# Machine Explanations and Human Understanding

Chacha Chen* Shi Feng*
Amit Sharma Chenhao Tan

FAccT 2023

TRANSACTIONS tmlr on ML RESEARCH

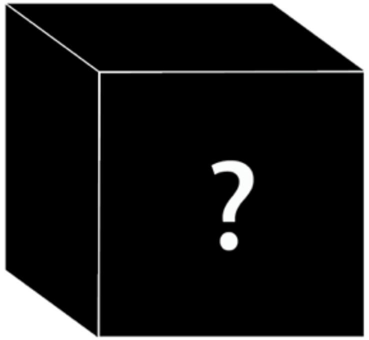
Chicago Human+AI

Microsoft Research

Understanding

Bias

Safety




Accountability

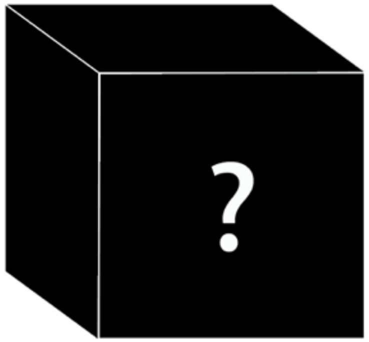
Ethics

Value-alignment

Understanding

Bias

Safety




Accountability

Ethics

Value-alignment

*Explanations* are hypothesized to improve human understanding of ML models in human-AI interaction

Empirical experiments found mixed and even conflicting results on the effect of explanations.

# Empirical experiments found mixed and even conflicting results on the effect of explanations.

Feature importance improve model debugging.
[Ribeiro et al., 2016]

# Empirical experiments found mixed and even conflicting results on the effect of explanations.

Feature importance improve model debugging.
[Ribeiro et al., 2016]

With feature importance, human + AI < AI.
[Lai & Tan, 2019]

# The definition of human understanding remains unclear.

The definition of human understanding remains unclear.

Under what conditions, explanation can improve human understanding, and in which way.

# How do we define human understanding?

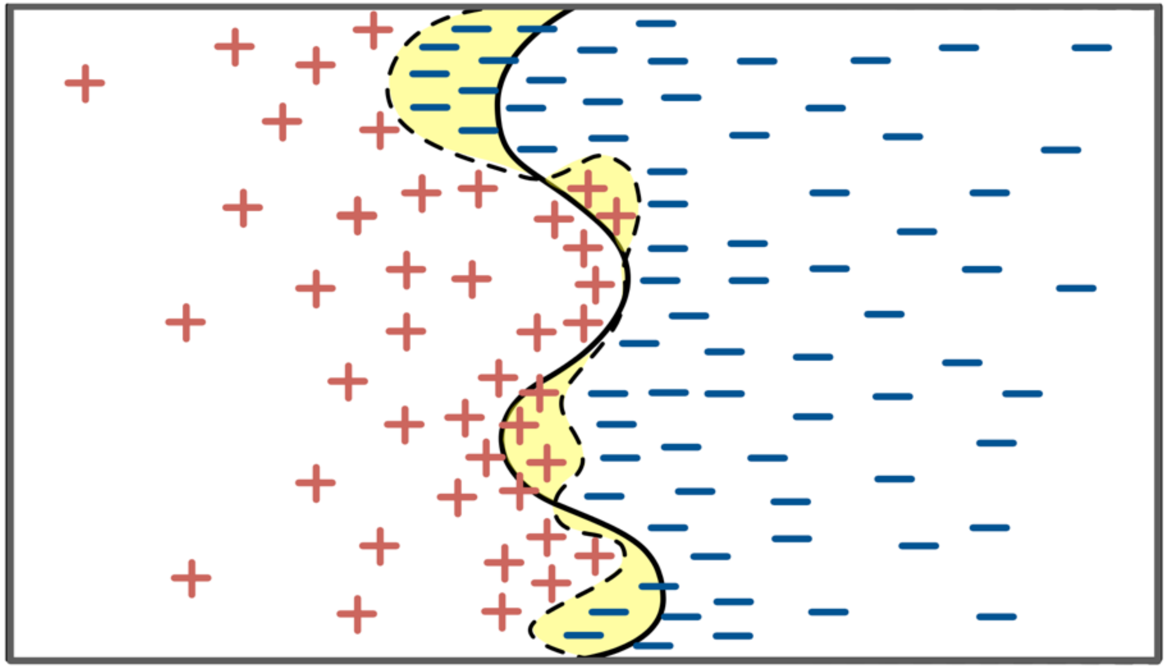
# Literature – quantifying human understanding

