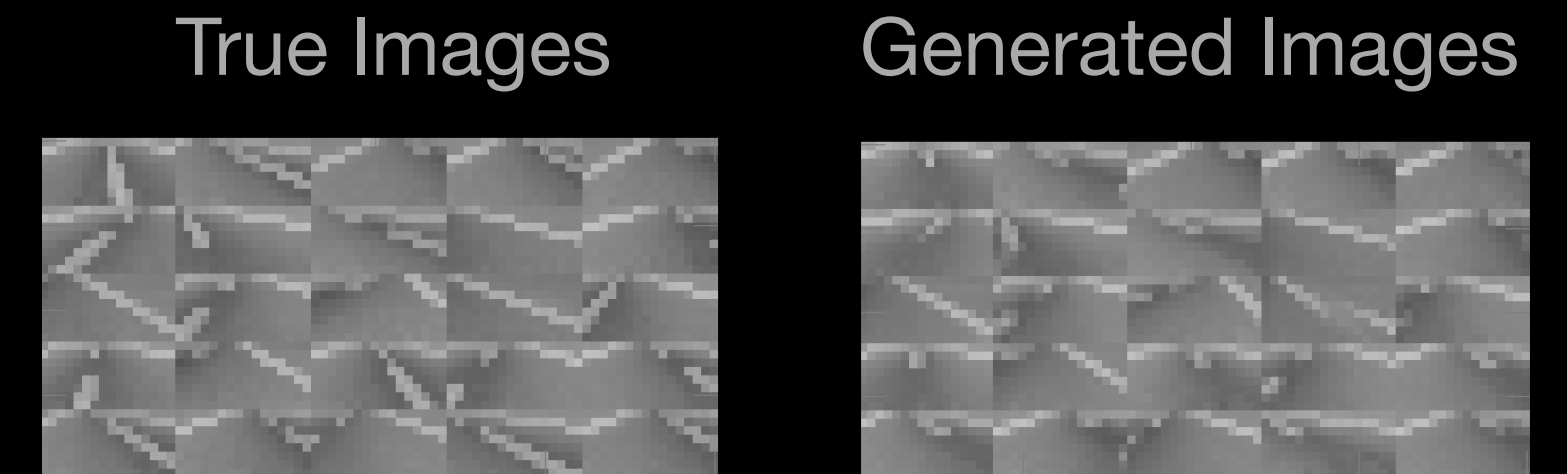
Safety Requirements

Risk-Driven Design

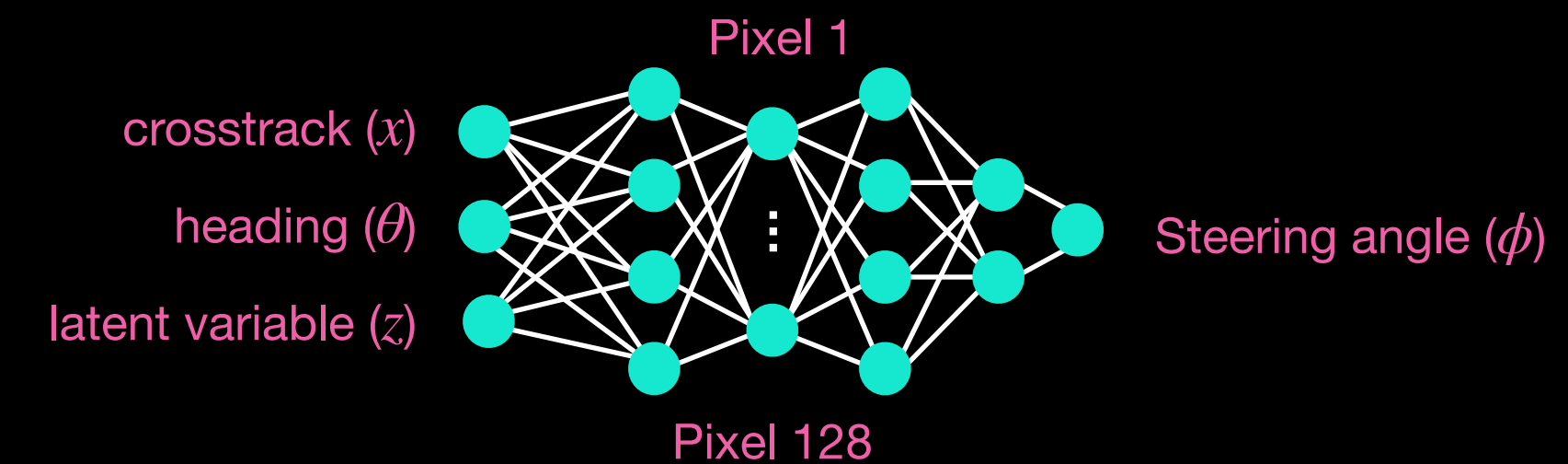
Image-Based Verification

Probabilistic Guarantees

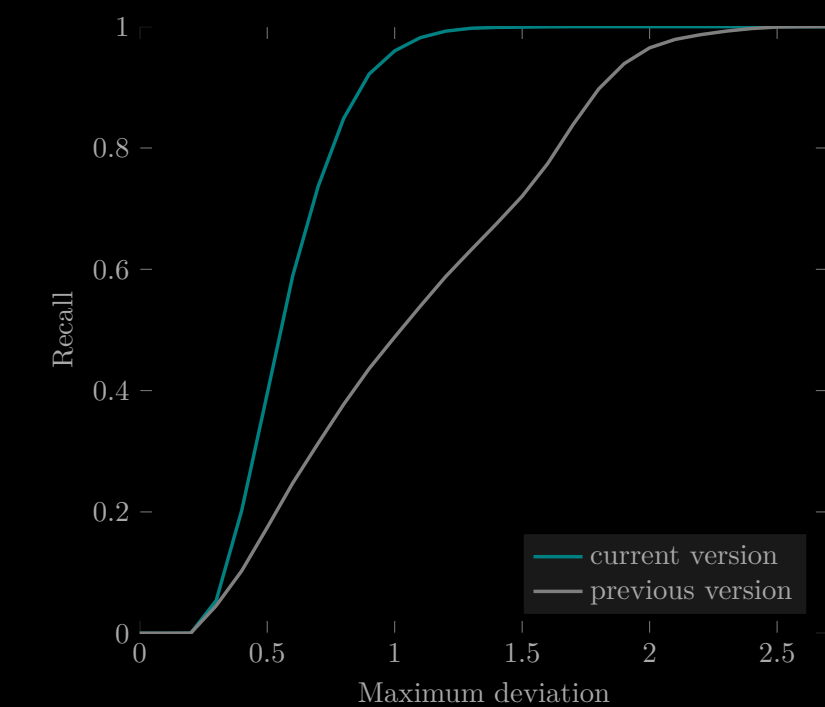
1 Train a **generative model** to approximate the set of plausible input images.

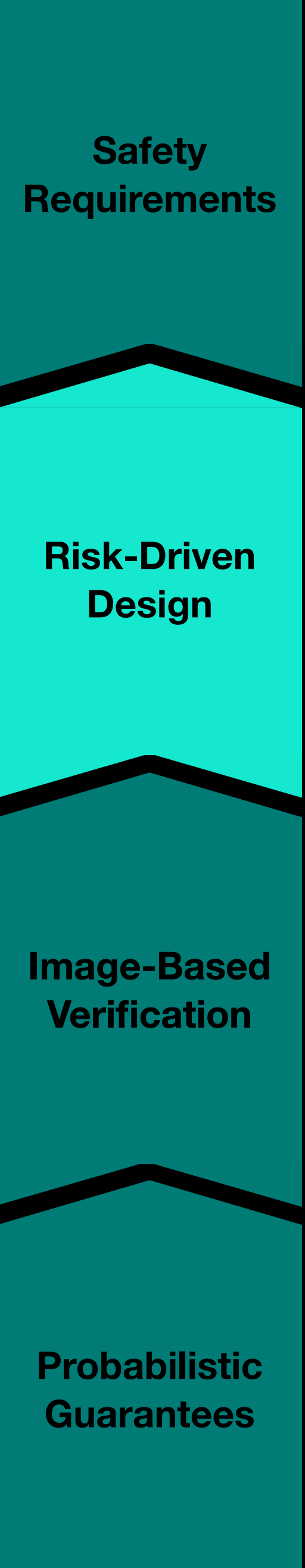


2 **Concatenate** the generative model and control network to **verify safety**.



3 **Validate** the generative model using a **recall metric** on the training data.

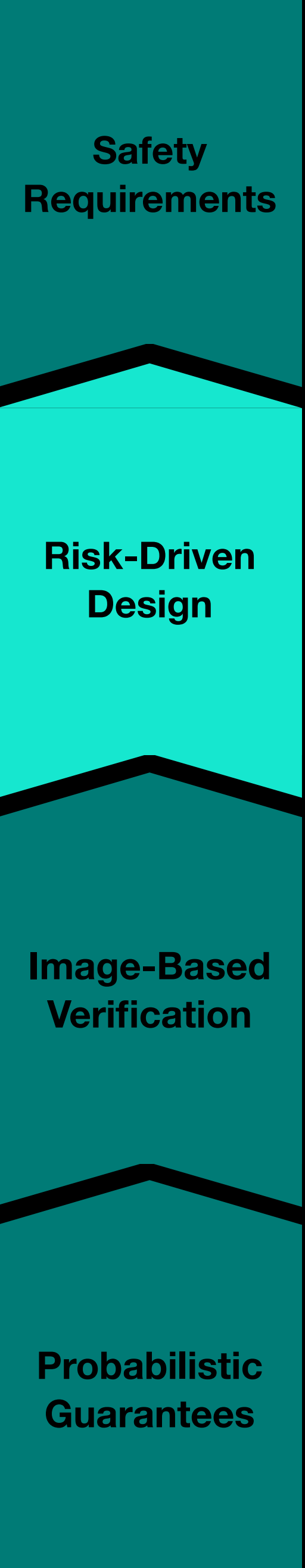




# Risk-Driven Design of Perception Systems

A. L. Corso\*, S. M. Katz\*, C. A. Innes, X. Du, S. Ramamoorthy, and M. J. Kochenderfer

Advances in Neural Information Processing Systems (2022)



