Trustworthy AI

Fadwa Idlahcen

Plan

-Motivation

-What’s AI trustworthiness?

-Robustness, generalization, explainability, transparency, reproducibility, fairness and privacy preservation : definitions, examples and widely used AI models

- AI lifecycle and where trustworthiness should be checked for

- Challenges and conclusion

Motivation

- Accuracy is the only metric for AI models
- Biased, lack in user privacy protection, vulnerable to attacks…

Insight - Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin

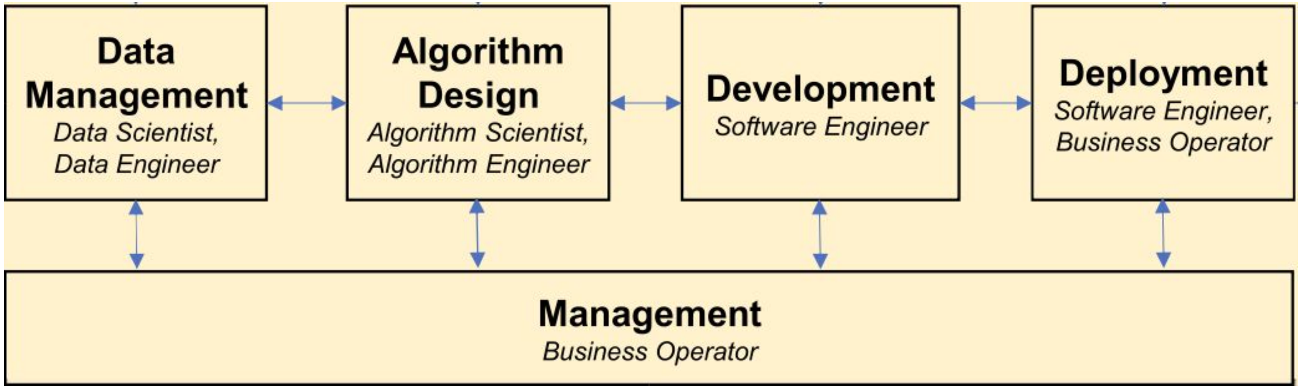
October 11, 2018 2:50 AM GMT+2 · Updated 5 years ago

Trustworthy AI

Practitioners need to address AI trustworthiness, including :

- Robustness, generalization, explainability, transparency, reproducibility, fairness and privacy preservation

Optimizing trustworthiness throughout AI lifecycle, rather than at each step.

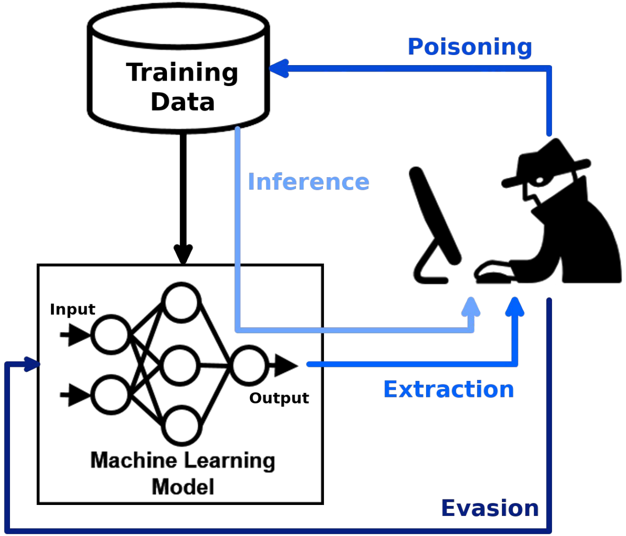


Robustness

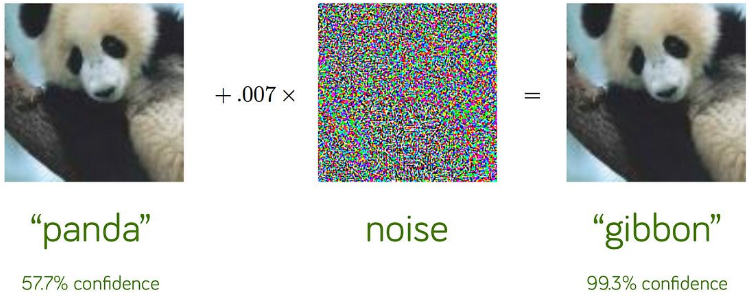
It’s not just about reliability against errors, erroneous inputs and unseen data but also safety.

3 levels :

- Data: when an AI model is trained without considering the diverse distributions of data in different scenarios.
- Algorithms : adversarial attacks
- Systems : the use of illegal inputs



Adversarial attacks



-A carefully computed example to be misclassified

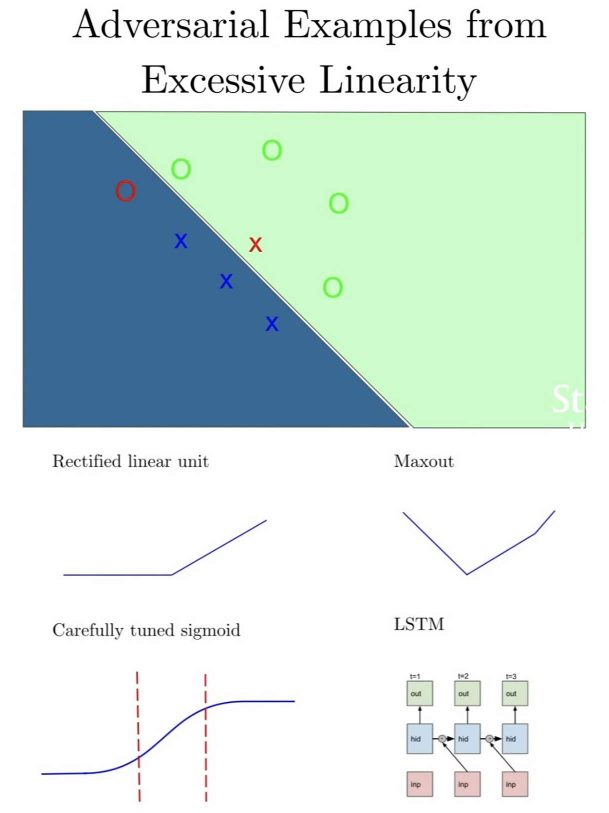
-Some use cases :

