CENG7880
Trustworthy and Responsible AI

Instructor: Sinan Kalkan

(https://ceng.metu.edu.tr/~skalkan)

For course logistics and materials:
https://metu-trai.github.io

MIDDLE EAST TECHNICAL UNIVERSITY
DEPARTMENT OF COMPUTER ENGINEERING

Today

- About the instructor
- Overview of the course
- Overview of the fundamental concepts in TRAI

About the instructor

Sinan Kalkan

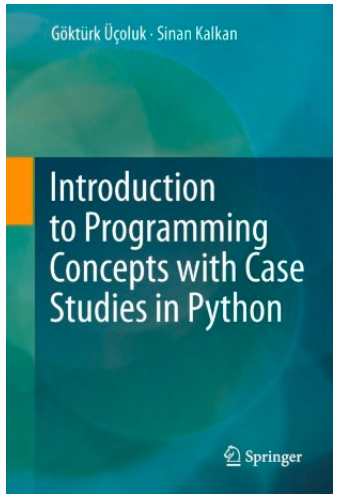
Twitter: @kalkansinan
Email: skalkan@metu.edu.tr
Web: ceng.metu.edu.tr/~skalkan

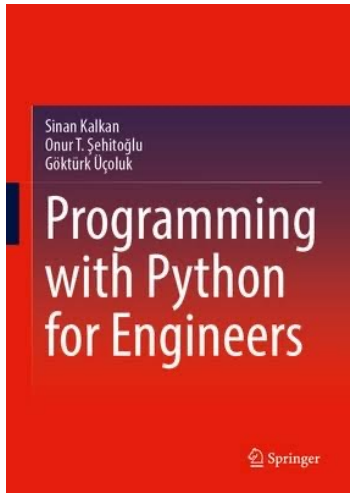
Education:

- BSc: Dept. of Computer Engineering, METU
- MSc: Dept. of Computer Engineering, METU
- PhD: Dept. of Informatics, University of Gottingen (Germany)

Research Interests:

- Computer Vision
- Machine Learning
- Cognitive Robotics





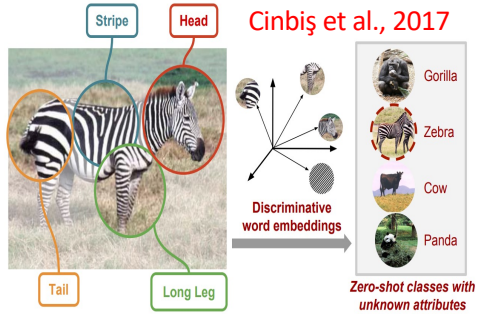
- Bilim Akademisi Genç Bilim İnsanı Ödülü (BAGEP), 2020.
- Outstanding Paper Award by IEEE Transactions on Cognitive and Developmental Systems, 2019.
- Yılın Tezi Ödülü, ODTÜ, 2018.
- Yılın Eğitimcisi Ödülü, ODTÜ, 2014.

ODTÜ METU ImageLab

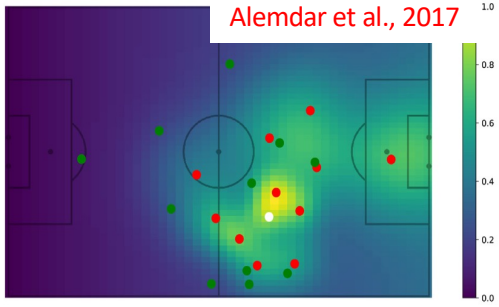
https://image.ceng.metu.edu.tr/

https://twitter.com/metu_imagelab

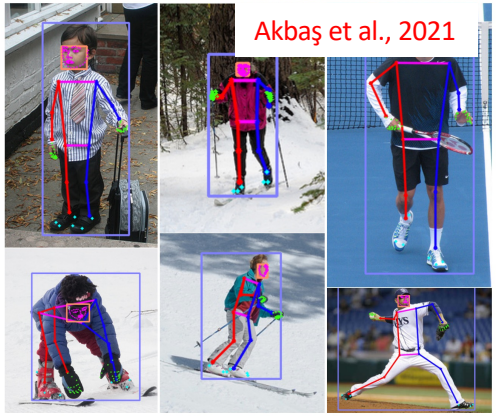
- Computer Vision
- Pattern Recognition
- Artificial Intelligence



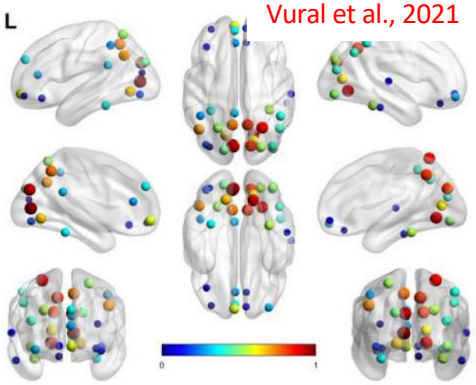
Cinbiş et al., 2017



Alemdar et al., 2017



Akbaş et al., 2021



Vural et al., 2021

Principal Investigators

Asst. Prof. Emre Akbas

Prof. Fatos Yarman Vural
Founder

Asst. Prof. Gokberk Cinbis

Asst. Prof. Hande Alemdar

Assoc. Prof. Sinan Kalkan

Graduate Students

Abdullah Alchihabi
PhD Student

Baran Baris Kivilcim
PhD Student

Baris Can Cam
MSc Student

Baris Nasir
PhD Student

Bedrettin Çetinkaya
MSc Student

Berkan Demirel
PhD Student

Duygu Kabakçı
MSc Student

Emel Varol
MSc Student

Erbil Yakiskan
MSc Student

Eren Sener
MSc Student

Erkin Eryol
PhD Student

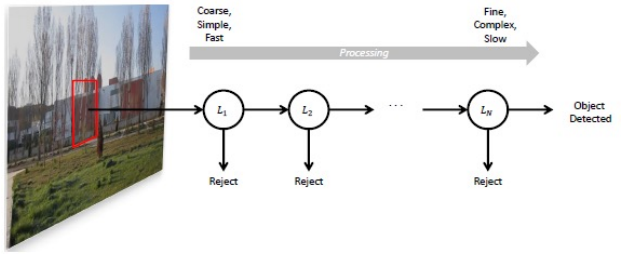
Erşan Vural Zorlu
MSc Student

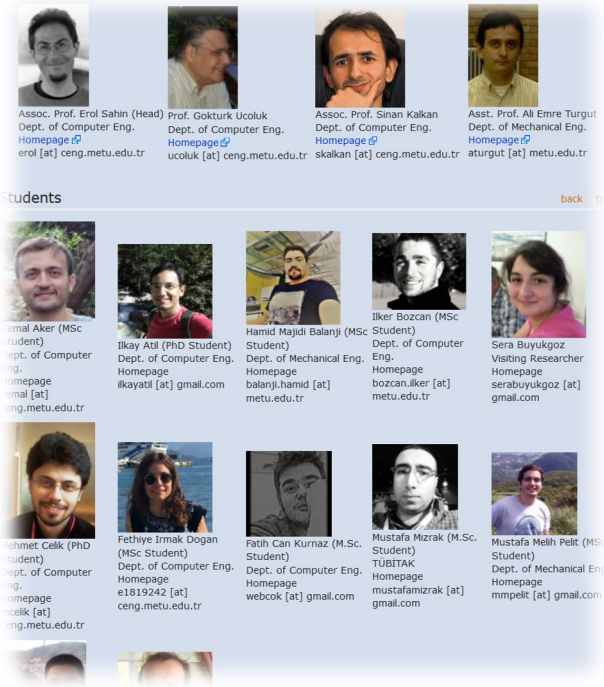
KOVAN researchlab

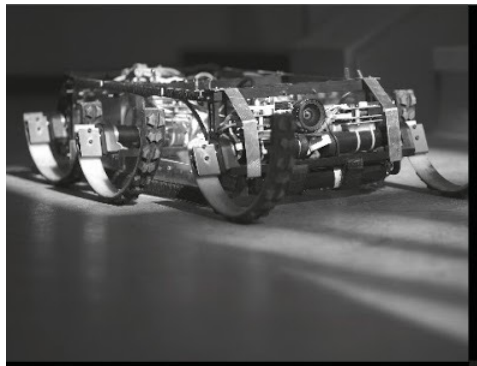
http://kovan.ceng.metu.edu.tr/

https://twitter.com/MetuKovan

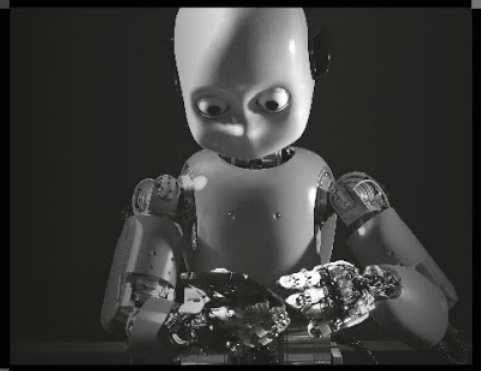
- Swarm Robotics
- Cognitive Robotics
- Computer Vision & Machine Learning







ODTÜ



ROMER

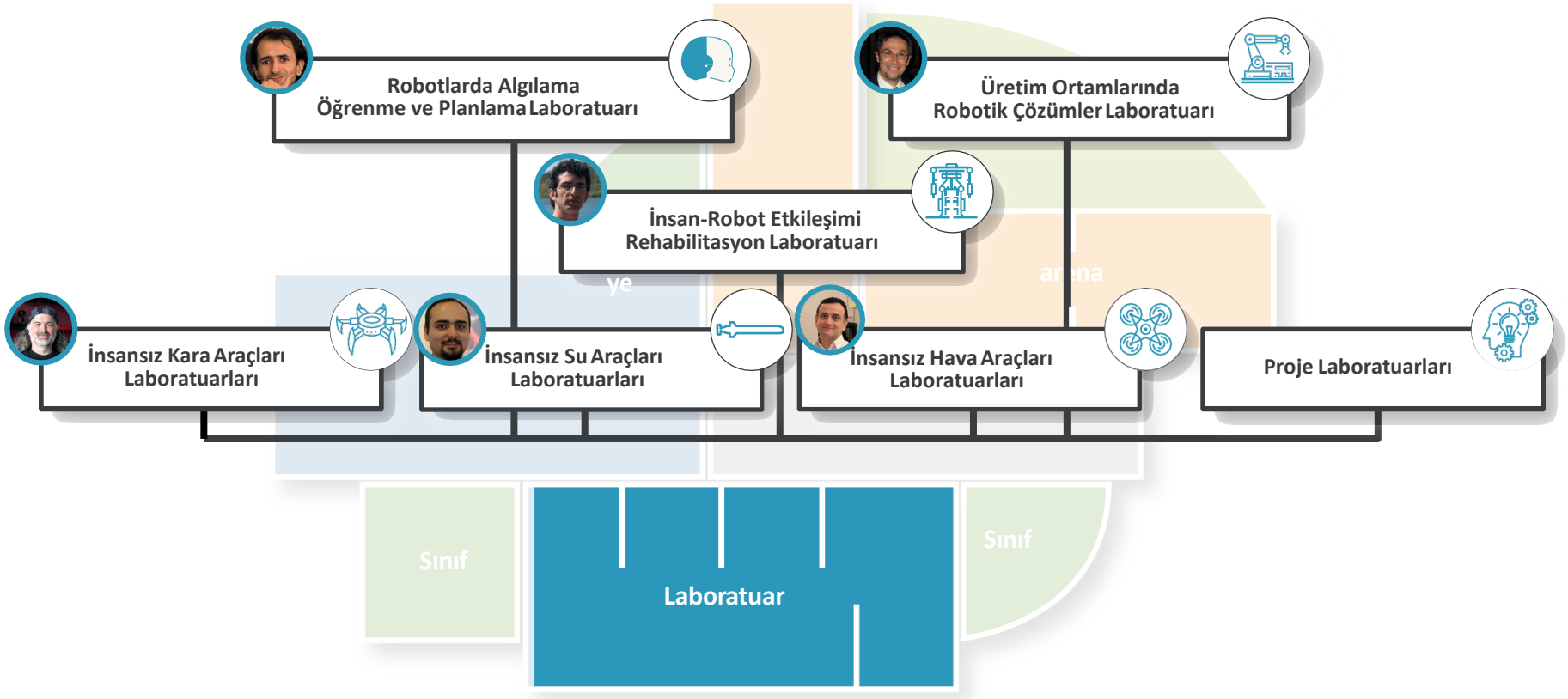
Robotik Teknolojileri Araştırma, Geliştirme ve Eğitim Merkezi

