End-to-End Robust and Trustworthy AI Solutions

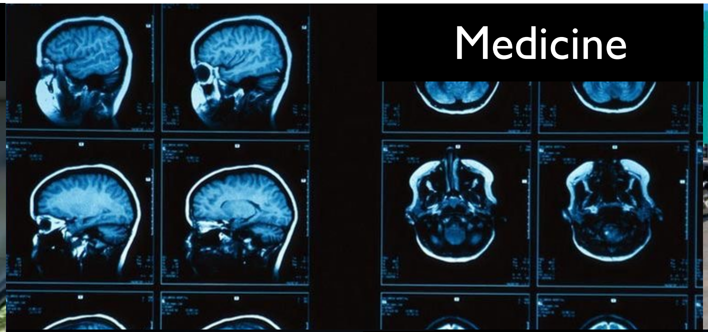
Alexander Amini

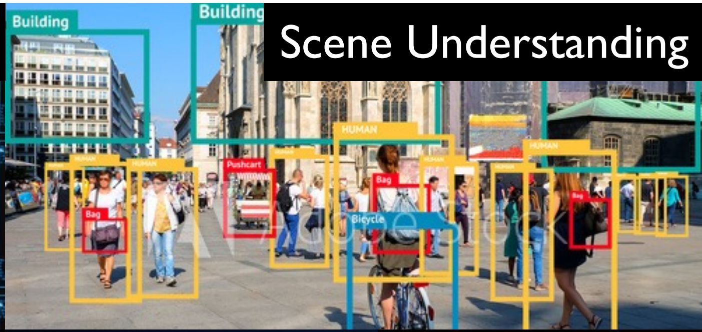
Chief Scientific Officer

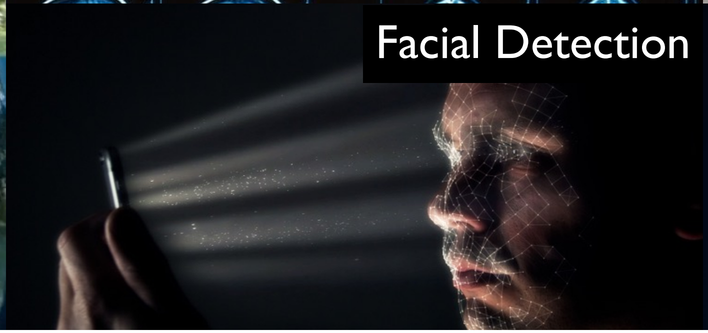
Themis AI

Artificial Intelligence in Safety Critical Applications

Robotics


Medicine


Scene Understanding


Facial Detection

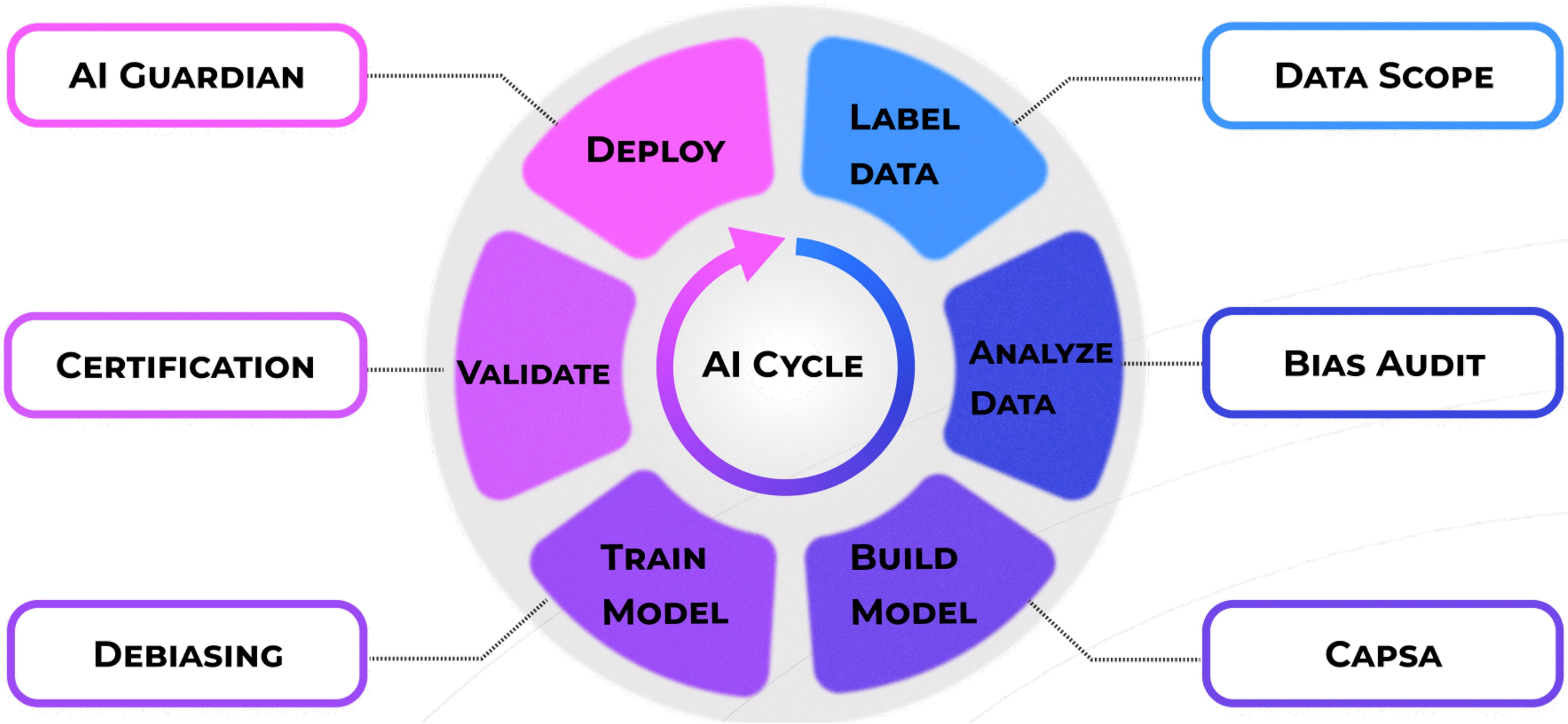
Security

Deep learning is being applied in many safety critical domains

Interacting with and making decisions in the presence of humans

Models must not propagate bias and reliably inform uncertainty

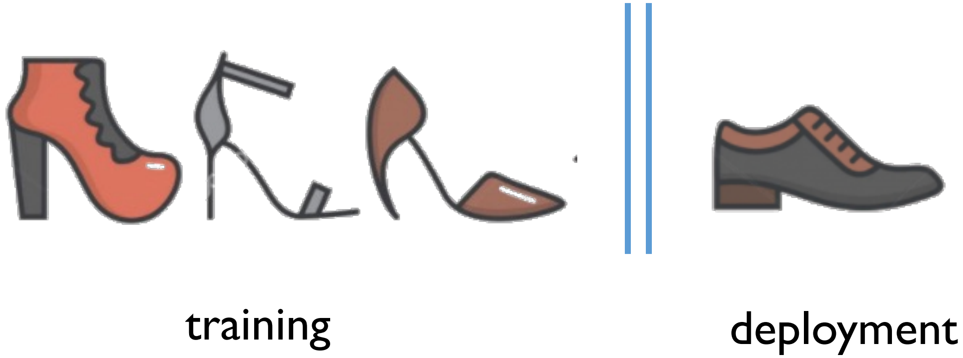
Themis AI: Empowering the world to create, advance, and deploy trustworthy AI



Bias and Uncertainty in Artificial Intelligence

Model Bias

Model decision changes if it exposed to additional "sensitive" feature inputs



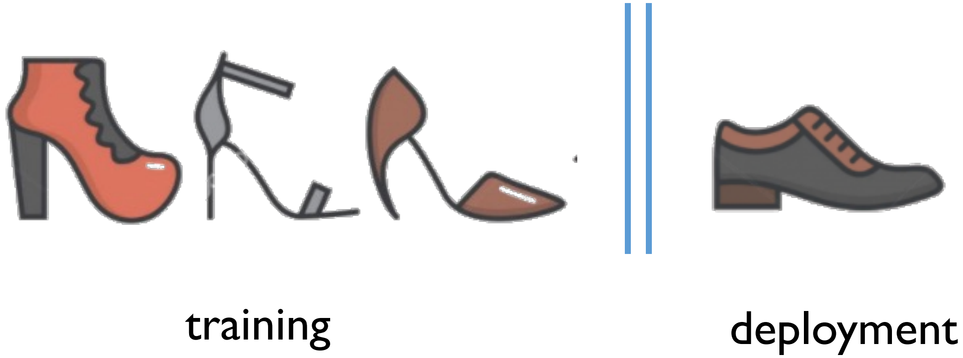
Uncertainty

Can we train models to understand when they don't know the answer?