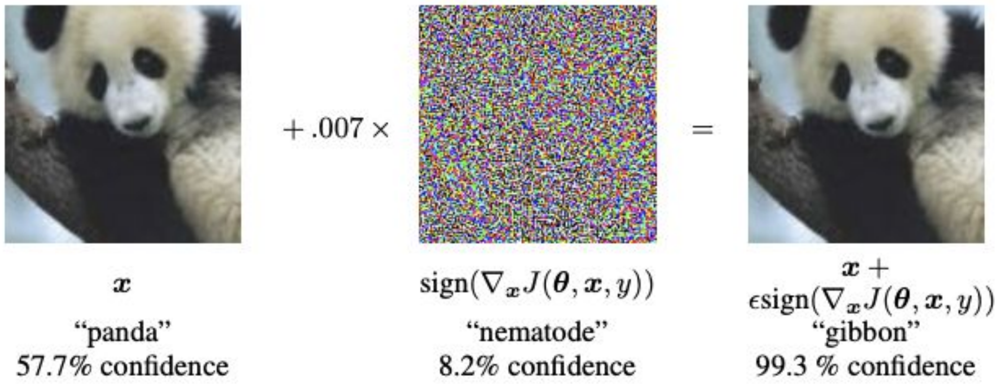
- Image classification: misclassification. E.g: autonomous vehicles, security systems, and medical imaging.

- Text generation: generating deceptive or misleading text that can manipulate sentiment analysis algorithms, spam filters, or automated content moderation systems.
- Malware evasion: designing malware that evades detection by antivirus or intrusion detection systems.

Adversarial attacks

Piecewise linearity of NN $\Longrightarrow$ Fast gradient sign method





Robustness

-Tests:

- Test datasets with various distributions are used
- Deriving a lower bound of the minimum distortion to an attack on an AI model

-Defense:

- defensive distillation: removing the gradient from the model to be protected;
- adversarial training (prone to overfitting)

⟹ Not effective!!!

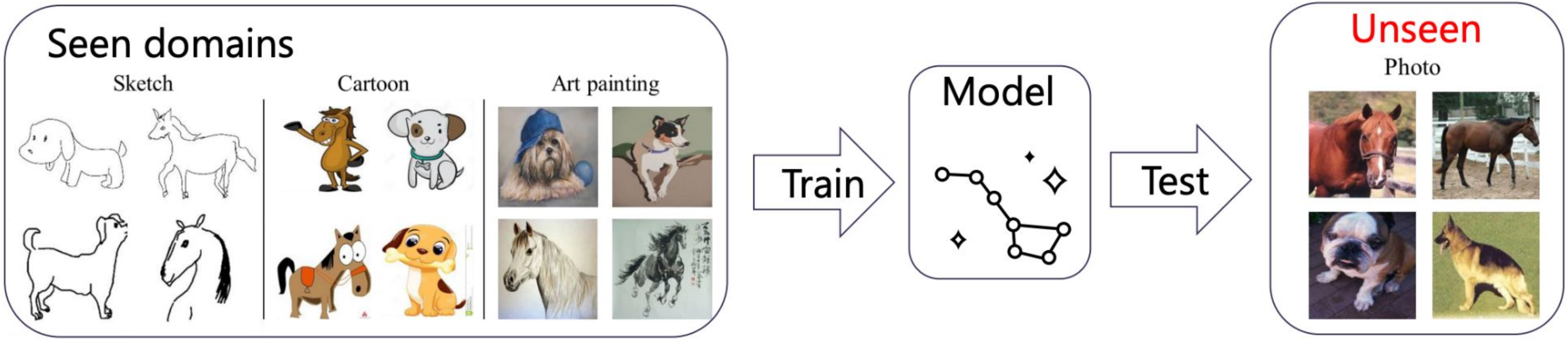
Akhilan Boopathy et al, CNN-Cert: An Efficient Framework for Certifying Robustness of Convolutional Neural Networks

Generalization

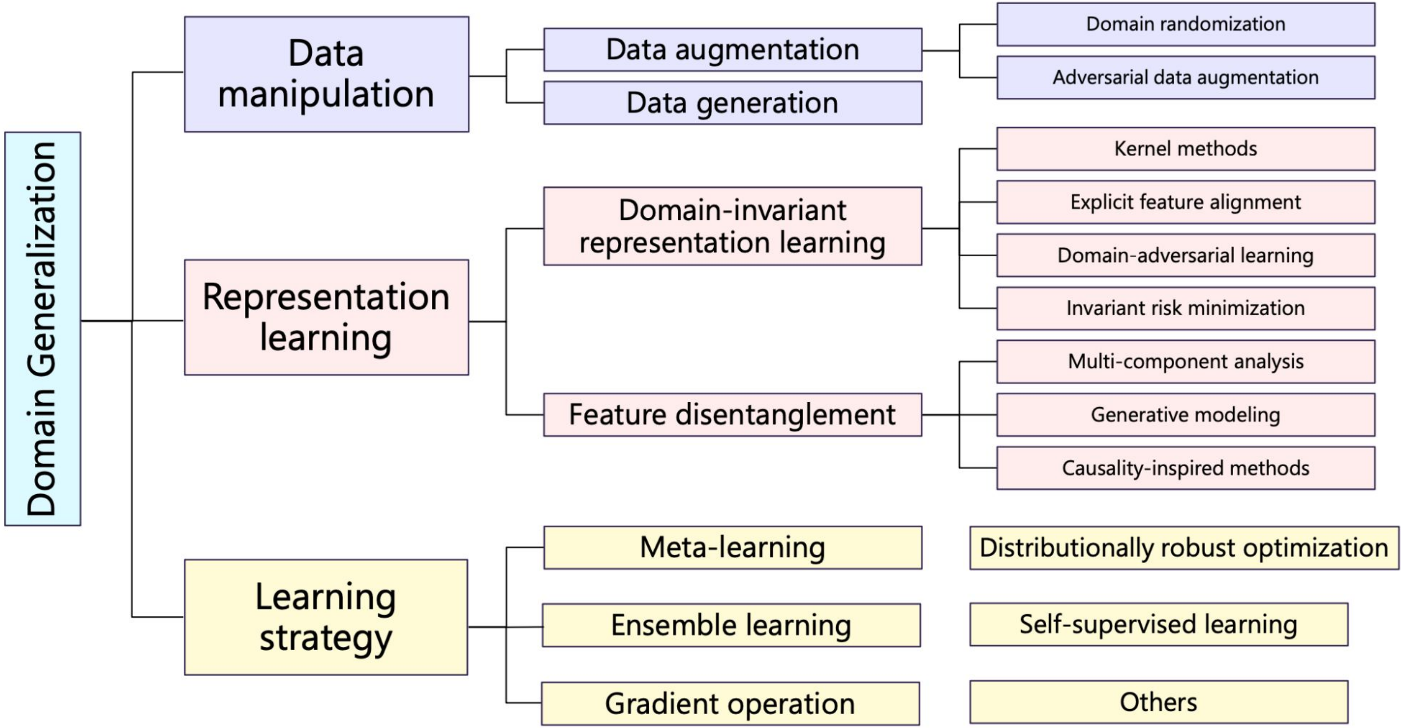
-From training limited data to accurate predictions on unseen data