30+ papers

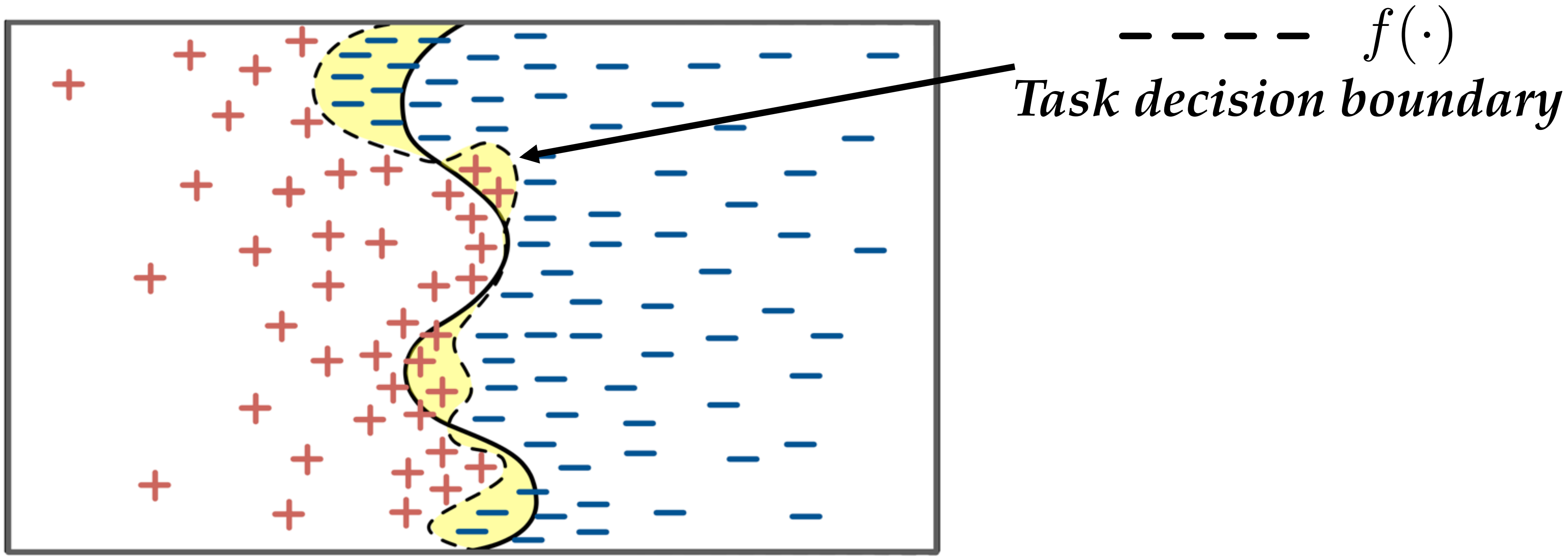
https://github.com/Chacha-Chen/Explanations-Human-Studies

