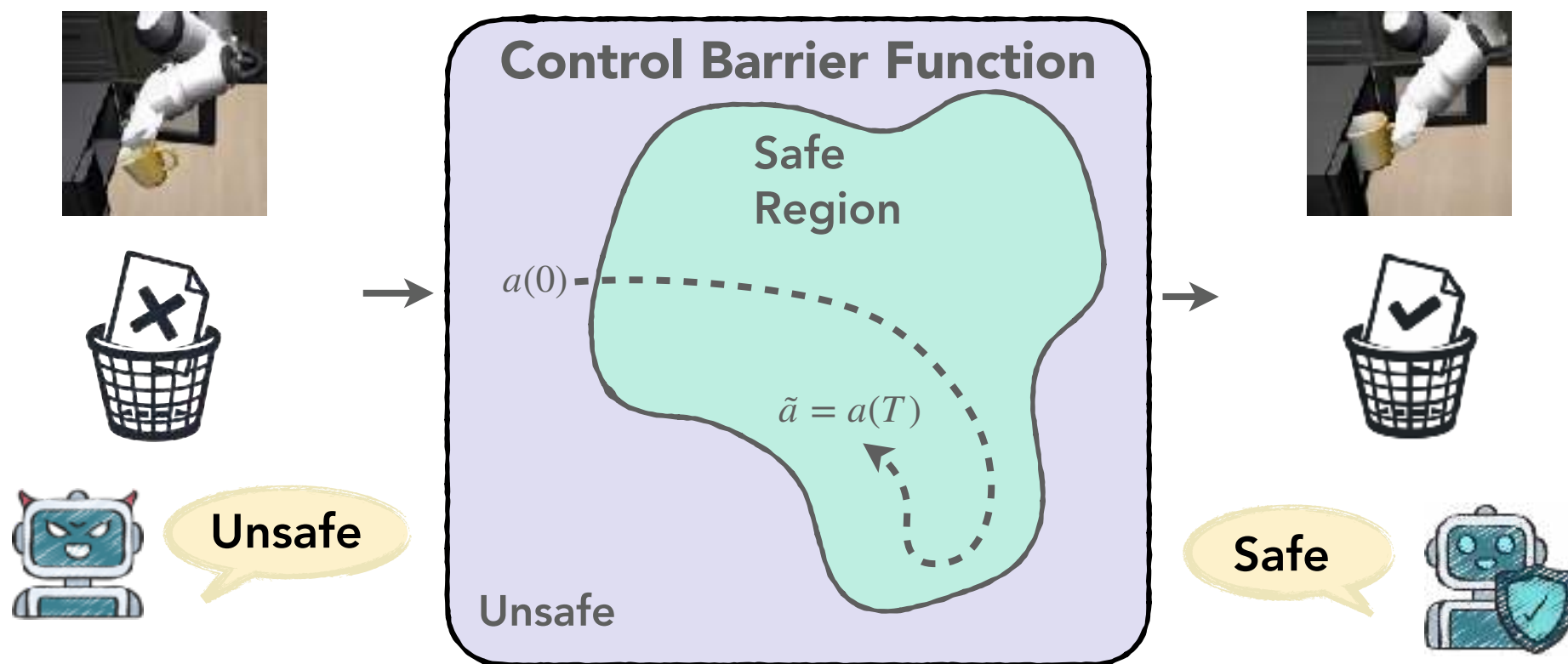


Diagnosing
Instability

Controlling
Behavior

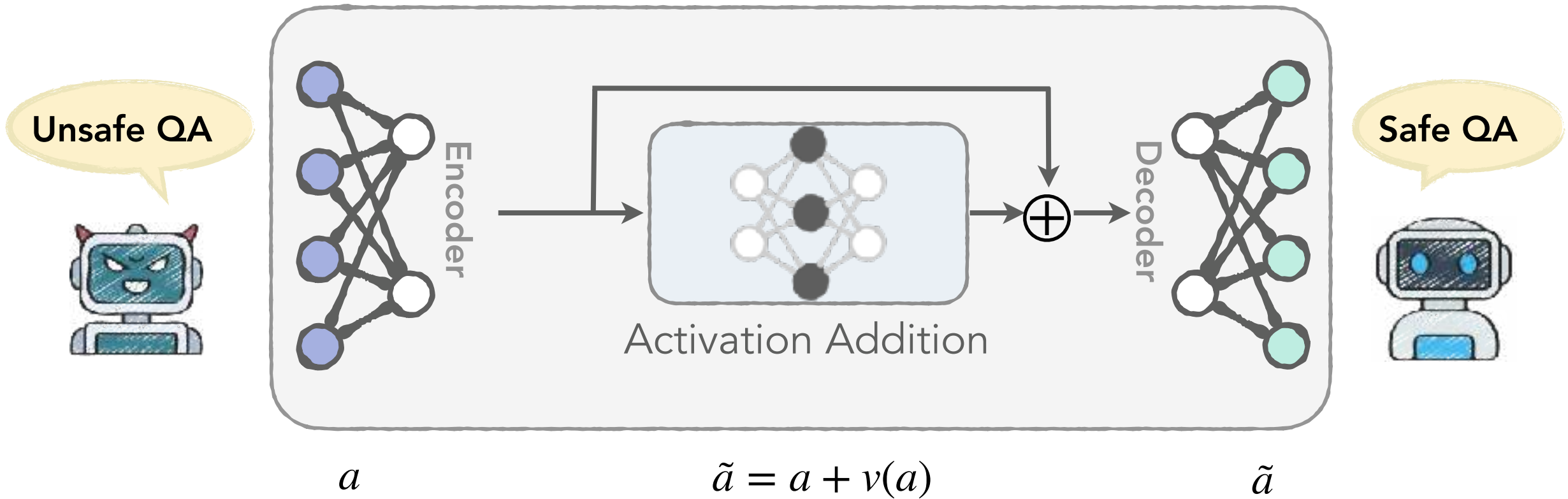
Challenge 2:

Static output → Agentic execution: Catastrophic Actions

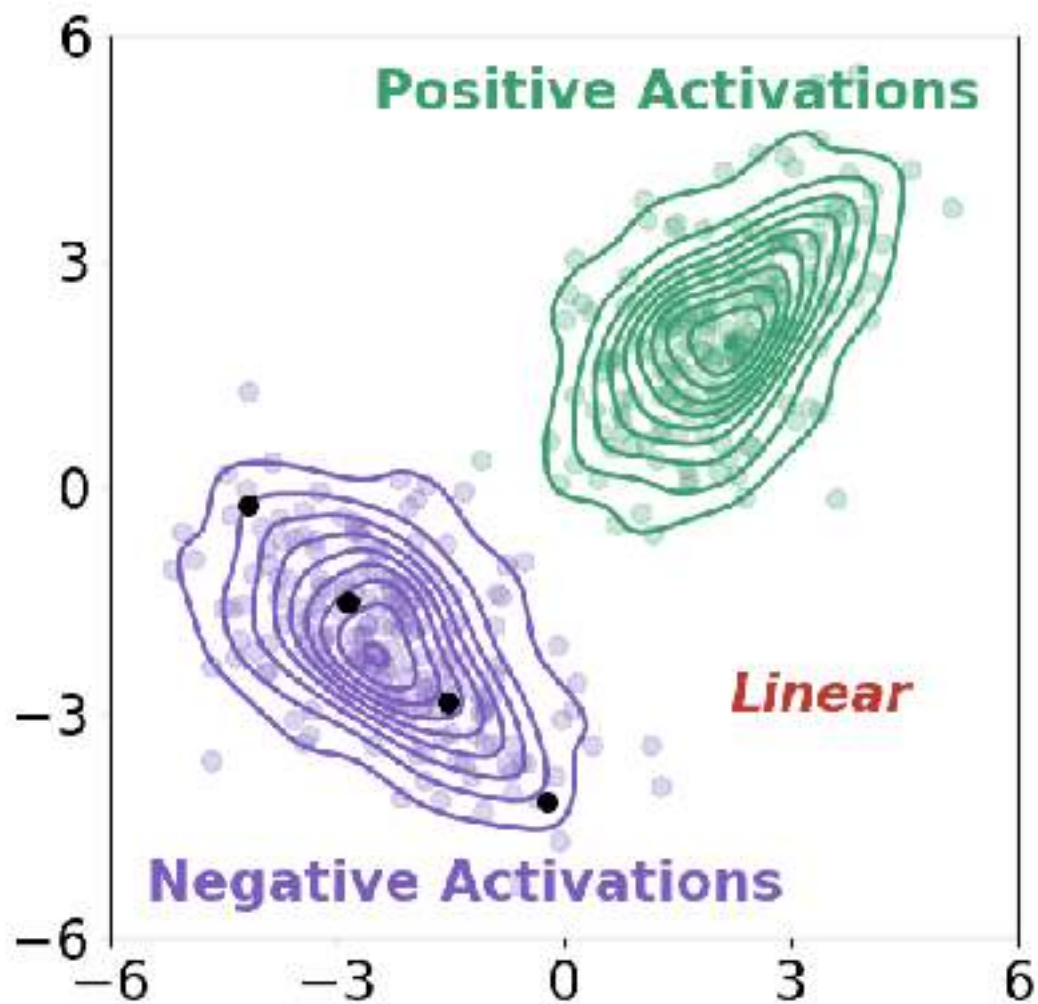


Safety Control via Steering towards Safe Region

Linear Steering

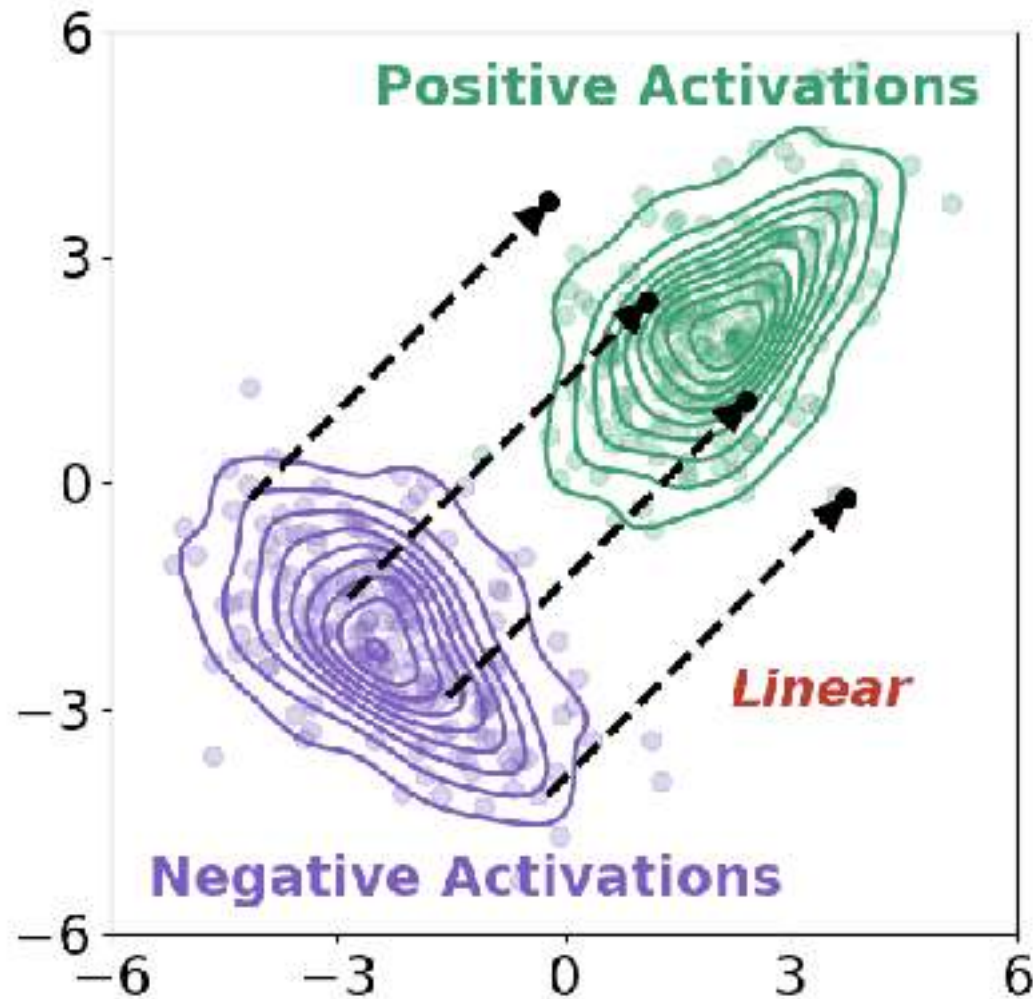


Linear Steering



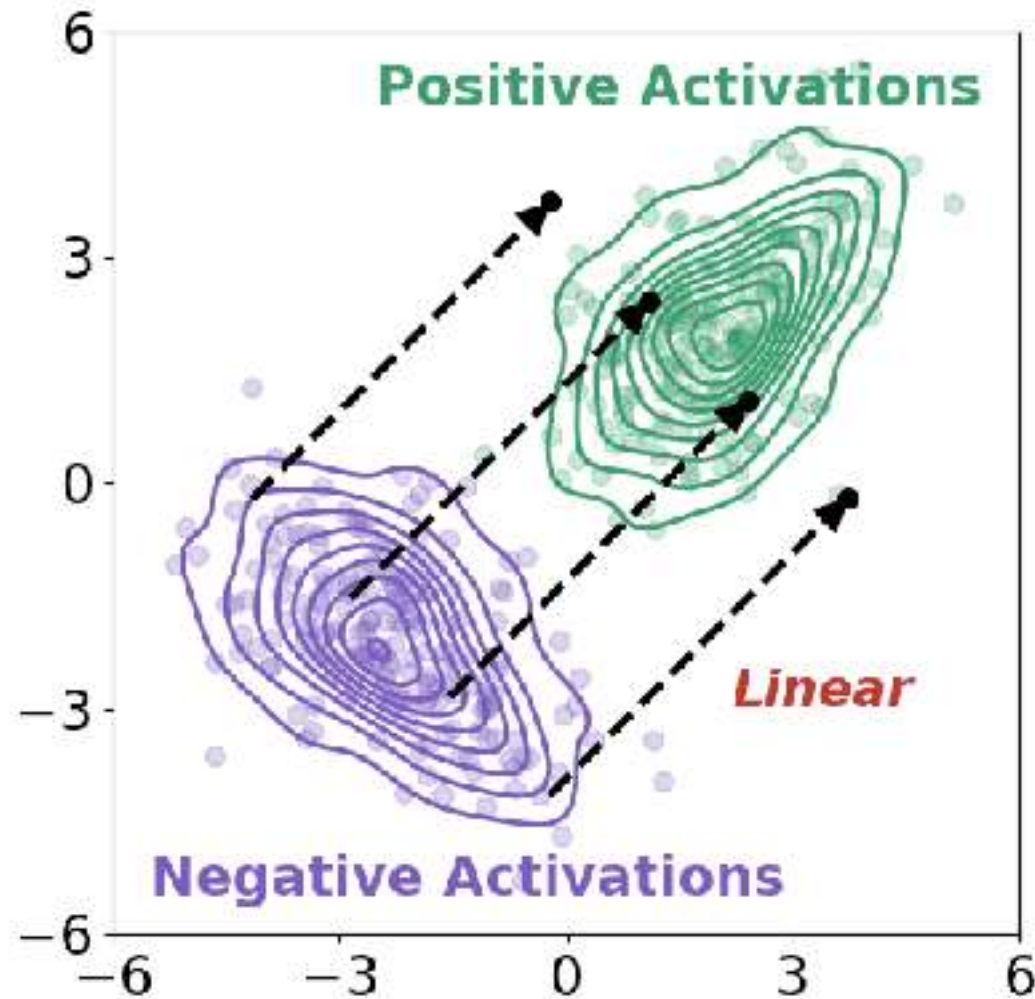
Linear Steering

ODE-Steering

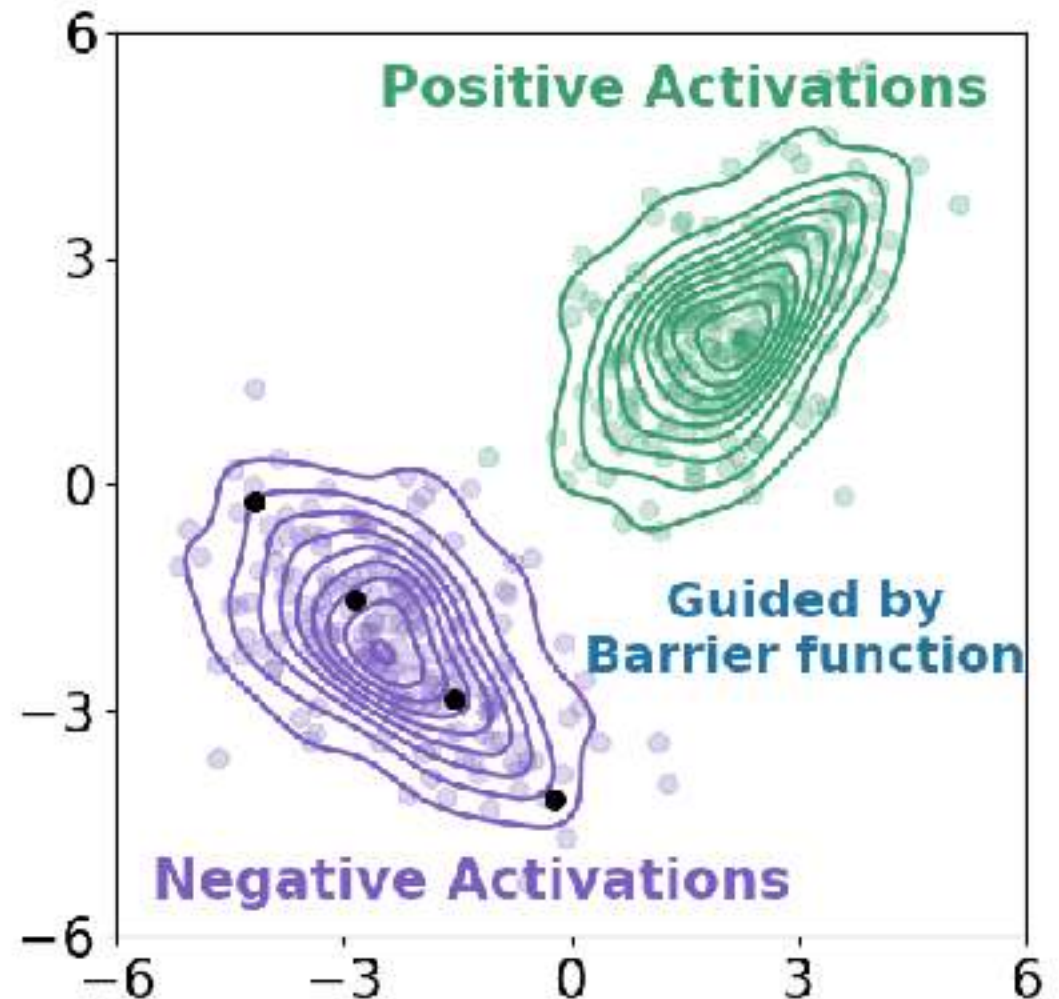


How to find a good steering direction?