Bias and Uncertainty in Artificial Intelligence

Model Bias

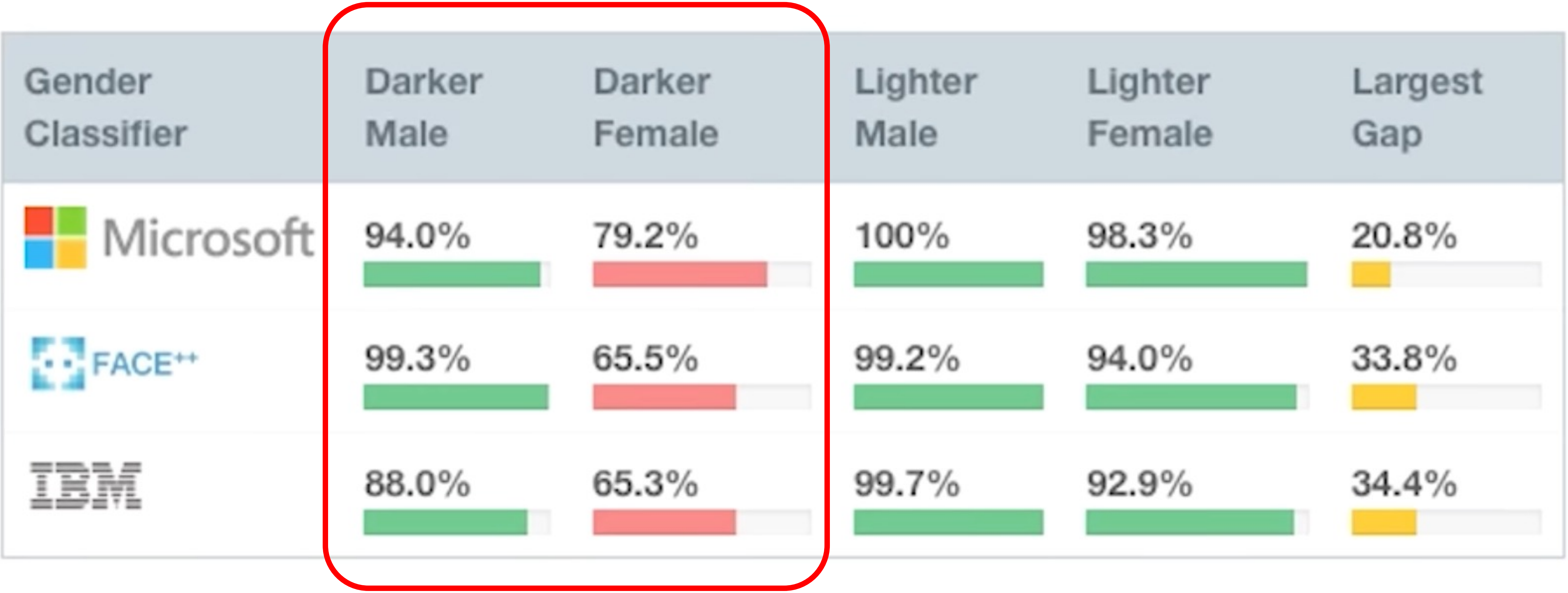
Model decision changes if it exposed to additional "sensitive" feature inputs



Uncertainty

Can we train models to understand when they don't know the answer?

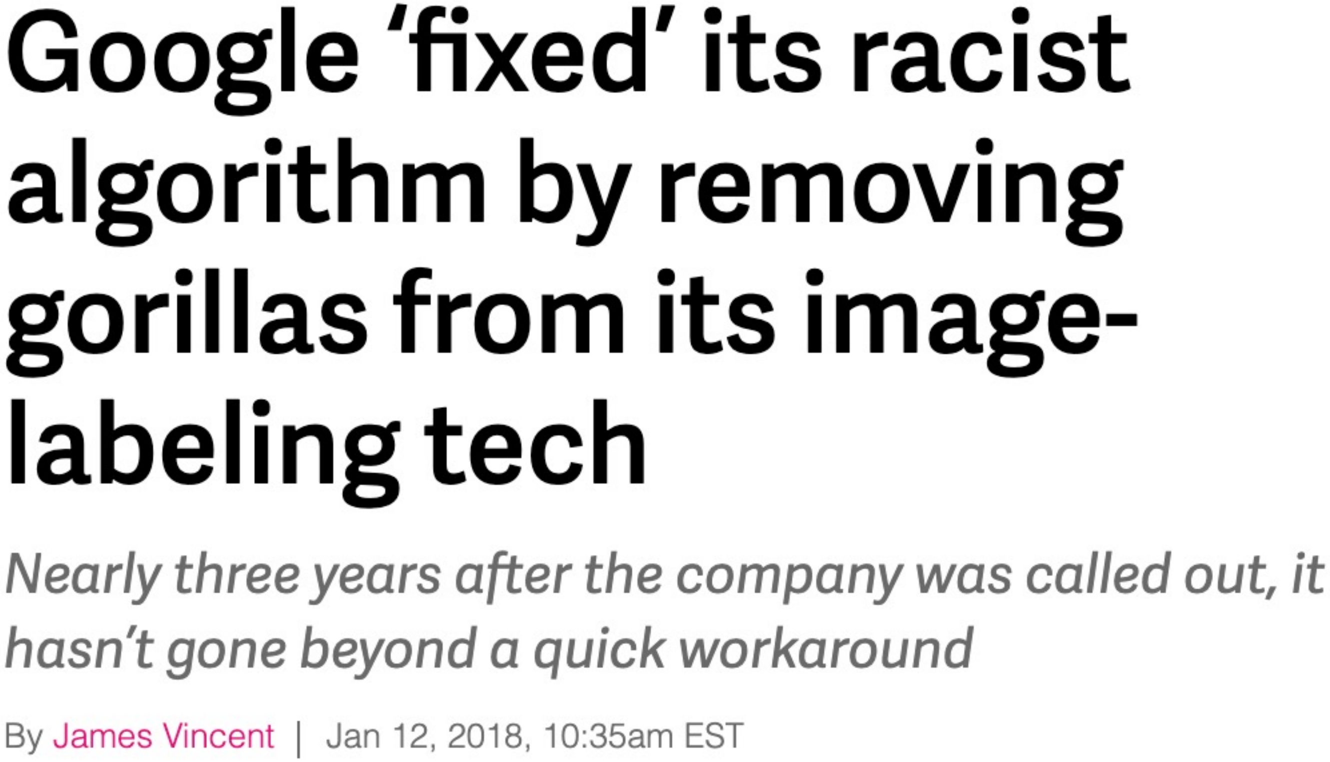
Bias in Facial Detection Systems



Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

Buolamwini et al. 2018

Google Photo’s: Image Labelling



Google ‘fixed’ its racist algorithm by removing gorillas from its image-labeling tech

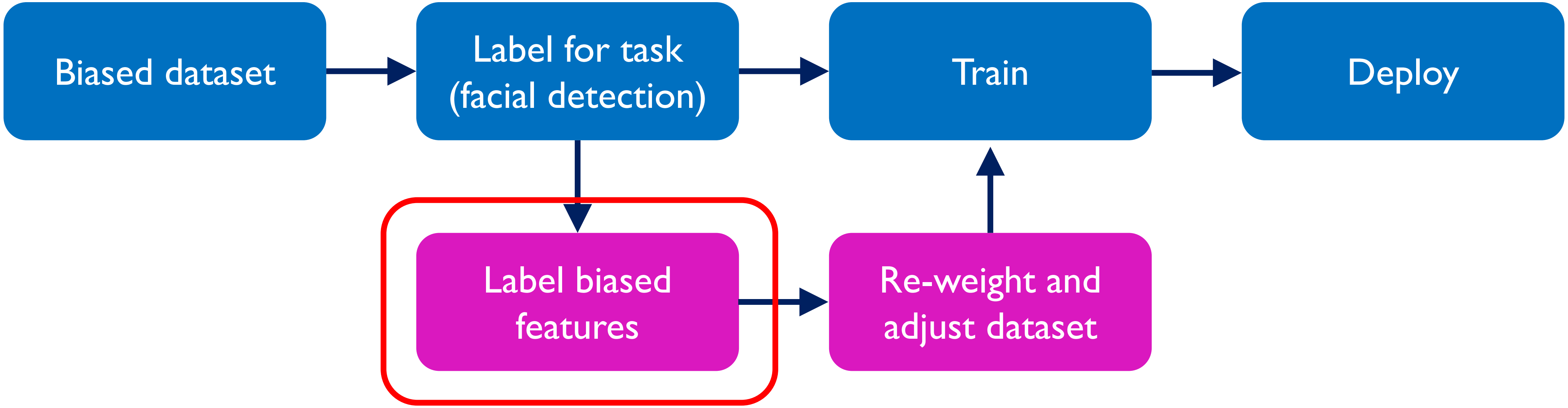
Nearly three years after the company was called out, it hasn’t gone beyond a quick workaround