-Robustness against distributional shifts = generalization problem

- Double effect : an algorithm that is robust against small perturbations has better generalization ≠ adversarial training may reduce the testing accuracy



Domain generalization



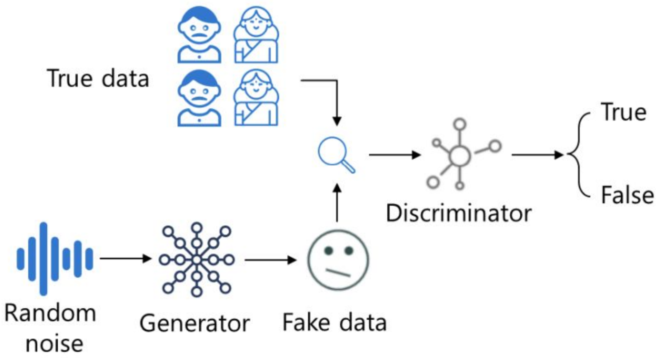
Wang et al. Generalizing to unseen domains: a survey on domain generalization. IEEE TKDE 2022.

Data manipulation

-Insufficient data in the target domain

-Adversarial data augmentation : via gradient training

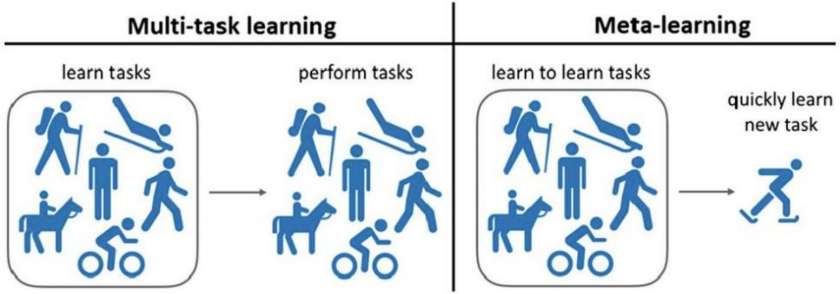
-Data generation : GANs



Learning strategy

-Meta learning : Divide domains into several tasks, then use meta-learning to learn general knowledge

- Learning to learn, or meta-learn the general knowledge
  - Instead of the original tasks, meta-learning wants to acquire knowledge about **new tasks**



Meta-knowledge acquisition

$$\phi^* = \arg\max_{\phi} \log P\left(\phi \mid \mathcal{D}_{\text{src}}\right)$$

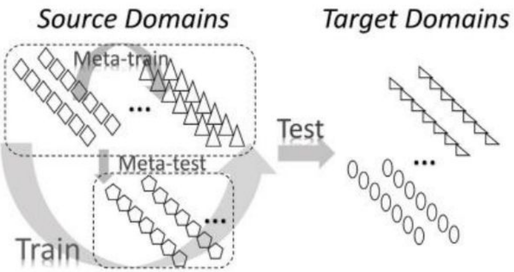
Meta-knowledge validation

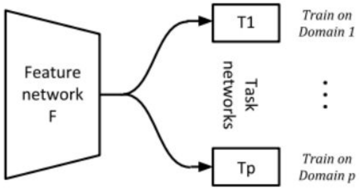
$$\theta^{*(i)} = \arg\max_{\theta} \log P\left(\theta \mid \phi^*, \mathcal{D}_{\text{tar}}^{\text{train }(i)}\right)$$

Huisman M, Van Rijn J N, Plaat A. A survey of deep meta-learning[J]. Artificial Intelligence Review, 2021, 54(6): 4483-4541.

Learning strategy

- How to adopt meta-learning for DG?
  - Key: Old tasks to new tasks in meta-learning → Old domains to new domains
- MLDG: Meta-learning for DG
- MetaReg: meta-learning for regularization





Algorithm 1 Meta-Learning Domain Generalization

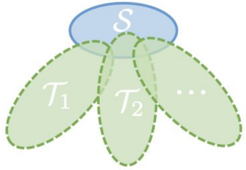
procedure MLDG
  Input: Domains $\mathcal{S}$
  Init: Model parameters Θ. Hyperparameters α, β, γ.
  for ite **in** iterations **do**
    Split: $\bar{\mathcal{S}}$ and $\breve{\mathcal{S}} \leftarrow \mathcal{S}$
    Meta-train: Gradients $\nabla_\Theta = \mathcal{F}'_\Theta(\bar{\mathcal{S}}; \Theta)$
    Updated parameters $\Theta' = \Theta - \alpha \nabla_\Theta$
    Meta-test: Loss is $\mathcal{G}(\breve{\mathcal{S}}; \Theta')$.
    Meta-optimization: Update Θ

$$\Theta = \Theta - \gamma \frac{\partial(\mathcal{F}(\bar{\mathcal{S}}; \Theta) + \beta \mathcal{G}(\breve{\mathcal{S}}; \Theta - \alpha \nabla_\Theta))}{\partial \Theta}$$

  end for
end procedure

- Li D, Yang Y, Song Y Z, et al. Learning to generalize: Meta-learning for domain generalization. AAAI 2018.
- Balaji Y, Sankaranarayanan S, Chellappa R. Metareg: Towards domain generalization using meta-regularization. NeurIPS 2018.