https://romer.metu.edu.tr/

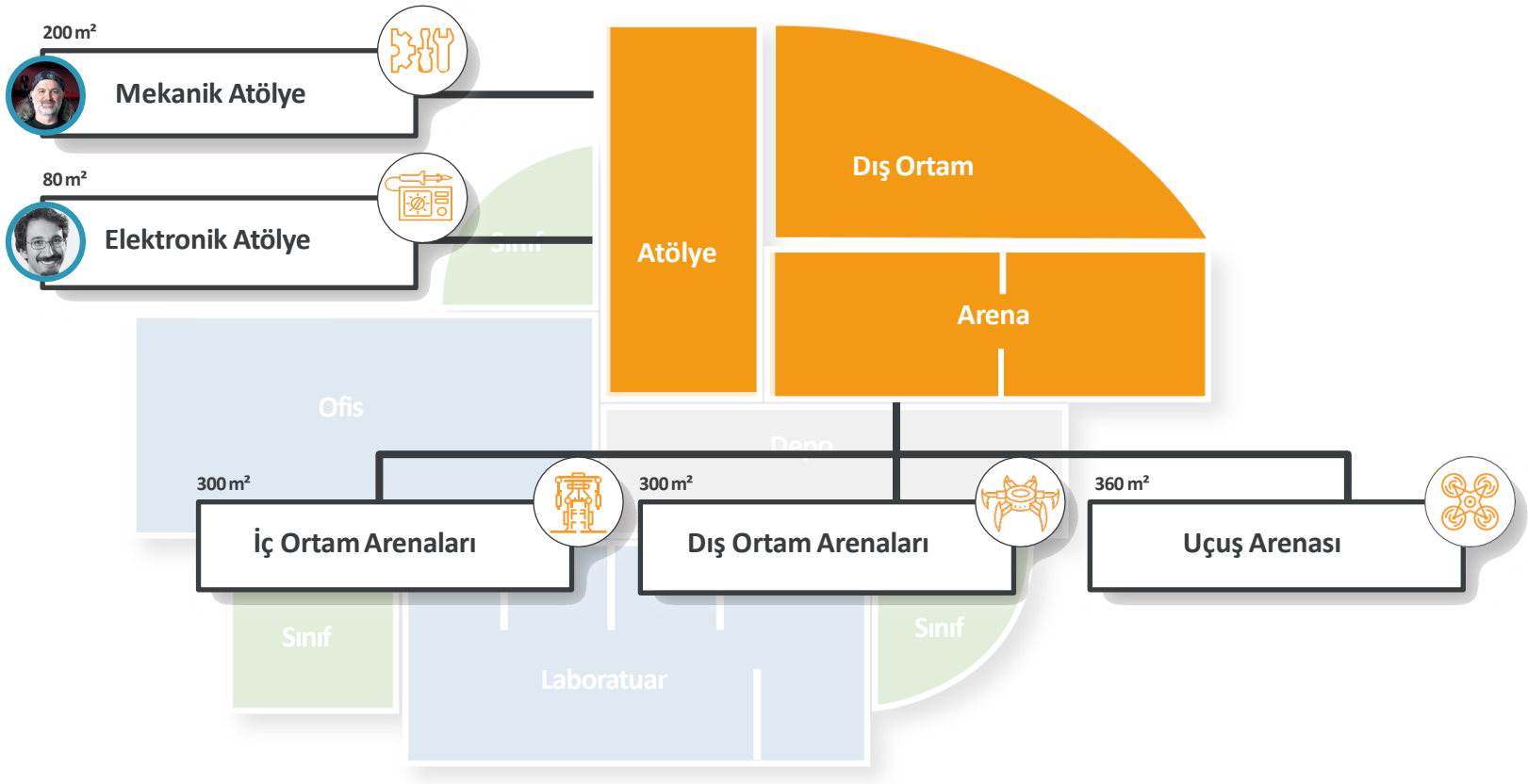
https://twitter.com/MetuRomer

12/2019 – 12/2022
~ 2 million euros

ROMER: Research Labs



ROMER: Workshops and Arenas



My Research:
Modeling Context

Ilker Bozcan | Irmak Dogan | Hande Celikkanat

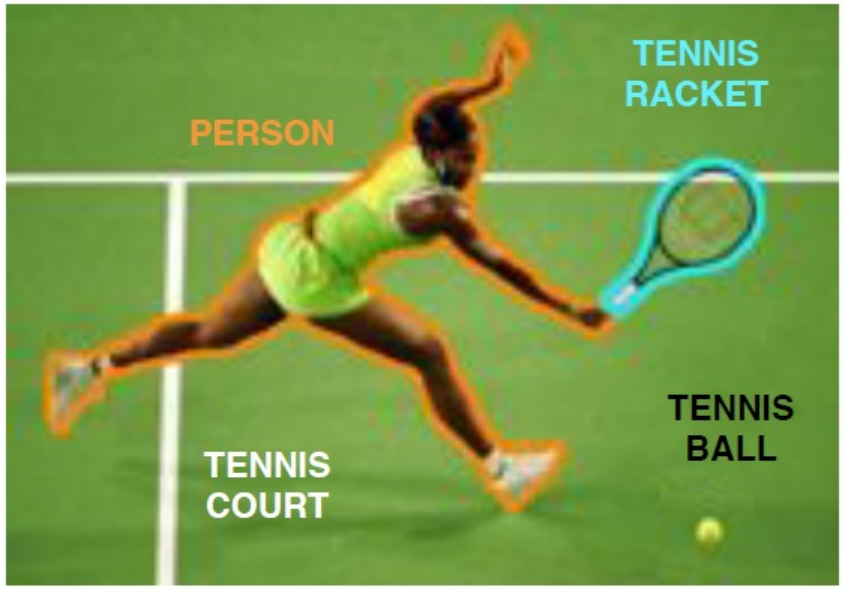


Figure from: Rabinovich, A., & Belongie, S. (2009). Scenes vs. objects: a comparative study of two approaches to context based recognition. CVPR.

Bozcan, I., & Kalkan, S. (2019). Cosmo: Contextualized scene modeling with boltzmann machines. Robotics and Autonomous Systems.

Doğan, F. I., Celikkanat, H., & Kalkan, S. (2018). A deep incremental boltzmann machine for modeling context in robots. IEEE International Conference on Robotics and Automation (ICRA).

Celikkanat, H., Orhan, G., Pugeault, N., Guerin, F., Şahin, E., & Kalkan, S. (2015). Learning context on a humanoid robot using incremental latent dirichlet allocation. IEEE Transactions on Cognitive and Developmental Systems, 8(1), 42-59.
* Outstanding paper award.

My Research:
Object Detection


Fehmi Kahraman


Baris Can Cam


Kemal Oksuz


Emre Akbas

Fig: https://analyticsprofile.com/machine-learning/object-detection-basic-tutorial-in-python/

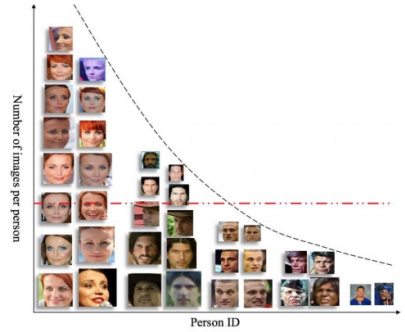
“Correlation Loss: Enforcing Correlation between Classification and Localization”, AAAI, 2023.
“One Metric to Measure them All: Localisation Recall Precision (LRP) for Evaluating Visual Detection Tasks", PAMI, 2022.
“Rank & Sort Loss for Object Detection and Instance Segmentation", ICCV, oral presentation, 2021.
"Imbalance Problems in Object Detection: A Review", PAMI, 43(10):3388-3415, 2021.
"A Ranking-based, Balanced Loss Function Unifying Classification and Localisation in Object Detection", NeurIPS, spotlight paper, 2020.

My Research:
Class Imbalance

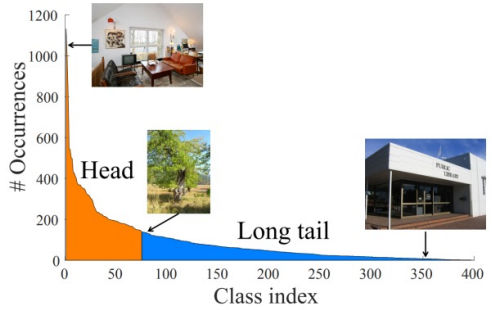
Project page: https://metu-balance.github.io/

Sonat Baltaci | Kemal Oksuz | Baris Can Cam | Emre Akbas

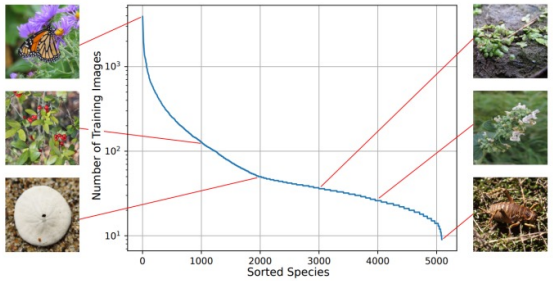
And:
- Selim Kuzucu
- Alpay Özkan
- Artun Özyeğin
- Kıvanç Tezören
- Feyza Yavuz



Faces [Zhang et al. 2017]



Places [Wang et al. 2017]



Species [Van Horn et al. 2019]

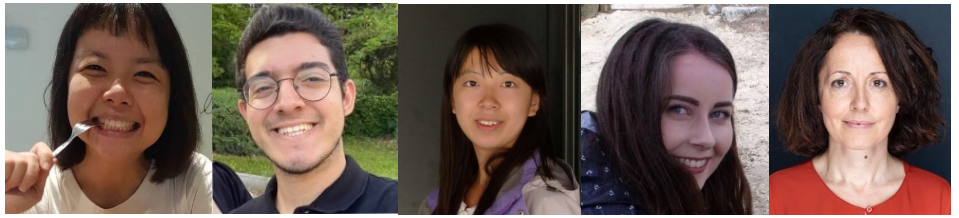
Actions [Zhang et al. 2019]