**Key Insight:** Not all perception errors will have an equal effect on **closed-loop performance**.

**Key Insight:** Not all perception errors will have an equal effect on **closed-loop performance**.

Safety  
Requirements

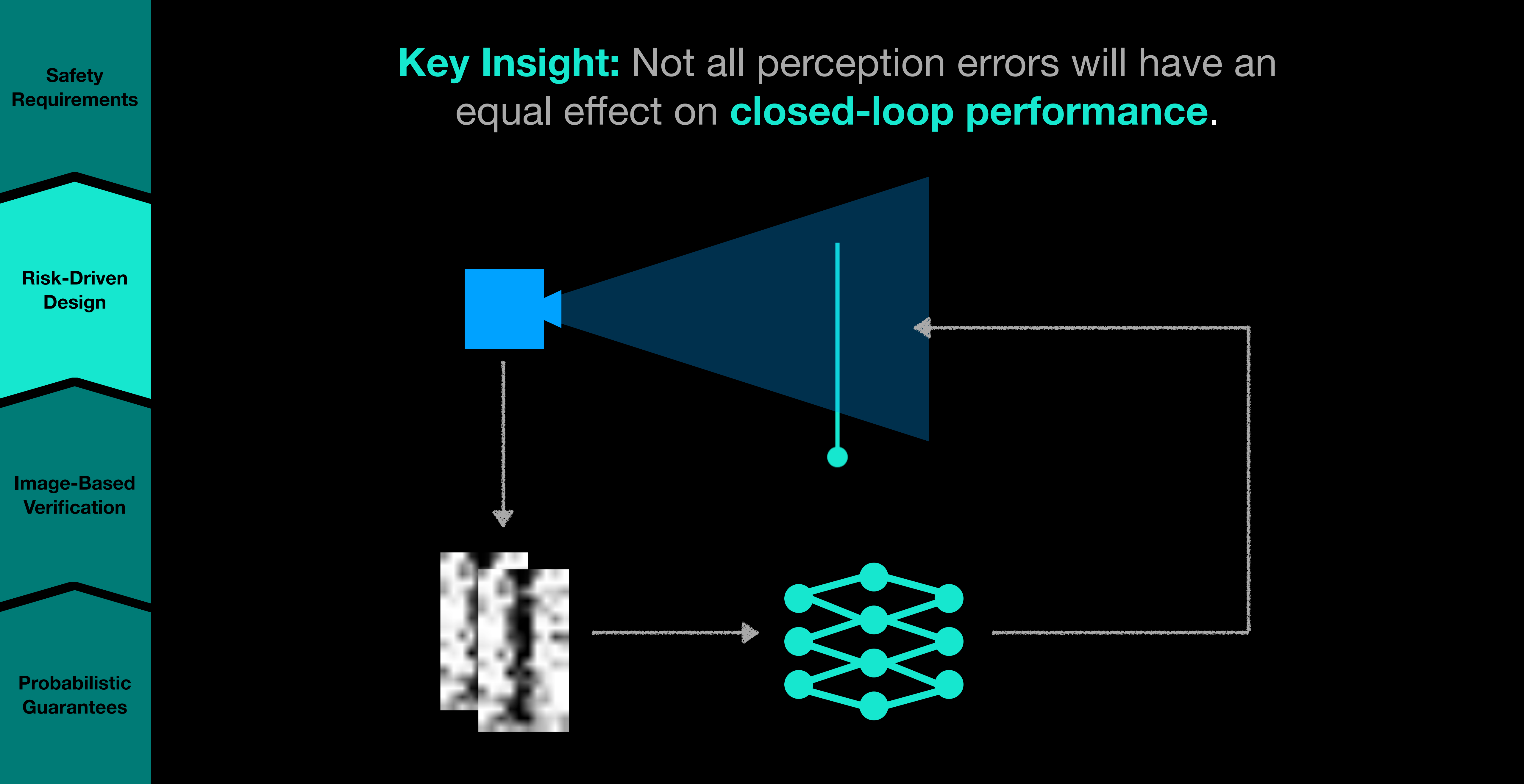
Risk-Driven  
Design

Image-Based  
Verification

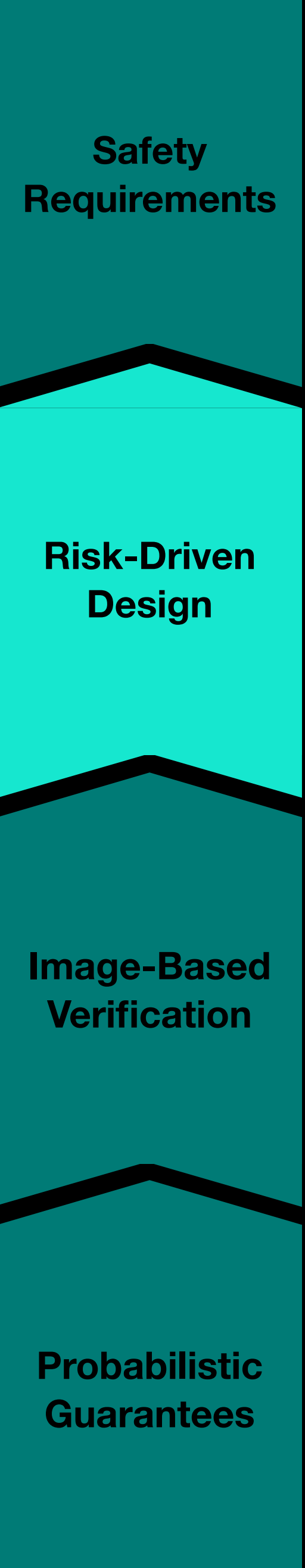
Probabilistic  
Guarantees



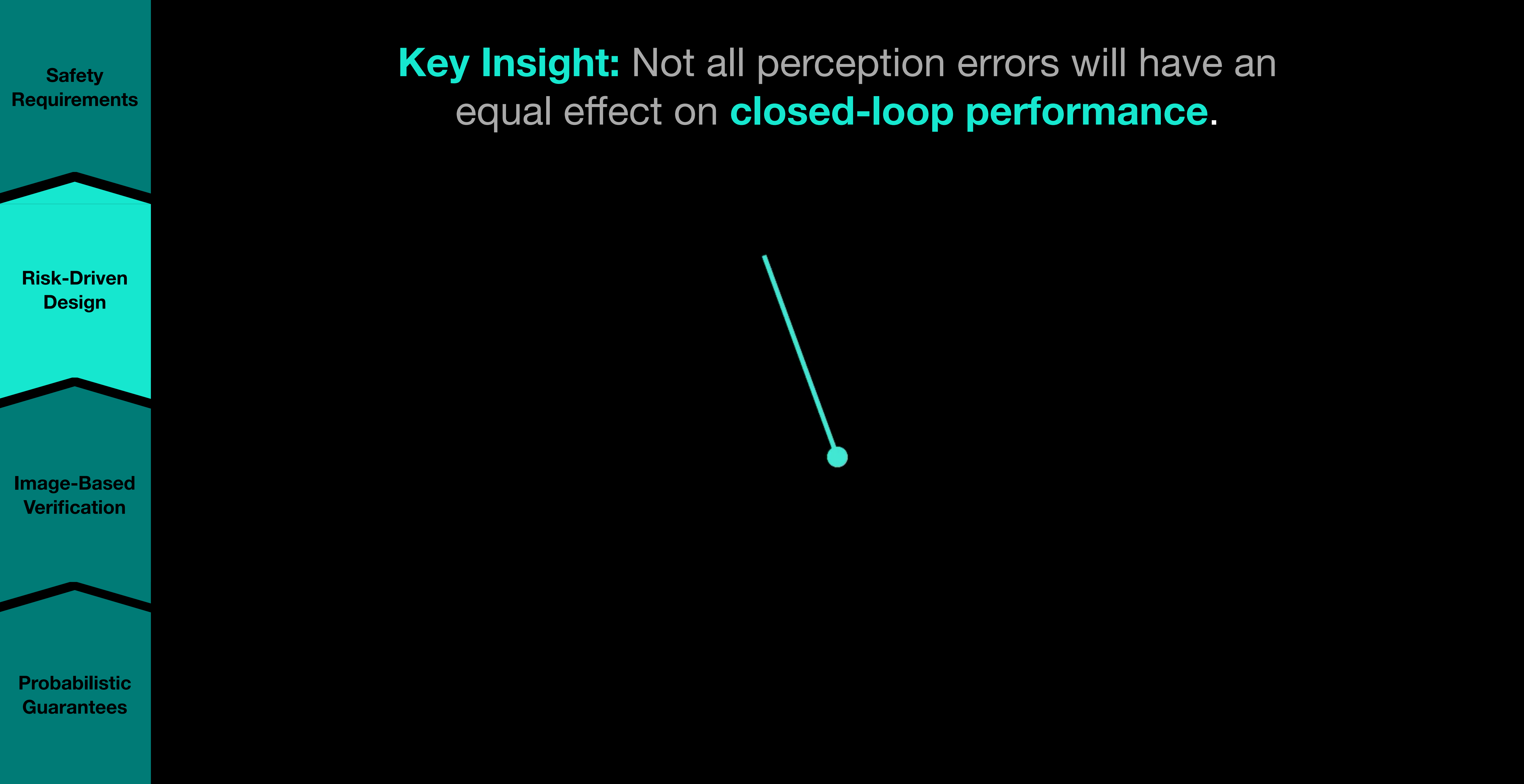
**Key Insight:** Not all perception errors will have an equal effect on **closed-loop performance**.



**Key Insight:** Not all perception errors will have an equal effect on **closed-loop performance**.



**Key Insight:** Not all perception errors will have an equal effect on **closed-loop performance**.



Safety Requirements

**Key Insight:** Not all perception errors will have an equal effect on **closed-loop performance**.

Risk-Driven Design

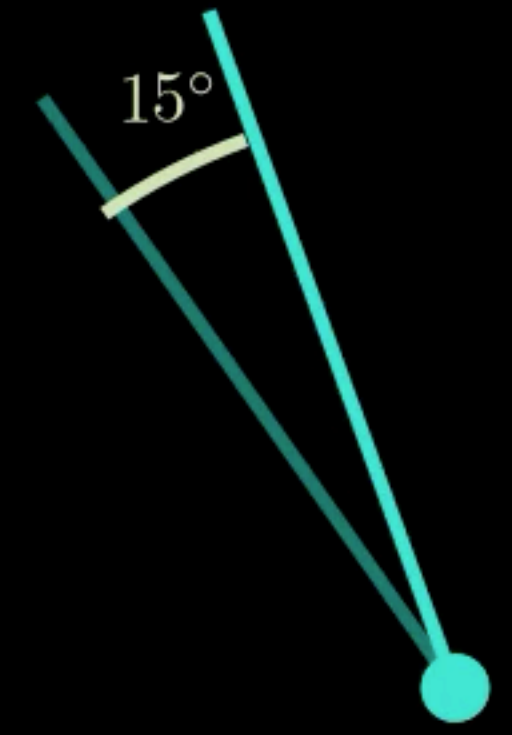


Image-Based Verification

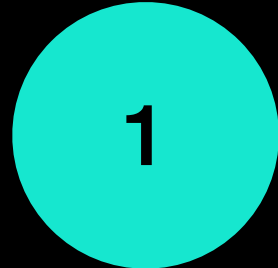
Probabilistic Guarantees



**Our Idea:** Account for the “**riskiness**” of making perception errors during the perception system **design process**.

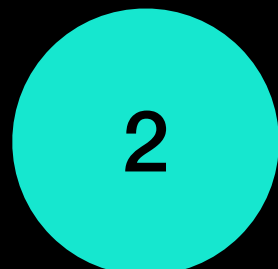


**Our Idea:** Account for the “**riskiness**” of making perception errors during the perception system **design process**.



Formulate a **risk function**.

$$\rho( , )$$



Use the risk function to **design a safer perception system**.



Risk-driven **training**

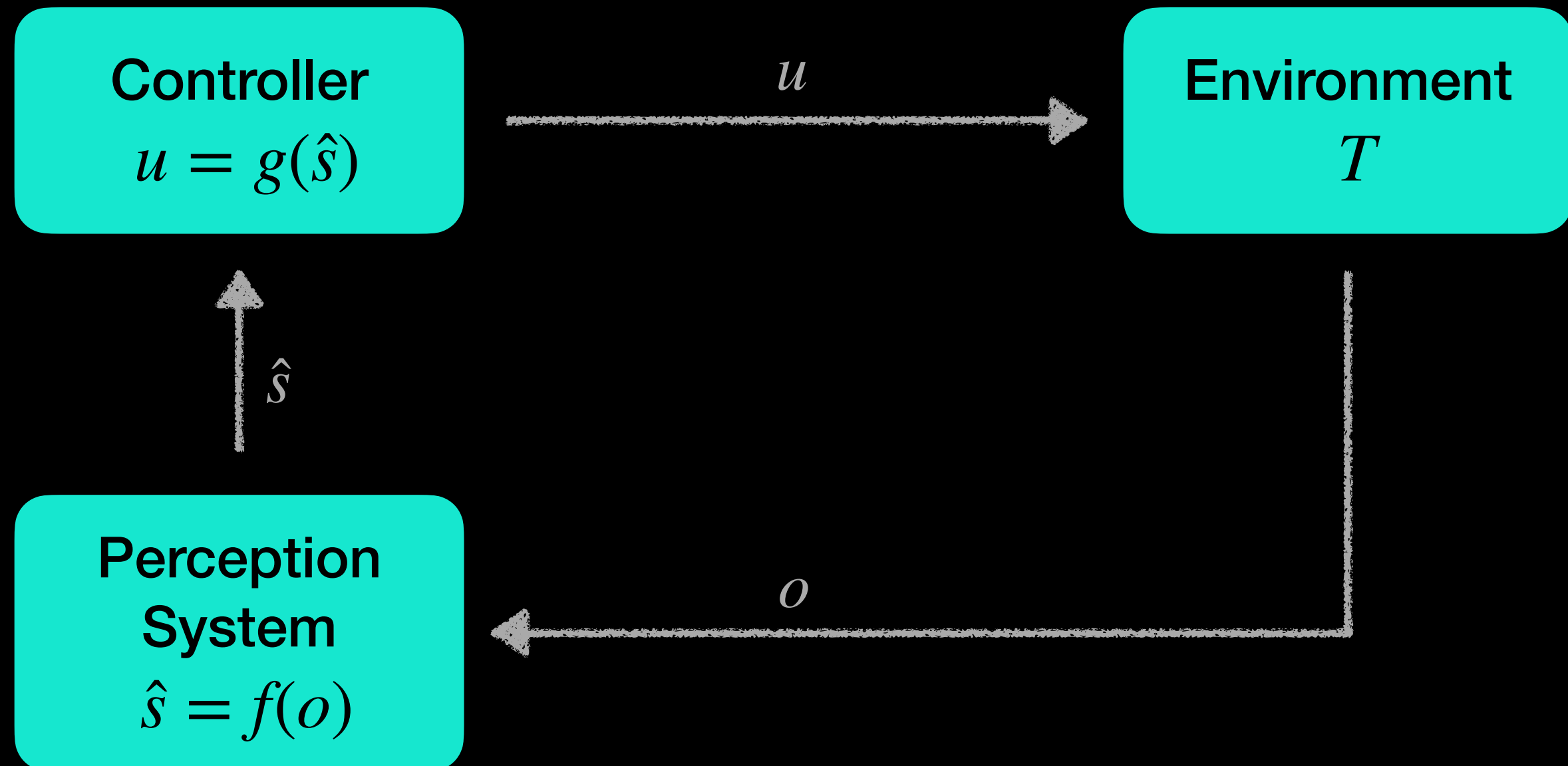


Risk-driven **data generation**



**Our Idea:** Account for the “**riskiness**” of making perception errors during the perception system **design process**.