By James Vincent | Jan 12, 2018, 10:35am EST

The Verge. 2018

Problems with Methods for Mitigating Bias

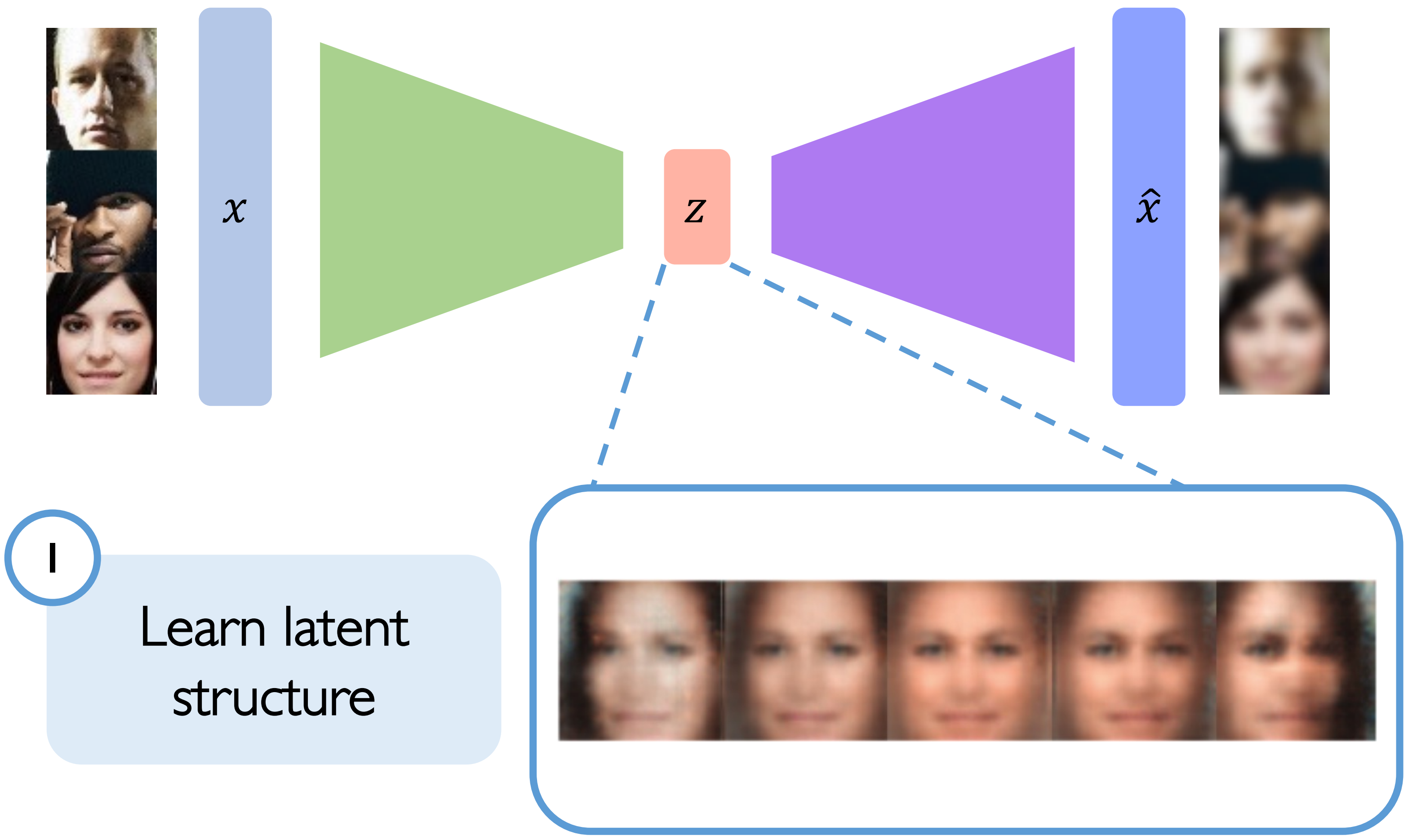
Knowing your dataset is biased is not enough, need algorithmic methods for de-biasing



Biased dataset
Label for task (facial detection)
Train
Deploy
Label biased features
Re-weight and adjust dataset

How can we know which labels to de-bias?

Mitigating Bias Through Learned Latent Structure

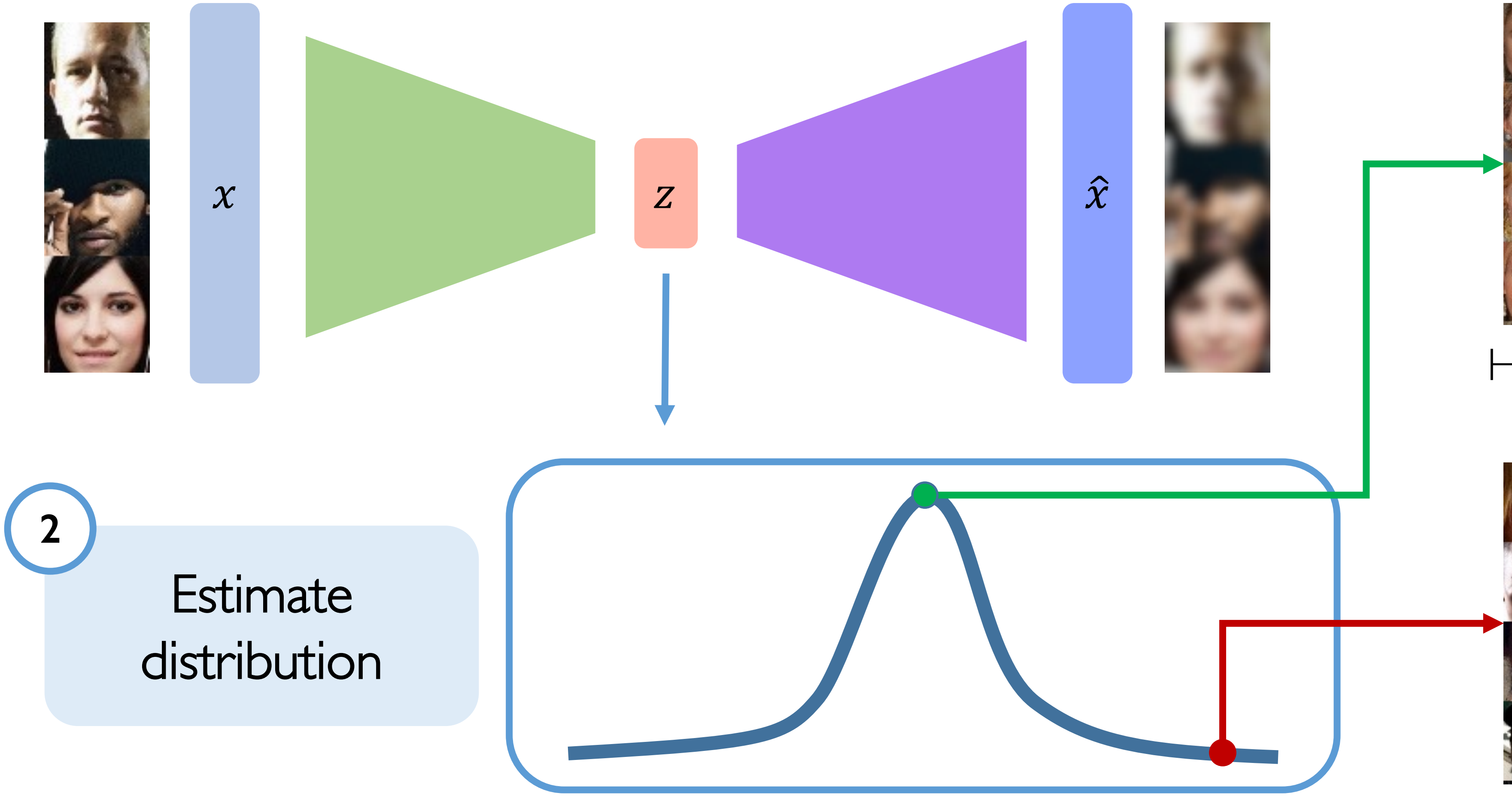


1

Learn latent structure

Amini/Soleimany et al., *AAAI/AIES* 2019.

Mitigating Bias Through Learned Latent Structure



2

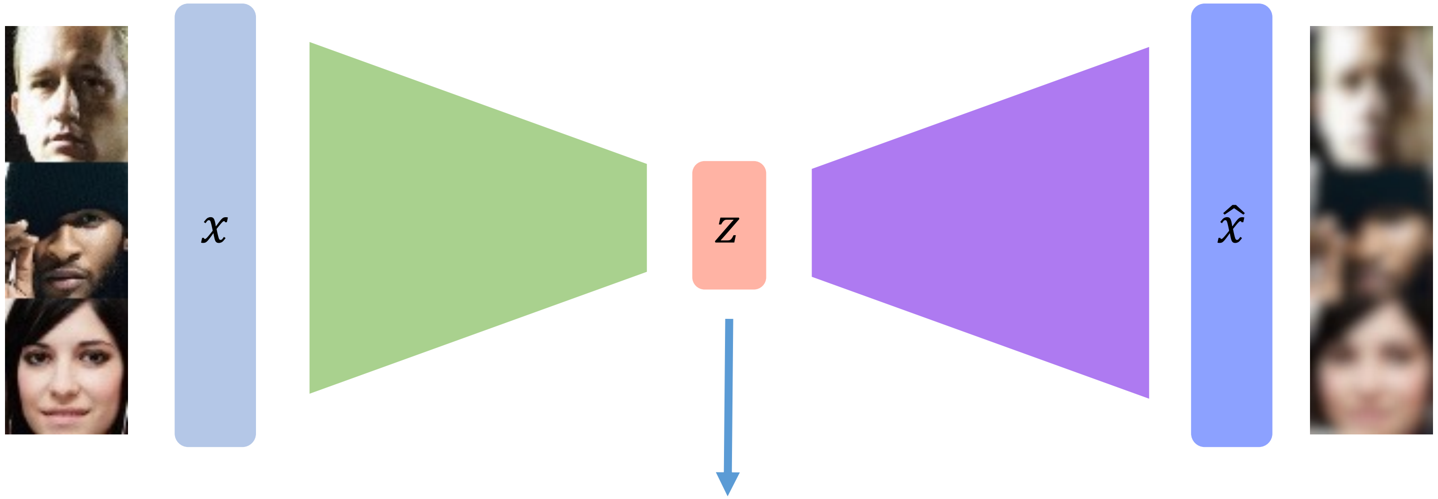
Estimate distribution

Homogeneous skin color, pose

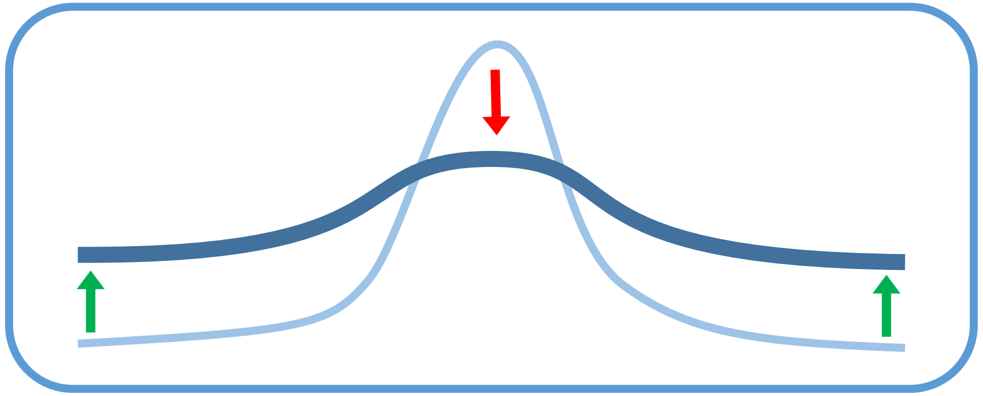
Diverse skin color, pose, illumination

Amini/Soleimany et al., *AAAI/AIES* 2019.