Source: https://liuziwei7.github.io/papers/longtail_slides.pdf

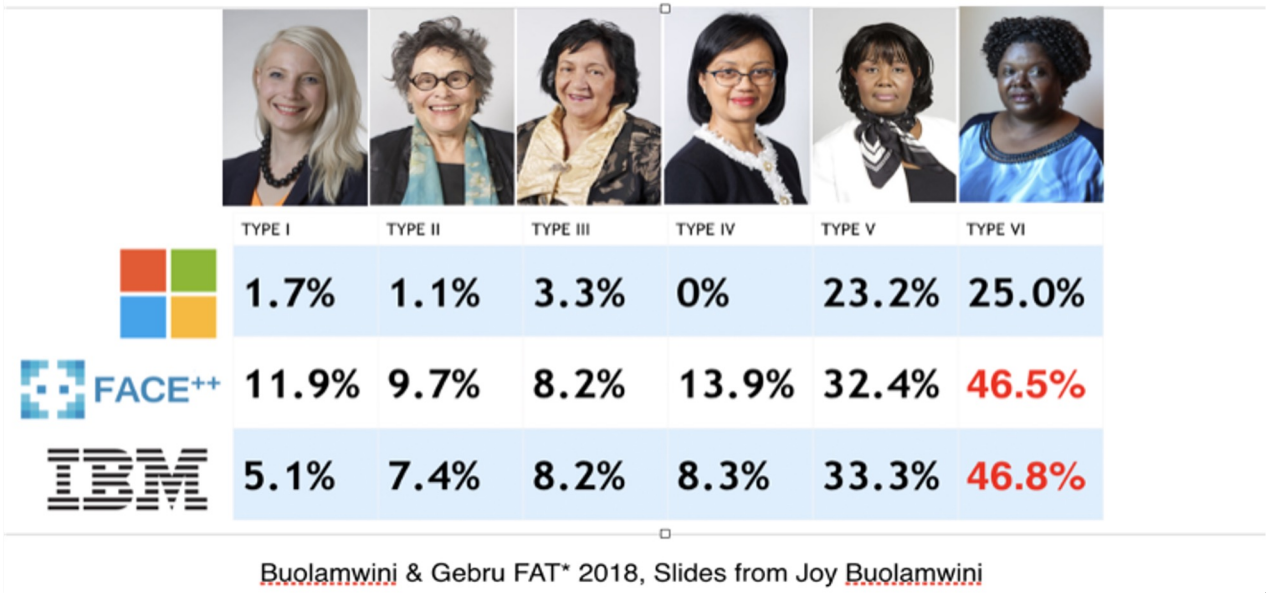
K. Oksuz, B. C. Cam, S. Kalkan, E. Akbas, "Imbalance Problems in Object Detection: A Review", PAMI, 2021.

Z. S. Baltaci, K. Oksuz, S. Kuzucu, K. Tezoren, B. K. Konar, A. Ozkan, E. Akbas*, S. Kalkan*, "Class Uncertainty: A Measure to Mitigate Class Imbalance", under review, 2023.

My Research:
Bias and Fairness



Jiaee Cheong | Selim Kuzucu | Tian Xu | Jennifer White | Hatice Güneş



	TYPE I	TYPE II	TYPE III	TYPE IV	TYPE V	TYPE VI
Microsoft	1.7%	1.1%	3.3%	0%	23.2%	25.0%
FACE++	11.9%	9.7%	8.2%	13.9%	32.4%	46.5%
IBM	5.1%	7.4%	8.2%	8.3%	33.3%	46.8%

Buolamwini & Gebru FAT* 2018, Slides from Joy Buolamwini

J. Cheong, S. Kuzucu, S. Kalkan, H. Gunes, "Towards Gender Fairness for Mental Health Prediction", 32nd Int. Joint Conf. on Artificial Intelligence (IJCAI), 2023.

Cheong, J., Kalkan, S., & Gunes, H. (2022). Counterfactual Fairness for Facial Expression Recognition, uECCV 2022 Workshop and Challenge on People Analysis: From Face, Body and Fashion to 3D Virtual Avatars, 2022.

Cheong, J., Kalkan, S., & Gunes, H. (2021). The Hitchhiker's Guide to Bias and Fairness in Facial Affective Signal Processing: Overview and techniques. IEEE Signal Processing Magazine, 38(6), 39-49.

Xu, T., White, J., Kalkan, S., & Gunes, H. (2020). Investigating Bias and Fairness in Facial Expression Recognition. ECCV 2020 Workshop on Fair Face Recognition and Analysis

My Research:
Robot Apprentice

Project page: https://metu-kalfa.github.io/

Yunus Terzioglu | Ozgur Aslan | Burak Bolat | Erol Sahin

O. Aslan, B. Bolat, B. Bal, T. Tumer, E. Sahin, S. Kalkan, “AssembleRL: Learning to Assemble Furniture from Their Point Clouds Only”, IROS, 2022.

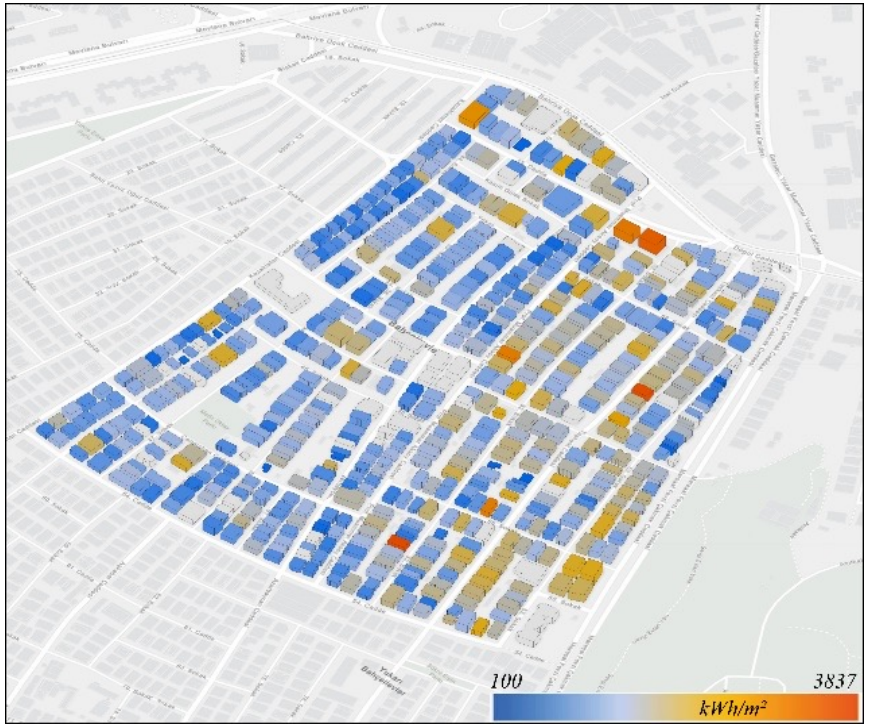
Terzioglu, O. Aslan, B. Bolat, B. Bal, T. Tumer, F. C. Kurnaz, S. Kalkan, E. Sahin, "APPRENTICE: Towards a Cobot Helper in Assembly Lines", ICRA2021 Workshop on Unlocking the Potential of HRC for Industrial Applications, 2021.

My Research:
ML for Energy

Project Page: https://metu-energy.github.io/

Help stakeholders by predicting

- Energy use of buildings
  - for heating or cooling
- Lighting costs of buildings
- Energy generation of solar panels
- Matching energy generation and demand



E. G. Halacli, I. Canli, O. K. Iseri, F. Yavuz, C. M. Akgul, S. Kalkan, I. G. Dino, "A Novel Graph Neural Network for Zone-Level Urban-Scale Building Energy Use Estimation", 10th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, 2023.

I. Canli, S. Kalkan, I. G. Dino, "Useful Daylight Illuminance Prediction Under Data Imbalance in an Urban Context", 41st Conference on Education and Research in Computer Aided Architectural Design in Europe, eCAADe 2023.

Overview of the course

About the course

Catalog Description: Ethical concepts and principles of trust and responsibility in AI; computational methods for dependability and robustness in AI; computational methods for explainability in AI; computational methods for bias and fairness in AI.

Background Requirements: Background in deep learning is a must. The students must have taken CENG403 (Introduction to Deep Learning) or CENG501 (Deep Learning).

Reference Material: (1) Lu, Q., Zhu, L., Whittle, J., & Xu, X. (2023). Responsible AI: Best practices for creating trustworthy AI systems. Addison-Wesley Professional. (2) Bartneck, C., Lütge, C., Wagner, A., & Welsh, S. (2021). An introduction to ethics in robotics and AI (p. 117). Springer Nature.

Syllabus

Webpage: https://metu-trai.github.io

Forum: ODTUclass page of the course.

Instructor: Sinan Kalkan, skalkan@metu.edu.tr (Office hours: by appointment)

Lectures: Wed: 9:40-12:30 [BMB-3]

Credits: METU: 3 Theoretical, 0 Laboratory; ECTS: 8.0

Weekly outline

Week & Date		Topic
1	29 Sep	**Course logistics and overview** [Intro to trust and responsibility in AI: Social, ethical and policy implications of AI]
2	6 Oct	**Overview** [Deep/machine learning concepts (fundamentals)]
3	13 Oct	**Overview** [Deep/machine learning concepts (recent trends)]
4	20 Oct	**Robust AI** [Fundamentals of Robust AI]
5	27 Oct	**Robust AI** [Adversarial Attacks & Corruptions]
6	3 Nov	**Robust AI** [Defense Mechanisms and Robust Training]
7	10 Nov	**Robust AI** [Quantifying Uncertainty and Calibration of Deep Networks]
8	17 Nov	**Explainable AI** [Fundamentals of Explainable AI]
9	24 Nov	**Explainable AI** [Model-specific Explainability Methods: Explainable shallow models, prototype-based models, concept bottleneck models, attention mechanisms for explainability]
10	1 Dec	**Explainable AI** [Post-Hoc Explanation Techniques: CAM, GradCAM, feature-based saliency estimation, counterfactual reasoning]
11	8 Dec	**Explainable AI** [Evaluating explainability, methods and metrics]
12	15 Dec	**Bias and Fairness in AI** [Sources of bias; definitions of fairness; measures of fairness]
13	22 Dec	**Bias and Fairness in AI** [Data-level and in-processing methods for fairness]
14	29 Dec	**Bias and Fairness in AI** [Regularization-based and post-processing methods for fairness]

Grading

Grading:

Quizzes	20%
Project	35%
Final Exam	45%

Warnings

(1) Participating in the final exam is subject to the following conditions: (i) Attending 60% of the quizzes. (ii) Completing half of the project. Not satisfying these will result in NA as the grade.

(2) The use of Generative AI tools (e.g., ChatGPT) is strictly forbidden.

Project

- Implement a conference paper without any implementation
  - Reproduce their results and provide a critical analysis
  - Produce a Github Repo with a detailed Readme file
- Papers can **only** be from the following top conferences:
  - ICLR, AAAI, NeurIPS, ICML, CVPR, ECCV or similar
- You can work in groups of two.
- Deadline for choosing papers: 13 October.
  - A link for submitting details about the papers will be provided.