Linear Steering

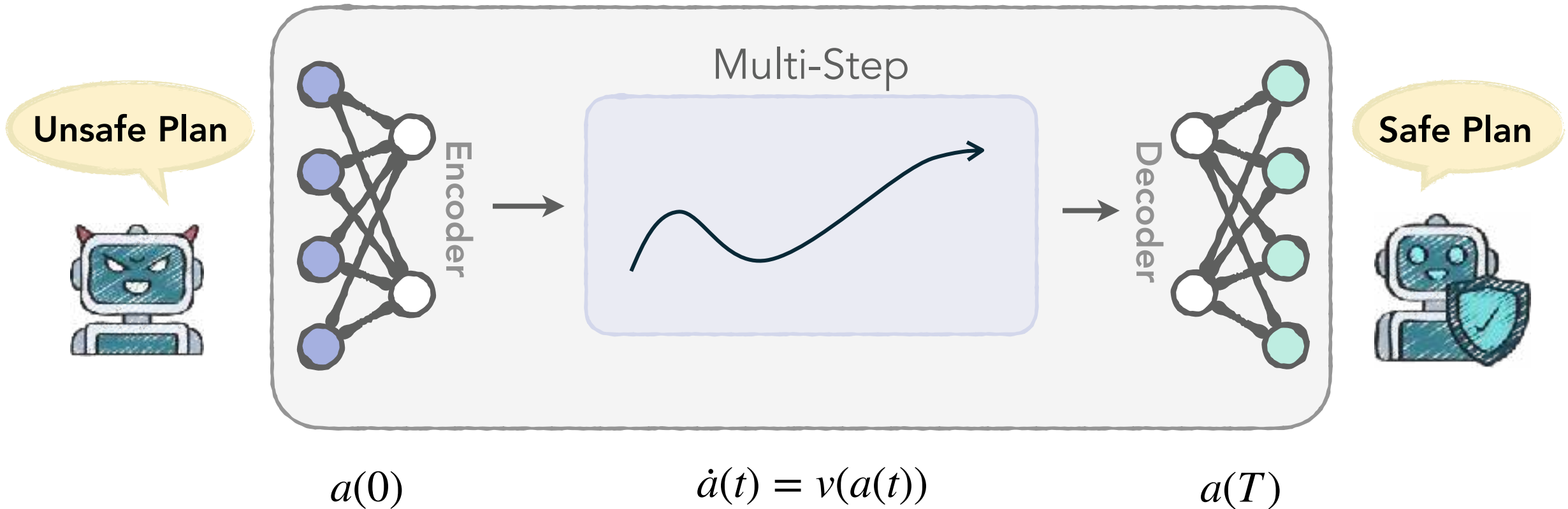


ODE-Steering



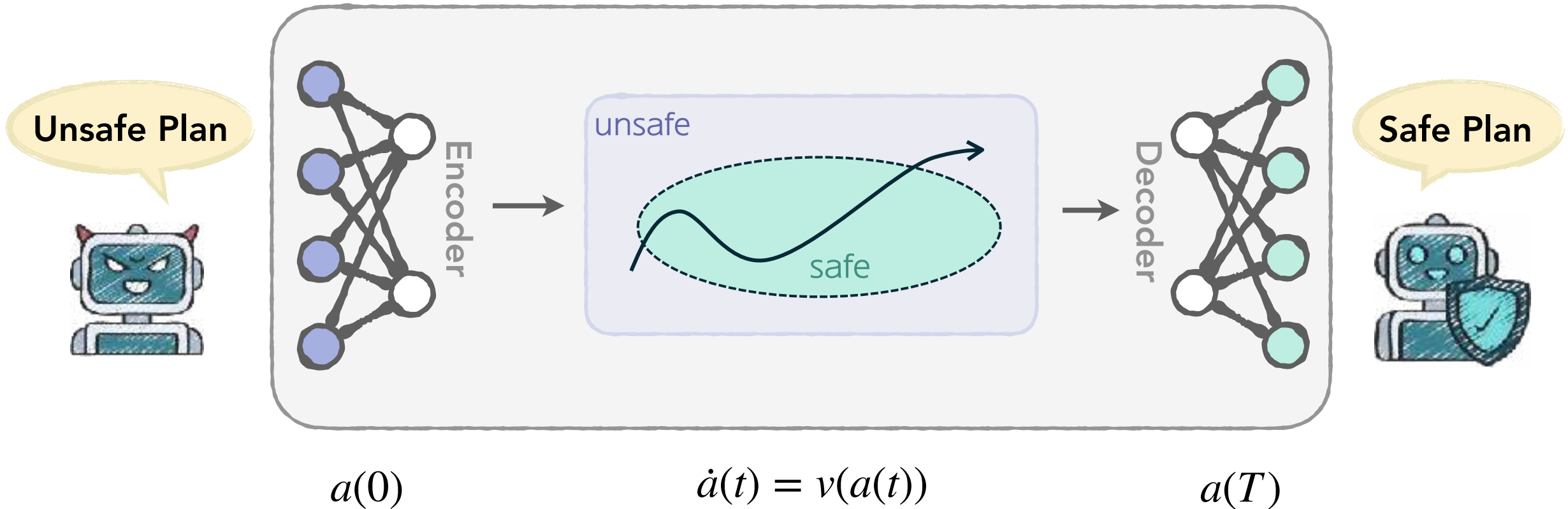
ODE-Steering

Modeling a as ODE



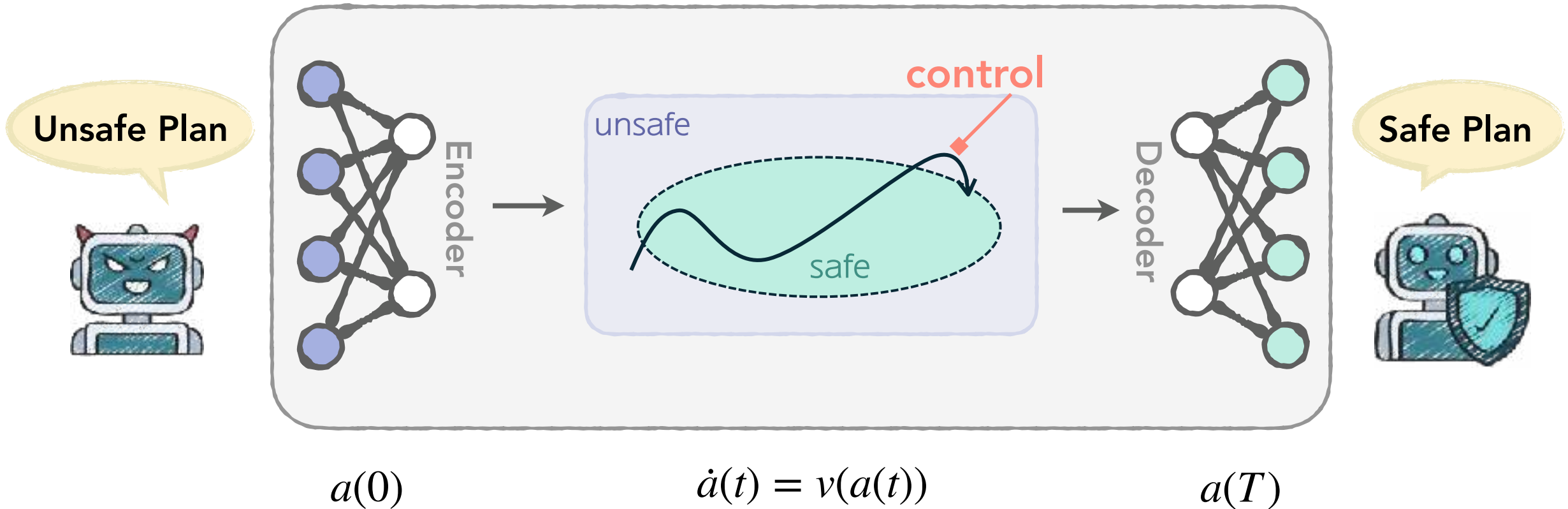
ODE-Steering: Control Barrier Function

If we can draw a **safe boundary** for a



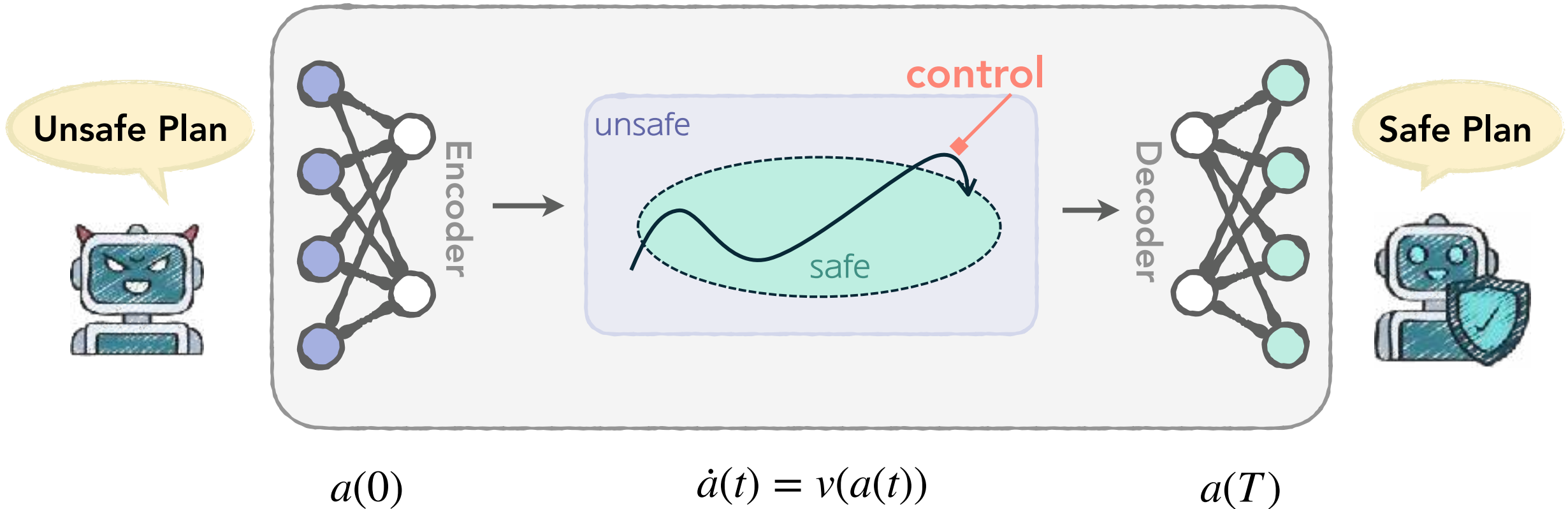
ODE-Steering: Control Barrier Function

If we can draw a **safe boundary** for a



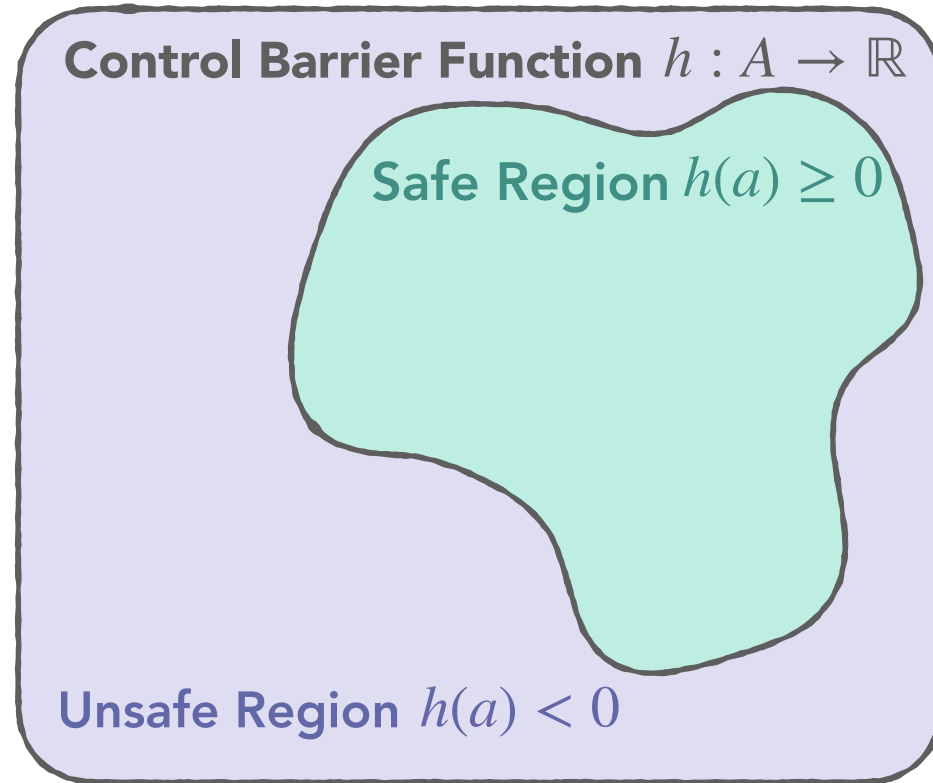
ODE-Steering: Control Barrier Function

How to learn the **safe boundary** h ?



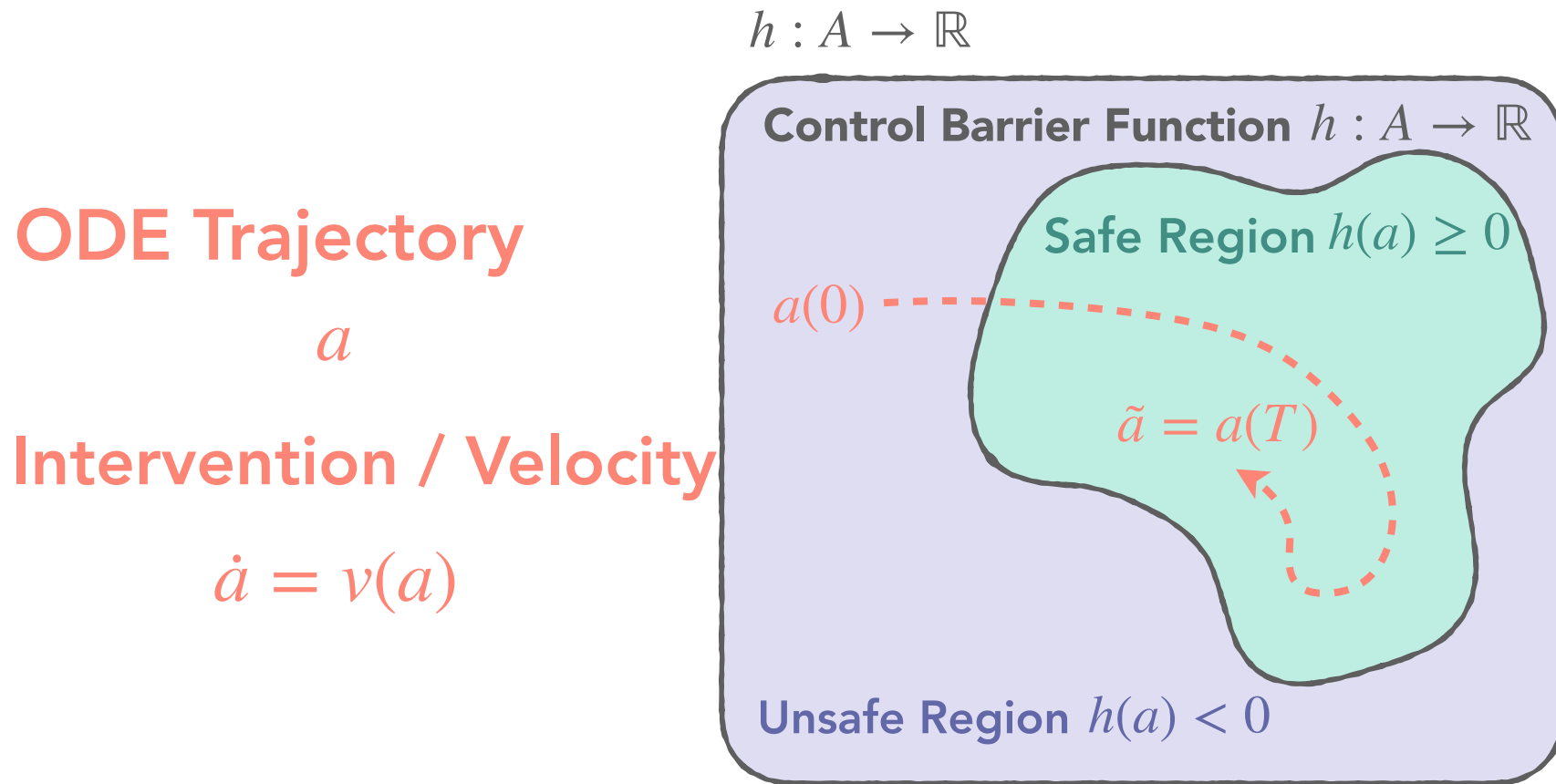
ODE-Steering: Control Barrier Function

How to learn the **safe boundary** h ?