Mitigating Bias Through Learned Latent Structure

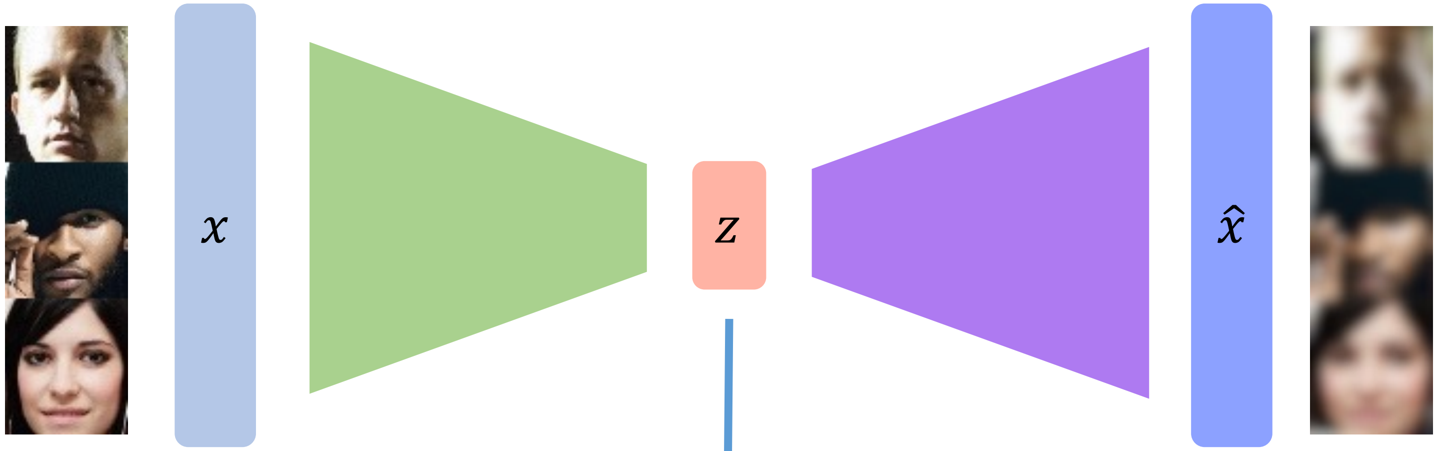


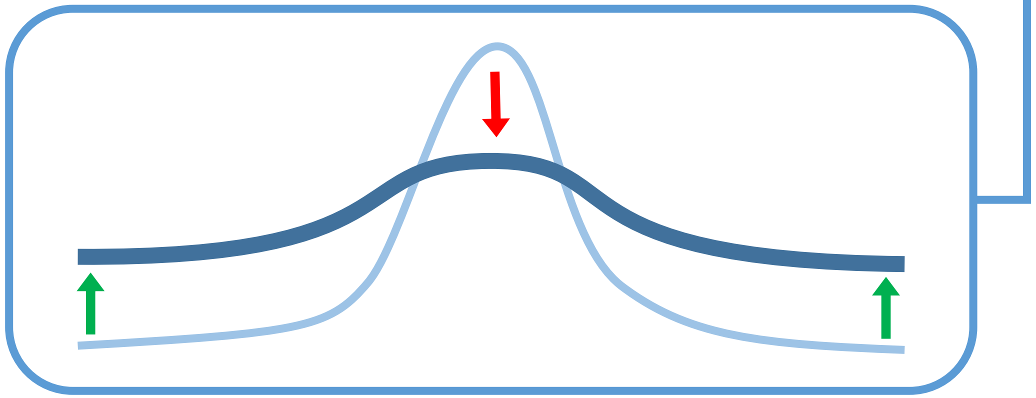
3 Adaptively guide learning



Amini/Soleimany et al., *AAAI/AIES* 2019.

Mitigating Bias Through Learned Latent Structure





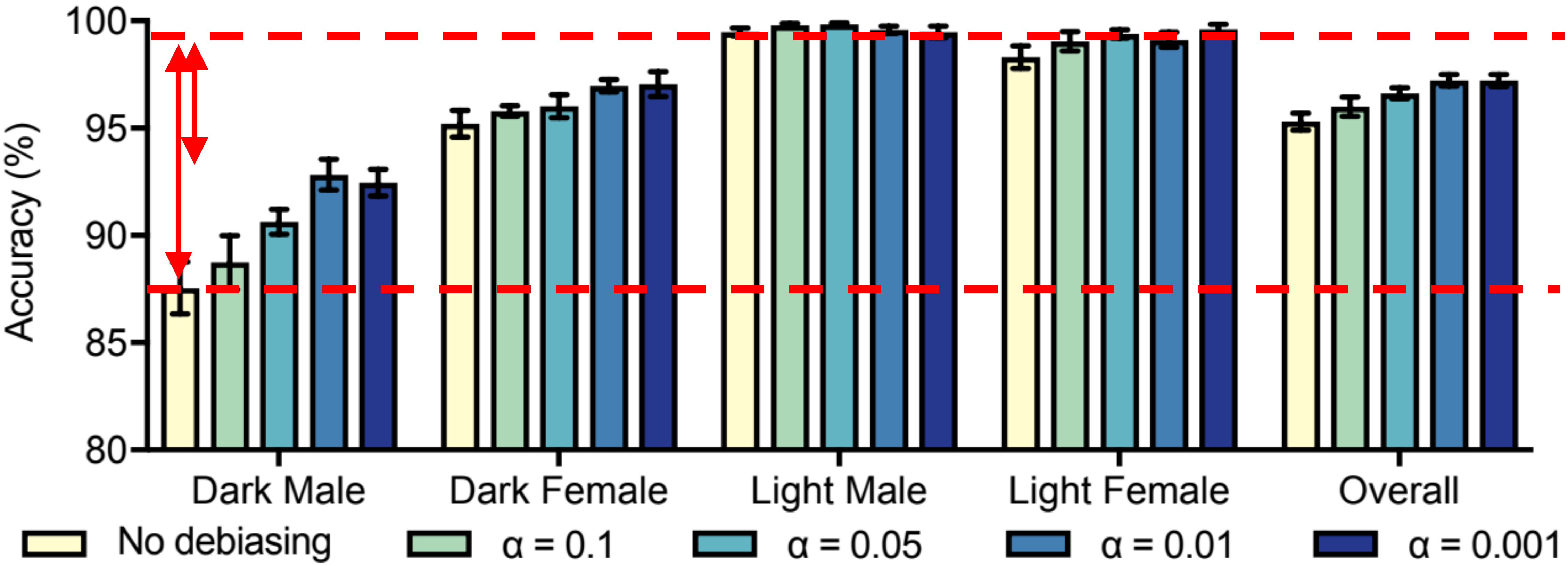
4

Learn from fair distributions

Latent distributions used to create fair and representative dataset

Amini/Soleimany et al., *AAAI/AIES* 2019.

Results: Increasing Strengths of Debiasing

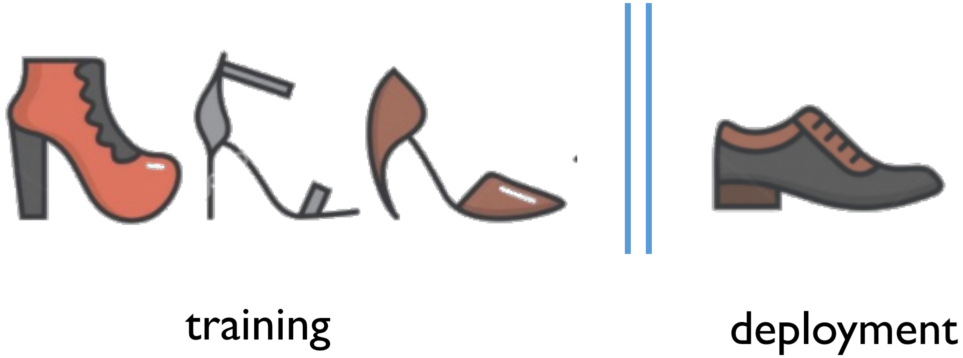


Amini/Soleimany et al., *AAAI/AIES* 2019.

Bias and Uncertainty in Artificial Intelligence

Model Bias

Model decision changes if it exposed to additional “sensitive” feature inputs



Uncertainty

Can we train models to understand when they don’t know the answer?