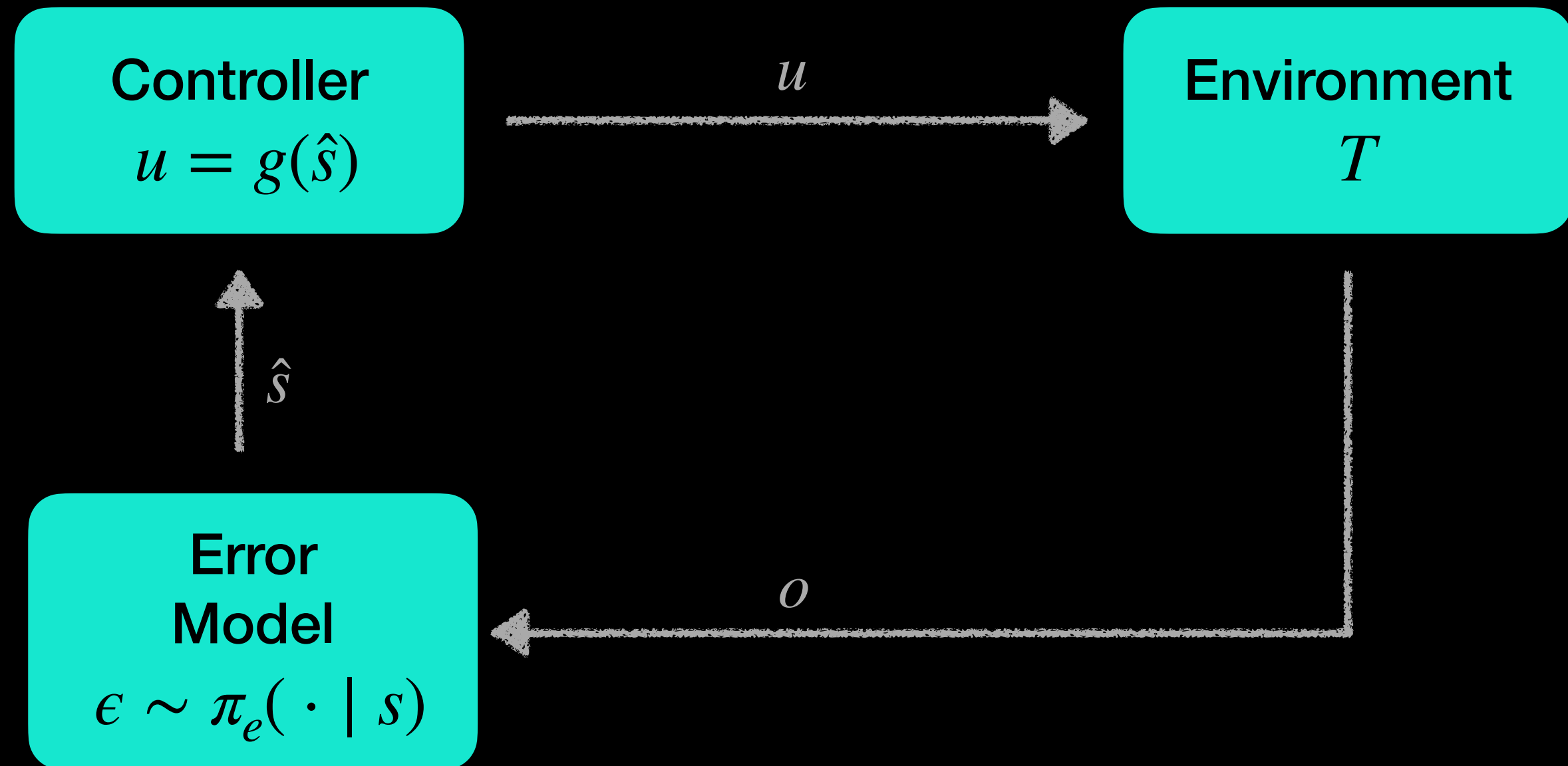
- 1 Formulate a **risk function**.





**Our Idea:** Account for the “**riskiness**” of making perception errors during the perception system **design process**.

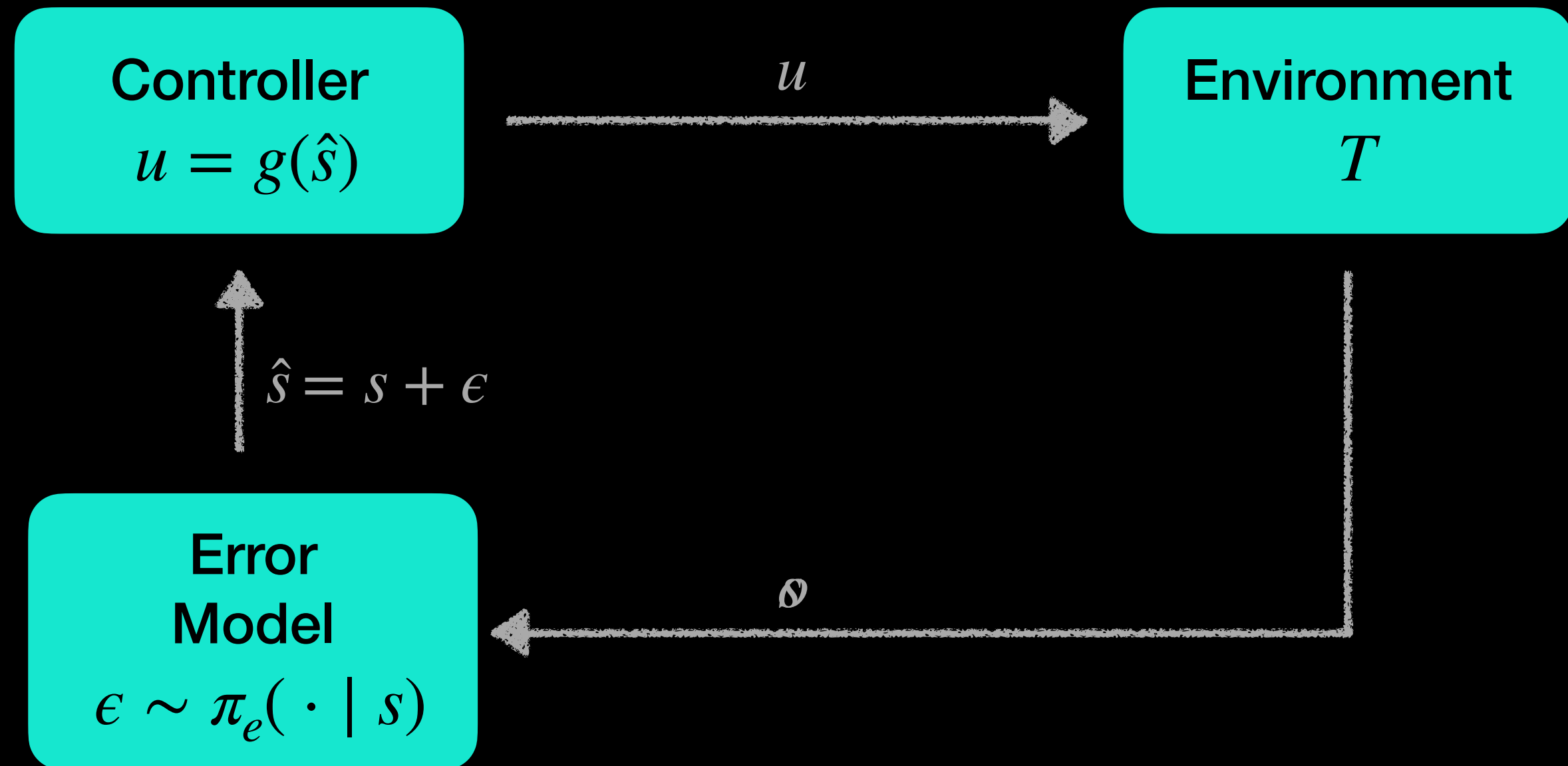
- 1 Formulate a **risk function**.





**Our Idea:** Account for the “**riskiness**” of making perception errors during the perception system **design process**.

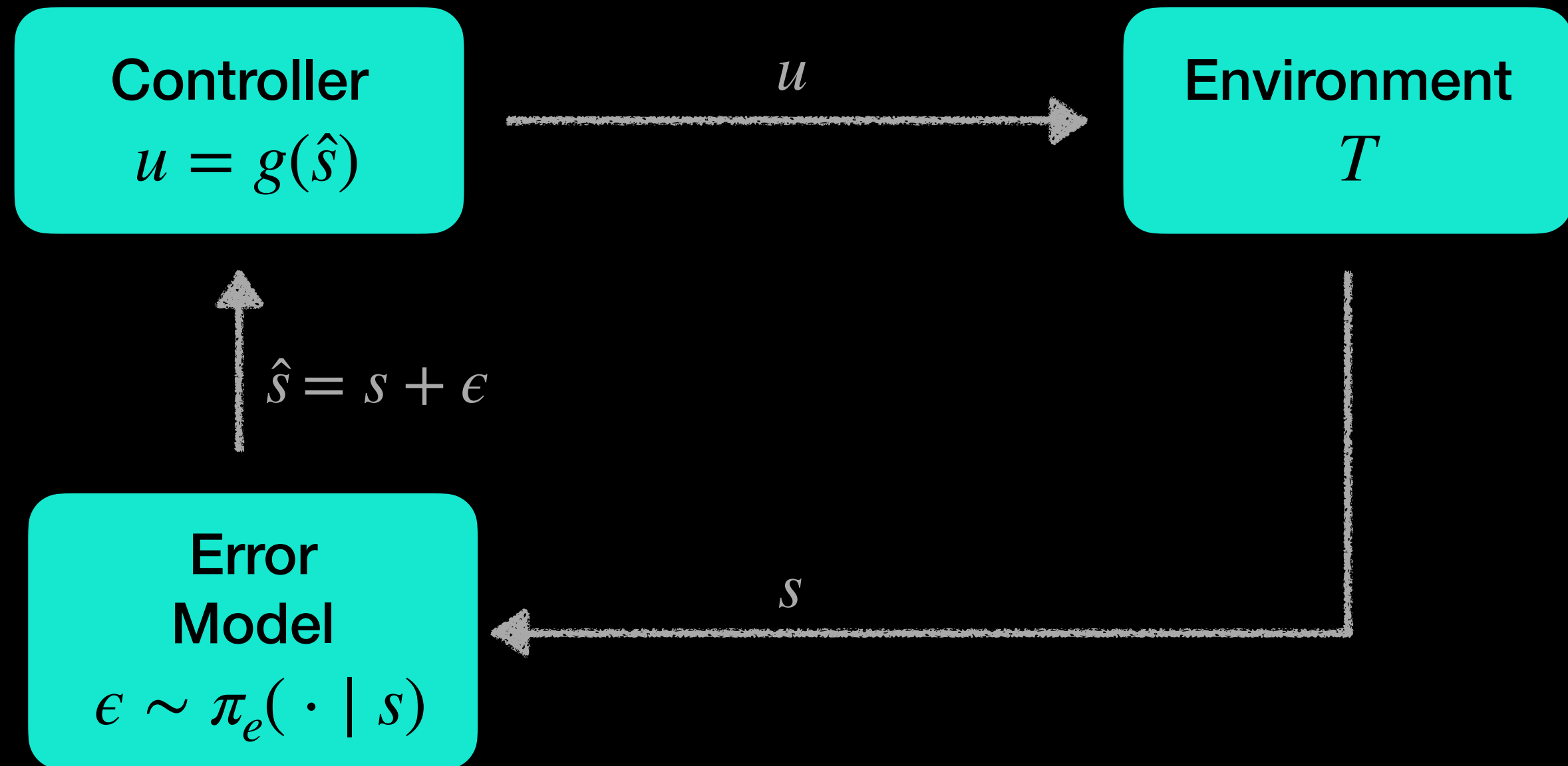
- 1 Formulate a **risk function**.





**Our Idea:** Account for the “**riskiness**” of making perception errors during the perception system **design process**.

- 1 Formulate a **risk function**.

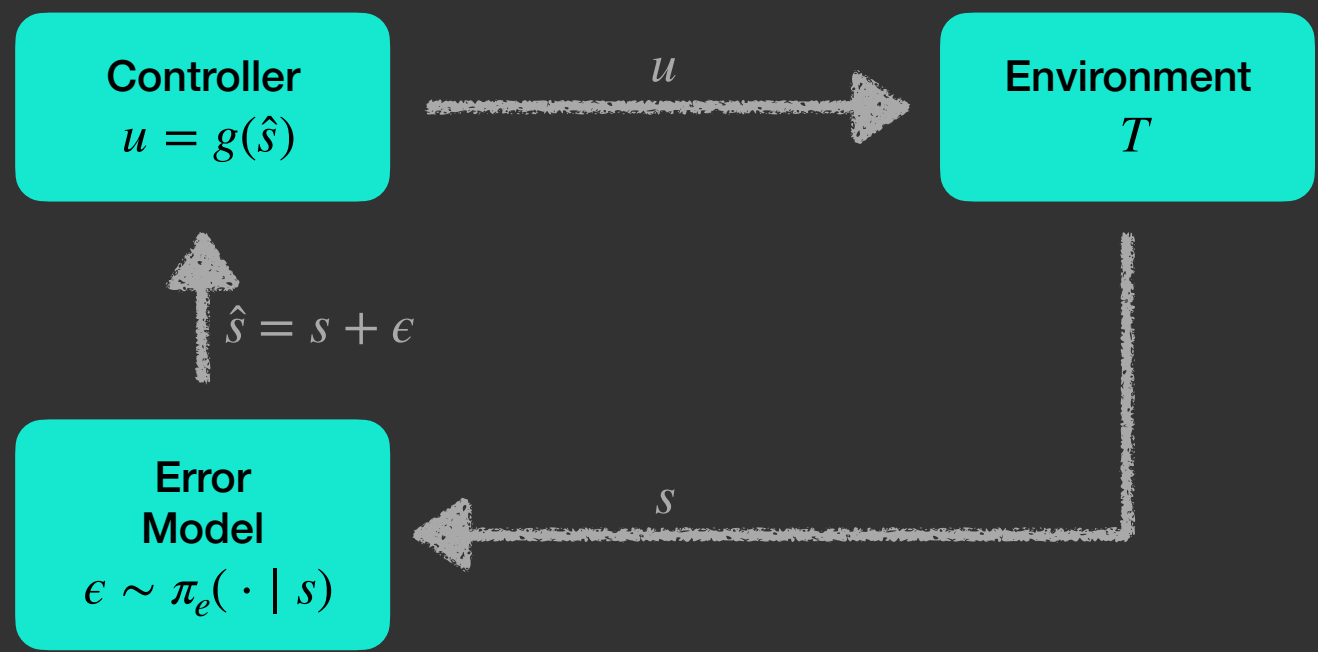




**Our Idea:** Account for the “**riskiness**” of making perception errors during the perception system **design process**.