ODE-Steering: Control Barrier Function

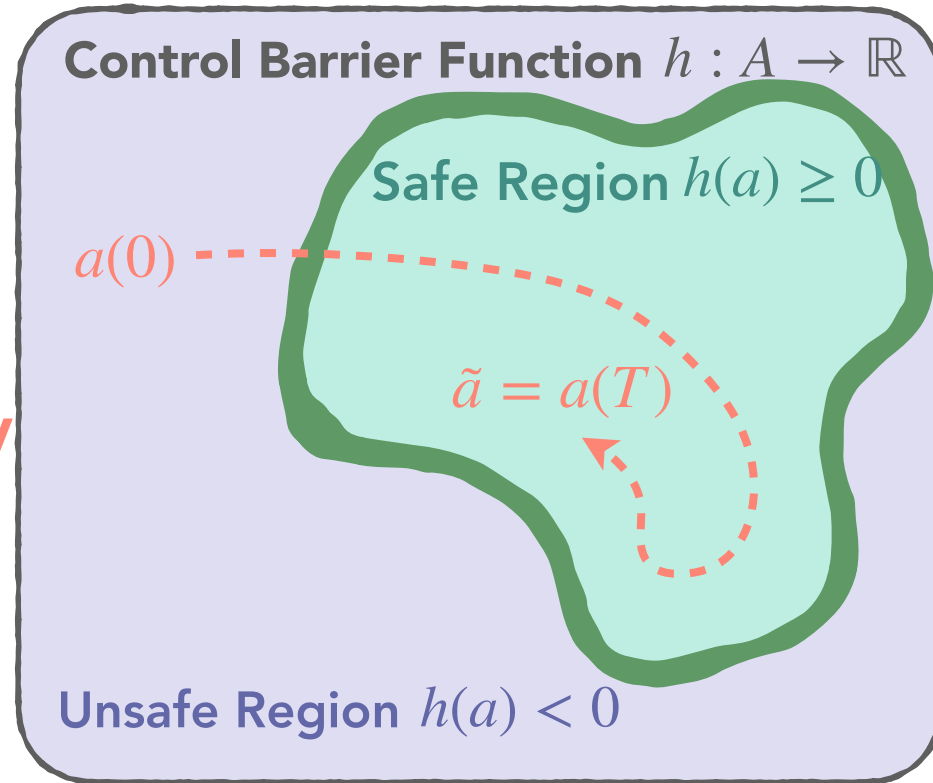
How to learn the **safe boundary** h ?



ODE-Steering: Control Barrier Function

Stay in the safe boundary h

$$h : A \rightarrow \mathbb{R}$$



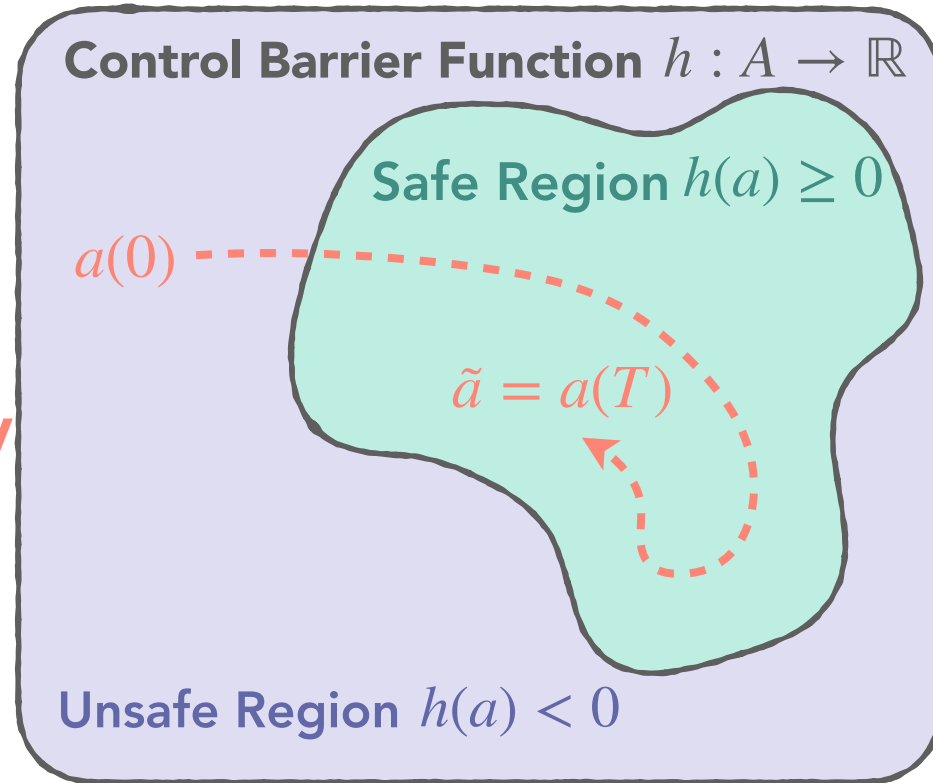
Value On Boundary:

$$h(a) = 0$$

ODE-Steering: Control Barrier Function

Stay in the safe boundary h

$$h : A \rightarrow \mathbb{R}$$



Value On Boundary:

$$h(a) = 0$$

Derivative:

$$\nabla h(a) v(a)$$

derivative

ODE Trajectory

a

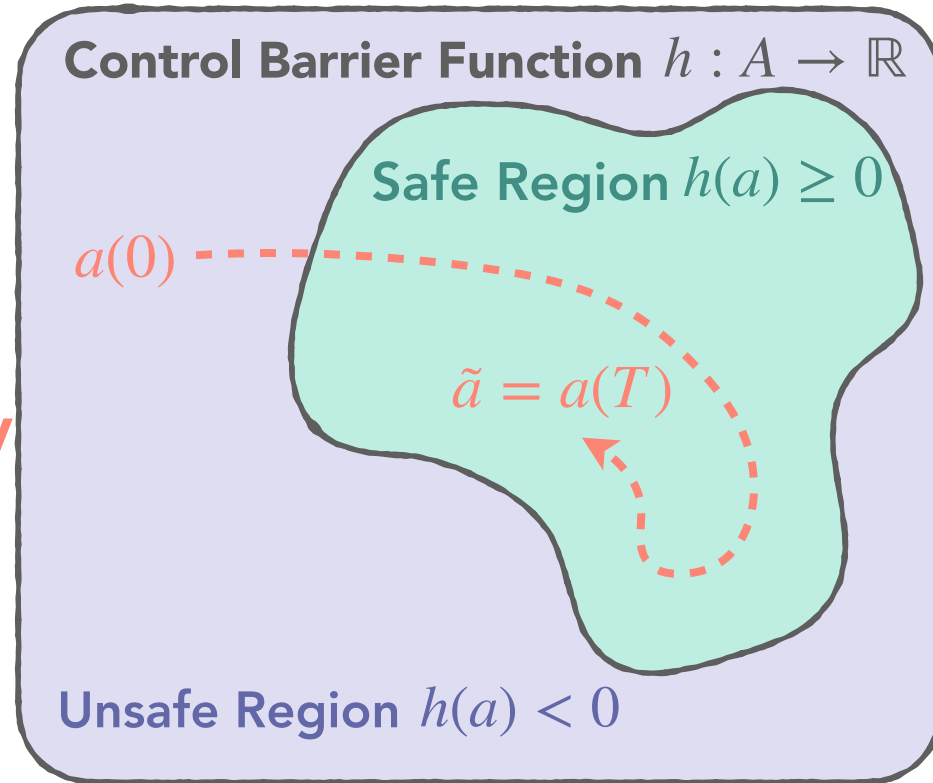
Intervention / Velocity

$$\dot{a} = v(a)$$

ODE-Steering: Control Barrier Function

Stay in the safe boundary h

$$h : A \rightarrow \mathbb{R}$$



Value On Boundary:

$$h(a) = 0$$

Derivative:

$$\nabla h(a) v(a)$$

derivative on velocity of a

ODE Trajectory

a

Intervention / Velocity

$$\dot{a} = v(a)$$

ODE-Steering: Control Barrier Function

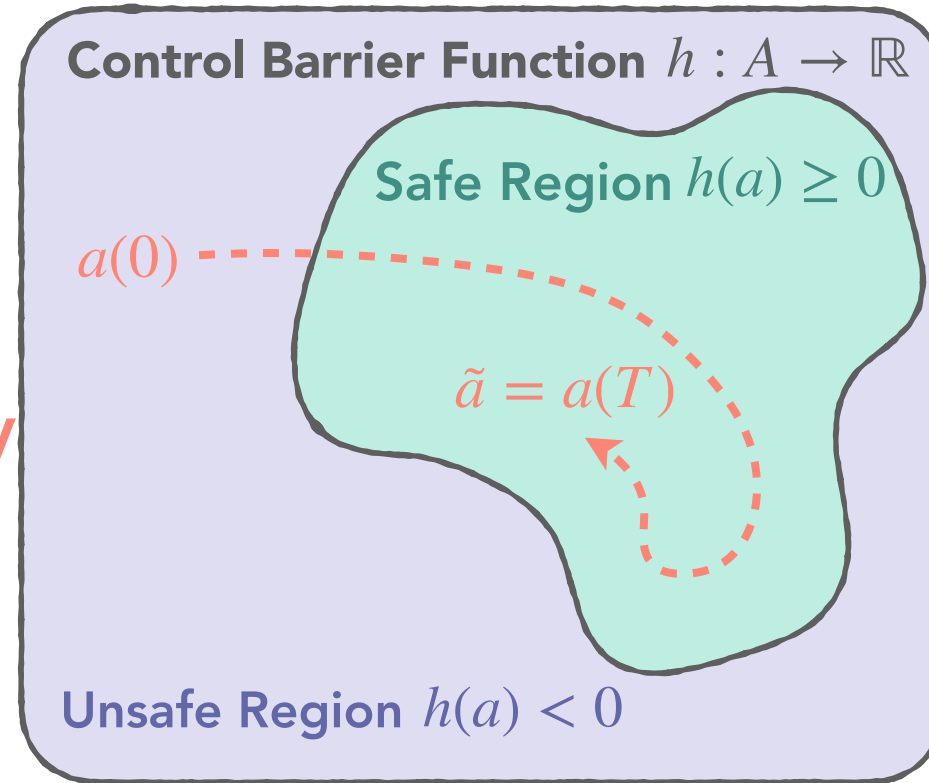
$$\nabla h(a) v(a) \geq 0$$

ODE Trajectory

a

Intervention / Velocity

$$\dot{a} = v(a)$$



Value On Boundary:

$$h(a) = 0$$

Derivative:

$$\nabla h(a) v(a)$$

ODE-Steering: Control Barrier Function

$$\nabla h(a) v(a) + \gamma h(a) \geq 0, \quad \gamma > 0$$

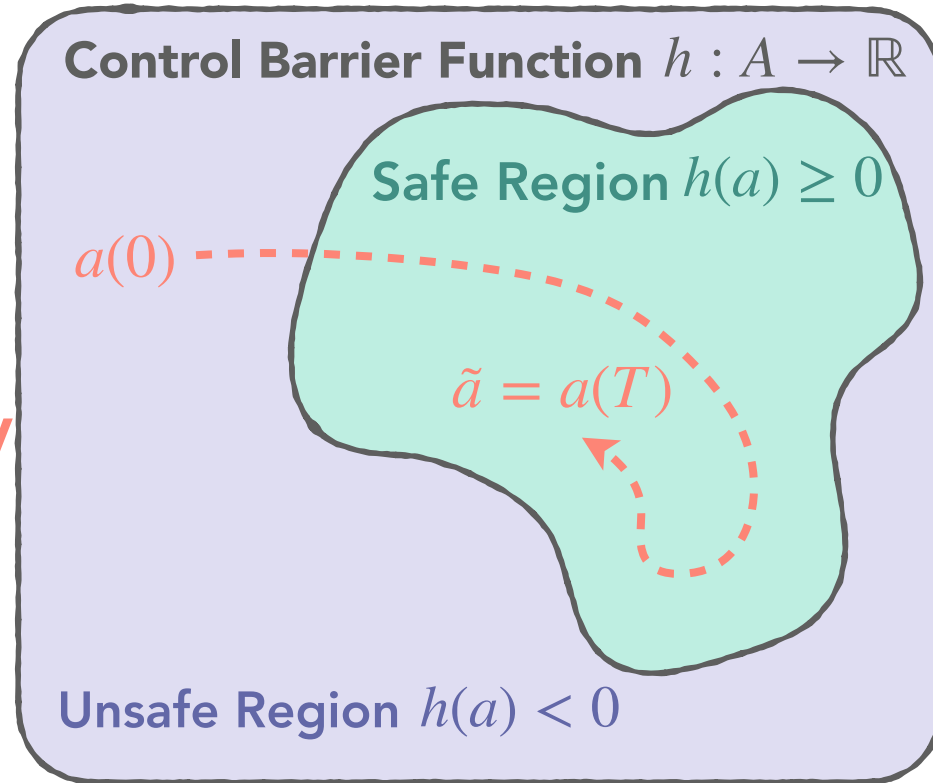
safe margin

ODE Trajectory

a

Intervention / Velocity

$$\dot{a} = v(a)$$



Value On Boundary:

$$h(a) = 0$$

Derivative:

$$\nabla h(a) v(a)$$

ODE-Steering: Control Barrier Function

$$\nabla h(a) v(a) + \gamma h(a) \geq 0, \quad \gamma > 0$$

What is to be learned?

min

$$v(a)$$

minimal intervention

s.t.

$$h(a) > 0, \quad \forall a \in C$$

positive region

$$h(a) < 0, \quad \forall a \notin C$$

negative region

$$\nabla h(a) v(a) + \gamma h(a) \geq 0, \quad \gamma > 0$$

derivative

Can it be learned via BloackBox Neural Network?

Neural Network parameterize Control Barrier Function

min

$$v(a)$$

minimal intervention

s.t.

$$h(a) > 0, \quad \forall a \in C$$

positive region

$$h(a) < 0, \quad \forall a \notin C$$

negative region

$$\nabla h(a) v(a) + \gamma h(a) \geq 0, \quad \gamma > 0$$

derivative

Neural Network parameterize Control Barrier Function

min

$v(a)$

minimal intervention

$$\mathcal{L}_{\text{ctrl}} = \frac{1}{N} \sum_{i=1}^N \left\| v_{\theta}(a^{(i)}) \right\|_2^2$$

s.t.

$h(a) > 0, \forall a \in C$

positive region

$h(a) < 0, \forall a \notin C$

$$\mathcal{L}_{\text{boundary}} = \frac{1}{N} \sum_{i=1}^N \text{ReLU} \left(\varepsilon - y^{(i)} h_{\phi}(a^{(i)}) \right)$$

$\nabla h(a) v(a) + \gamma h(a) \geq 0, \quad \gamma > 0$

derivative

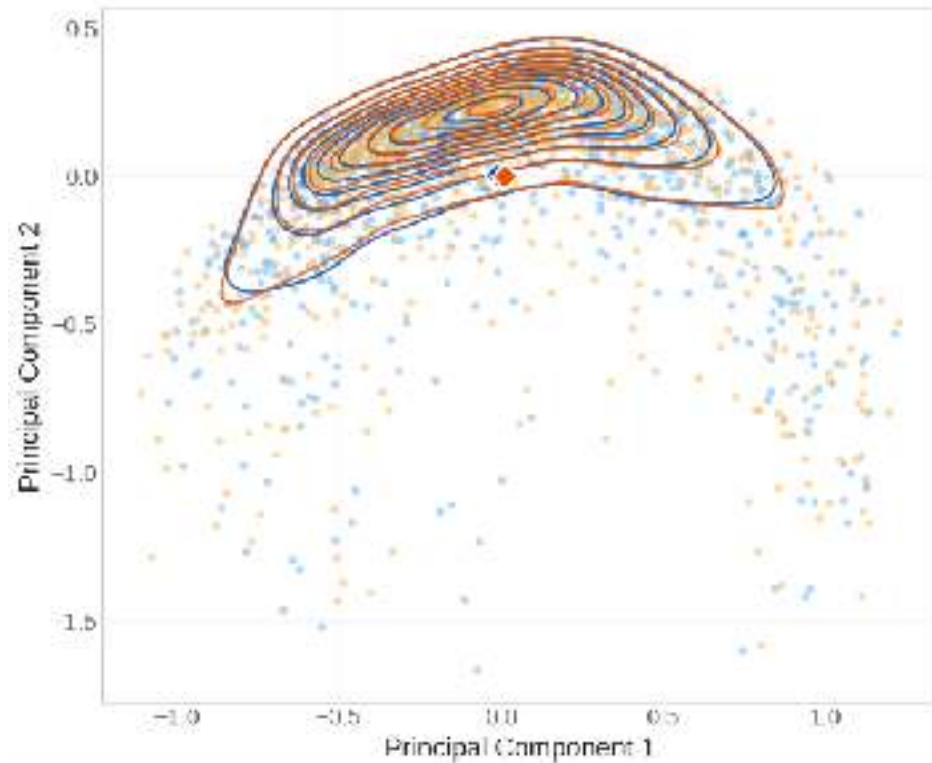
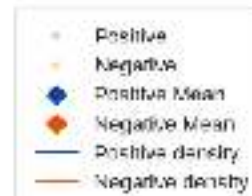
$$\mathcal{L}_{\text{deriv}} = \frac{1}{N} \sum_{i=1}^N \text{ReLU} \left(\varepsilon - \nabla h_{\phi}(a^{(i)}) v_{\theta}(a^{(i)}) - \alpha h_{\phi}(a^{(i)}) \right)$$

Brute-force training Neural Network cannot work

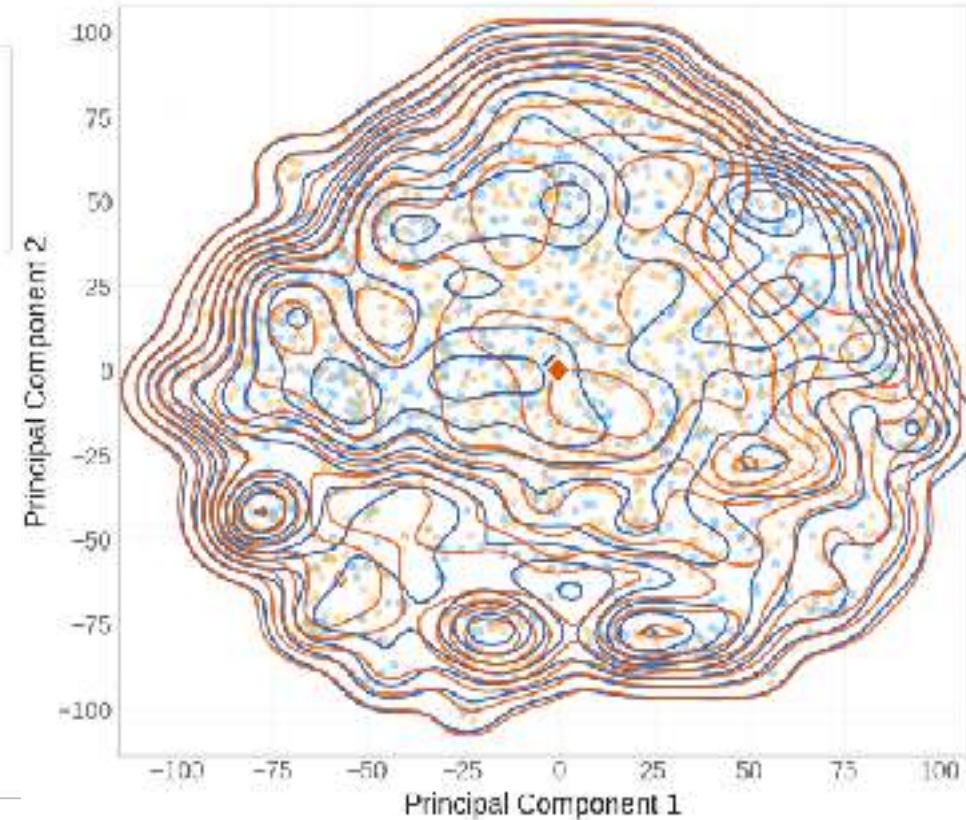
Not separable



2D Visualization of SafeRLHF using PCA



2D Visualization of SafeRLHF using TSNE

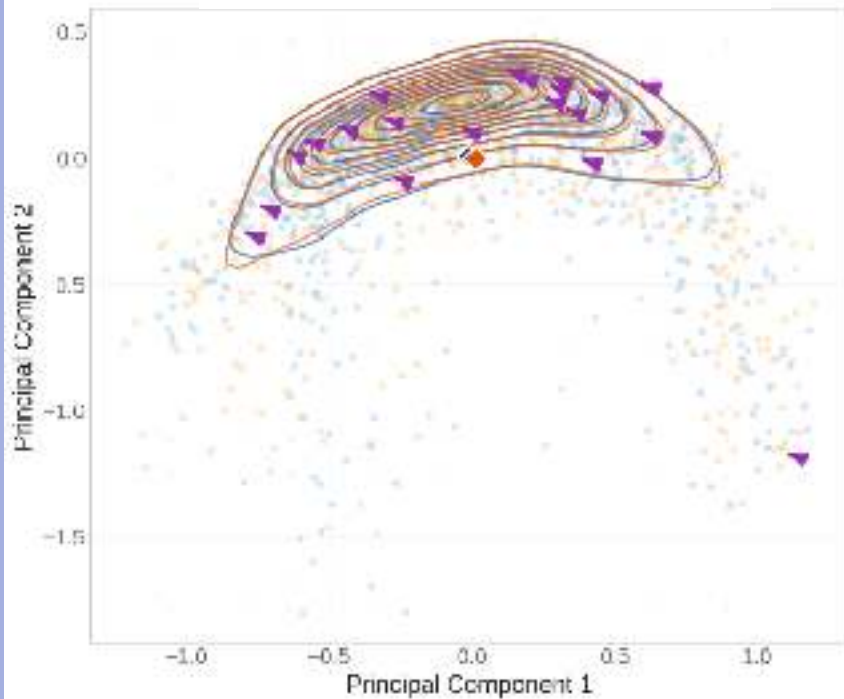


Dataset: PKU-SafeRLHF (Classifier Accuracy: ~60%)

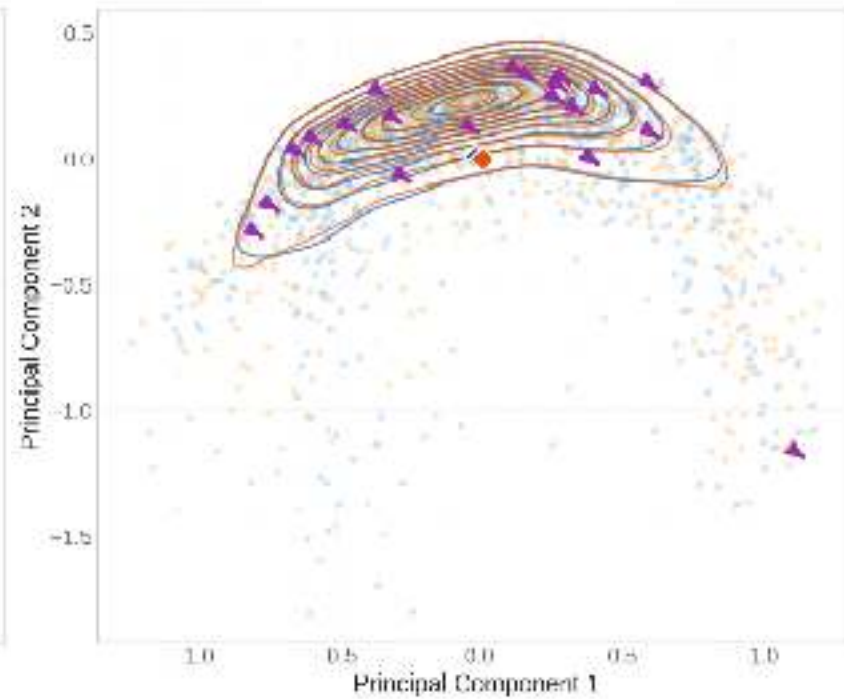
Brute-force training Neural Network cannot work

What are Control Barrier Function parameterize?