Bias and Uncertainty in Artificial Intelligence

Model Bias

Model decision changes if it exposed to additional "sensitive" feature inputs

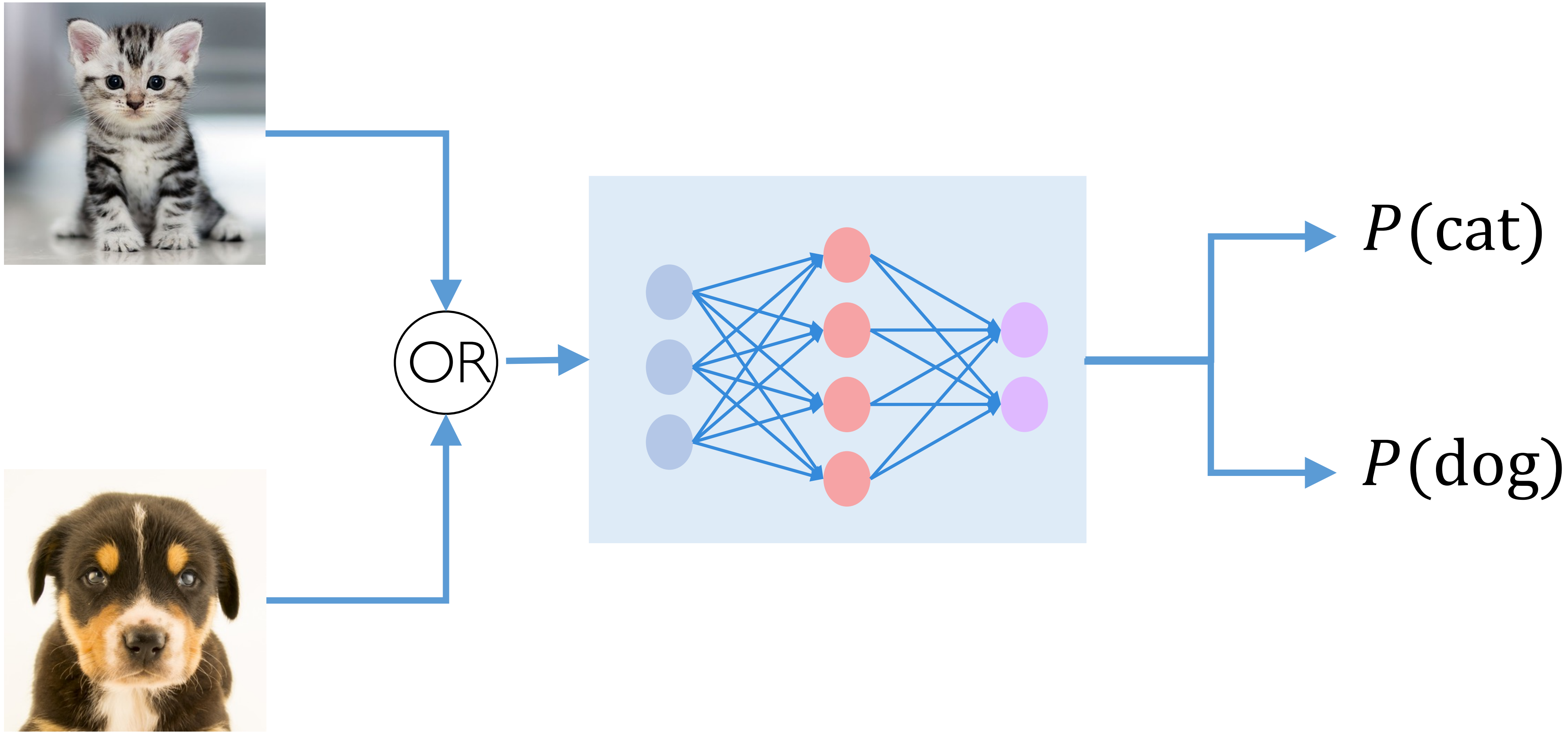
training

deployment

Uncertainty

Can we train models to understand when they don't know the answer?

Why Care About Uncertainty?



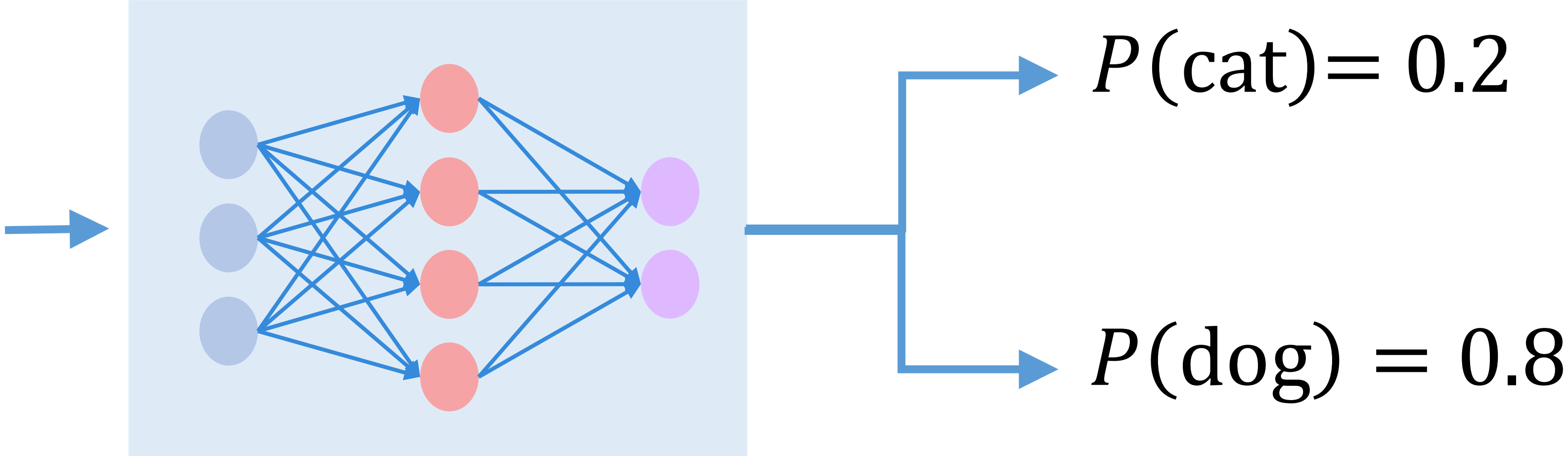
OR

$P(\text{cat})$

$P(\text{dog})$

Why Care About Uncertainty?

We need **uncertainty** metrics to assess the network's **confidence** in its predictions.



$P(\text{cat}) = 0.2$

$P(\text{dog}) = 0.8$

Remember: $P(\text{cat}) + P(\text{dog}) = 1$

Deep Evidential Learning

View learning as an **evidence acquisition** process

More evidence → increased predictive confidence

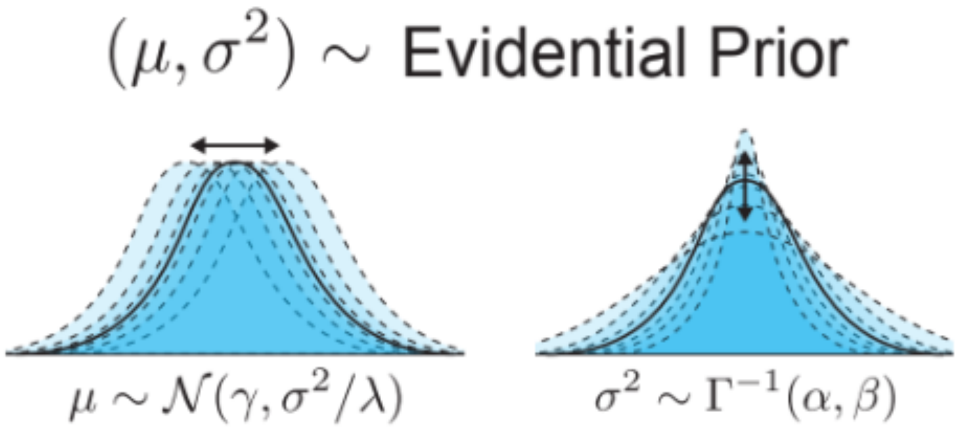
1. Assume data is drawn from a Gaussian with unknown mean and unknown variance

$$(y_1, \ldots, y_N) \sim \mathcal{N}(\mu, \sigma^2)$$

$$\mu \sim \mathcal{N}(\gamma, \sigma^2 \upsilon^{-1}) \qquad \sigma^2 \sim \Gamma^{-1}(\alpha, \beta).$$

2. Place prior over distributional parameters to probabilistically learn them

$$p(\underbrace{\mu, \sigma^2}_{\boldsymbol{\theta}} \,|\, \underbrace{\gamma, \upsilon, \alpha, \beta}_{\boldsymbol{m}}) = \frac{\beta^\alpha \sqrt{\upsilon}}{\Gamma(\alpha)\sqrt{2\pi\sigma^2}} \left(\frac{1}{\sigma^2}\right)^{\alpha+1} \exp\left\{-\frac{2\beta + \upsilon(\gamma - \mu)^2}{2\sigma^2}\right\}.$$


$(\mu, \sigma^2) \sim$ Evidential Prior
$\mu \sim \mathcal{N}(\gamma, \sigma^2/\lambda)$
$\sigma^2 \sim \Gamma^{-1}(\alpha, \beta)$