Taking the course

Backpropagation

For each output unit c, calculate its grad term δ_c^o:

$$\delta_{ic}^o = \frac{\partial L_i}{\partial net_{ic}^o} = \frac{\partial L_i}{\partial \hat{y}_{ic}} \frac{\partial \hat{y}_{ic}}{\partial net_{ic}^o} = (\hat{y}_{ic} - y_{ic})\hat{y}_{ic}(1 - \hat{y}_{ic})$$

For each hidden unit j, calculate its grad term δ_j^h:

$$\delta_{ij}^h = \frac{\partial L_i}{\partial net_{ij}^h} = \frac{\partial L_i}{\partial h_{ij}} \frac{\partial h_{ij}}{\partial net_{ij}^h} = \left(\sum_{c \in C} \frac{\partial L_i}{\partial net_{ic}^o} \frac{\partial net_{ic}^o}{\partial h_{ij}}\right) h_{ij}(1 - h_{ij})$$
$$= \left(\sum_{c \in C} \delta_{ic}^o w_{cj}\right) h_{ij}(1 - h_{ij})$$

Update weight w_{jk}^o in the output layer:

$$w_{jk}^o = w_{jk}^o - \eta \delta_{ij}^o h_{ik}$$

Update weight w_{jk}^h in the hidden layer:

$$w_{jk}^h = w_{jk}^h - \eta \delta_{ij}^h x_{ik}$$

The Model

Hidden activations: $h_{ij} = \sigma(\mathbf{w}_j^h \cdot \mathbf{x}_i) = \sigma(net_{ij}^h)$

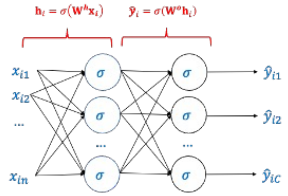
Output layer: $\hat{y}_{ic} = \sigma(\mathbf{w}_c^o \cdot \mathbf{h}_i) = \sigma(net_{ic}^o)$

The loss function:

$$L(\boldsymbol{\theta}) = \frac{1}{2} \sum_{i=1}^{N} \sum_{c \in C} (\hat{y}_{ic} - y_{ic})^2$$

- For one sample:

$$L_i(\boldsymbol{\theta}) = \frac{1}{2} \sum_{c \in C} (\hat{y}_{ic} - y_{ic})^2$$



Triplet Loss: Schroff *et al.* [17] proposed Triplet loss as an augmentation over Contrastive loss [3]. Triplet loss jointly minimizes the distances between the feature embeddings of a given sample (anchor) and another sample of the same class (positive) while maximizing the distance of the embeddings of a suitable sample of a different class (negative) to the anchor. The loss is defined as below:

$$\mathcal{L} = \sum_{a,p,n \subset N} \left[\|\boldsymbol{f}_a - \boldsymbol{f}_p\|^2 - \|\boldsymbol{f}_a - \boldsymbol{f}_n\|^2 + \alpha \right]_+ \quad (1)$$

The terms $\boldsymbol{f}_a, \boldsymbol{f}_p, \boldsymbol{f}_n$ correspond to feature embeddings for the anchor, positive and negative samples, where a, p, n are sampled from the training dataset N. α defines the margin enforced between the anchor-negative embedding dis-

similarities amongst the positive samples and the negative samples in conjunction with the self-similarity measure to handle all three forms of similarities available. The loss is derived from the binomial deviance loss and is formulated as:

$$\mathcal{L} = \frac{1}{m} \sum_{i=1}^{m} \left\{ \frac{1}{\alpha} \log \left[1 + \sum_{p \in \mathcal{P}_i} e^{-\alpha(S_{ip} - \lambda)} \right] + \frac{1}{\beta} \log \left[1 + \sum_{n \in \mathcal{N}_i} e^{\beta(S_{in} - \lambda)} \right] \right\} \quad (3)$$

The first *log* term deals with the similarity scores S_{ip} for the positive samples $p \in \mathcal{P}_i$ which comprises the set of posi-

```
dout_row = dout[index].reshape(C, outH*outW)
neuron = 0
for i in range(0, H-PH+1, stride):
    for j in range(0, W-PW+1, stride):
        pool_region = x[index,:,i:i+PH,j:j+PW].reshape(C,PH*PW)
        max_pool_indices = pool_region.argmax(axis=1)
        dout_cur = dout_row[:,neuron]
        neuron += 1
        # pass gradient only through indices of max pool
        dmax_pool = np.zeros(pool_region.shape)
        dmax_pool[np.arange(C),max_pool_indices] = dout_cur
        dx[index,:,i:i+PH,j:j+PW] += dmax_pool.reshape(C,PH,PW)
```

Basic DL Concepts

Ensure that you know the following:

- Why does DL work now?
- End-to-end learning
- Distributed representations
- Advantages and disadvantages of DL

Basic ML Concepts

Ensure that you know the following:

- Supervised vs. unsupervised learning
- Discriminative vs. generative learning
- Model selection, cross validation
- Overfitting, memorization, bias-variance trade-off

Taking the course

- Background
  - Programming, Python
  - Data structures and algorithms
  - Linear algebra, Calculus, Statistics
  - Fundamental deep learning models: MLP, CNN, RNN, Transformers, ..

- Fill out the following form until 3 Oct, midnight:

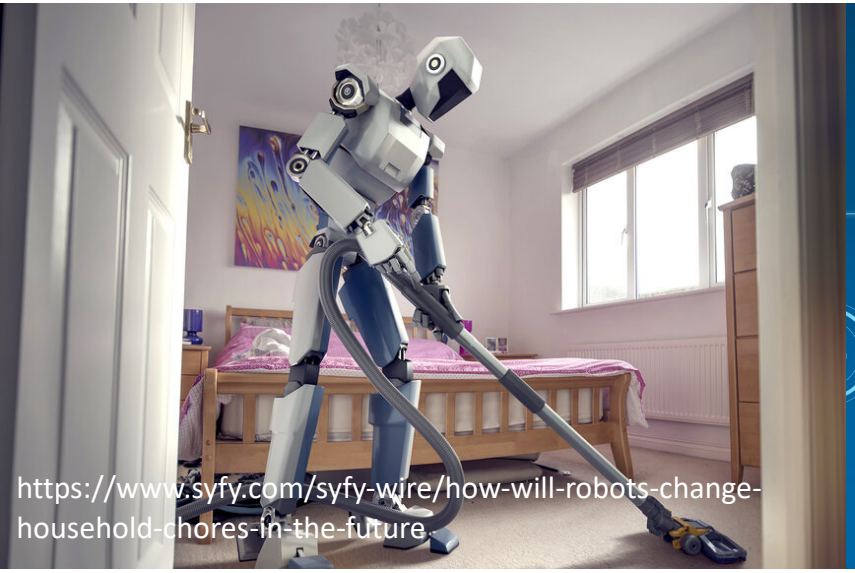
https://tinyurl.com/ceng7880

(long URL: https://forms.gle/goKjPGkHn7oWomkW6)

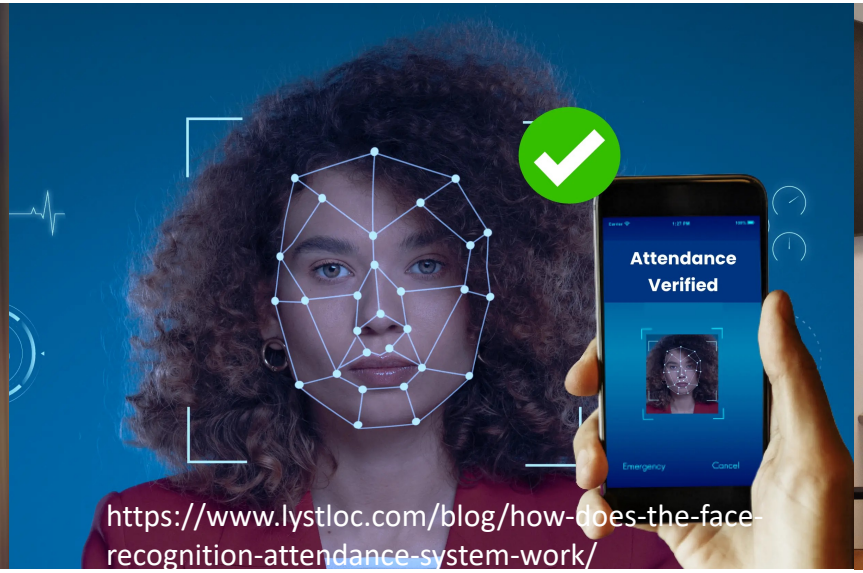
Overview of the fundamental concepts in TRAI

Social, ethical and policy implications of AI

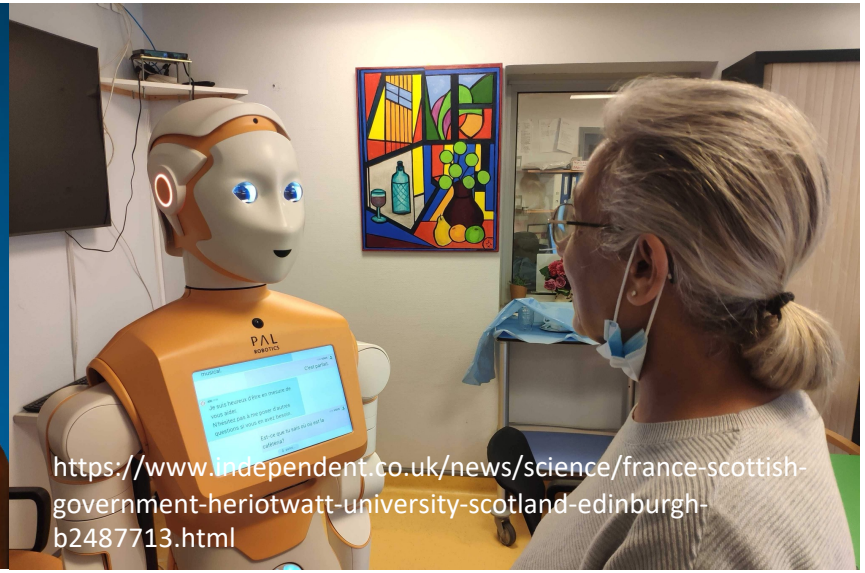
Trust and Responsibility in AI Systems



https://www.syfy.com/syfy-wire/how-will-robots-change-household-chores-in-the-future

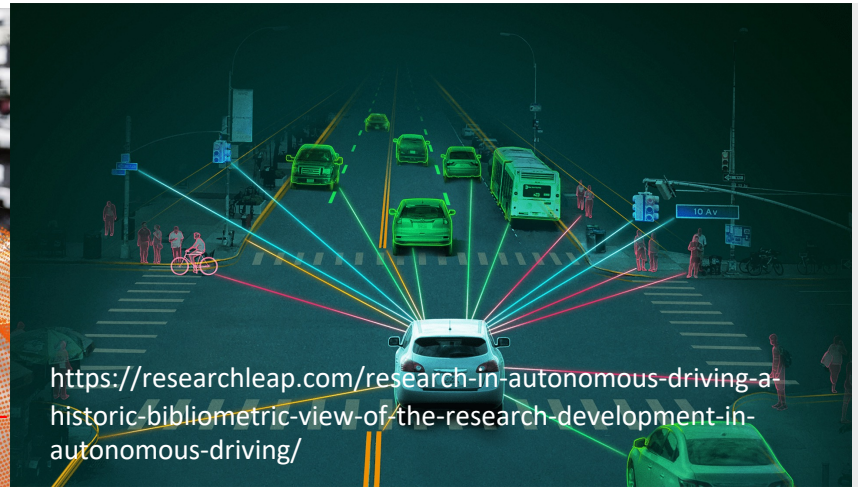


https://www.lystloc.com/blog/how-does-the-face-recognition-attendance-system-work/



https://www.independent.co.uk/news/science/france-scottish-government-heriotwatt-university-scotland-edinburgh-b2487713.html