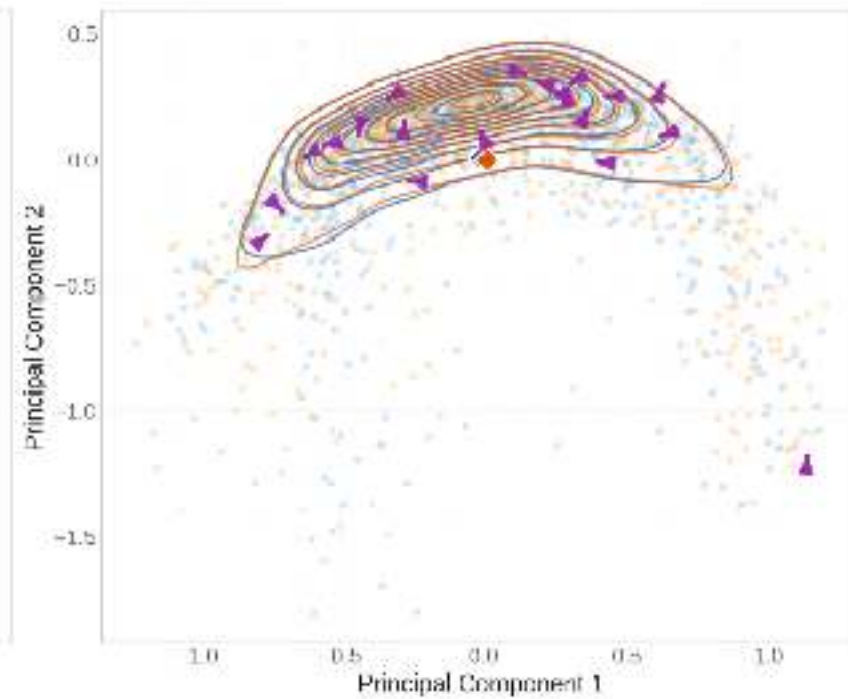
Difference-in-Means [1]



Linear Probe [2]



Neural Network



Dataset: PKU-SafeRLHF (Classifier Accuracy: ~60%)

Solution: **Blackbox Parameterization** → **Principled Solution**

What is to be learned?

$v(a)$

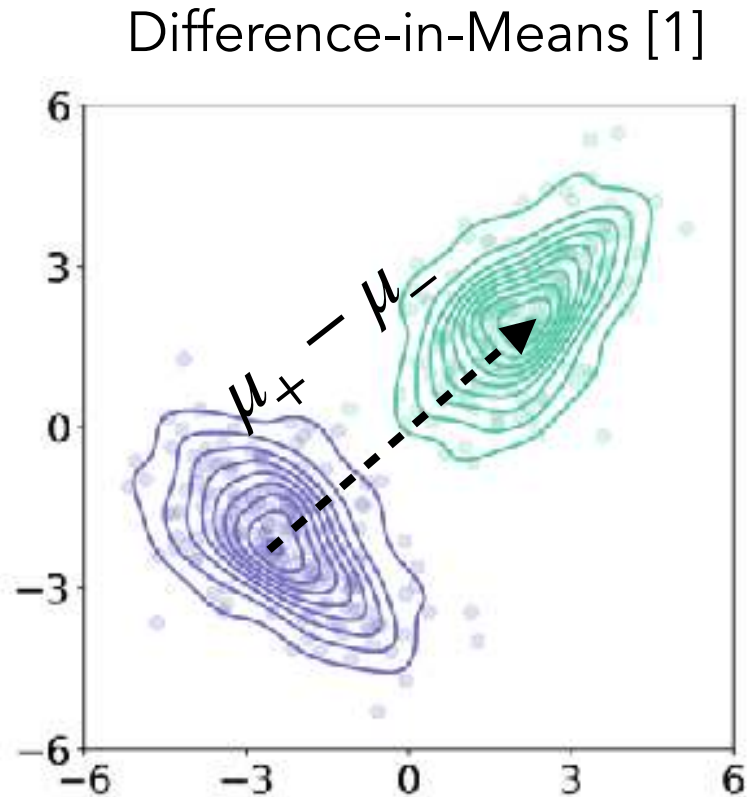
$h(a)$

What is to be learned?

Steering Vector $v(a) = \nabla h(a)$

Barrier Function $h(a)$

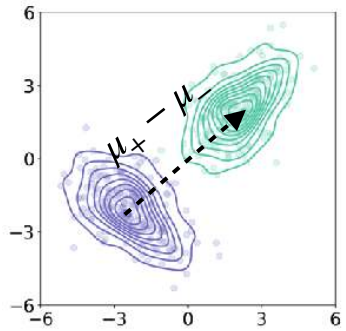
The umbrella: **All Steering is Barrier Function Design**



The umbrella: **All Steering is Barrier Function Design**

Steering Vector $v(a) = \nabla h(a)$

Barrier Function $h(a)$



Difference-in-Means [1]

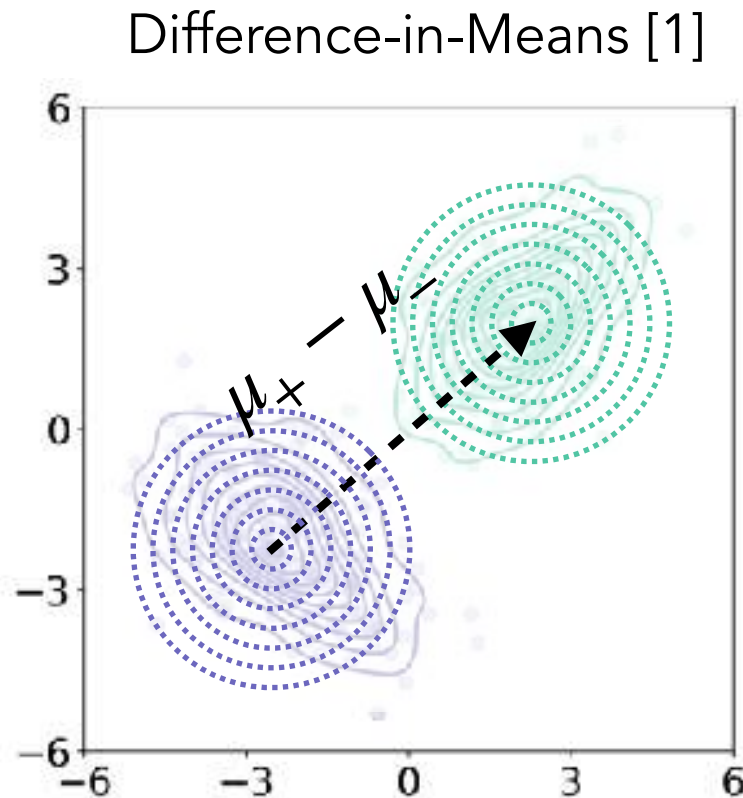
$$\mu_+ - \mu_-$$

The umbrella: **All Steering is Barrier Function Design**

Barrier Function $h(\mathbf{a})$

Log-density ratio $h(\mathbf{a}) = \log \frac{p_+(\mathbf{a})}{p_-(\mathbf{a})}$

Gaussian Assumption $p_{\pm}(\mathbf{a}) = \mathcal{N}(\mu_{\pm}, \mathbf{I})$



The umbrella: **All Steering is Barrier Function Design**

Steering Vector $v(a) = \nabla h(a)$

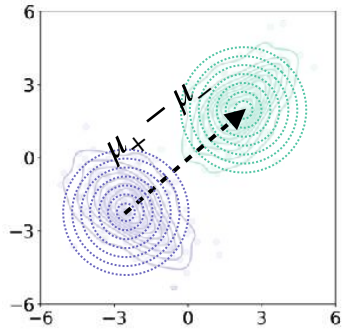
Barrier Function $h(a)$

Difference-in-Means [1]

$$\mu_+ - \mu_-$$

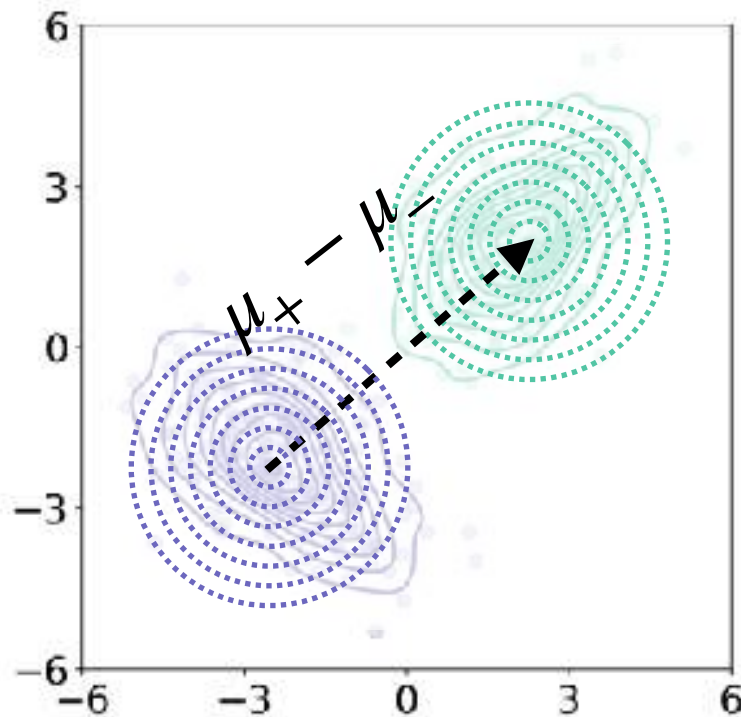
Log-density ratio $h(\mathbf{a}) = \log \frac{p_+(\mathbf{a})}{p_-(\mathbf{a})}$

Gaussian Assumption $p_{\pm}(\mathbf{a}) = \mathcal{N}(\mu_{\pm}, \mathbf{I})$

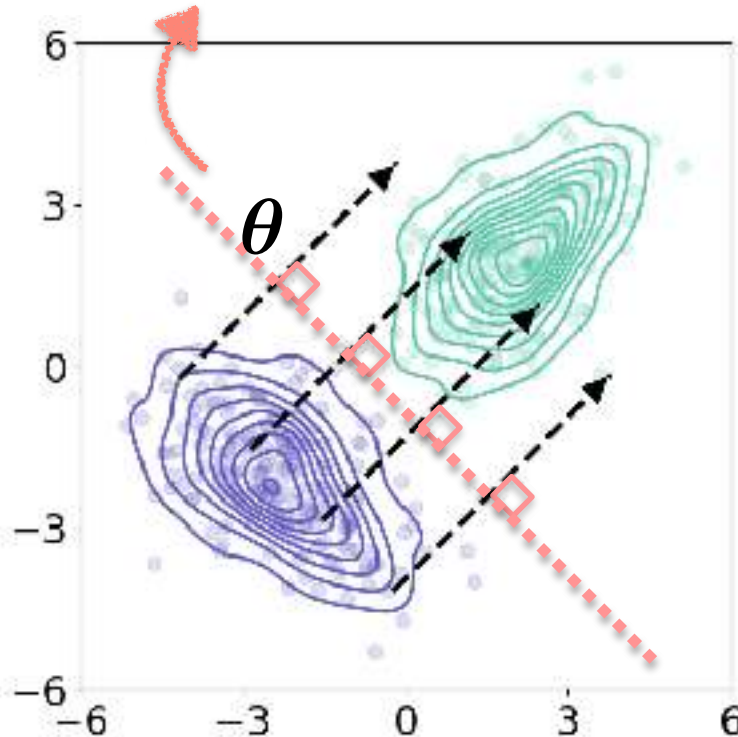


The umbrella: **All Steering is Barrier Function Design**

Difference-in-Means [1]