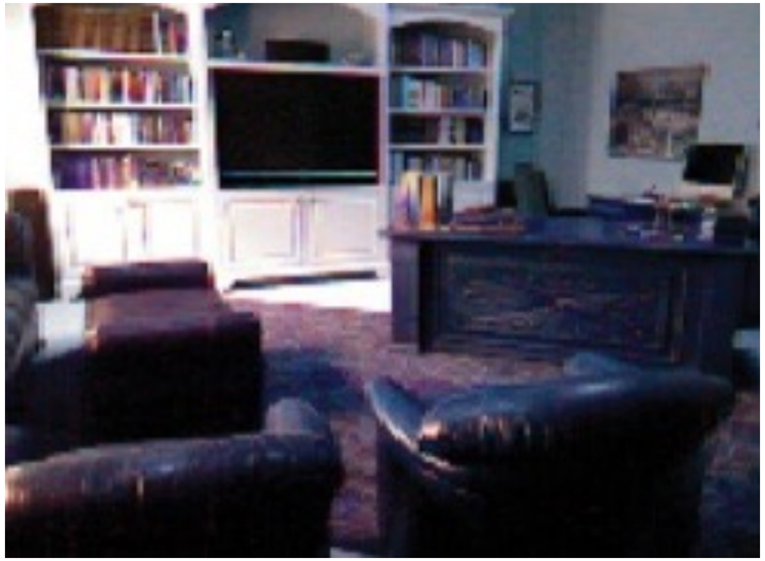
Amini et al. "Deep evidential regression" NeurIPS 2020

Monocular Depth Estimation

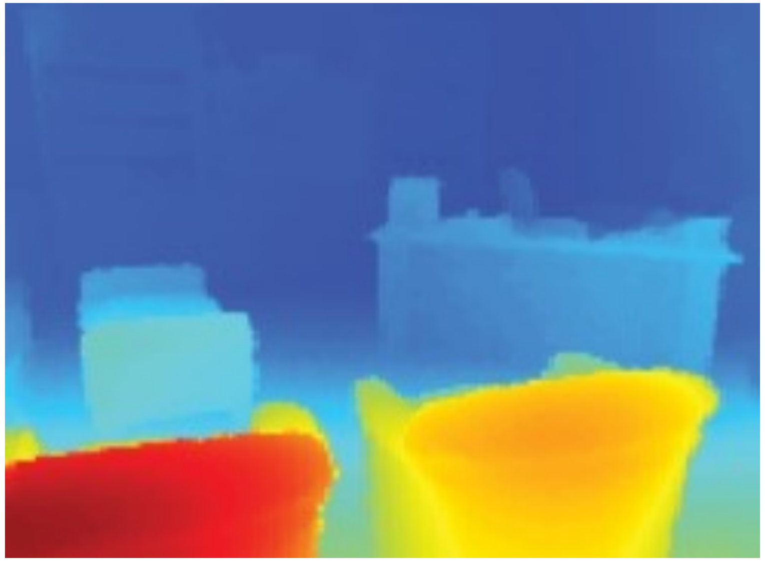
Task: Given a monocular RGB image, predict the depth of every pixel