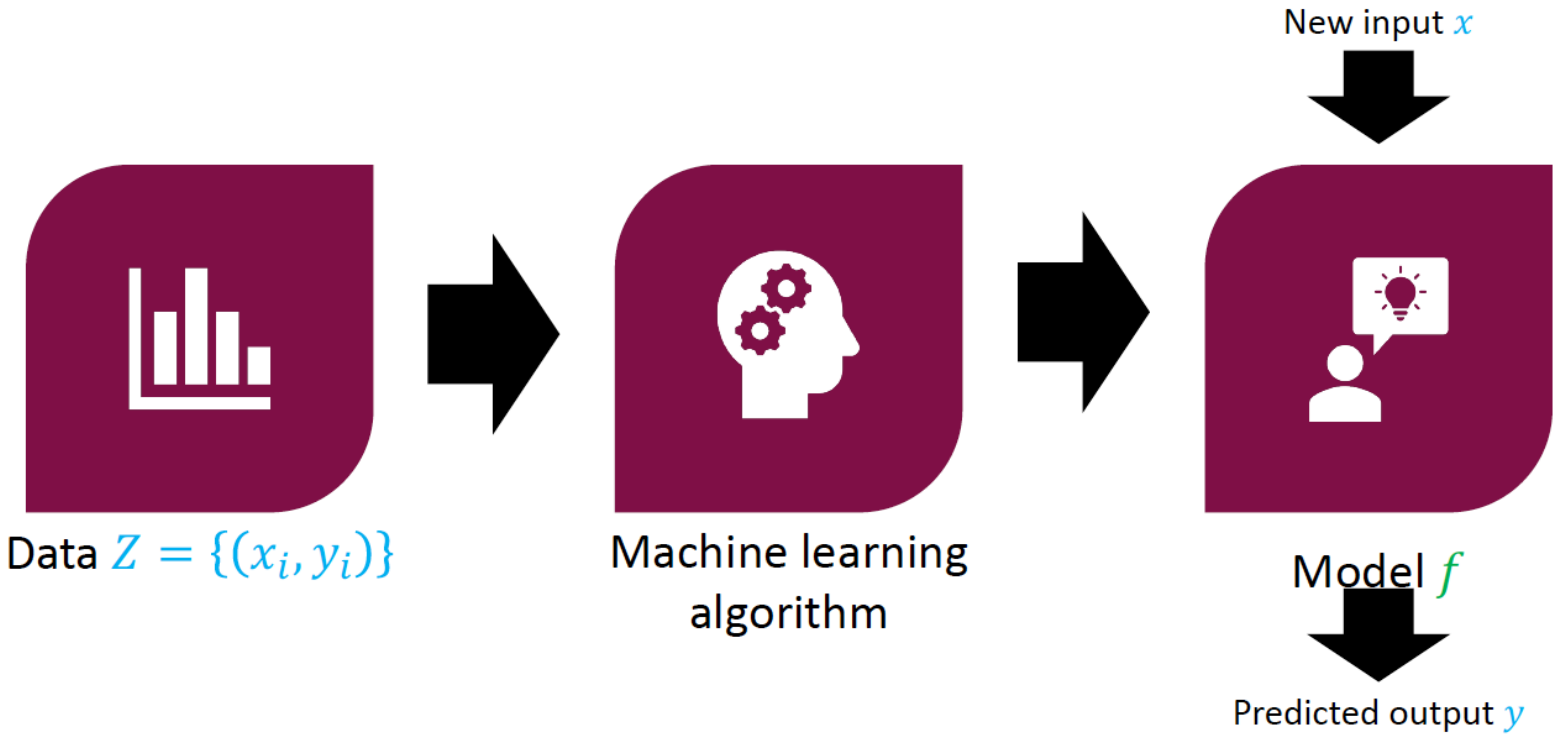
https://www.kuka.com/en-us/future-production/human-robot-collaboration



https://researchleap.com/research-in-autonomous-driving-a-historic-bibliometric-view-of-the-research-development-in-autonomous-driving/

https://techcrunch.com/2021/10/14/actuator-8/

Beyond accuracy



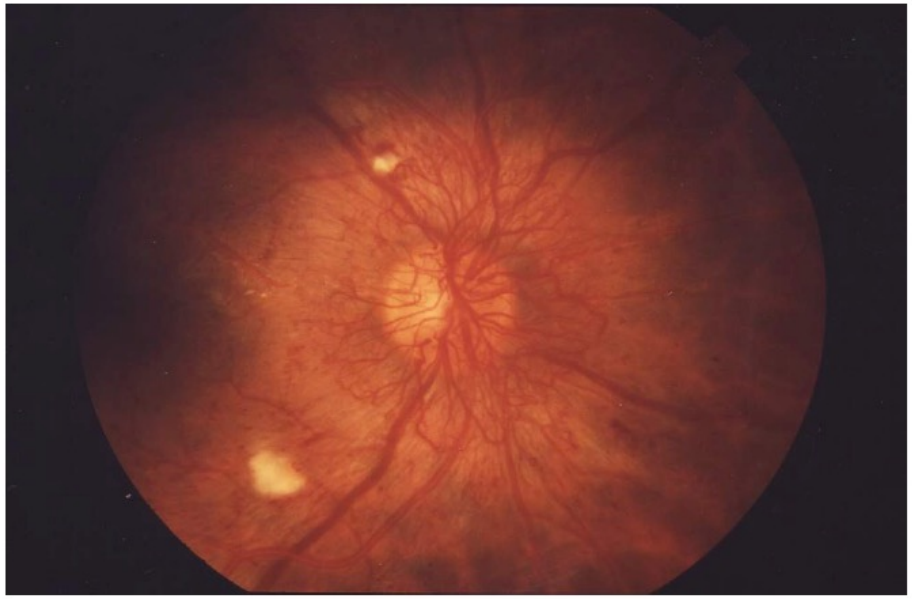
Goal: Maximize performance (e.g., accuracy, MSE, etc.) on new predictions

Is this enough?

Slide credit: Osbert Bastani and Rajeev Alur, CIS 7000 Trustworthy ML, UPenn.

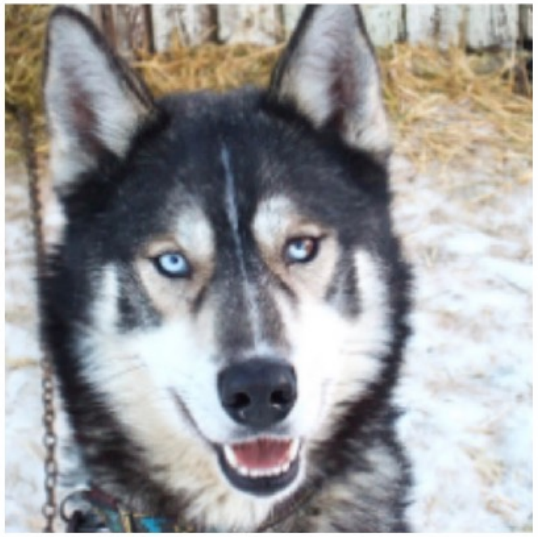
Beyond accuracy

- **Example:** Help a doctor determine whether a patient has diabetic retinopathy
- Does the doctor trust the prediction? (interpretability)
- Should the doctor double check the prediction? (uncertainty quantification)



Slide credit: Osbert Bastani and Rajeev Alur, CIS 7000 Trustworthy ML, UPenn.

Beyond accuracy



(a) Husky classified as wolf

(b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

Ribeiro et al., "Why Should I Trust You? Explaining the Predictions of Any Classifier", 2016

Slide credit: Osbert Bastani and Rajeev Alur, CIS 7000 Trustworthy ML, UPenn.

Beyond accuracy

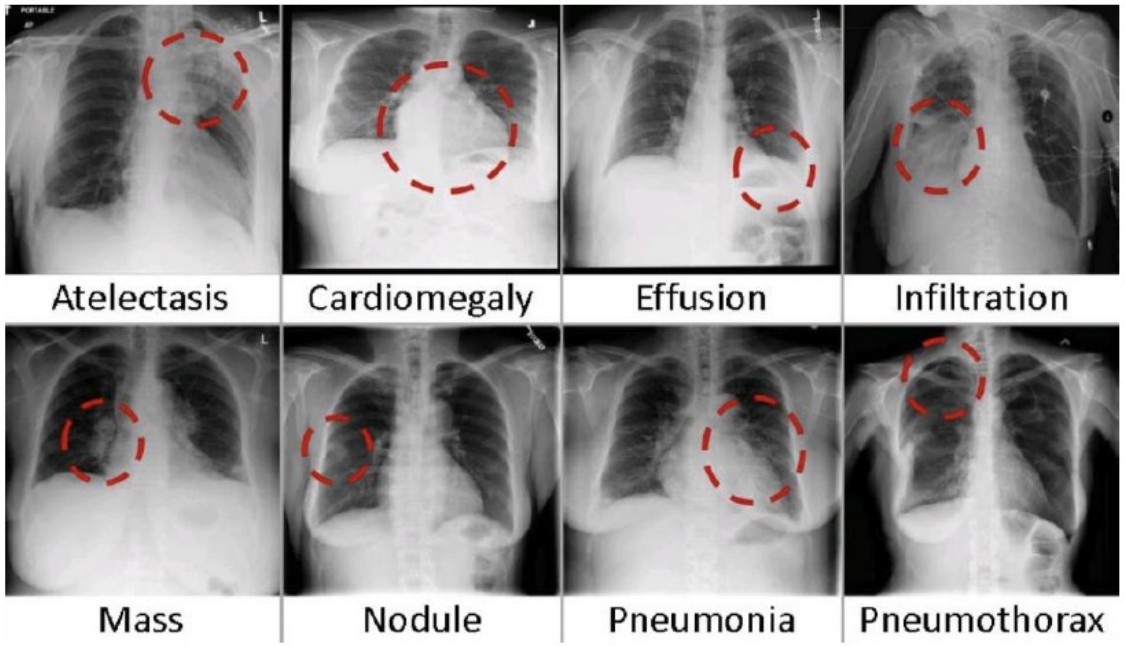
Right for the Wrong Reason: Can Interpretable ML Techniques Detect Spurious Correlations?

Susu Sun[1], Lisa M. Koch[2,3], and Christian F. Baumgartner[1]

[1] Cluster of Excellence – ML for Science, University of Tübingen, Germany
[2] Hertie Institute for AI in Brain Health, University of Tübingen, Germany
[3] Institute of Ophthalmic Research, University of Tübingen, Germany
{susu.sun,lisa.koch,christian.baumgartner}@uni-tuebingen.de

Abstract. While deep neural network models offer unmatched classification performance, they are prone to learning spurious correlations in the data. Such dependencies on confounding information can be difficult to detect using performance metrics if the test data comes from the same distribution as the training data. Interpretable ML methods such as post-hoc explanations or inherently interpretable classifiers promise to identify faulty model reasoning. However, there is mixed evidence whether many of these techniques are actually able to do so. In this paper, we propose a rigorous evaluation strategy to assess an explanation technique's ability to correctly identify spurious correlations. Using this strategy, we evaluate five post-hoc explanation techniques and one inherently interpretable method for their ability to detect three types of artificially added confounders in a chest x-ray diagnosis task. We find that the post-hoc technique SHAP, as well as the inherently interpretable Attri-Net provide the best performance and can be used to reliably identify faulty model behavior.