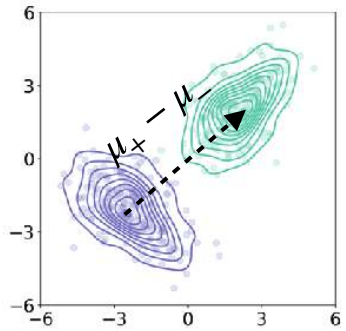
Linear Probe [2]



The umbrella: **All Steering is Barrier Function Design**

Steering Vector $v(a) = \nabla h(a)$

Barrier Function

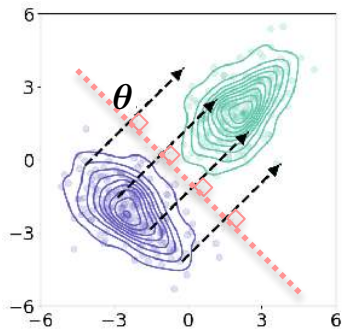


Difference-in-Means [1]

$$\mu_+ - \mu_-$$

Log-density ratio $h(\mathbf{a}) = \log \frac{p_+(\mathbf{a})}{p_-(\mathbf{a})}$

Gaussian Assumption $p_{\pm}(\mathbf{a}) = \mathcal{N}(\mu_{\pm}, \mathbf{I})$



Linear Probe [2]

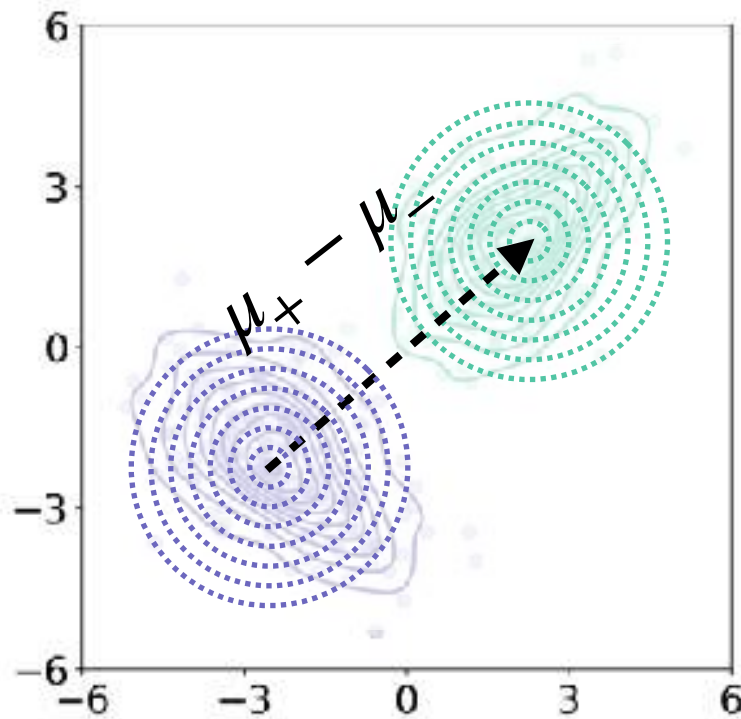
$$\theta$$

Log-density ratio $h(\mathbf{a}) = \log \frac{p_+(\mathbf{a})}{p_-(\mathbf{a})}$

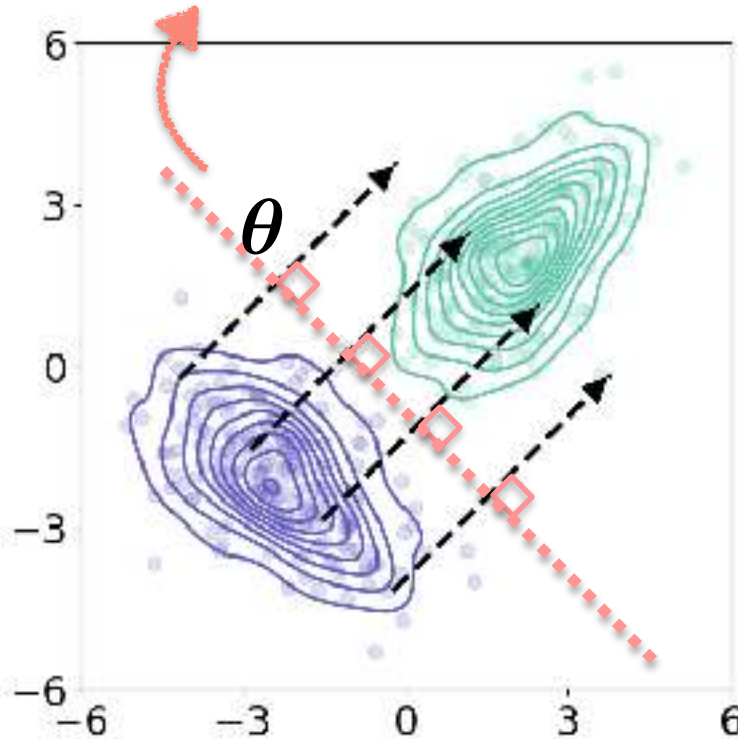
Logistic Regression $p_{\theta}(\mathbf{a}) = \text{sigmoid}(\theta^{\top} \mathbf{a})$.

The umbrella: **All Steering is Barrier Function Design**

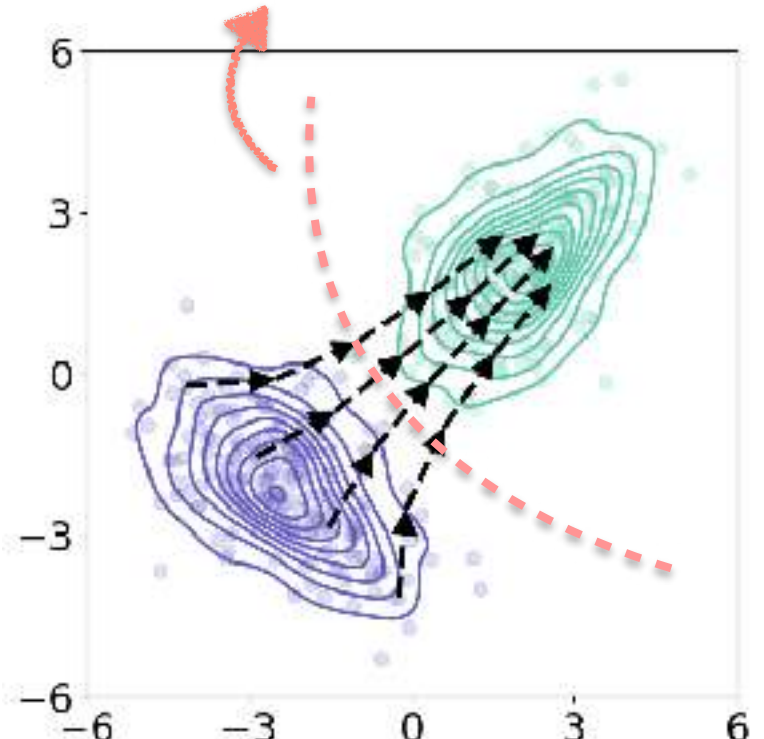
Difference-in-Means [1]



Linear Probe [2]



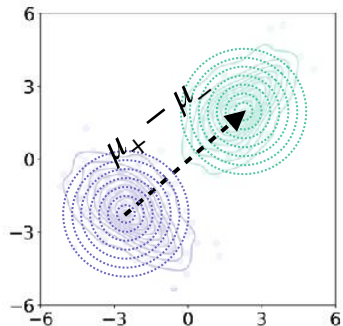
Ours (Control Barrier Function)



Steering Vector $v(a) = \nabla h(a)$

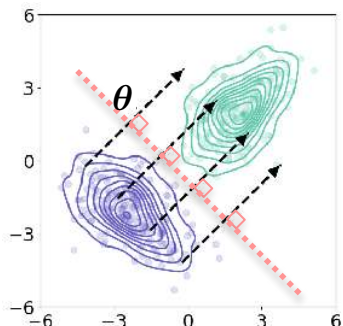
Difference-in-Means [1]

$$\mu_+ - \mu_-$$

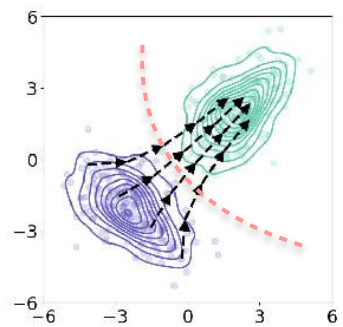


Linear Probe [2]

$$\theta$$



$$\nabla h(a) = J_\phi(a)^\top w$$



Barrier Function $h(a)$

Log-density ratio $h(a) = \log \frac{p_+(a)}{p_-(a)}$

Gaussian Assumption $p_\pm(a) = \mathcal{N}(\mu_\pm, I)$

Log-density ratio $h(a) = \log \frac{p_+(a)}{p_-(a)}$

Logistic Regression $p_\theta(a) = \text{sigmoid}(\theta^\top a)$.

Log-density ratio $h(a) = \log \frac{p_+(a)}{p_-(a)}$

$$= w^\top \phi(a) + b$$

nonlinear feature map

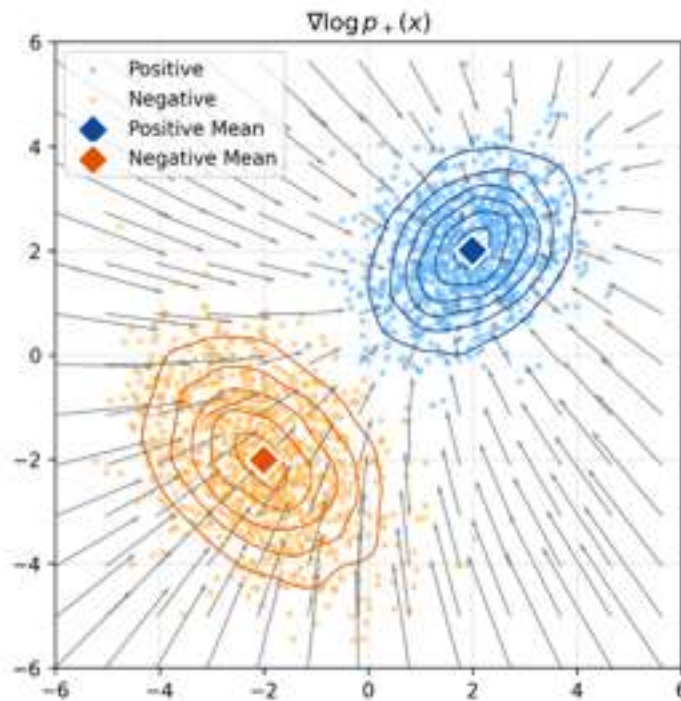
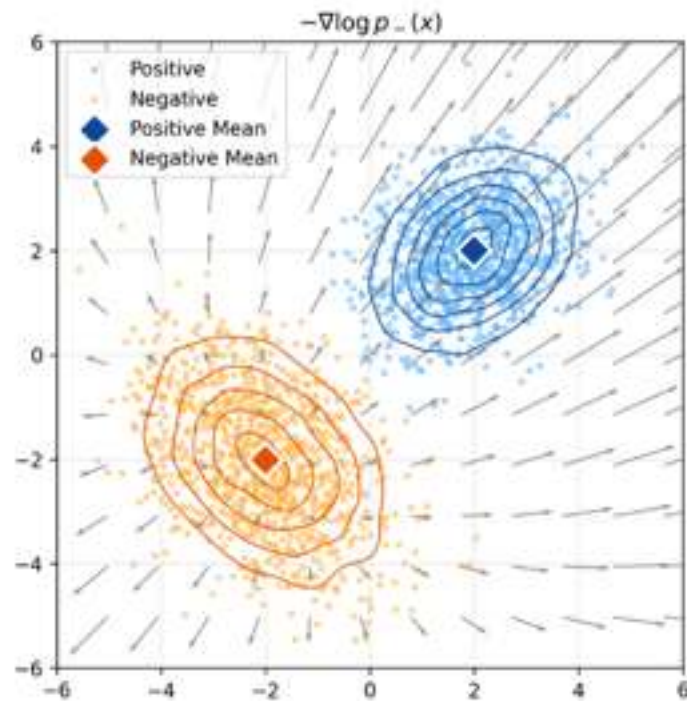
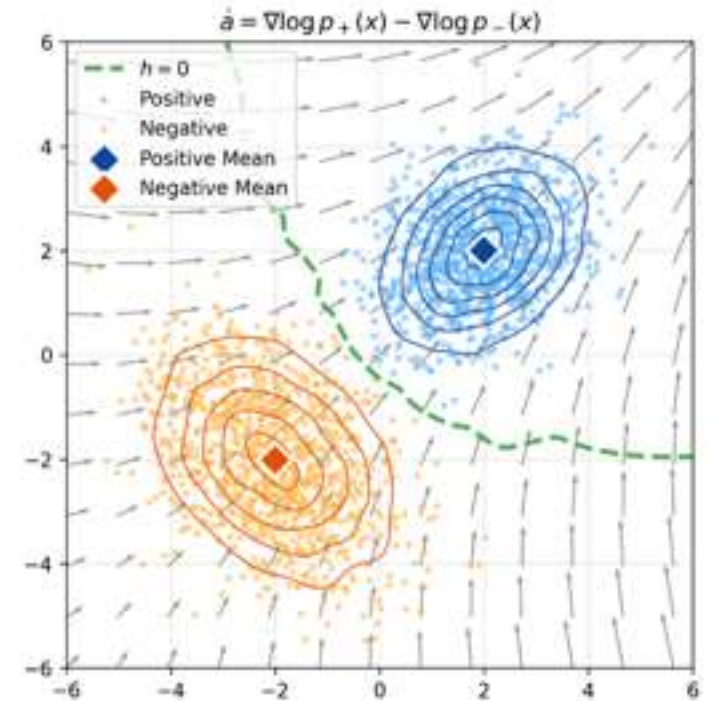
Solution: **Blackbox Parameterization** → **Principled Solution**