Applications in autonomous vehicles, home and industrial robots

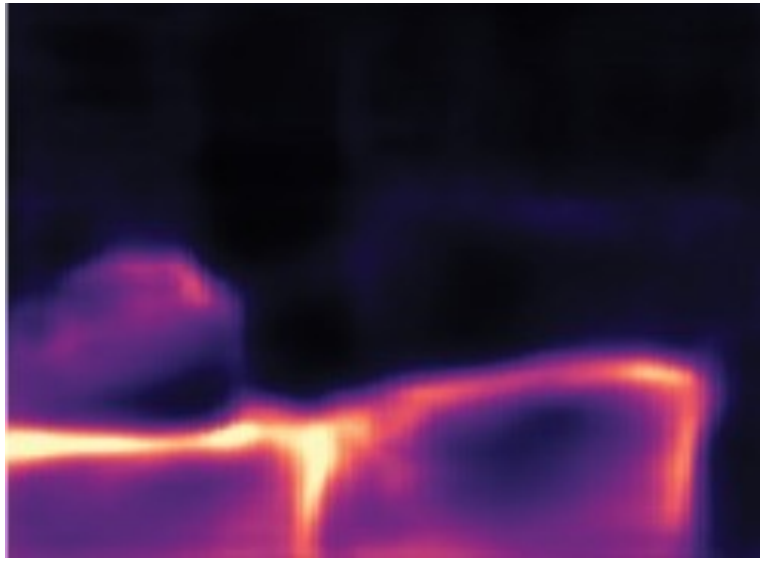
Input Image



Predicted Depth

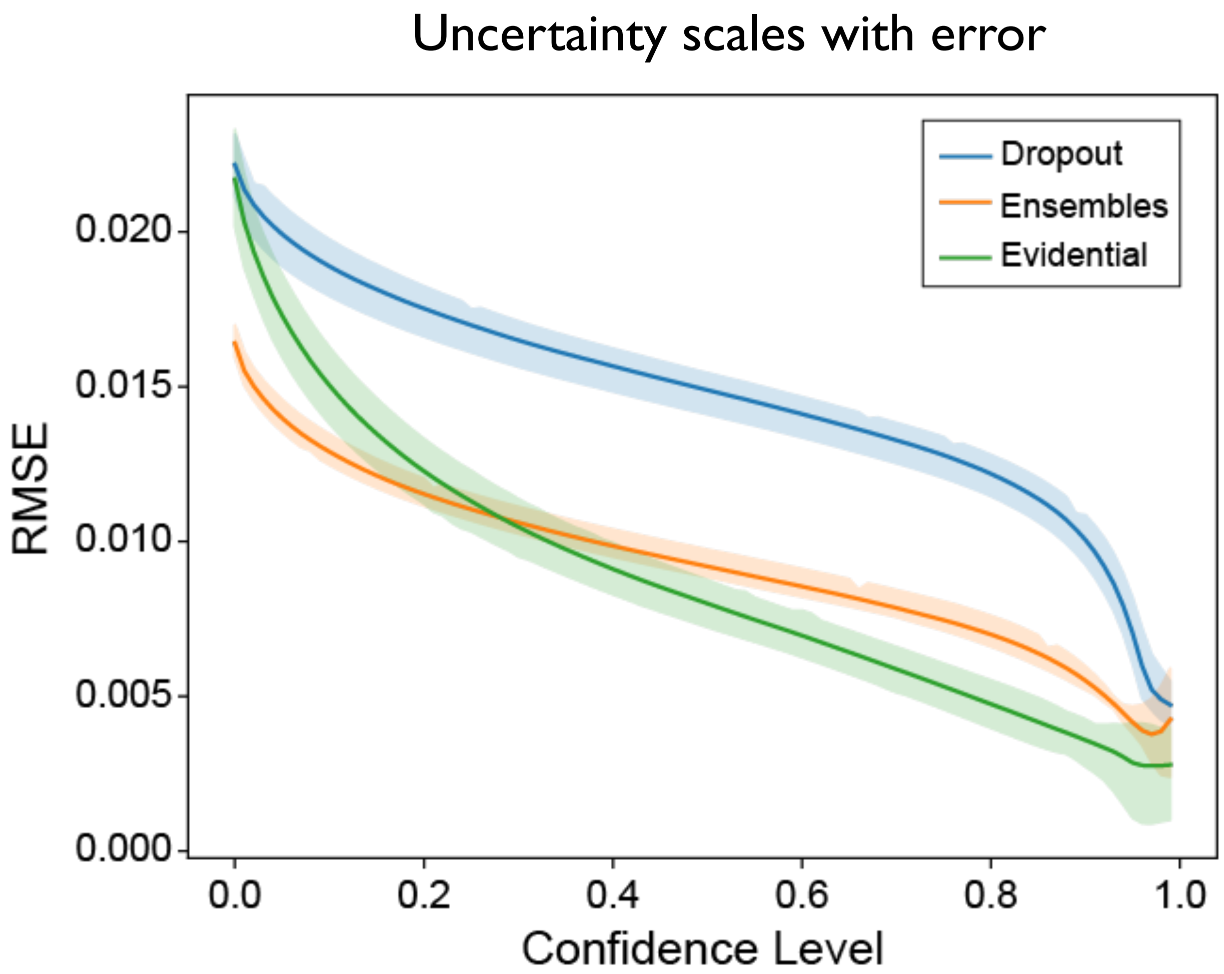


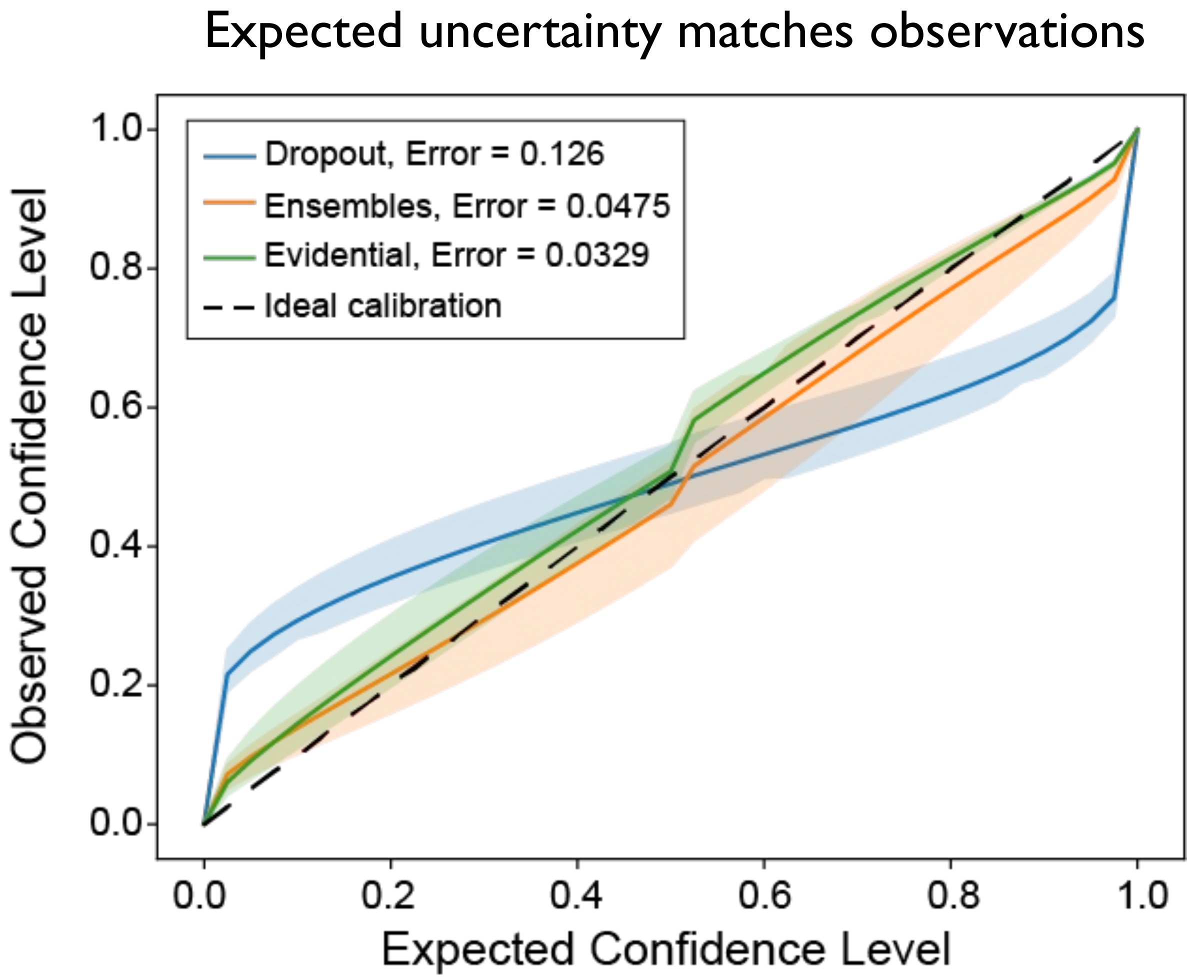
Evidential Uncertainty



Amini et al. "Deep evidential regression"
NeurIPS 2020

Evidential uncertainty is well calibrated to errors

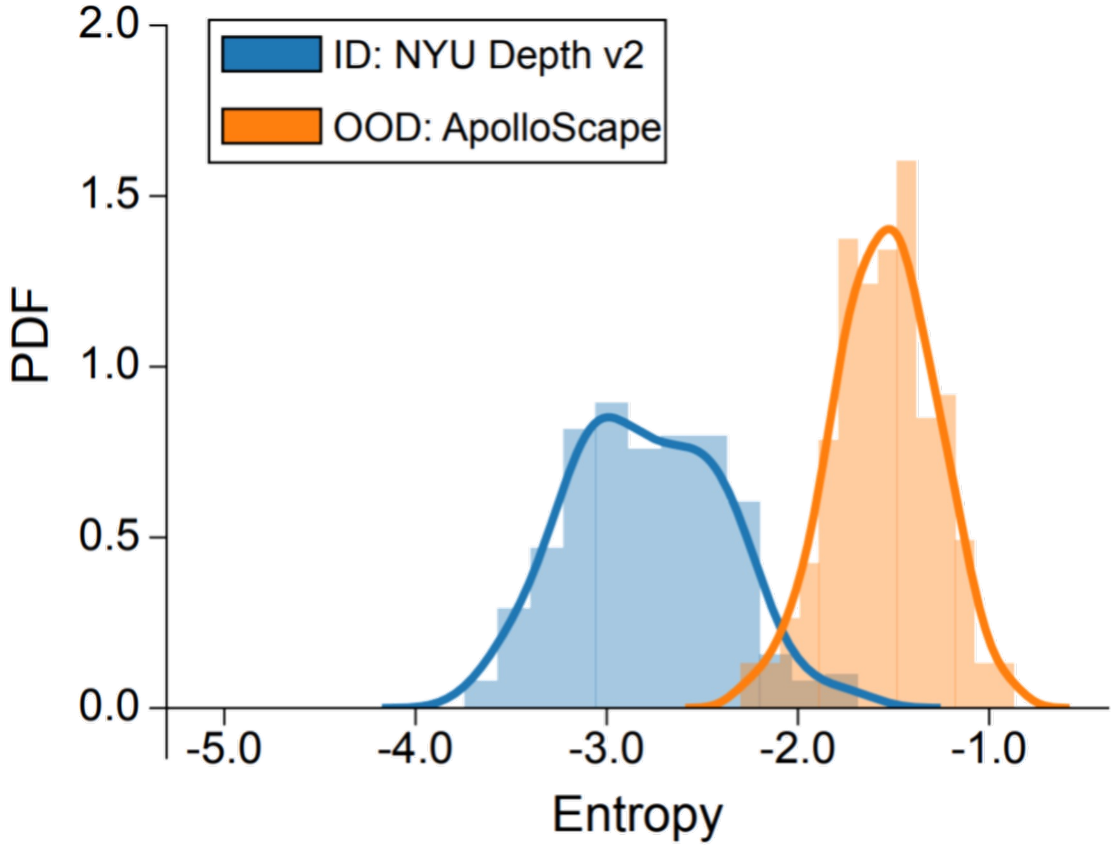


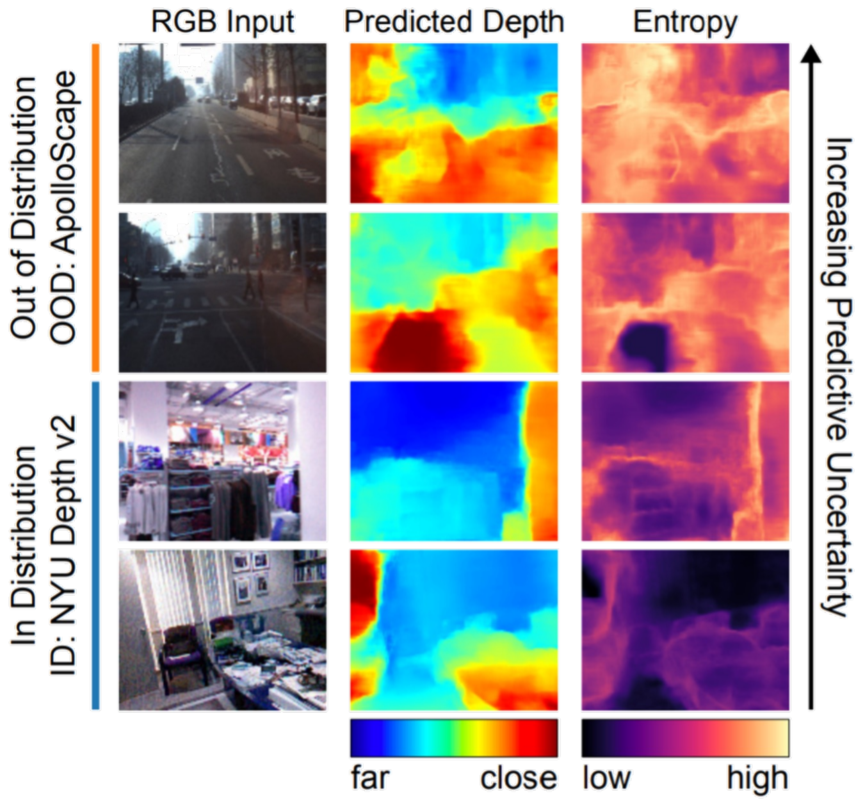


Amini et al."Deep evidential regression"
NeurIPS 2020

Calibration to errors and out-of-distribution data

Strong increase in predictive uncertainty on **out-of-distribution data**





Amini et al. "Deep evidential regression"
NeurIPS 2020