1 Formulate a **risk function**.

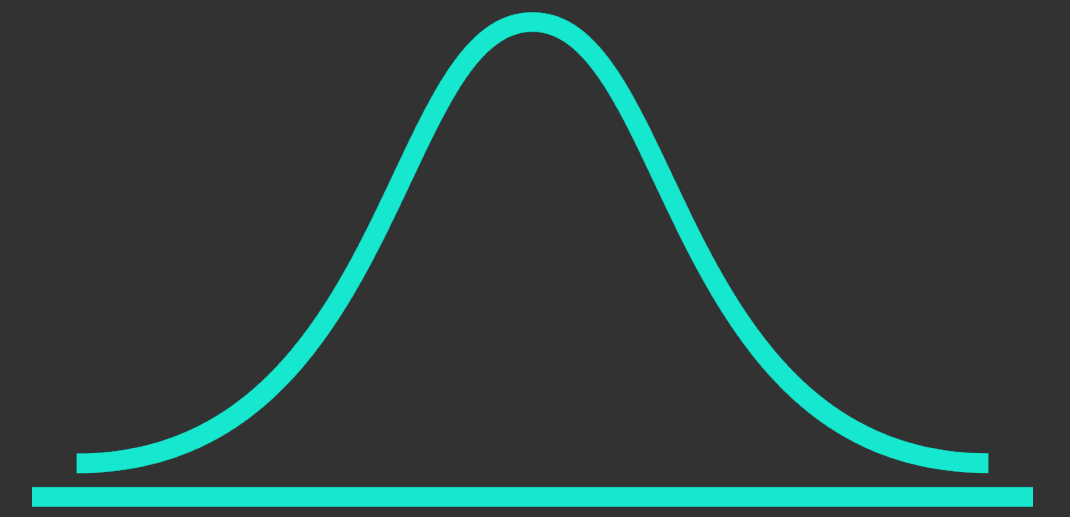
Markov Decision Process



Distributional Dynamic Programming

Distribution of Future Costs

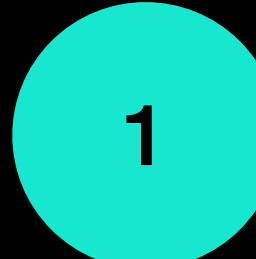
$$Z(s, \epsilon)$$



$$\rho(s, \epsilon) = \text{CVaR} [ \quad ]$$

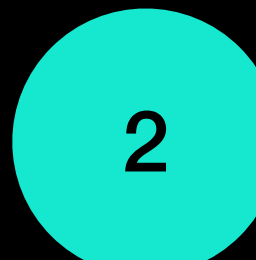


**Our Idea:** Account for the “**riskiness**” of making perception errors during the perception system **design process**.



Formulate a **risk function**.

$$\rho(s, \epsilon) = \text{CVaR}[Z(s, \epsilon)]$$



Use the risk function to **design a safer perception system**.

➤ Risk-driven **training**

➤ Risk-driven **data generation**



**Our Idea:** Account for the “**riskiness**” of making perception errors during the perception system **design process**.

- 1 Formulate a **risk function**.  $\rho(s, \epsilon) = \text{CVaR}[Z(s, \epsilon)]$
- 2 Use the risk function to **design a safer perception system**.

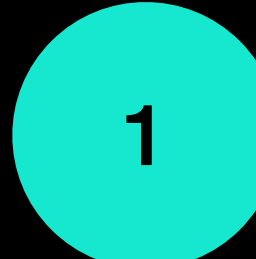
✓ Risk-driven **training**

$$\mathcal{L}(s, \epsilon) = \frac{1}{n} \sum \epsilon^2 +$$

➤ Risk-driven **data generation**

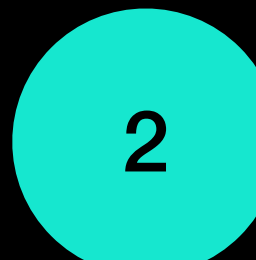


**Our Idea:** Account for the “**riskiness**” of making perception errors during the perception system **design process**.



Formulate a **risk function**.

$$\rho(s, \epsilon) = \text{CVaR}[Z(s, \epsilon)]$$



Use the risk function to **design a safer perception system**.

➤ Risk-driven **training**

➤ Risk-driven **data generation**



**Our Idea:** Account for the “**riskiness**” of making perception errors during the perception system **design process**.

- 1 Formulate a **risk function**.  $\rho(s, \epsilon) = \text{CVaR}[Z(s, \epsilon)]$
- 2 Use the risk function to **design a safer perception system**.

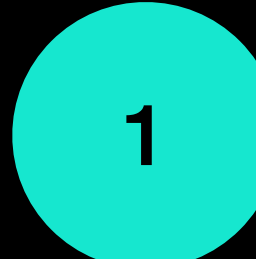
➤ Risk-driven **training**

✓ Risk-driven **data generation**

$$w(s) = \max_{\epsilon} \rho(s, \epsilon) - \rho(s, 0)$$

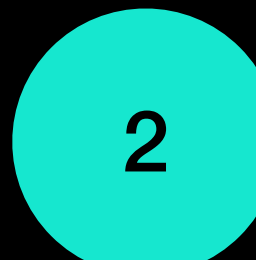


**Our Idea:** Account for the “**riskiness**” of making perception errors during the perception system **design process**.



Formulate a **risk function**.

$$\rho(s, \epsilon) = \text{CVaR}[Z(s, \epsilon)]$$



Use the risk function to **design a safer perception system**.

➤ Risk-driven **training**

➤ Risk-driven **data generation**

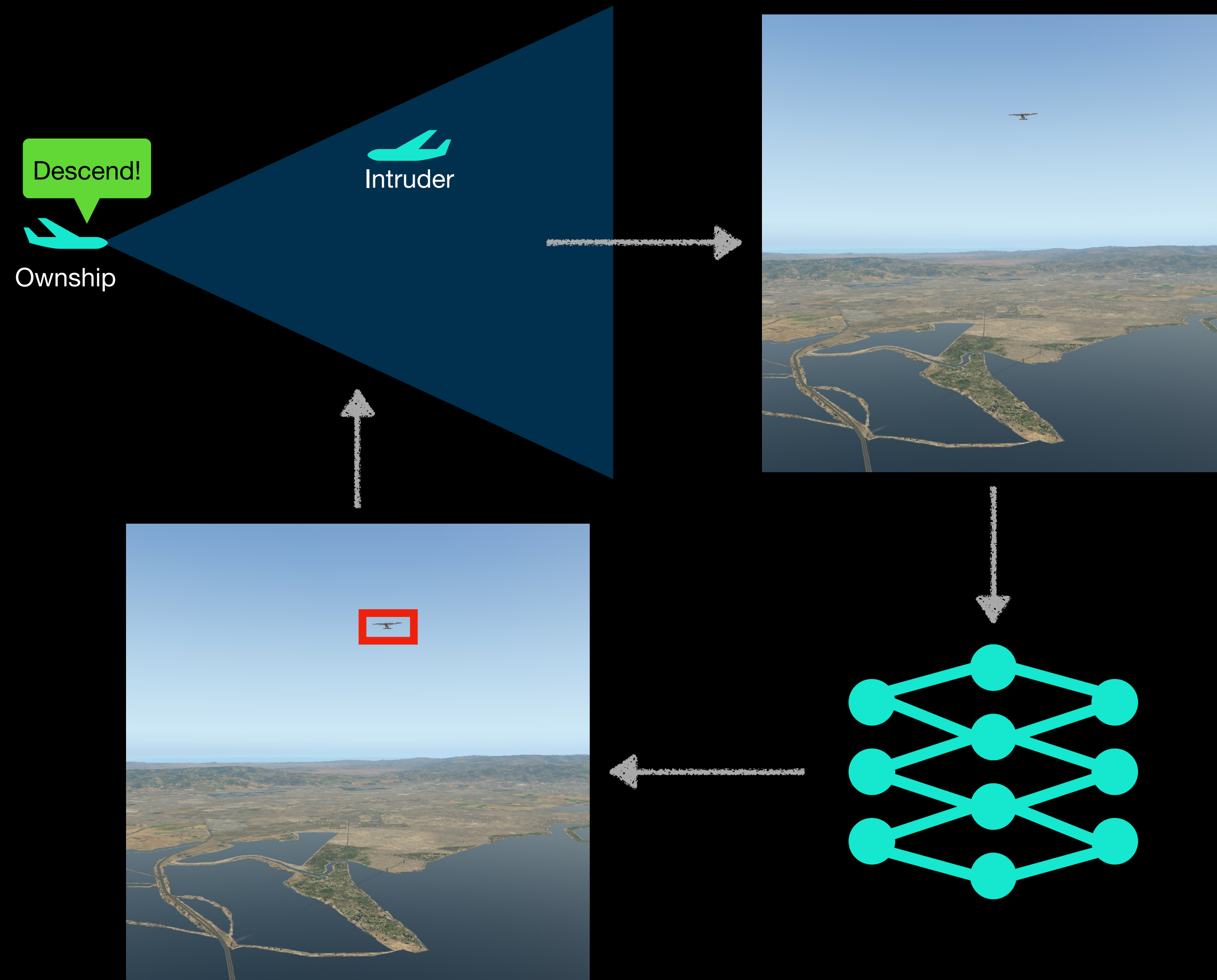
# Vision-based Aircraft Detect and Avoid Example