Data manipulation - Meta-learning

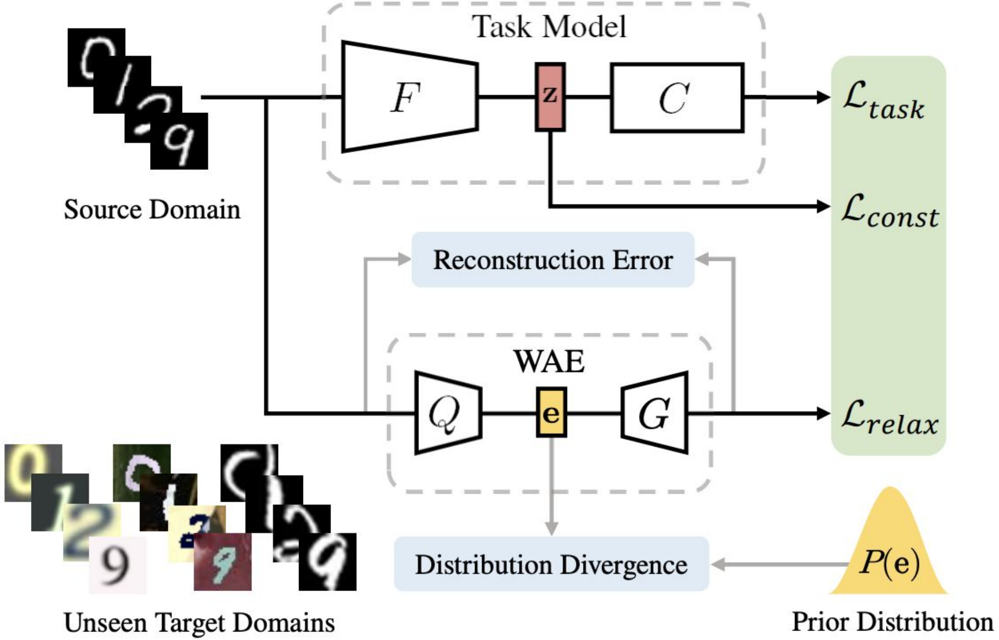


$\mathcal{S} \cap \mathcal{T}_i < \mathcal{T}_i$
(c) $\mathcal{T}_i \cup \mathcal{T}_j > \mathcal{T}_{i/j}$

$\mathcal{S}$: Source domain(s)
$\mathcal{T}$: Target domain(s)

-VAE for data generation

F : feature extractor; C : classifier ; z : latent representation of x ; L_task : classification loss; L_const : worst-case loss; L_relax : relaxation loss



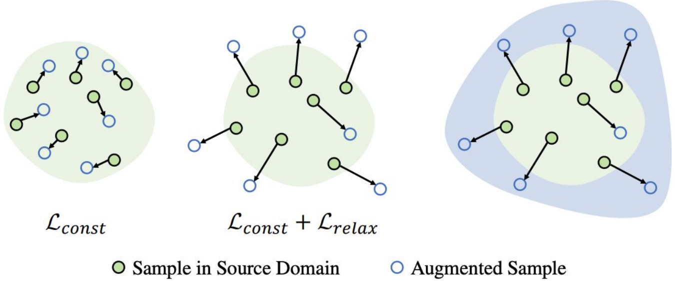


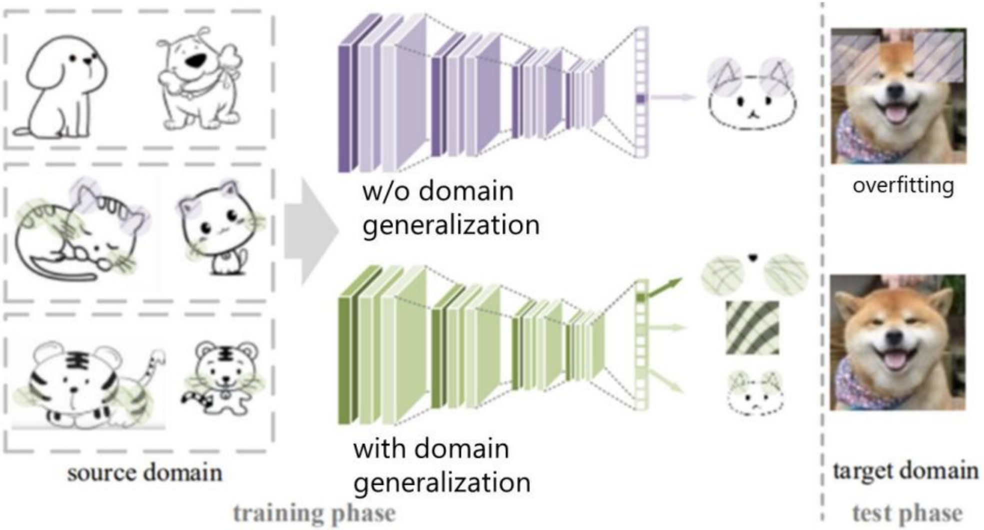
Figure 3. Motivation of $\mathcal{L}_{\text{relax}}$. **Left:** The augmented samples may be close to the source domain if applying $\mathcal{L}_{\text{const}}$. **Middle:** We expect to create out-of-domain augmentations by incorporating $\mathcal{L}_{\text{relax}}$. **Right:** This would yield an enlarged training domain.

Qiao et al. Learning to Learn Single Domain Generalization. CVPR 2020.

Representation learning

-Learning invariant features

$$\min_{f,g} \mathbb{E}_{\mathbf{x},y}\ell(\underbrace{f}_{\text{Classifier}}(\underbrace{g}_{\text{Feature}}(\mathbf{x})), y) + \lambda\underbrace{\ell_{\text{reg}}}_{\text{Regularization}}$$



Representation learning

How?