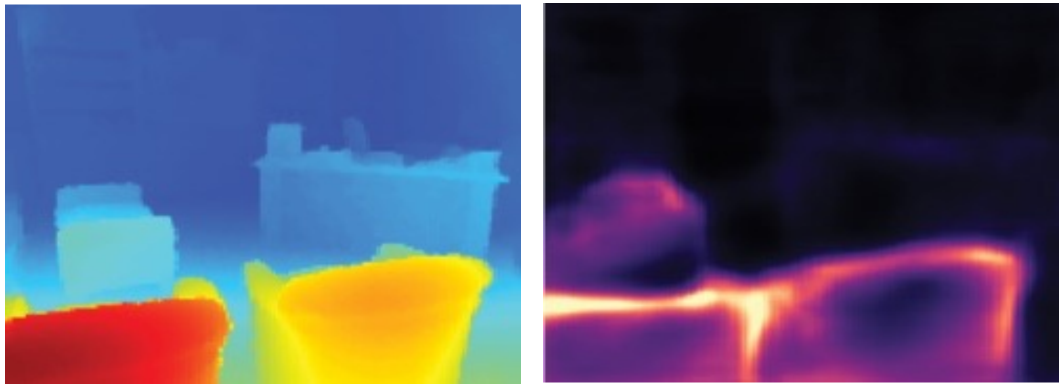
Bias and uncertainty in deep learning

Model Bias

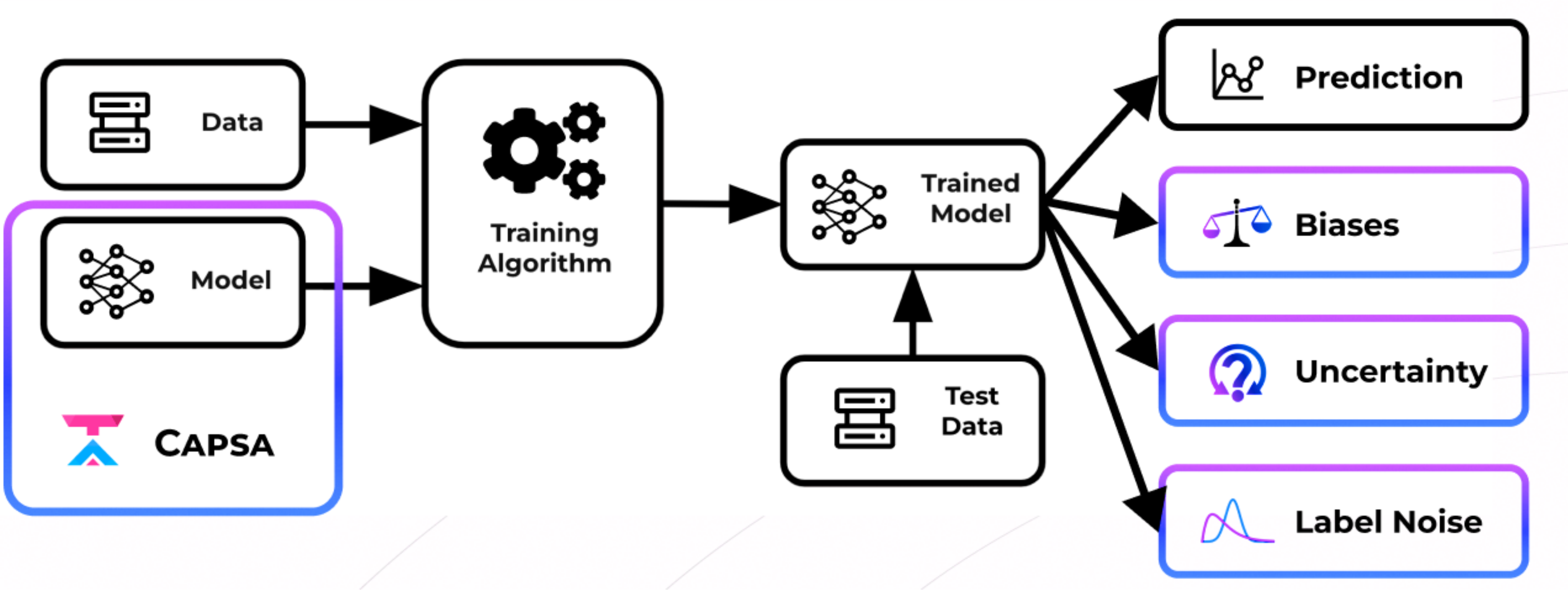
Model decision changes if it exposed to additional "sensitive" feature inputs

Uncertainty

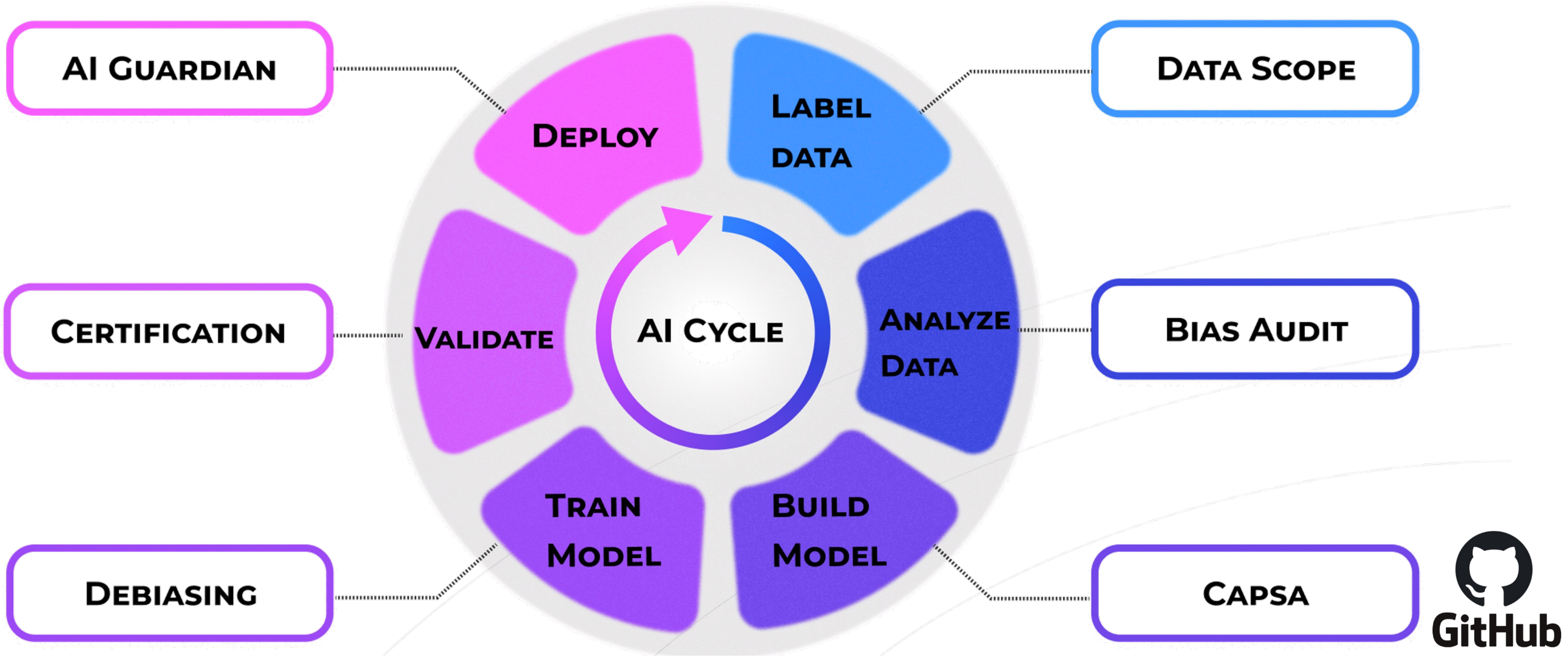
Can we train models to understand when they don't know the answer?



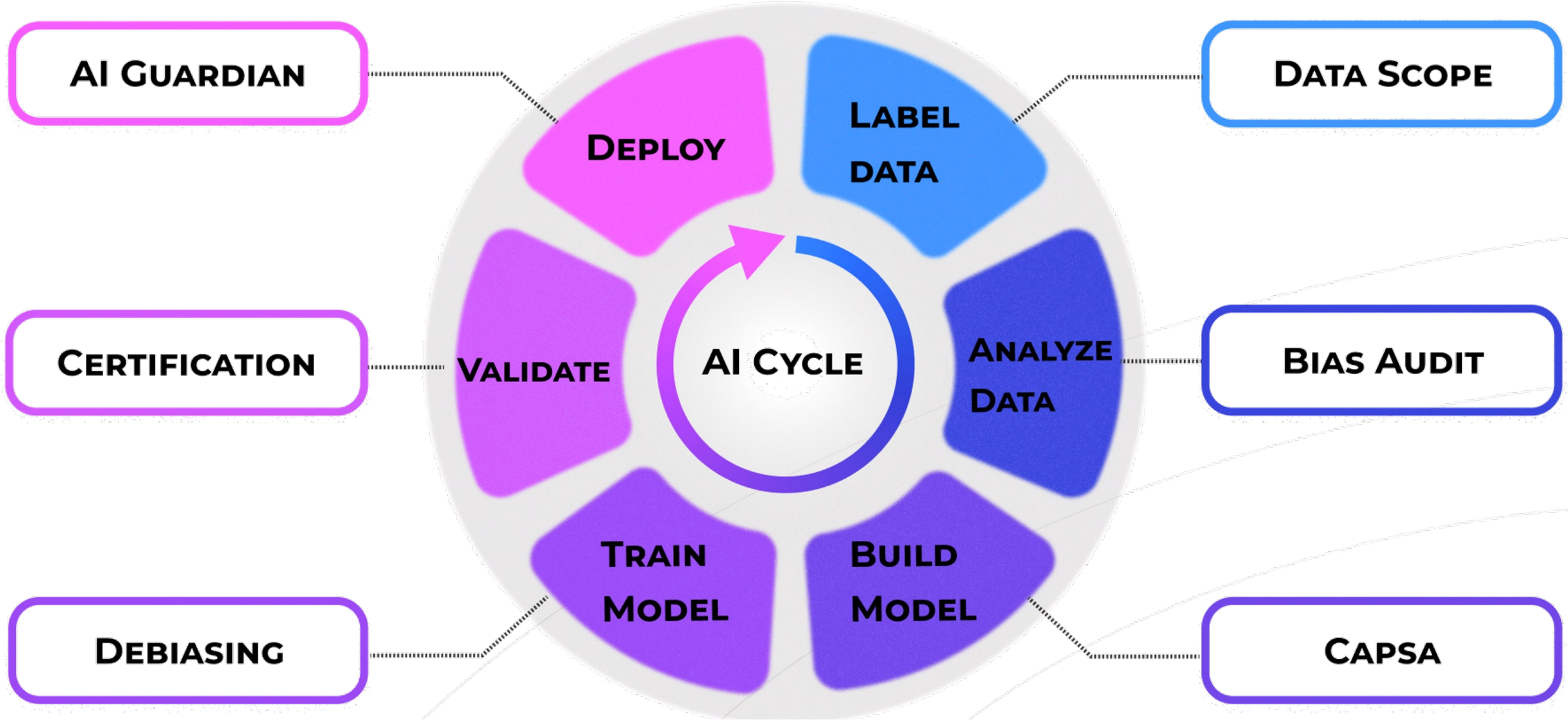
Capsa: automatically transform AI models for risk-aware learning and deployment



Themis AI: Empowering the world to create, advance, and deploy trustworthy AI



Themis AI: Empowering the world to create, advance, and deploy trustworthy AI



GitHub

We are releasing capsa FREE to the public!
Signup here:

bit.ly/themisai