Safety Requirements

Risk-Driven Design

Image-Based Verification

Probabilistic Guarantees



# Vision-based Aircraft Detect and Avoid Example

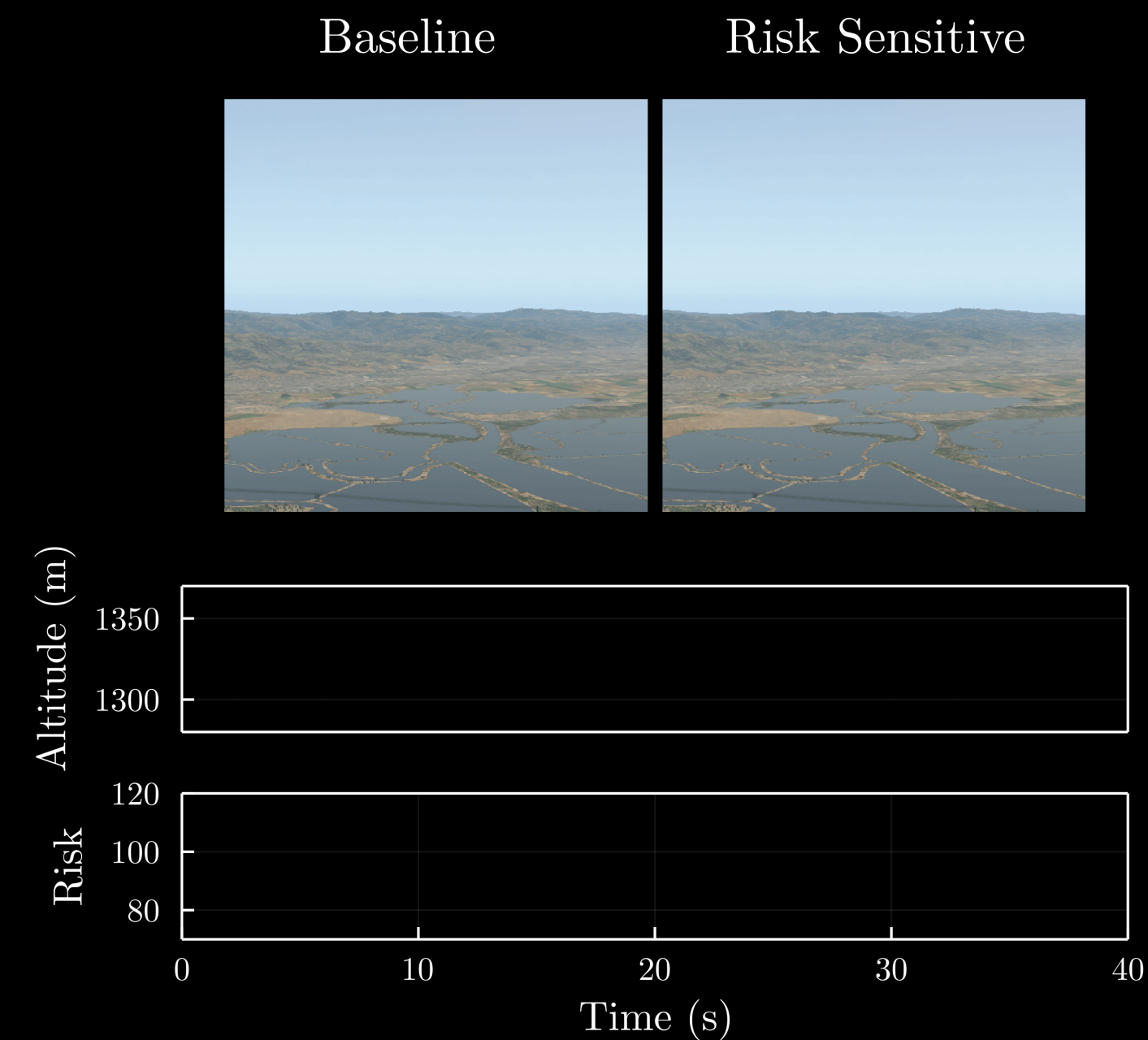
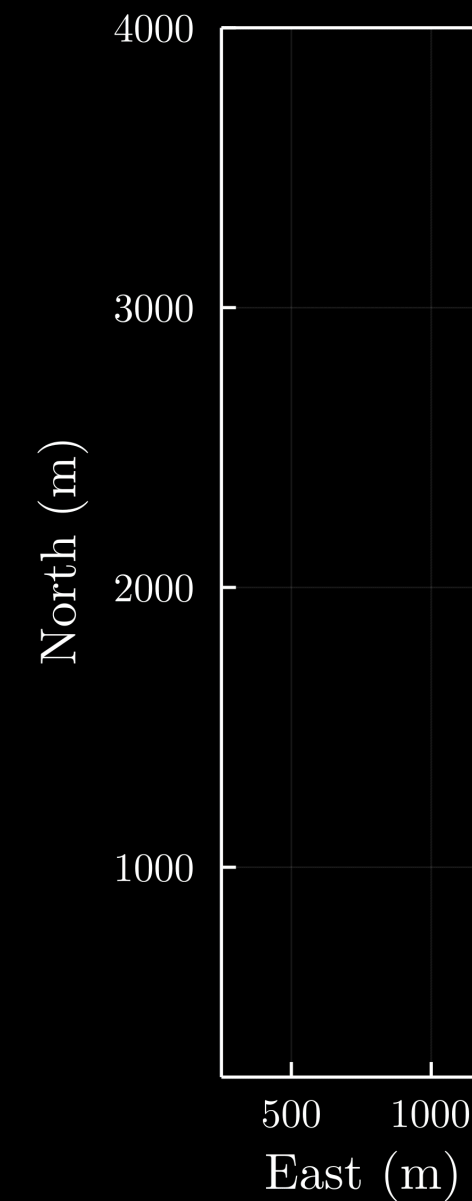
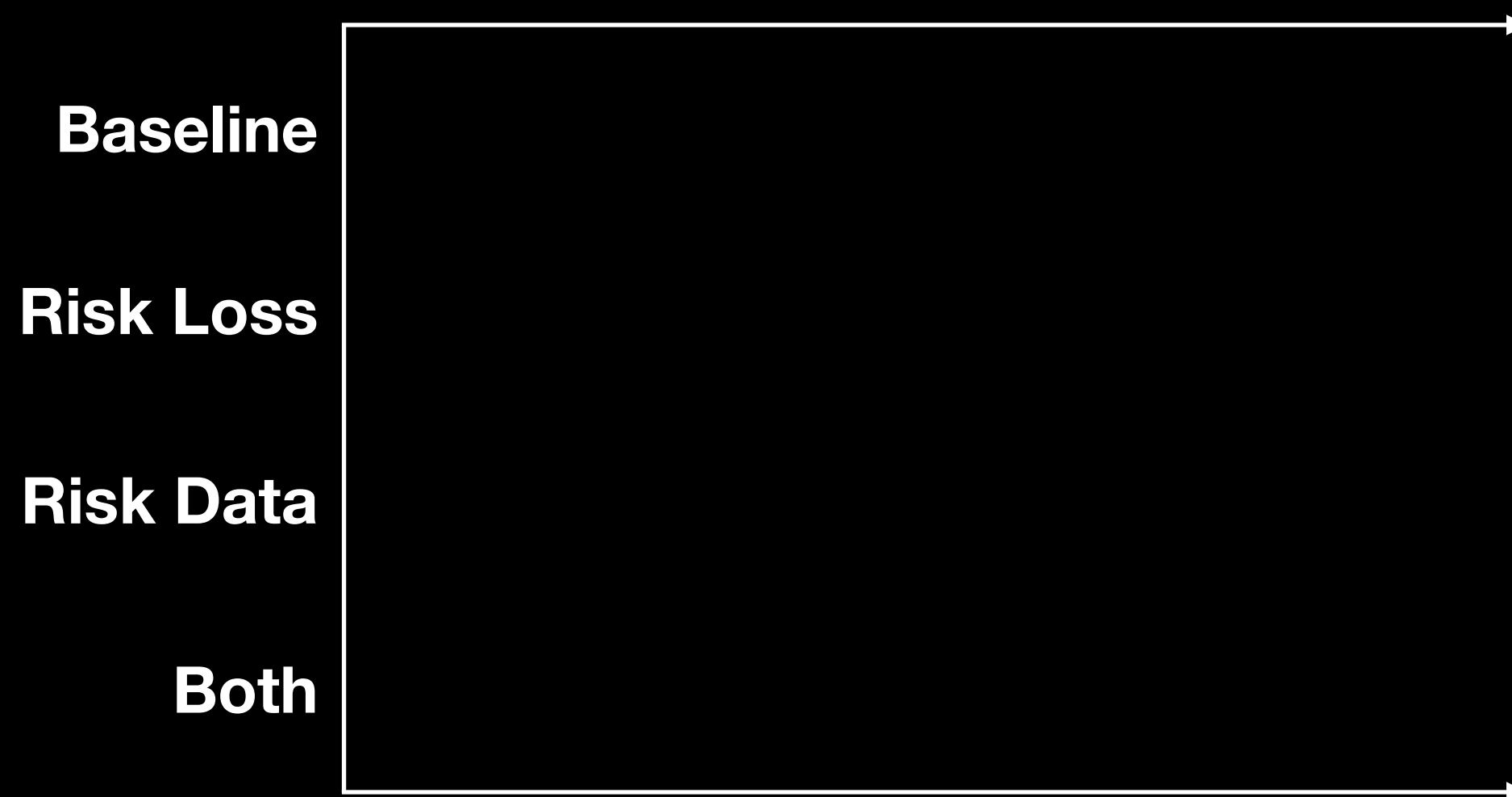
Safety Requirements

Risk-Driven Design

Image-Based Verification

Probabilistic Guarantees

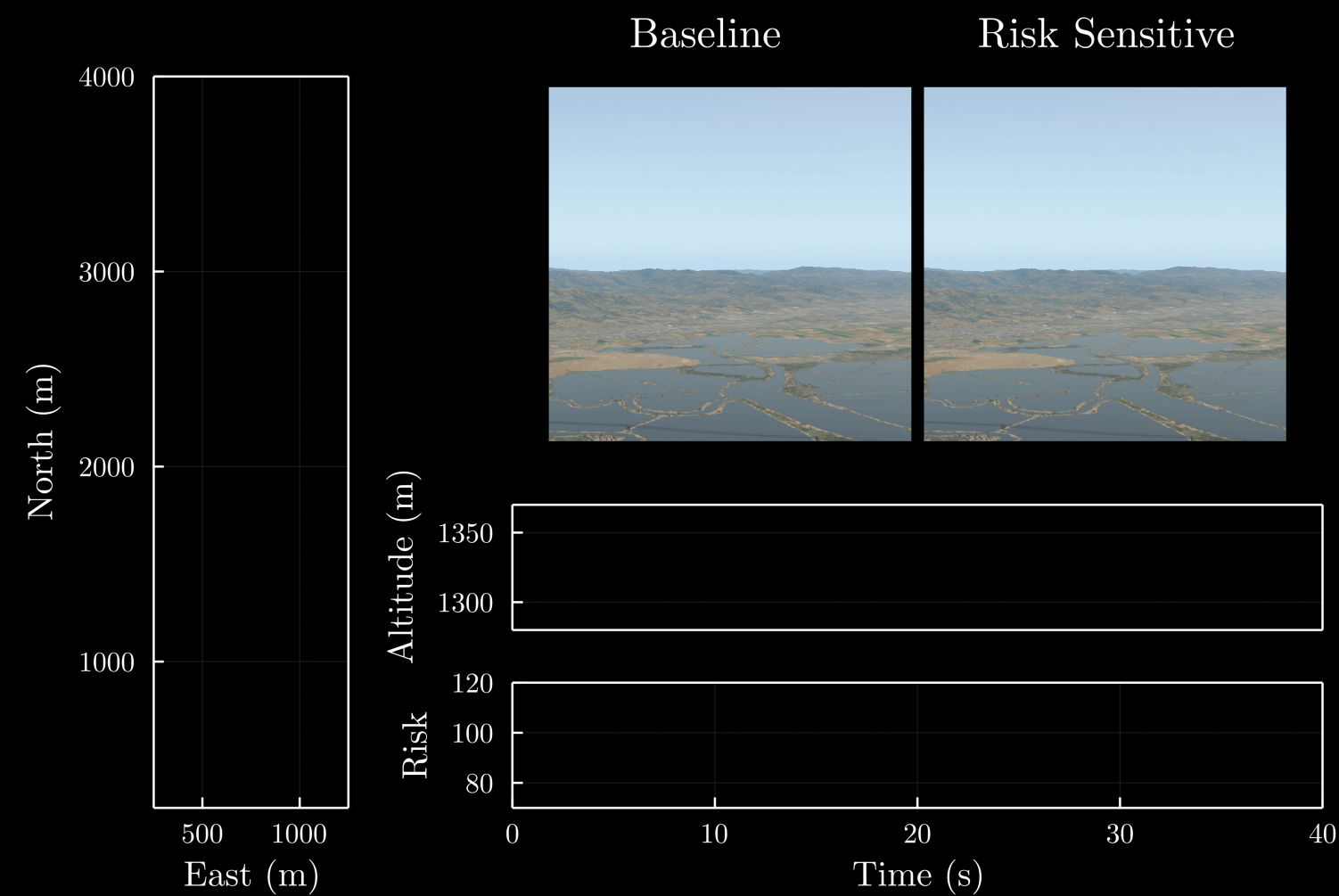
Number of Near Mid-Air Collisions

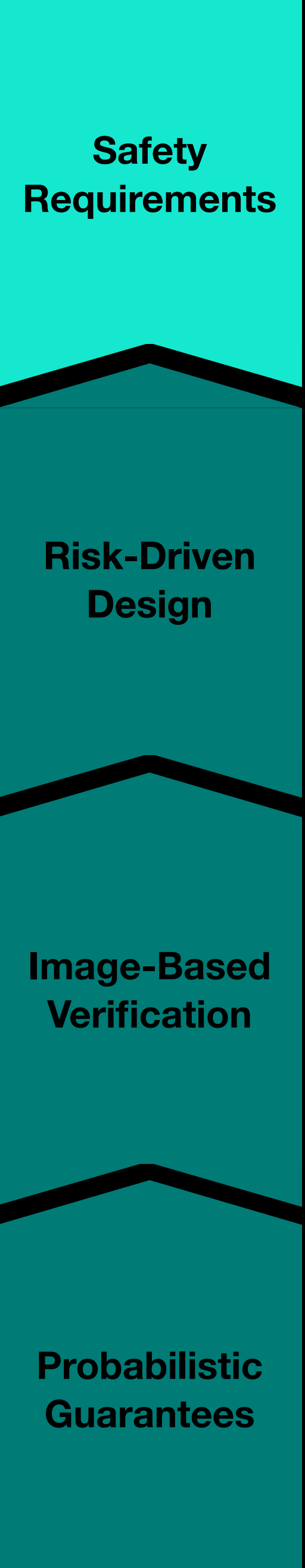


Up to 37% reduction in collision risk!

# Risk-Driven Design of Perception Systems

- 1 Formulate a **risk function**.  $\rho(s, \epsilon) = \text{CVaR}[Z(s, \epsilon)]$
- 2 Use the risk function to **design a safer perception system**.