Defining Barrier Function

$$h(\mathbf{a}) = \log \left(\frac{p_+(\mathbf{a})}{p_-(\mathbf{a})} \right) = \log p_+(\mathbf{a}) - \log p_-(\mathbf{a})$$

Positive Embedding
Probability Density

Negative Embedding
Probability Density



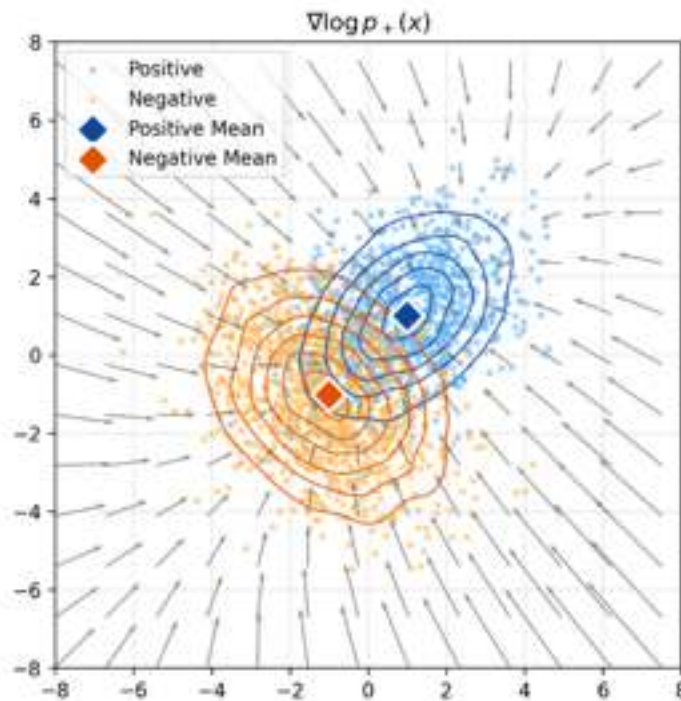
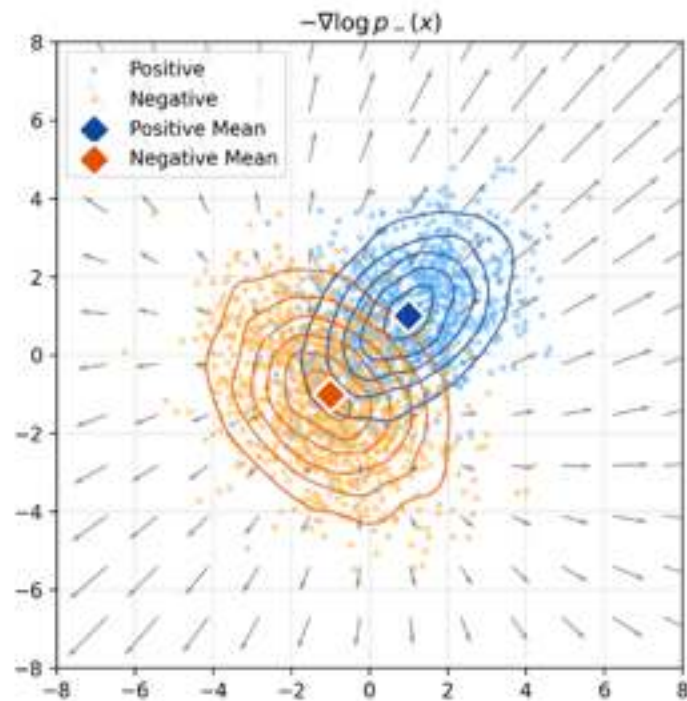
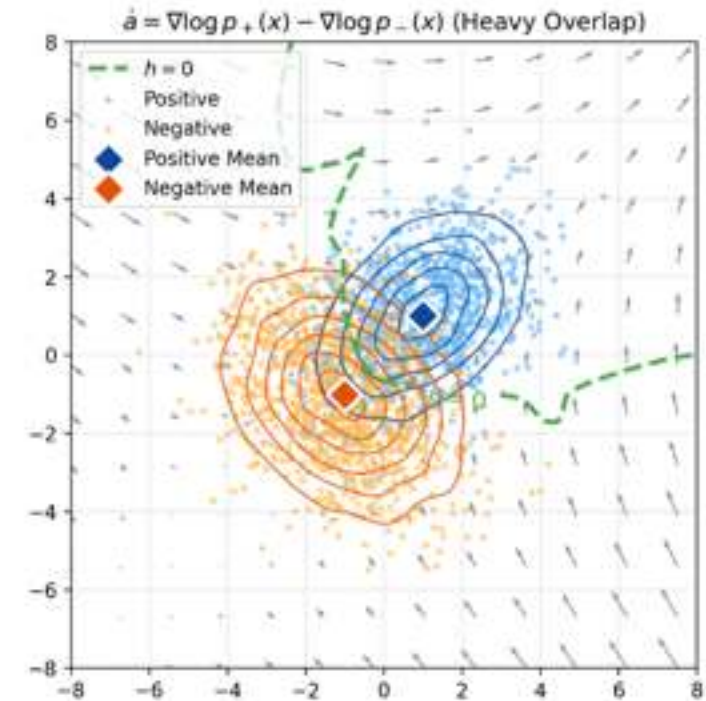
Solution: **Blackbox Parameterization** → **Principled Solution**

Defining Barrier Function

$$h(\mathbf{a}) = \log \left(\frac{p_+(\mathbf{a})}{p_-(\mathbf{a})} \right) = \log p_+(\mathbf{a}) - \log p_-(\mathbf{a})$$

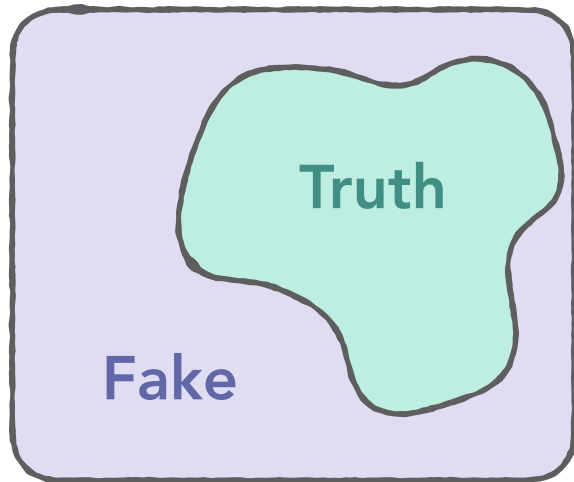
Positive Embedding
Probability Density

Negative Embedding
Probability Density

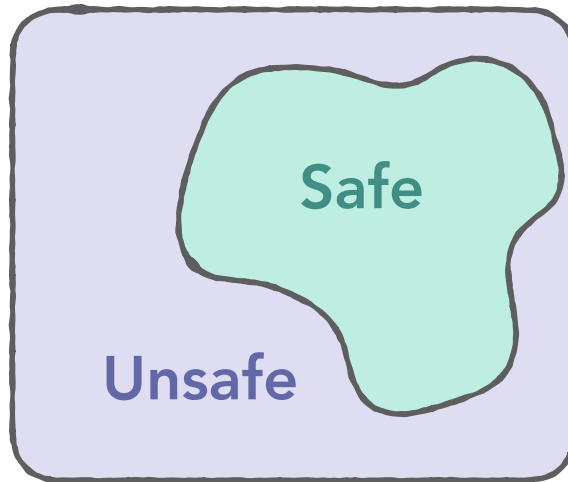


ODE-Steering: Control Barrier Function

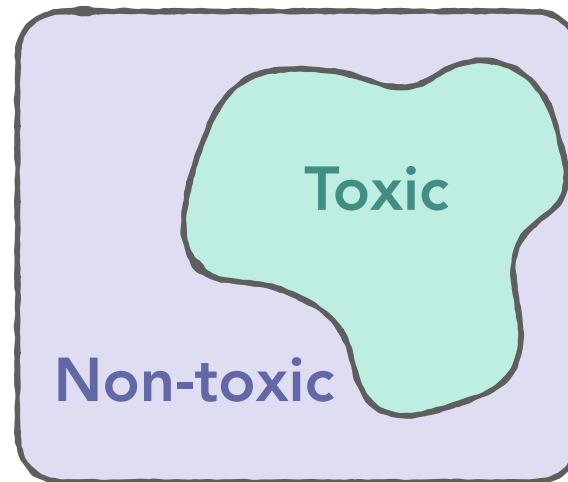
From “static models” to “dynamic control”



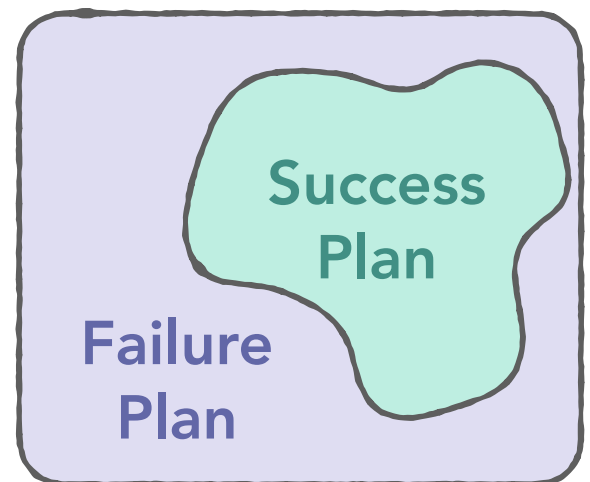
Truthful QA
+5.7%



SafeRLHF
+2.5%



RealToxicityPrompts
+2.4%



OpenVLA

Results towards Safe VLA



before steering



after steering