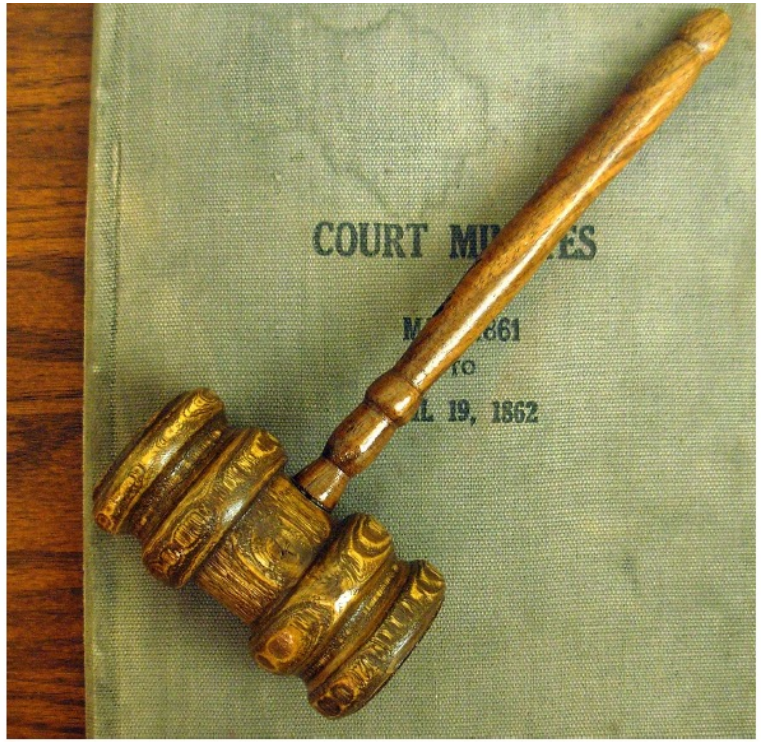
Keywords: Interpretable machine learning · Confounder detection



Slide credit: Osbert Bastani and Rajeev Alur, CIS 7000 Trustworthy ML, UPenn.

Beyond accuracy

- **Example:** Help a judge decide whether to give a defendant bail
- Does the judge trust the prediction? (interpretability)
- Does the algorithm discriminate against minorities? (fairness)



Slide credit: Osbert Bastani and Rajeev Alur, CIS 7000 Trustworthy ML, UPenn.

Beyond accuracy

Algorithms were supposed to make Virginia judges fairer. What happened was far more complicated.

Analysis by Andrew Van Dam
Staff writer | + Follow

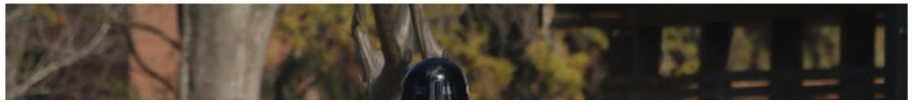
November 19, 2019 at 7:00 a.m. EST

The Accomack County Courthouse in February of this year. (Timothy C. Wright for the Washington Post)

Slide credit: Osbert Bastani and Rajeev Alur, CIS 7000 Trustworthy ML, UPenn.

Beyond accuracy

- **Example:** Deploy on a self-driving car to classify obstacles from LIDAR point clouds



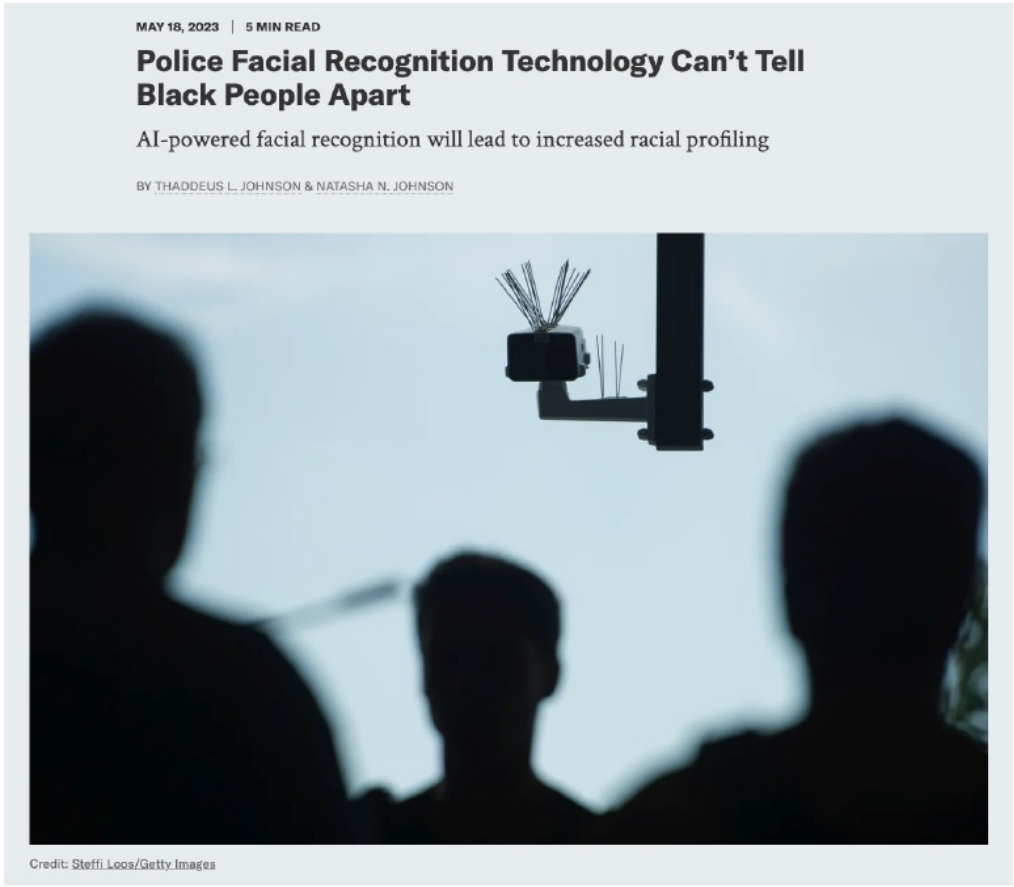
- Should the car act m cautiously? (uncertai quantification)

- What if the car is driv city? In the snow? (r

Slide credit: Osbert Bastani and Rajeev Alur, CIS 7000 Trustworthy ML, UPenn.

Beyond accuracy

- **Example:** Facial recognition based logi
- What if so algorithm
- Does the a racial subg



MAY 18, 2023 | 5 MIN READ

Police Facial Recognition Technology Can't Tell Black People Apart

AI-powered facial recognition will lead to increased racial profiling

BY THADDEUS L. JOHNSON & NATASHA N. JOHNSON

Credit: Steffi Loos/Getty Images

Slide credit: Osbert Bastani and Rajeev Alur, CIS 7000 Trustworthy ML, UPenn.

The General-Purpose AI Code of Practice

PAGE CONTENTS

The Code of Practice helps industry comply with the AI Act legal obligations on safety, transparency and copyright of general-purpose AI models.

The General-Purpose AI (GPAI) Code of Practice is a voluntary tool, prepared by independent experts in a multi-stakeholder process, designed to help industry comply with the AI Act's obligations for providers of general-purpose AI models. Read more about the timeline and the drafting process of the code.

The code was published on July 10, 2025. It is complemented by Commission guidelines on key concepts related to general-purpose AI models. The Commission and the AI Board have confirmed that the code is an **adequate voluntary tool** for providers of GPAI models to demonstrate compliance with the AI Act.

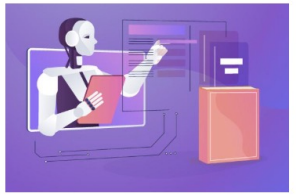
Following the endorsement, AI model providers who voluntarily sign it can show they comply with the AI Act by adhering to the code. This will **reduce their administrative burden** and give them **more legal certainty** than if they proved compliance through other methods. Find more information on the questions and answers (Q&A) about the code of practice for General-Purpose AI.

Providers of general-purpose AI models may sign the code by completing the Signatory Form and sending the signed form to EU-AIOFFICE-CODE-SIGNATURES@ec.europa.eu. Potential Signatories may also email this

Share

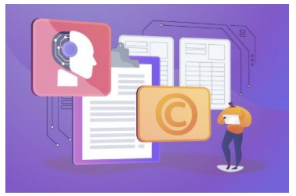
Quick links

- Drawing-up a General-Purpose AI Code of Practice
- Signing the General-Purpose AI Code of Practice – Questions & Answers
- AI Office invites providers to sign the GPAI Code of Practice
- Commission Assessment of the GPAI Code of Practice
- General-Purpose AI Models in the AI Act – Questions & Answers
- The code of practice for General-Purpose AI – Questions & Answers



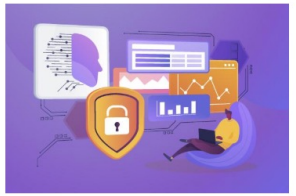
Transparency

The Transparency chapter (PDF) offers a user-friendly Model Documentation Form (DOCX) which allows providers to easily document the information necessary to comply with the AI Act obligation to on model providers to ensure sufficient transparency.



Copyright

The Copyright chapter (PDF) offers providers practical solutions to meet the AI Act's obligation to put in place a policy to comply with EU copyright law.



Safety and Security

The Safety and Security chapter (PDF) outlines concrete state-of-the-art practices for managing systemic risks, i.e. risks from the most advanced models. Providers can rely on this chapter to comply with the AI Act obligations for providers of general-purpose AI models with systemic risk.

https://ec.europa.eu/commission/presscorner/detail/en/ip_25_1787

Trust and Fairness in AI Systems: Ethical Principles

Items from Bartneck et al., “An Introduction to Ethics in Robotics and AI”, 2021.

- Trustworthy systems
  - Dependability, false alarm rate, transparency, task complexity
  - Performance and reliability

Trust and Fairness in AI Systems: Ethical Principles

Floridi et al., "AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations", 2018.