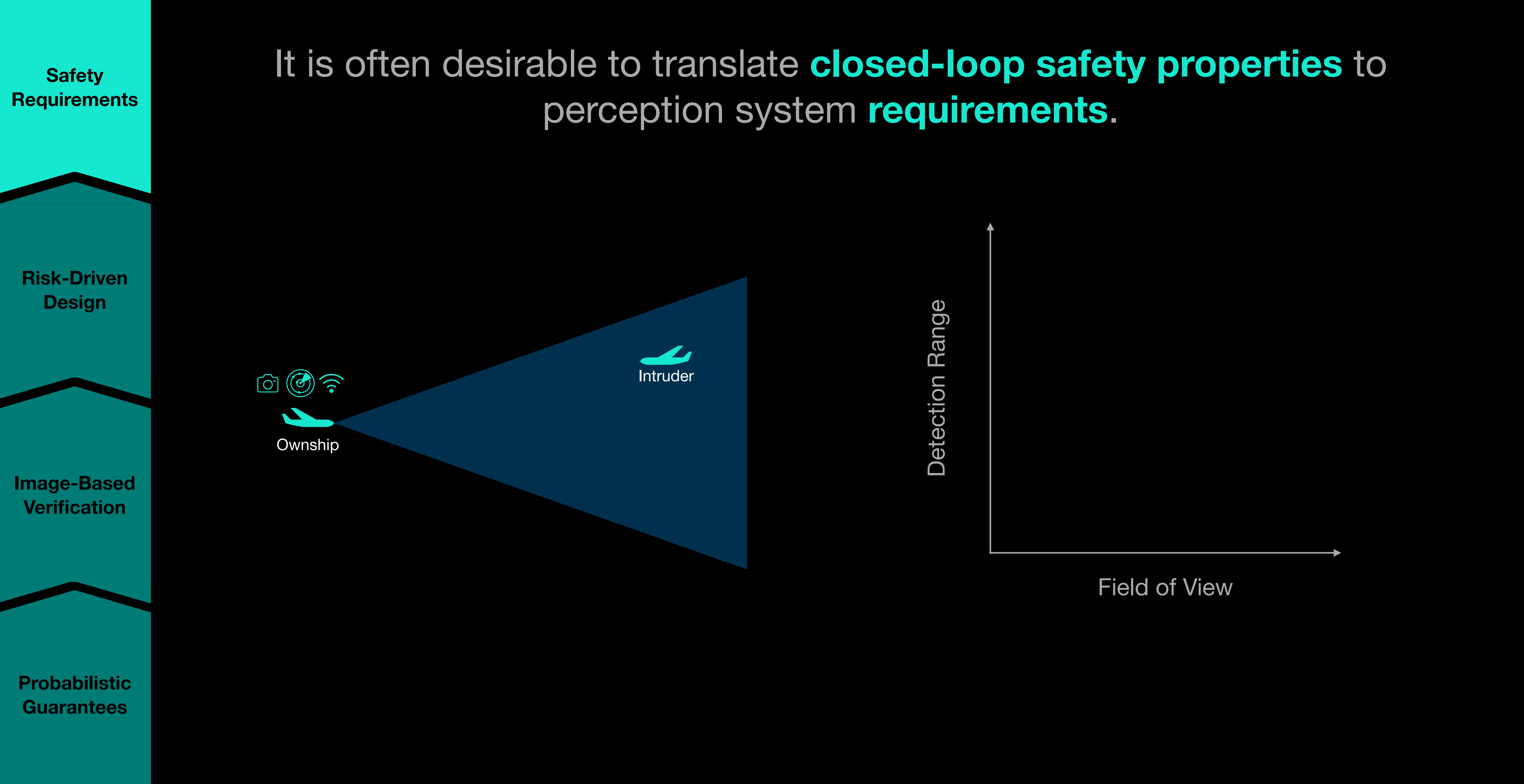


# Efficient Determination of Safety Requirements for Perception Systems

S. M. Katz, A. L. Corso, E. Yel, and M. J. Kochenderfer

Digital Avionics Systems Conference (2023)

It is often desirable to translate **closed-loop safety properties** to perception system **requirements**.



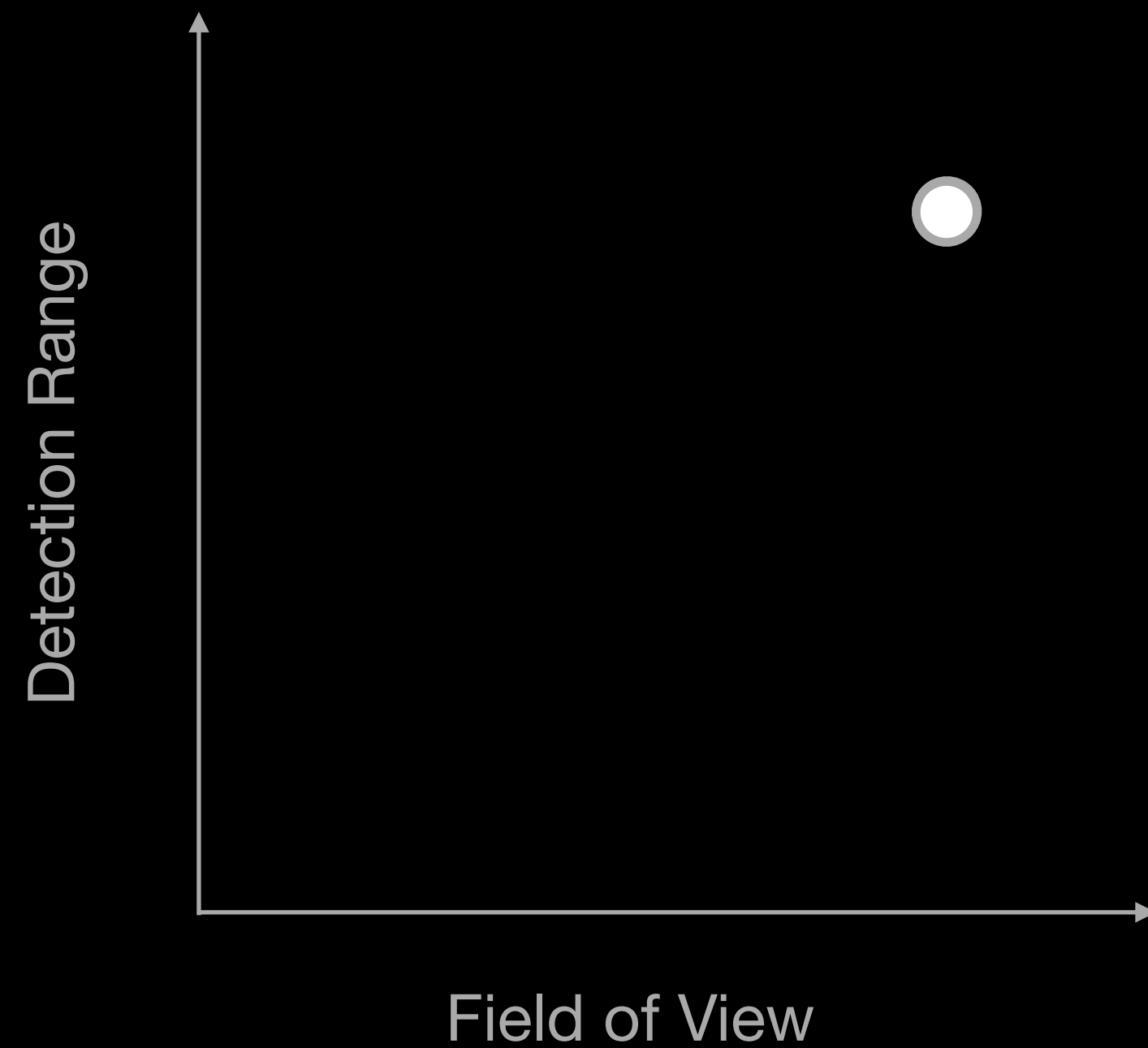
It is often desirable to translate **closed-loop safety properties** to perception system **requirements**.

Safety Requirements

Risk-Driven Design

Image-Based Verification

Probabilistic Guarantees



Monte Carlo Simulations

Black Box Simulator

$P_{fail}$



$\stackrel{?}{<} \gamma$

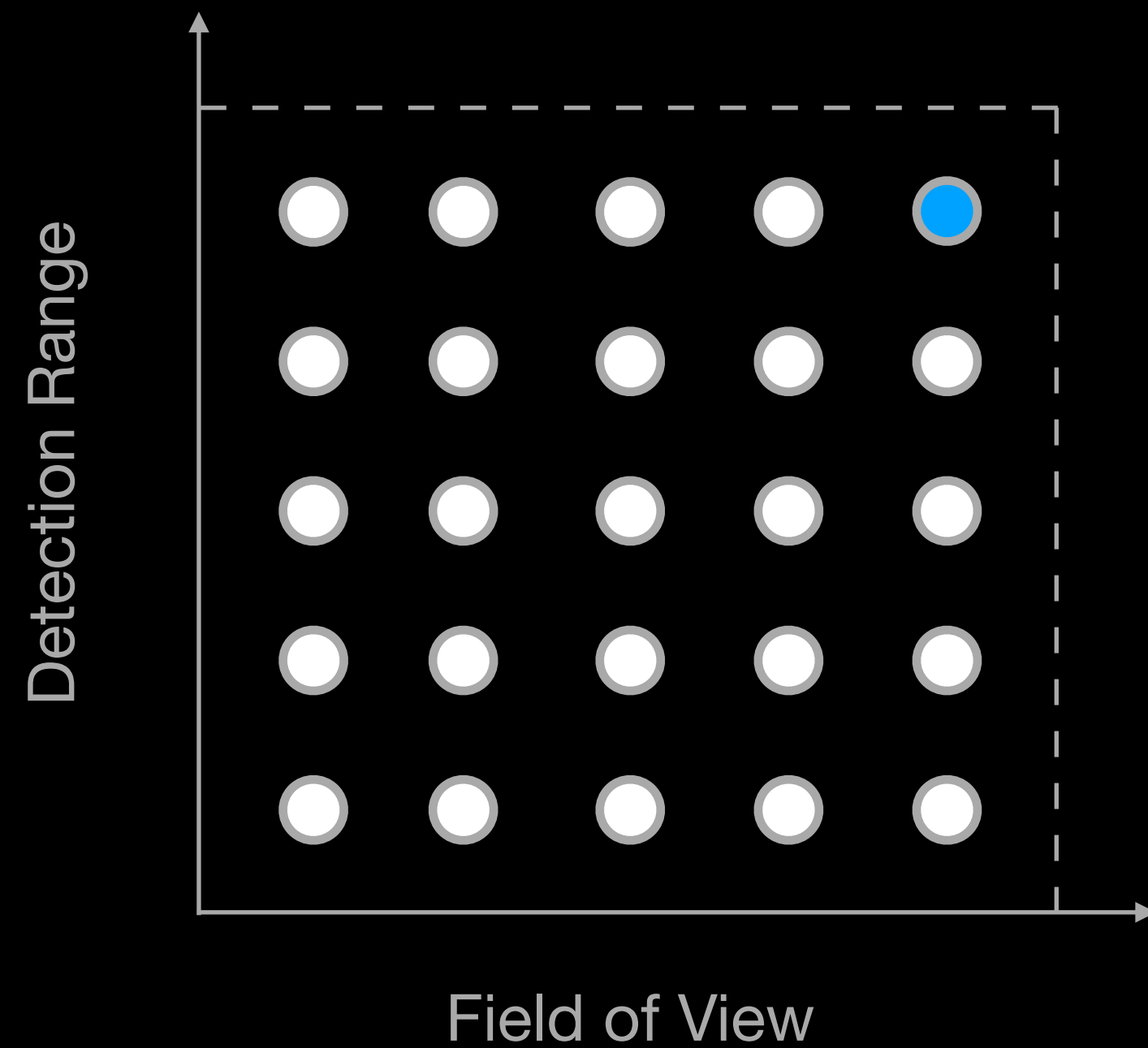
It is often desirable to translate **closed-loop safety properties** to perception system **requirements**.

Safety Requirements

Risk-Driven Design

Image-Based Verification

Probabilistic Guarantees



Monte Carlo Simulations

Black Box Simulator

$P_{fail}$



$\stackrel{?}{<} \gamma$

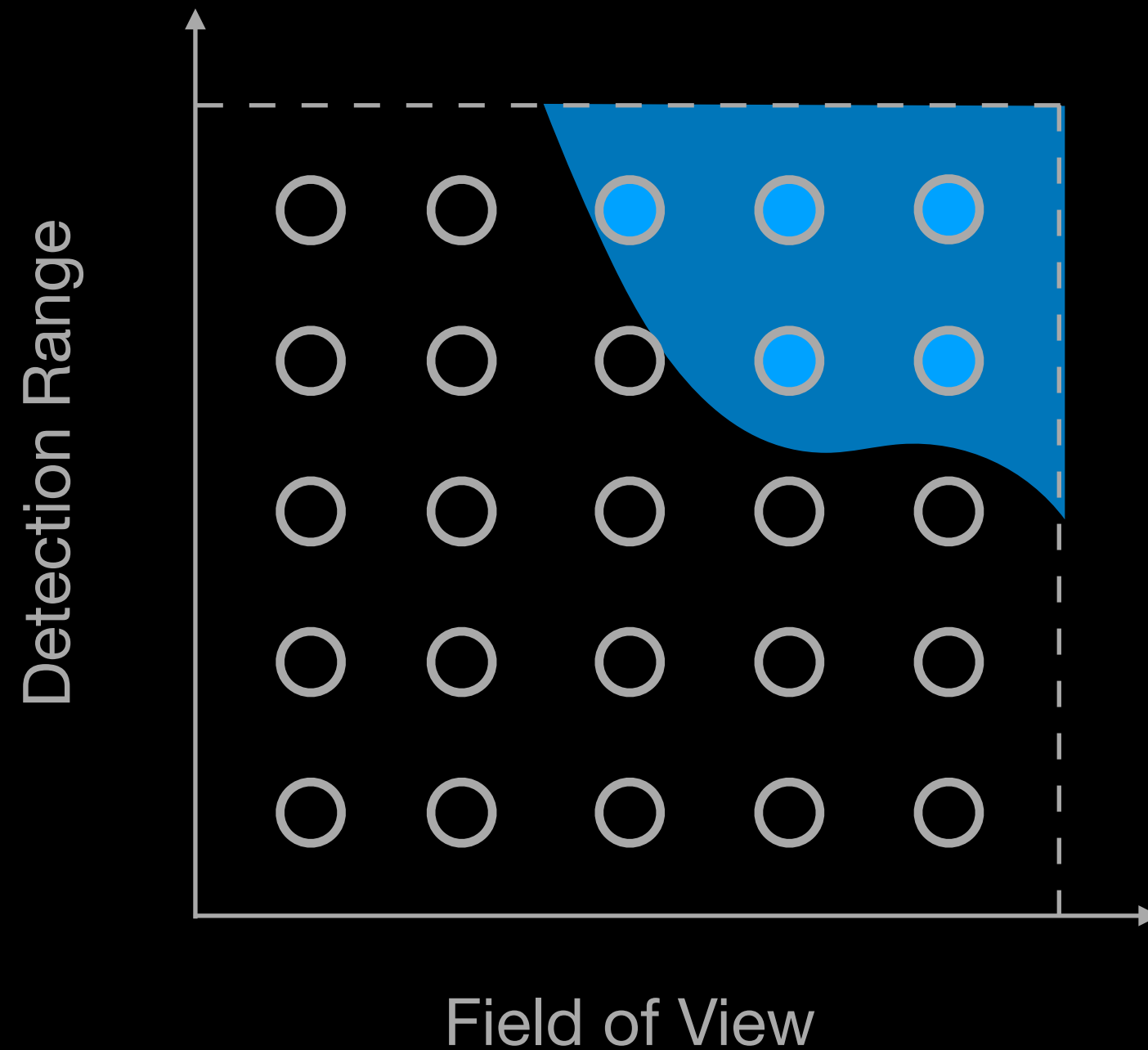
It is often desirable to translate **closed-loop safety properties** to perception system **requirements**.