- kernel based models:

  transfer component analysis

- Domain adversarial learning

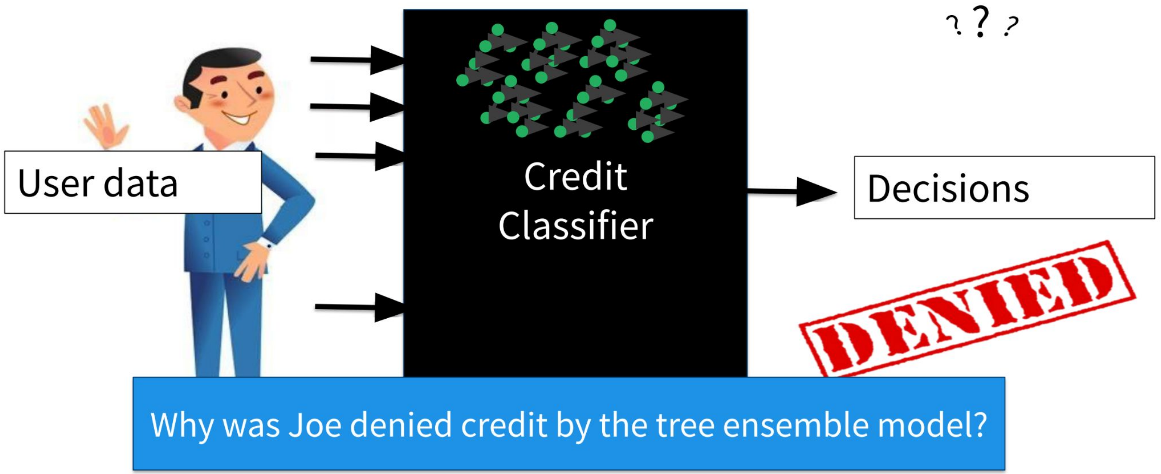
…

Explainability

-Explainability: understanding how an AI model makes decisions

- Model explainability by design
- Post hoc model explainability

-Explainability vs Interpretability



Explainable ML

Design

-self-explainable models :

KNN, linear/logistic regression, decision trees/rules, and probabilistic graphical models

⇒ Complex structure = unexplainable

-Hybrid combinations of self-explainable models and black-box models

Post hoc

-Explainer approximation

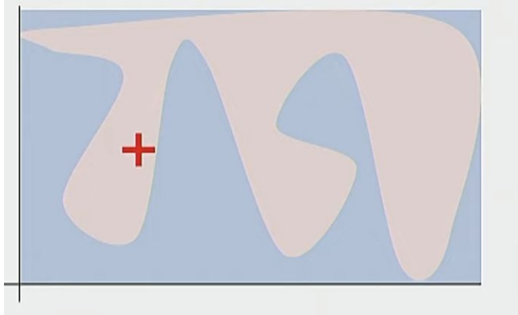
-Feature based explanation

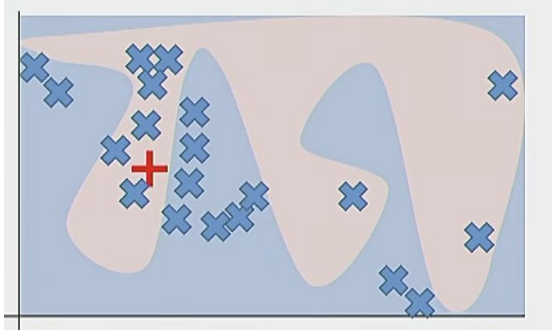
-Two types : local and global

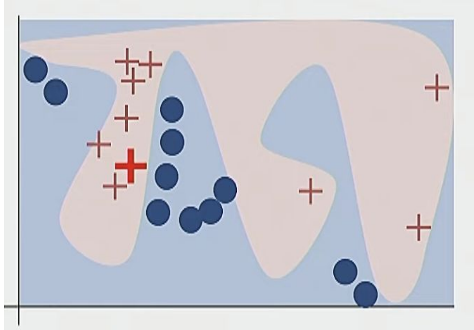
Local explainability

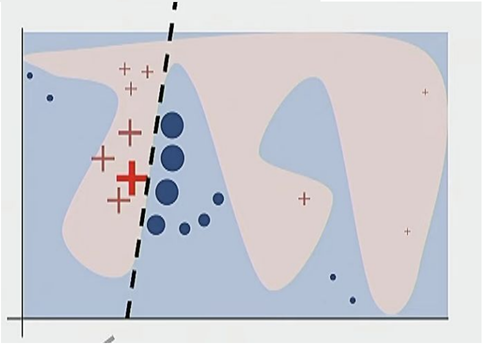
-Feature importance: LIME : explaining individual predictions at a time

- sample points around xi
- use model to predict labels for each sample
- weigh samples according to distance to xi
- learn simple linear model on weighted samples
- use simple linear model to explain





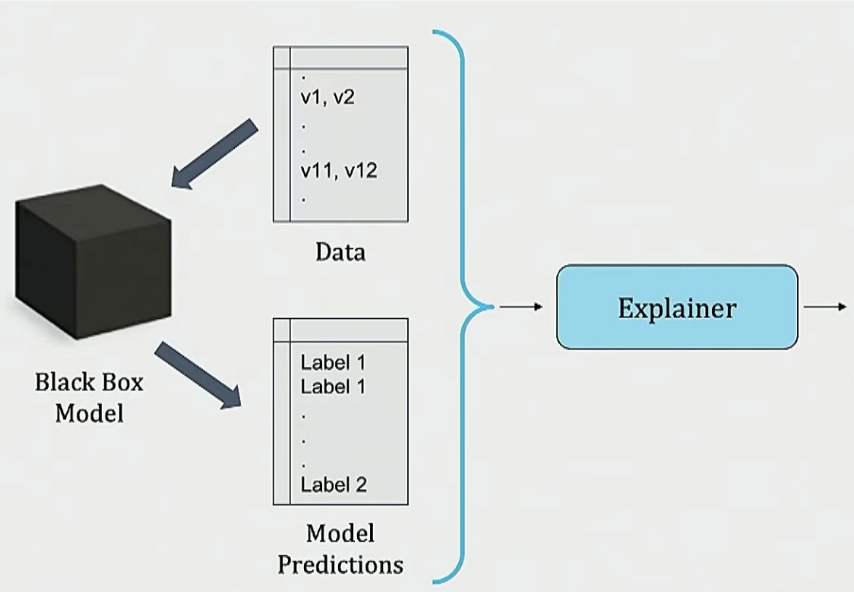




Global explainability

-Combination of local explanations

-Model Distillation



If Exercise =Yes **and** Smoking =No:

- **If** Rapid-Weight-Gain =Yes **and** Tiredness =Yes **and** Melancholy =Yes **and** Insomnia =Yes **and** Age <50, **then** Depression
- **If** Tiredness =Yes **and** Melancholy =Yes **and** Age ≥ 50, **then** Depression
- **If** Tiredness =No **and** Melancholy =No, **then** Healthy

If Smoking =Yes:

- **If** Rapid-Weight-Gain =Yes **and** Melancholy =Yes, **then** Depression
- **If** Tiredness =No **and** Insomnia =No **and** Melancholy =No **and** Rapid-Weight-Gain =No, **then** Healthy
- **If** Insomnia =Yes **and** Past-Depression =Yes **and** Tiredness =Yes, **then** Depression

Default:

- **If** Past-Depression =Yes **and** Tiredness =Yes **and** Melancholy =Yes, **then** Depression
- **If** Past-Depression =No **and** Rapid-Weight-Gain =Yes **and** Tiredness =No **and** Melancholy =Yes, **then** Depression
- **If** Family-Depression =Yes **and** Age ≥ 50 **and** Male =No **and** Tiredness =Yes, **then** Depression

Stanford

Transparency

-Transparency considers AI as a software system, and seeks to disclose information regarding its entire lifecycle

Documentation

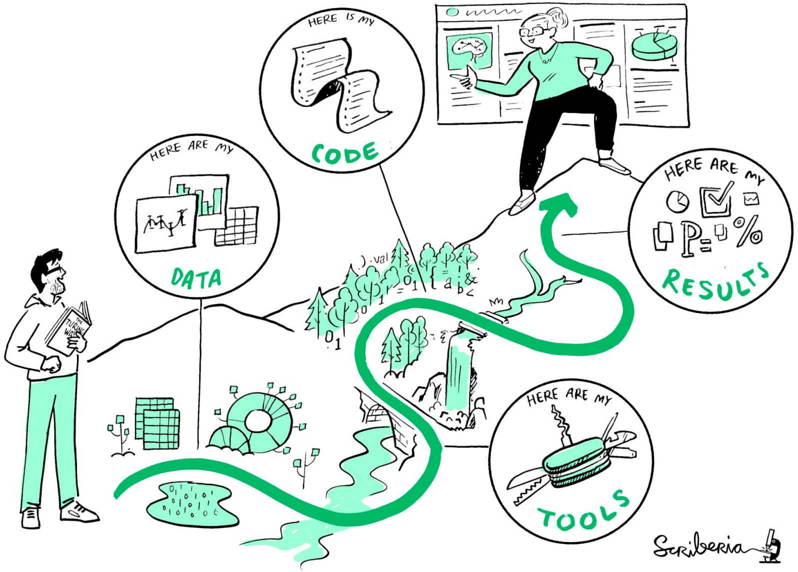
Auditing

Information Sharing

Reproducibility

-Essential step to verify AI research

-Should be considered over the entire lifecycle (data, methods, and experiments)



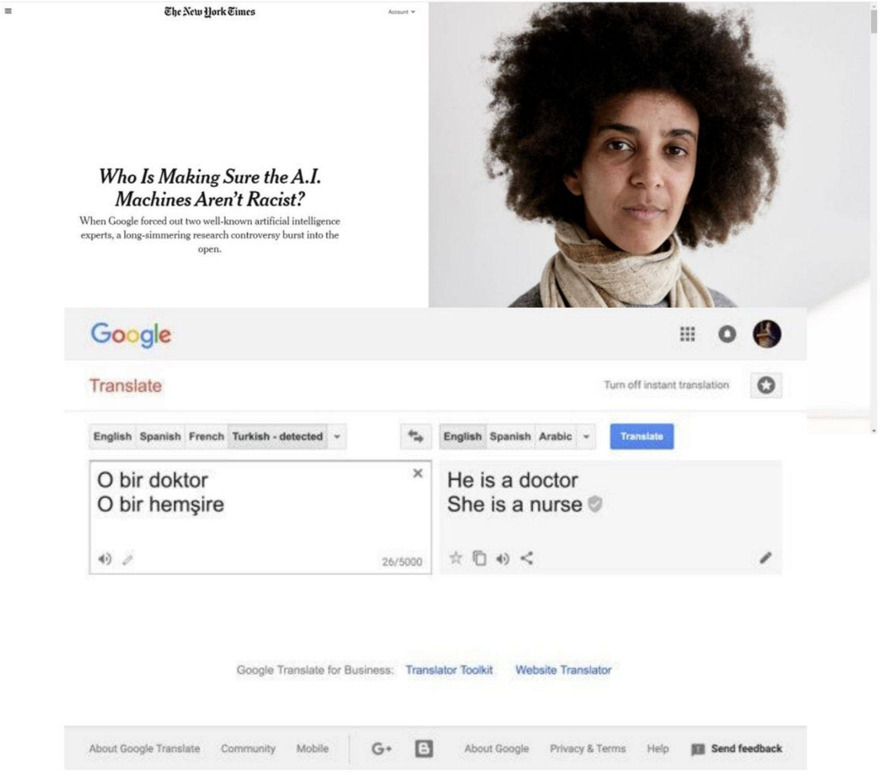
-ChatGPT-3 and reproducibility issues

Fairness

-Hiring

-Financial risk assessment

-Face identification



https://web.stanford.edu/class/cs329t/2022/slides/part1-week1-overview.pptx.pdf

Common Biases

MIT 6.S191: AI Bias and Fairness Lecture

Improving Fairness

-Bias mitigation : removing problematic signal

-Inclusion: adding signal for desired features (counterfactual data augmentation)