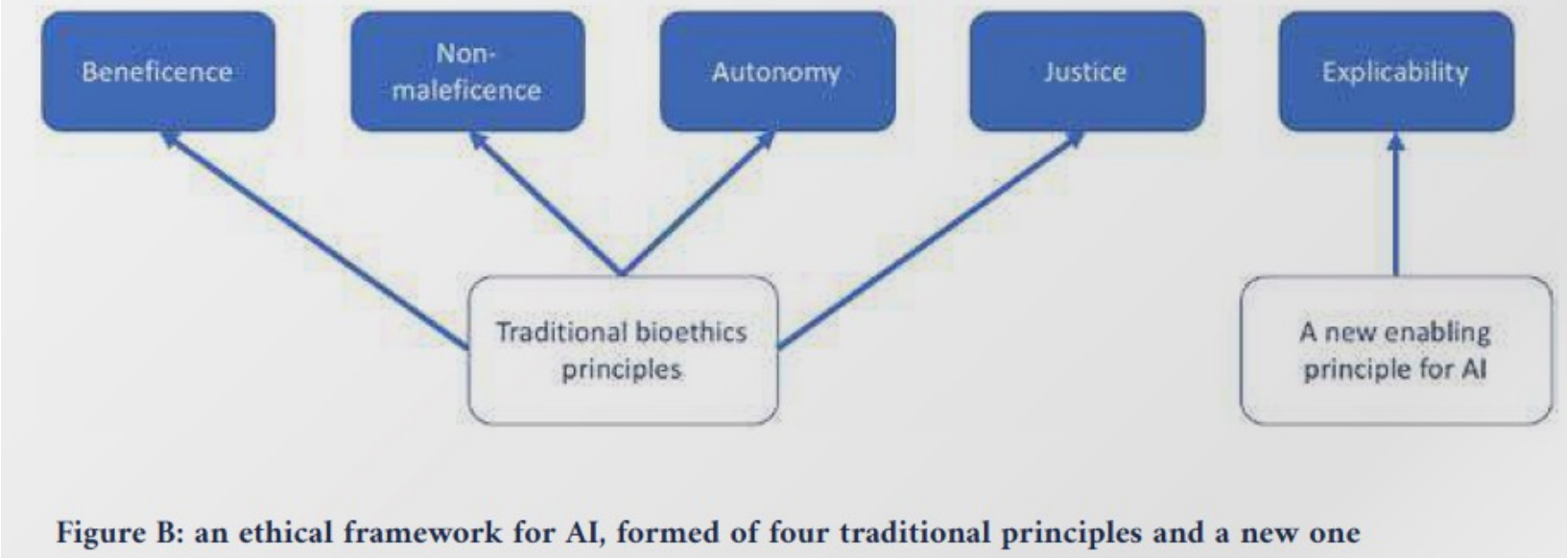


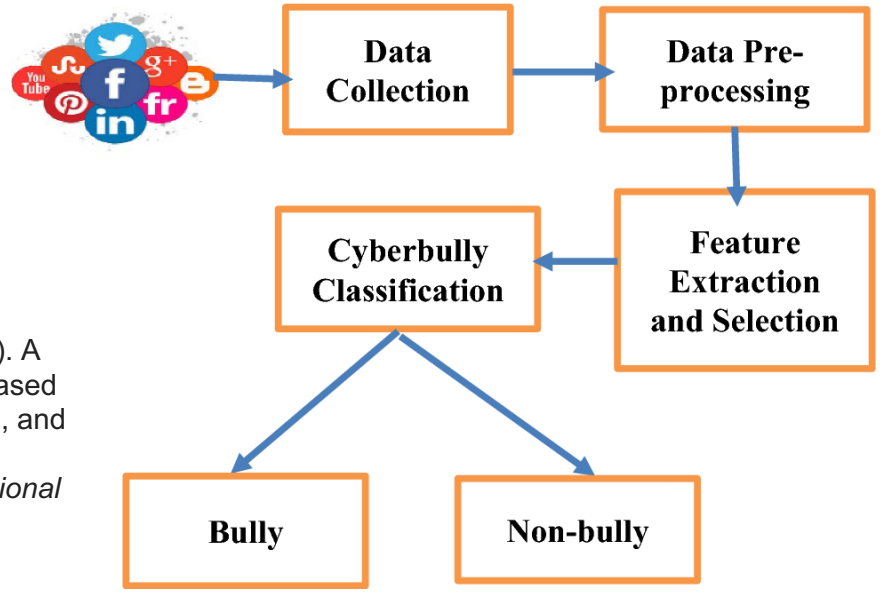
Figure B: an ethical framework for AI, formed of four traditional principles and a new one

Trust and Fairness in AI Systems: Ethical Principles

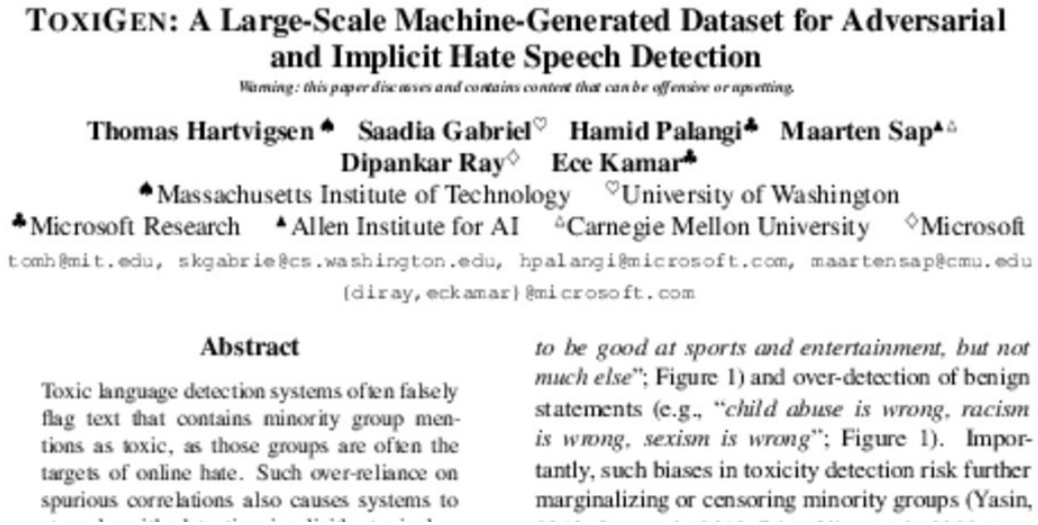
Items from Bartneck et al., "An Introduction to Ethics in Robotics and AI", 2021.

Non-maleficence: "AI shall not harm people"

- Social Media Bullying and Harassment
- Hate speech



Kumar, R., & Bhat, A. (2022). A study of machine learning-based models for detection, control, and mitigation of cyberbullying in online social media. *International Journal of Information Security*, *21*(6), 1409-1431.



TOXIGEN: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection

Warning: this paper discusses and contains content that can be offensive or upsetting.

Thomas Hartvigsen ♠ Saadia Gabriel ♡ Hamid Palangi ♣ Maarten Sap ▲△
Dipankar Ray ◇ Ece Kamar ♣
♠Massachusetts Institute of Technology ♡University of Washington
♣Microsoft Research ▲Allen Institute for AI △Carnegie Mellon University ◇Microsoft
tomh@mit.edu, skgabrie@cs.washington.edu, hpalangi@microsoft.com, maartensap@cmu.edu
{diray,eckamar}@microsoft.com

Abstract

Toxic language detection systems often falsely flag text that contains minority group mentions as toxic, as those groups are often the targets of online hate. Such over-reliance on spurious correlations also causes systems to

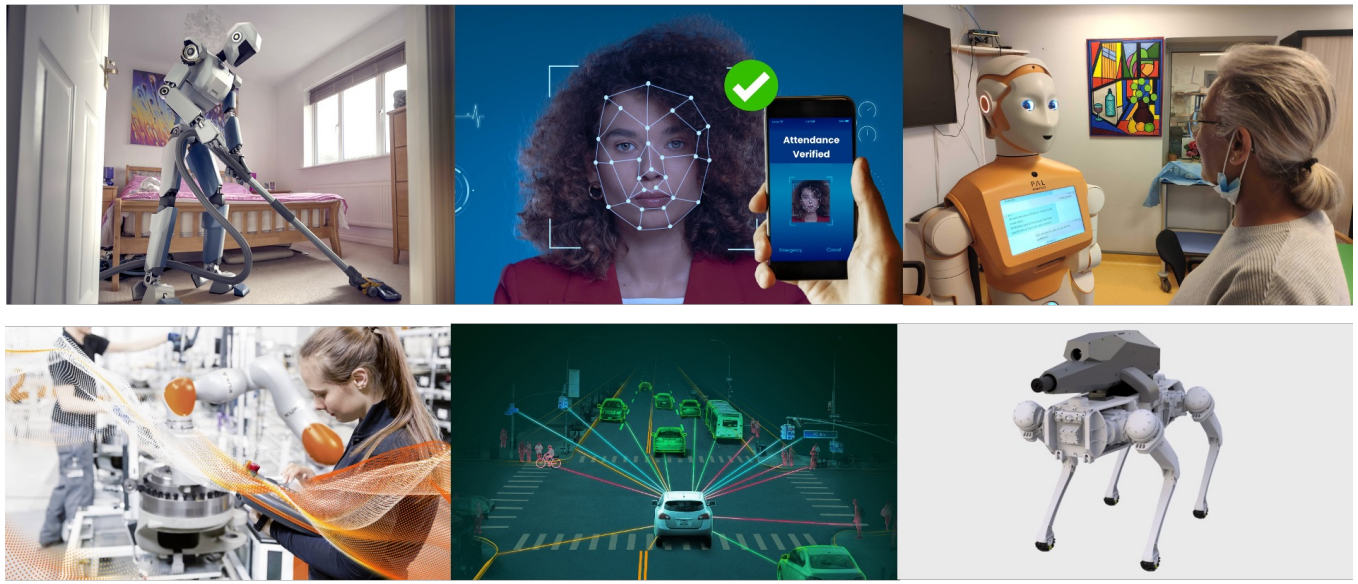
to be good at sports and entertainment, but not much else"; Figure 1) and over-detection of benign statements (e.g., "*child abuse is wrong, racism is wrong, sexism is wrong*"; Figure 1). Importantly, such biases in toxicity detection risk further marginalizing or censoring minority groups (Yasin,

Items from Bartneck et al., “An Introduction to Ethics in Robotics and AI”, 2021.

Trust and Fairness in AI Systems: Ethical Principles

Beneficence: “AI shall do good”

- Health & medicine
- Automation
- Education
- Security
- Environment



* See Slide 27 for references to image sources.

Items from Bartneck et al., “An Introduction to Ethics in Robotics and AI”, 2021.

Trust and Fairness in AI Systems: Ethical Principles

Autonomy: “AI shall respect people’s goals and wishes”

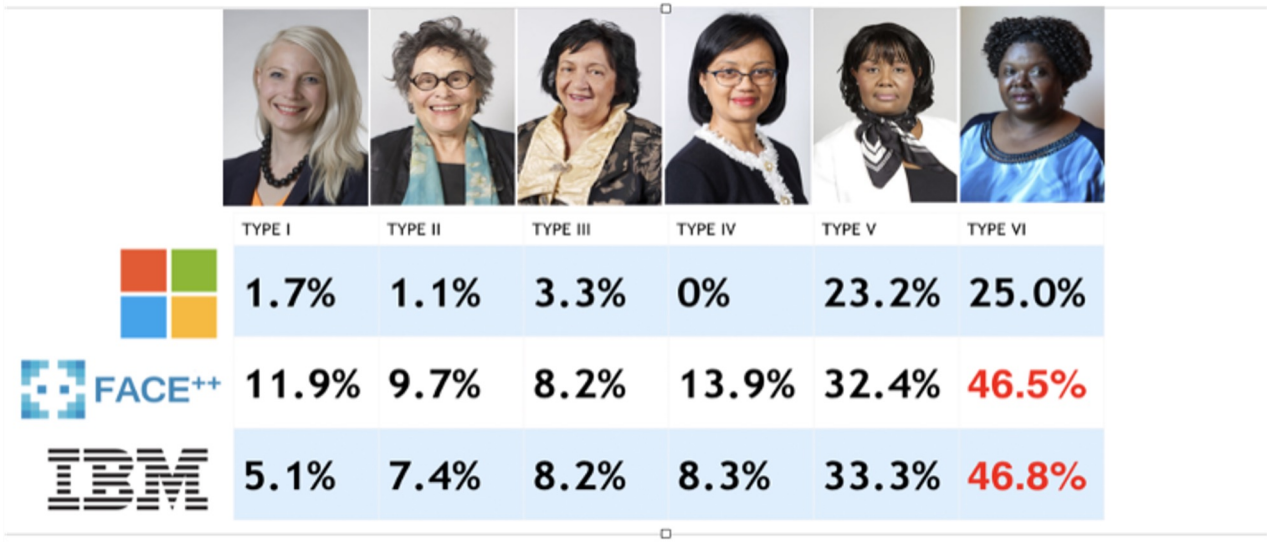
- “autonomy refers to the ability of a person to make decisions”

- What if the owner gives an order harmful that can be harmful to others?
- Robots (e.g. police robots) designed to use force against humans

Items from Bartneck et al., “An Introduction to Ethics in Robotics and AI”, 2021.