Safety Requirements

Risk-Driven Design

Image-Based Verification

Probabilistic Guarantees



Monte Carlo Simulations

Black Box Simulator

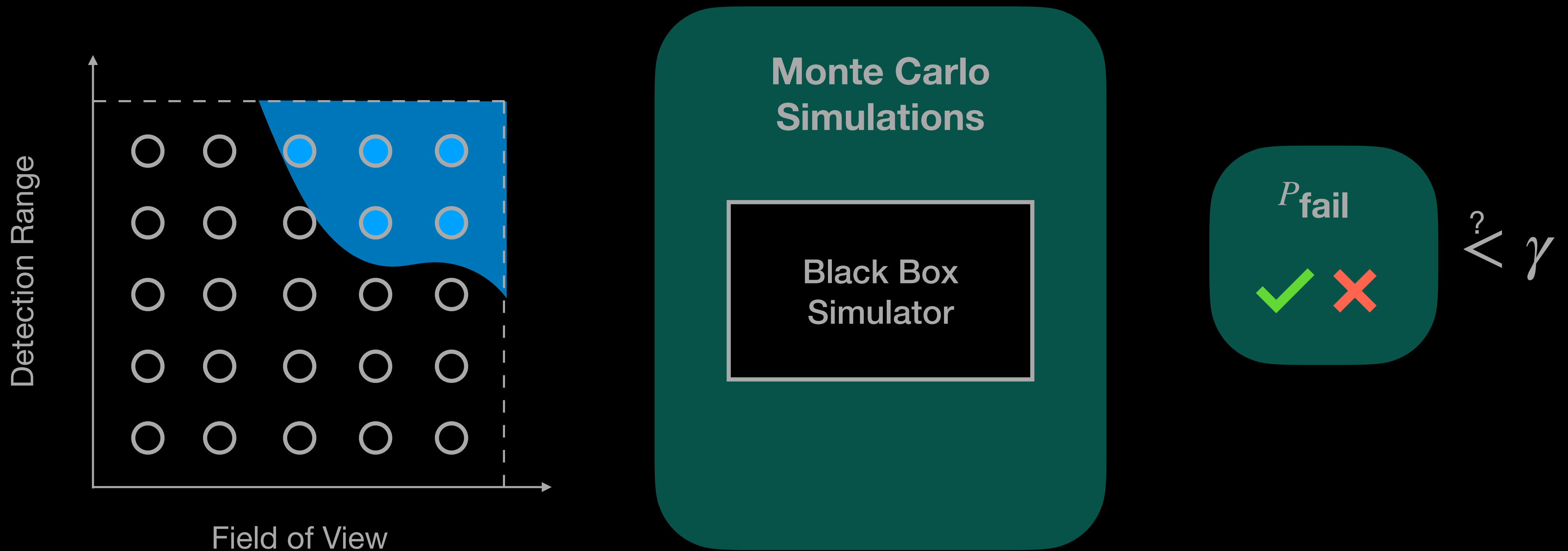
$P_{fail}$



$\stackrel{?}{<} \gamma$

**Goal:** determine sets of safe performance characteristics.

It is often desirable to translate **closed-loop safety properties** to perception system **requirements**.



**Goal:** efficiently determine sets of safe performance characteristics.

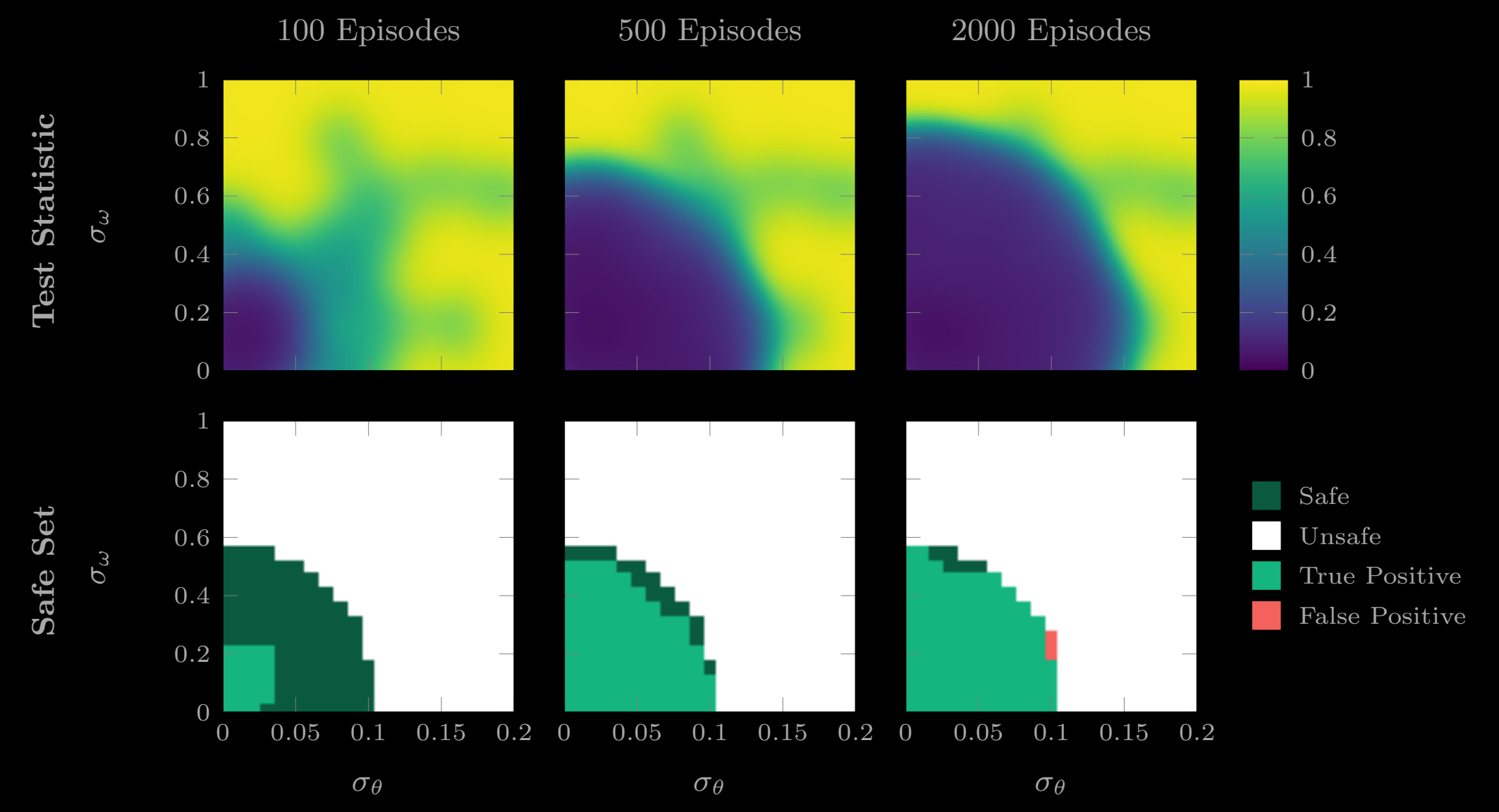
**Goal:** efficiently determine sets of safe performance characteristics.

1 Account for **spatial correlation**.

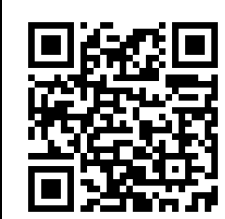
> Level-set estimation using Gaussian processes

2 Account for **multi-level nature** of the estimation problem.

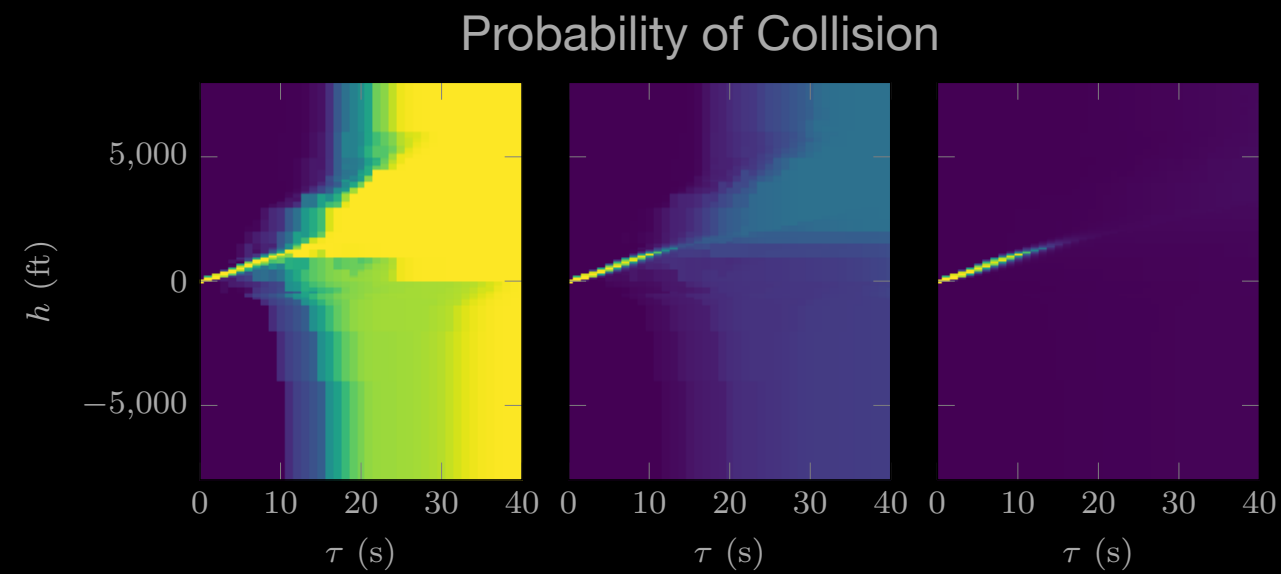
> Threshold bandits



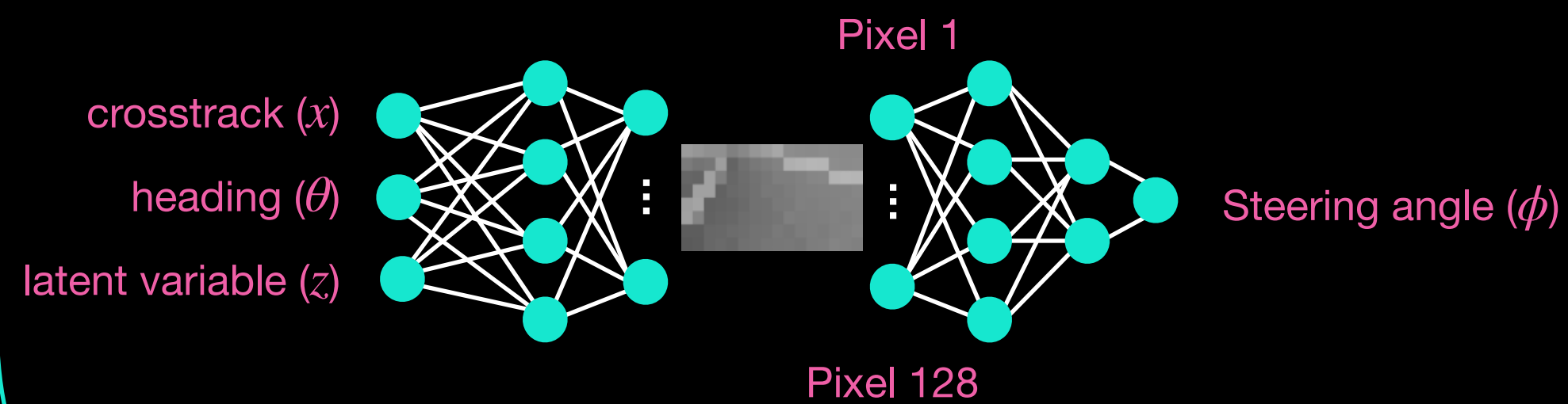
# Safety Verification



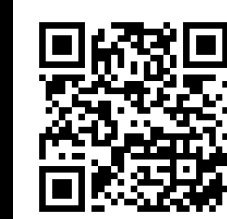
Generating Probabilistic Safety Guarantees for Neural Network Controllers