Example : Language modelling

Hidden features; counterfactual effect of adding a bias

Adversarial Multi-Task Learning to Mitigate Bias

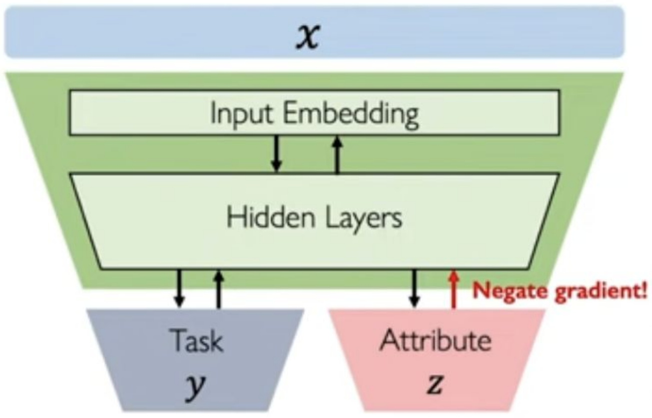
Setup: specify attribute z for which we seek to mitigate bias. Jointly predict output y and z.

Two discriminator output heads:

1. Target / class label y
2. Sensitive attribute z

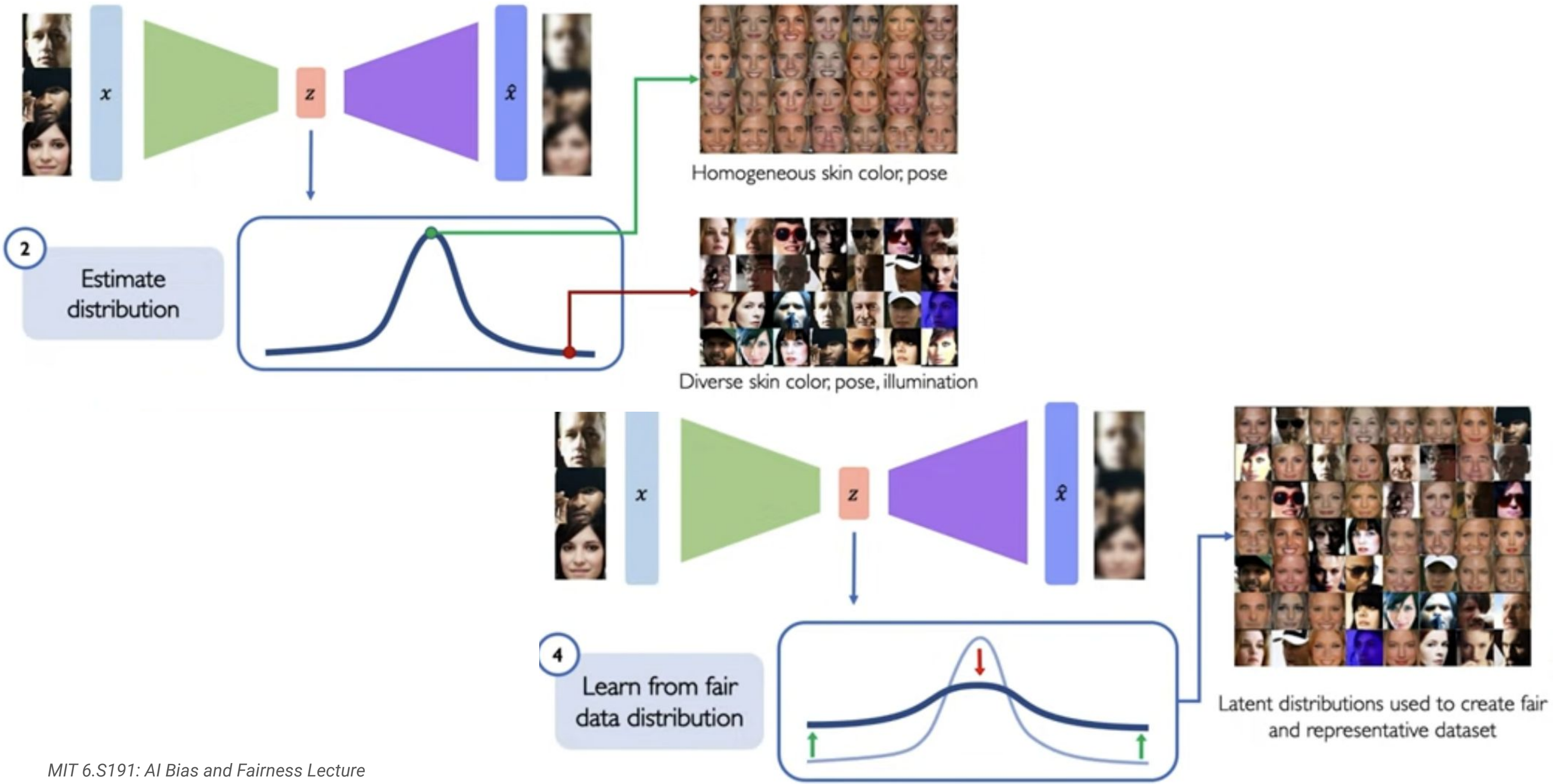
Train adversarially:

1. Predict sensitive attribute z
2. Negate gradient for z head
3. "Remove" effect of z on task decision



Jointly predict output label y and sensitive attribute z to remove from decision

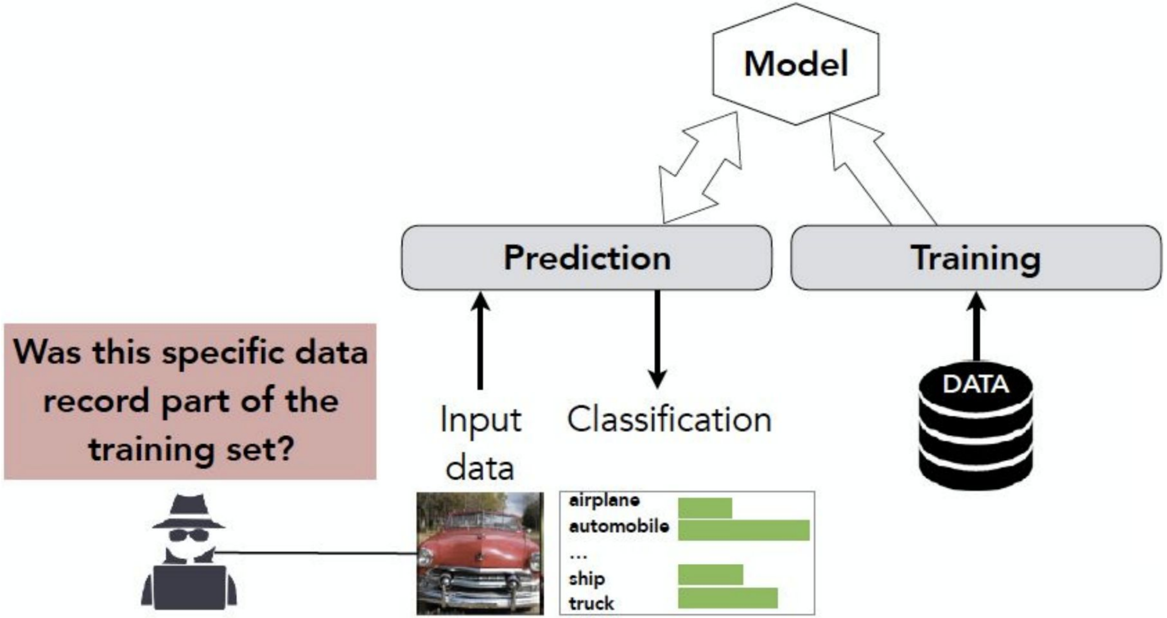
Mitigating Bias through Learned Latent Structure



Privacy Protection

-Protection against information leakage

Membership inference attack



https://web.stanford.edu/class/cs329t/2022/slides/part1-week1-overview.pptx.pdf

AI Lifecycle

Lifecycle	Approaches
Data Preparation	Data Collection
	Data Preprocessing
Algorithm Design	Adversarial Robustness
	Explainability ML
	Model Generalization
	Algorithmic Fairness
	Privacy Computing
Development	Functional Testing
	Performance Benchmarking
	Simulation
	Formal Verification
Deployment	Anomaly Monitoring
	Human-AI Interaction
	Fail-Safe Mechanism
	Hardware Security
Management	Documentation
	Auditing
	Cooperation
Workflow	MLOps

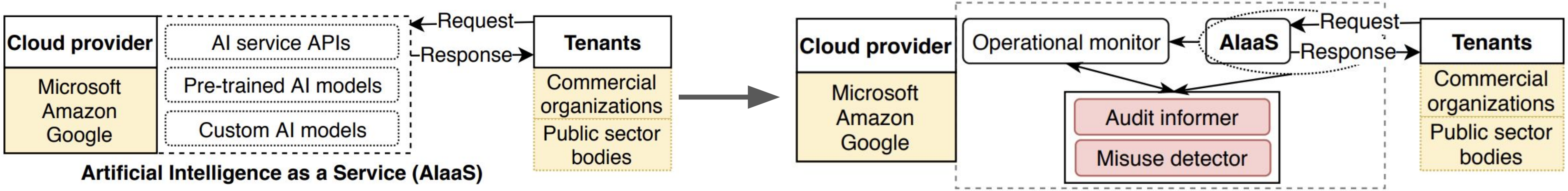
Deployment

-Anomaly monitoring :

police use of facial recognition in Hong Kong

‘Deepfakes’

-Microsoft limits the request rate to their Face API, Amazon prevents more than 100 faces from being detected in single image



Technical indicator for vision and face services	*Potential implications*
High request rate for face detection	Population surveillance
Large number of faces in an image/video	Population surveillance
Large number of different faces are analysed	Population surveillance
Large number of identification attempts for particular individual(s)	Privacy threats to an individual
Detection of ‘black-listed’ objects	Controversial application

Seyyed Ahmad Javadi et al. Monitoring Misuse for Accountable ‘Artificial Intelligence as a Service’,AIES '20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2020

Deployment

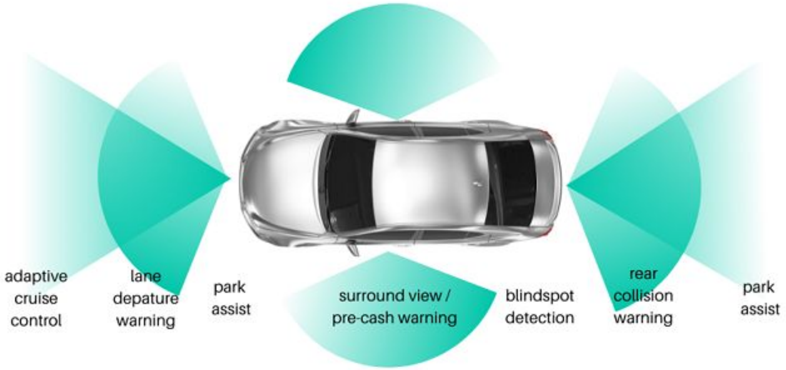
-Human-AI interaction :

User interface: visualization of ML models and interactive parameter-tuning

Human intervention :

- participating in decisions : advanced driver-assistance system (ADAS)
- monitoring failure

-Fail-Safe Mechanisms : if an AI system is causing harm



MLOps

-similar to DevOps

-A start point to build the workflow for trustworthy AI

-Properties:

- Aligned principles of trustworthiness
- Extensive management of artifacts
- Continuous feedback loops
- Close collaboration between interdisciplinary roles

Challenges and Conclusion

-Shift of focus from performance-driven AI to trust-driven AI

-The good and the bad : Large-scale Pre-trained Models

-Limitations in current evaluations of trustworthiness

-End-User Awareness of the Importance of AI Trustworthiness

-Inter-disciplinary and International Cooperation

Thank you.