Trust and Fairness in AI Systems: Ethical Principles

- Justice (Fairness): “AI shall act in a just and unbiased way”
  - Determining Creditworthiness
  - Use in Court



Buolamwini & Gebru FAT* 2018, Slides from Joy Buolamwini

Items from Bartneck et al., “An Introduction to Ethics in Robotics and AI”, 2021.

Trust and Fairness in AI Systems: Ethical Principles

- Explicability: “AI shall be able to explain why it arrived at a certain conclusion or result”

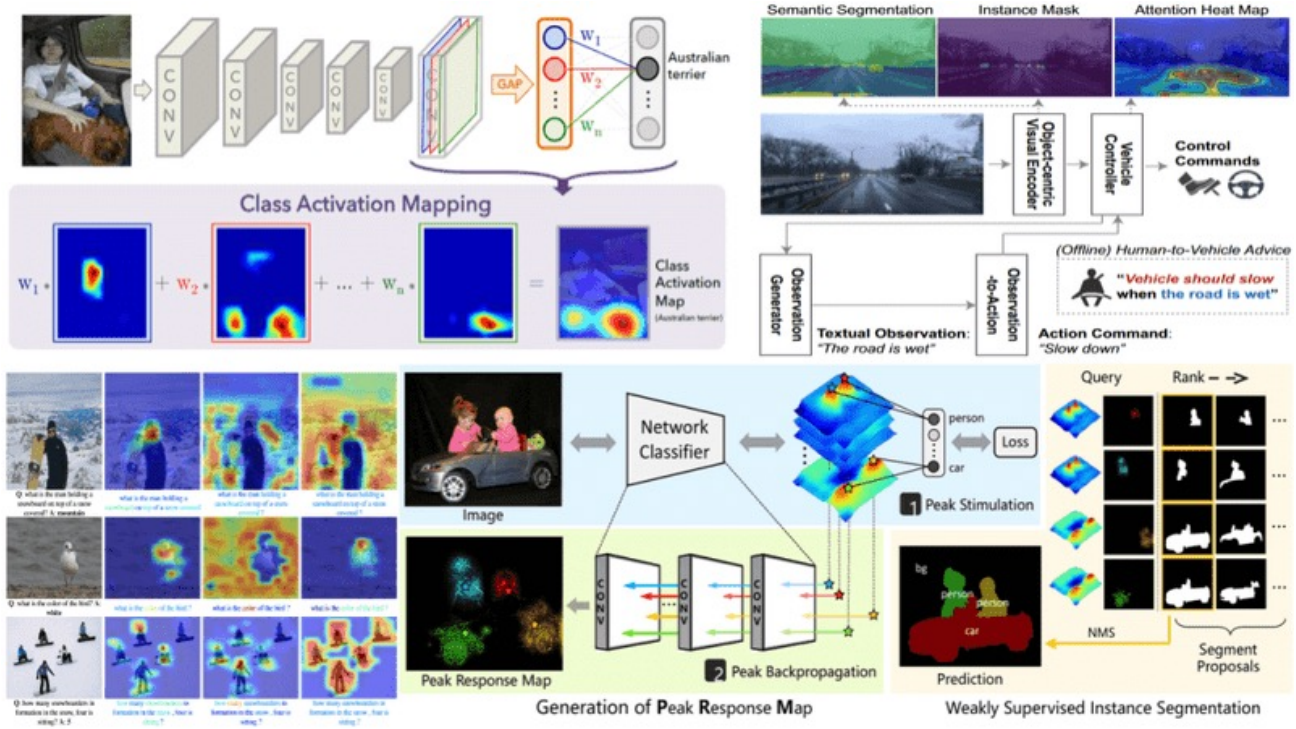


Fig: https://theaisummer.com/xai/

Privacy Issues of AI

- Why AI needs data?
- Private data collection and its dangers
  - Persistence surveillance
    - Smart cameras/toys/devices to collect data
  - Usage of private data for non-intended purposes
    - Manipulation by advertising systems
    - Data/systems to affect our choices (e.g., political decisions)
    - Impersonate people
    - Possible to infer personal information (genetics, physical/mental limitations, ..)

Privacy Issues of AI

- Private data collection and its dangers
  - Auto insurance discrimination:
    - Insurance companies use data to estimate customer risks
    - Can include bias
  - The Chinese social credit system:
    - Credit scoring based on how good a citizen is (using also web/social/cloud data)

Course Coverage

Robustness

- "ML Model robustness denotes the capacity of a model to sustain stable predictive performance in the face of **variations and changes in the input data**"

https://arxiv.org/pdf/2404.00897v2

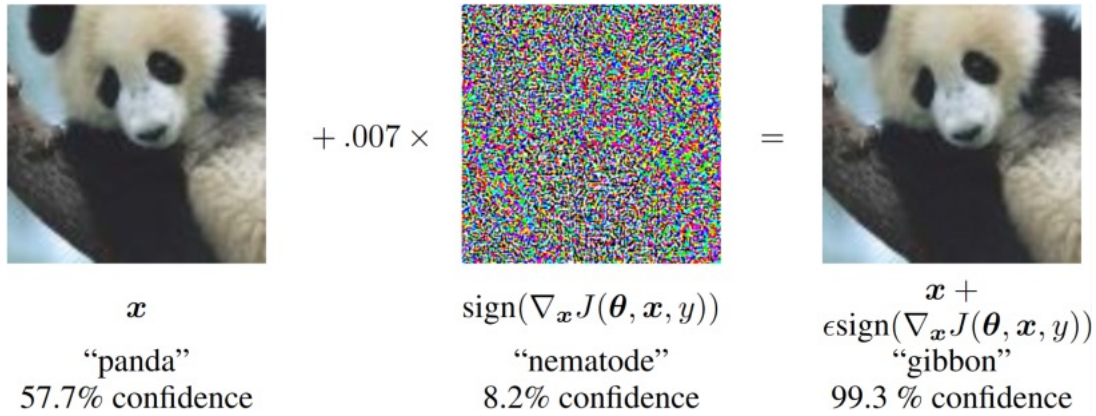
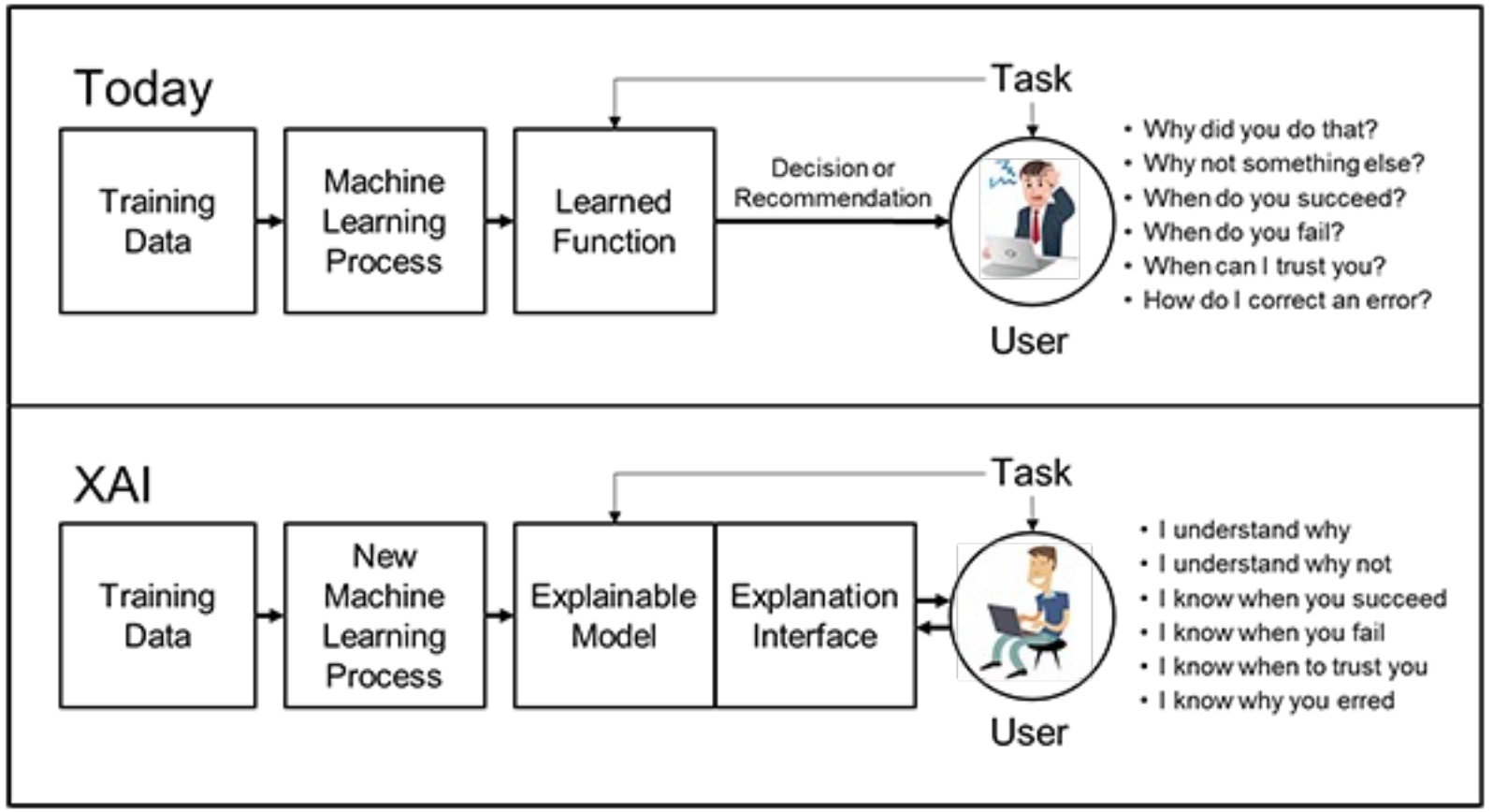


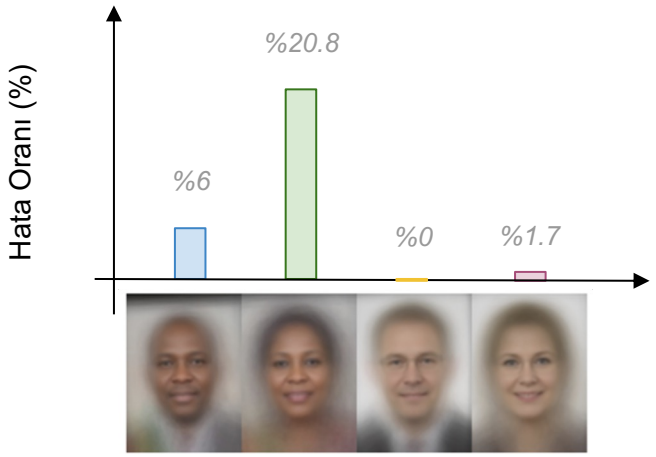
Fig: https://arxiv.org/pdf/1412.6572

Explainability



https://www.darpa.mil/research/programs/explainable-artificial-intelligence

Bias and Fairness



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, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on fairness, accountability and transparency. pp. 77-91 (2018).