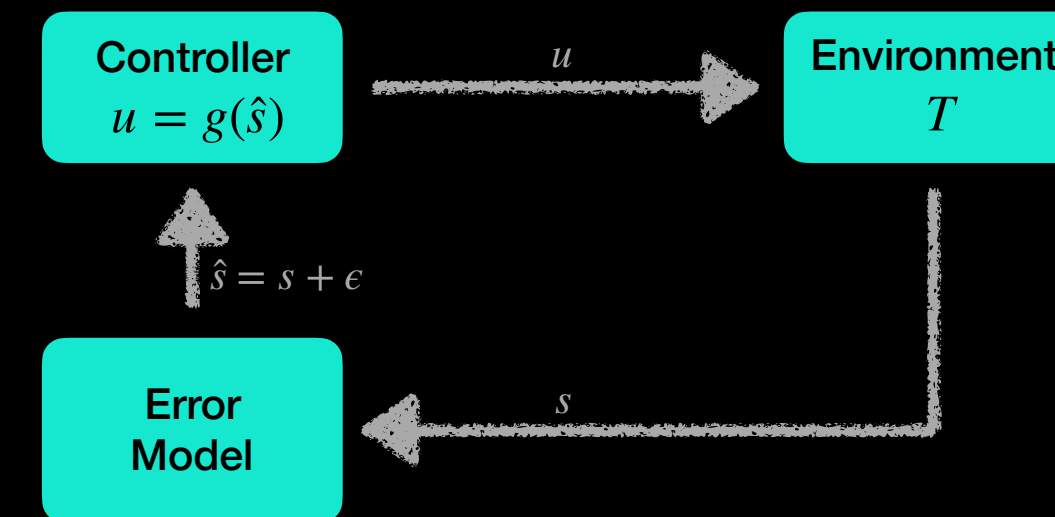
Verification of Image-Based Neural Network Controllers Using Generative Models



# Safe Design

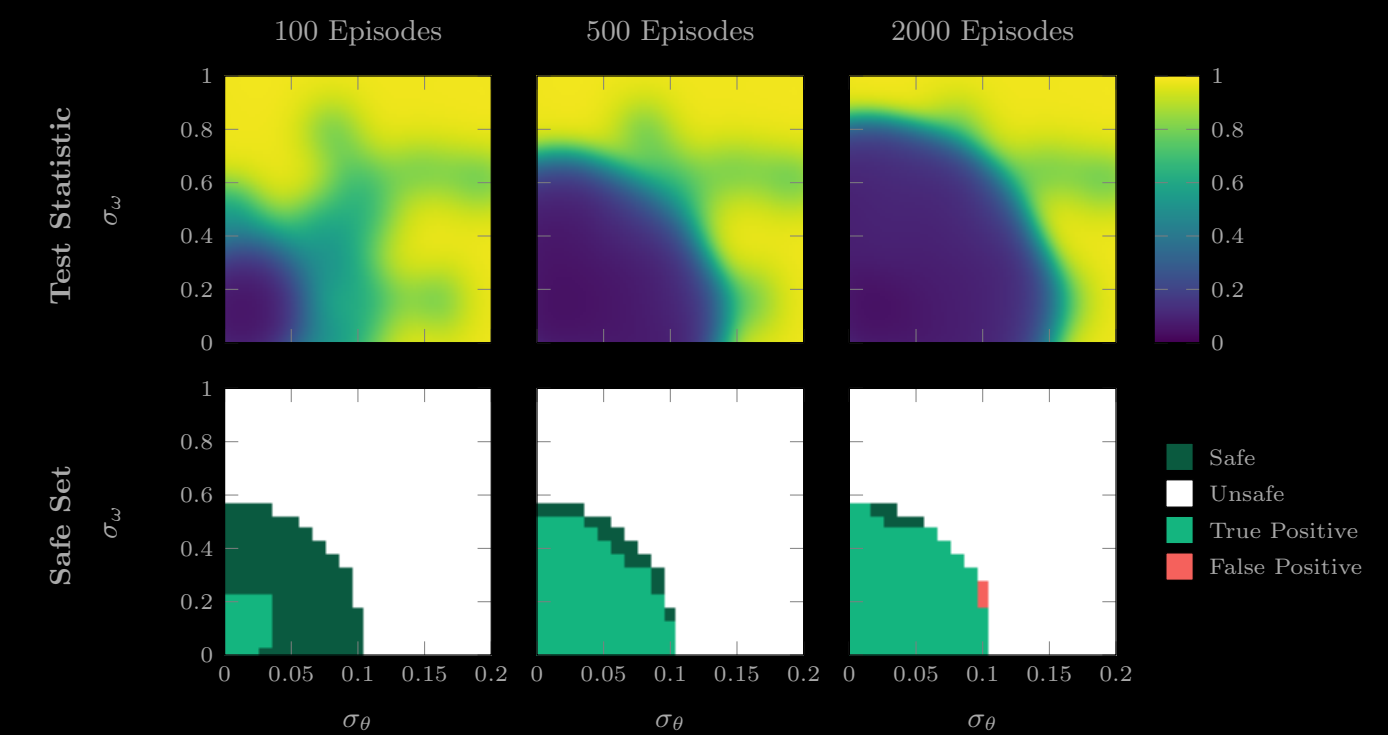


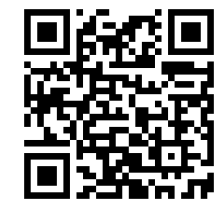
Risk-Driven Design of Perception Systems



WIP

Efficient Determination of Safety Requirements for Perception Systems





Generating Probabilistic Safety Guarantees for Neural Network Controllers



Risk-Driven Design of Perception Systems



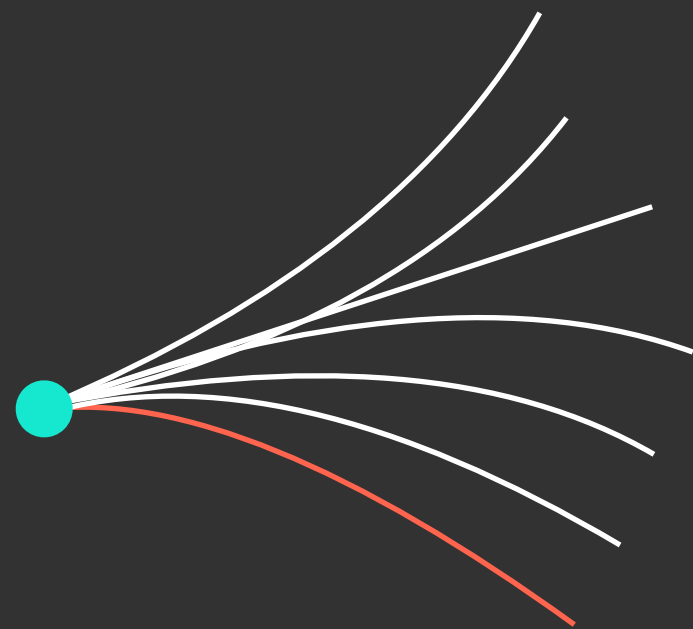
Verification of Image-Based Neural Network Controllers Using Generative Models



Efficient Determination of Safety Requirements for Perception Systems

## Future Work

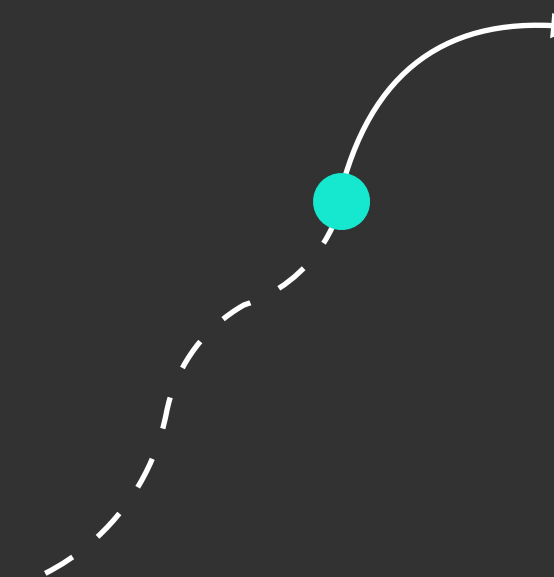
Falsification

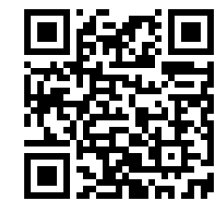


Runtime Monitoring



History Dependence





Generating Probabilistic Safety Guarantees for Neural Network Controllers



Risk-Driven Design of Perception Systems



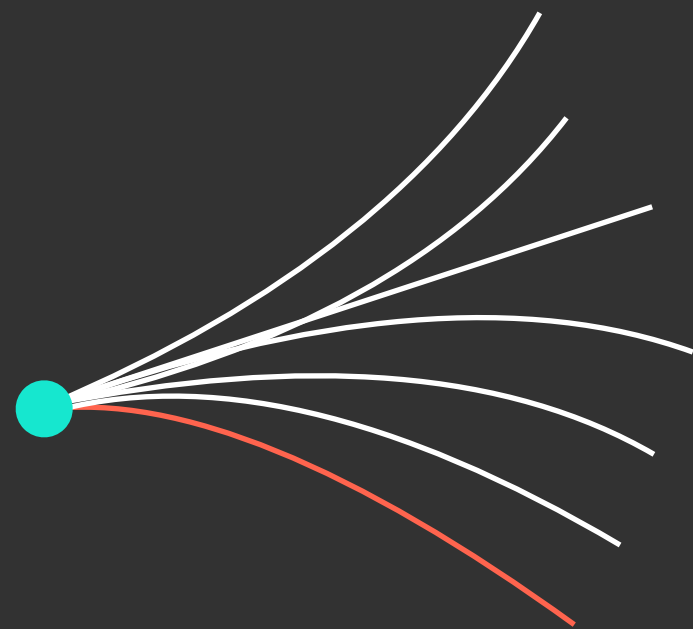
Verification of Image-Based Neural Network Controllers Using Generative Models

WIP

Efficient Determination of Safety Requirements for Perception Systems

## Future Work

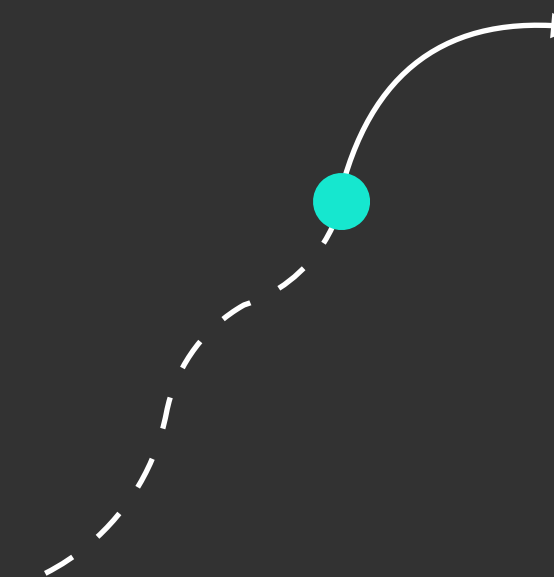
Falsification



Runtime Monitoring



History Dependence



# Publications

N. Rober, **S. M. Katz**, C. Sidrane, E. Yel, M. Everett, M. J. Kochenderfer, and J. P. How, “Backward reachability analysis of neural feedback loops: Techniques for linear and nonlinear systems,” *IEEE Open Journal of Control Systems*, 2023

A. L. Corso\*, **S. M. Katz\***, C. A. Innes, X. Du, S. Ramamoorthy, and M. J. Kochenderfer, “Risk-driven design of perception systems,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022

**S. M. Katz\***, A. L. Corso\*, C. A. Strong\*, and M. J. Kochenderfer, “Verification of image-based neural network controllers using generative models,” *Journal of Aerospace Information Systems*, 2022

**S. M. Katz**, L. E. Alvarez, M. Owen, S. Wu, M. W. Brittain, A. Das, and M. J. Kochenderfer, “Collision risk and operational impact of x change advisories as aircraft collision avoidance maneuvers,” in *AIAA AVIATION Forum*, 2022

C. A. Strong, **S. M. Katz**, A. L. Corso, and M. J. Kochenderfer, “ZoPE: A fast optimizer for ReLU networks with low-dimensional inputs,” in *NASA Formal Methods Symposium (NFM)*, 2022

C. Sidrane, **S. M. Katz**, A. L. Corso, and M. J. Kochenderfer, “Verifying Inverse Model Neural Networks,” *ArXiv*, iss. 2202.02429, 2022

**S. M. Katz\***, A. L. Corso\*, C. A. Strong\*, and M. J. Kochenderfer, “Verification of image-based neural network controllers using generative models,” in *Digital Avionics Systems Conference (DASC)*, 2021

**S. M. Katz**, K. D. Julian, C. A. Strong, and M. J. Kochenderfer, “Generating probabilistic safety guarantees for neural network controllers,” *Machine Learning*, iss. 2103.01203, 2021

**S. M. Katz\***, A. Maleki\*, E. Biyik, and M. J. Kochenderfer, “Preference-based learning of reward function features,” *ArXiv*, iss. 2103.02727, 2021

**S. M. Katz**, A. LeBihan, and M. J. Kochenderfer, “Learning an urban air mobility encounter model from expert preferences,” in *Digital Avionics Systems Conference (DASC)*, 2019.