| Paper | Model | Prediction | Explanations | $g$ | $z$ | $f$ |
|---|---|---|---|---|---|---|
| Colin et al. (2022) | InceptionV1, ResNet | Hidden | Local feature importance (Saliency,Gradient Input, Integrated Gradients, Occlusion (OC), SmoothGrad (SG) and Grad-CAM) | ✓ | ✗ | ✗ |
| Taesiri et al. (2022) | ResNet, kNN, other deep learning models | Shown | Confidence score, example-based methods (nearest neighbors) | ✗ | ✓ | ✓ |
| Kim et al. (2022) | CNN, BagNet, ProtoPNet, ProtoTree | Mixed | Example-based methods (ProtoPNet, ProtoTree), local feature importance (GradCAM, BagNet) | ✓ | ✓ | ✓ |
| Nguyen et al. (2021) | ResNet | Shown | Model uncertainty (classification confidence (or probability)); Local feature importance (gradient-based, salient-object detection model); Example-based methods (prototypes) | ✗ | ✓ | ✓ |
| Buçinca et al. (2021) | Wizard of Oz | Shown | Model uncertainty (classification confidence (or probability)) | ✗ | ✓ | ✓ |
| Chromik et al. (2021) | Decision trees/random forests | Shown | Local feature importance (perturbation-based SHAP) | ✓ | ✗ | ✗ |
| Nourani et al. (2021) | Other deep learning models | Shown | Local feature importance (video features) | ✓ | ✓ | ✓ |
| Liu et al. (2021) | Support-vector machines (SVMs) | Shown | Local feature importance (coefficients) | ✓ | ✓ | ✓ |
| Wang & Yin (2021) | Logistic regression | Shown | Example-based methods (Nearest neighbor or similar training instances); Counterfactual explanations (counterfactual examples); Global feature importance (permutation-based); | ✓ | ✓ | ✗ |
| Poursabzi-Sangdeh et al. (2021) | Linear regression | Shown | Presentation of simple models (linear regression); Information about training data (input features or information the model considers) | ✓ | ✓ | ✓ |
| Bansal et al. (2020) | RoBERTa; Generalized additive models | Shown | Model uncertainty (classification confidence (or probability)); Local feature importance (perturbation-based (LIME)); Natural language explanations (expert-generated rationales); | ✗ | ✓ | ✓ |
| Zhang et al. (2020) | Decision trees/random forests | Shown | Model uncertainty (classification confidence (or probability)); Local feature importance (perturbation-based SHAP); Information about training data (input features or information the model considers) | ✗ | ✓ | ✓ |
| Abdul et al. (2020) | Generalized additive models | Shown | Global feature importance (shape function of GAMs) | ✓ | ✗ | ✗ |
| Lucic et al. (2020) | Decision trees/random forests | Hidden | Counterfactual explanations (contrastive or sensitive features) | ✓ | ✗ | ✗ |
| Lai et al. (2020) | BERT; Support-vector machines | Shown | Local feature importance (attention); Model performance (accuracy); Global example-based explanations (model tutorial) | ✗ | ✓ | ✓ |
| Alqaraawi et al. (2020) | Convolution Neural Networks | Hidden | Local feature importance (propagation-based (LRP), perturbation-based (LIME)) | ✓ | ✗ | ✗ |
| Carton et al. (2020) | Recurrent Neural Networks | Shown | Local feature importance (attention) | ✗ | ✓ | ✓ |
| Hase & Bansal (2020) | Other deep learning models | Shown | Local feature importance (perturbation-based (LIME)); Rule-based explanations (anchors); Example-based methods (Nearest neighbor or similar training instances); Partial decision boundary (traversing the latent space around a data input) | ✓ | ✗ | ✗ |
| Buçinca et al. (2020) | Wizard of Oz | Mixed | Example-based methods (Nearest neighbor or similar training instances) | ✓ | ✓ | ✓ |
| Kiani et al. (2020) | Other deep learning models | Shown | Model uncertainty (classification confidence (or probability)); Local feature importance (gradient-based) | ✗ | ✓ | ✓ |
| Gonzalez et al. (2020) | Other deep learning models | Shown | extractive evidence | ✗ | ✓ | ✓ |

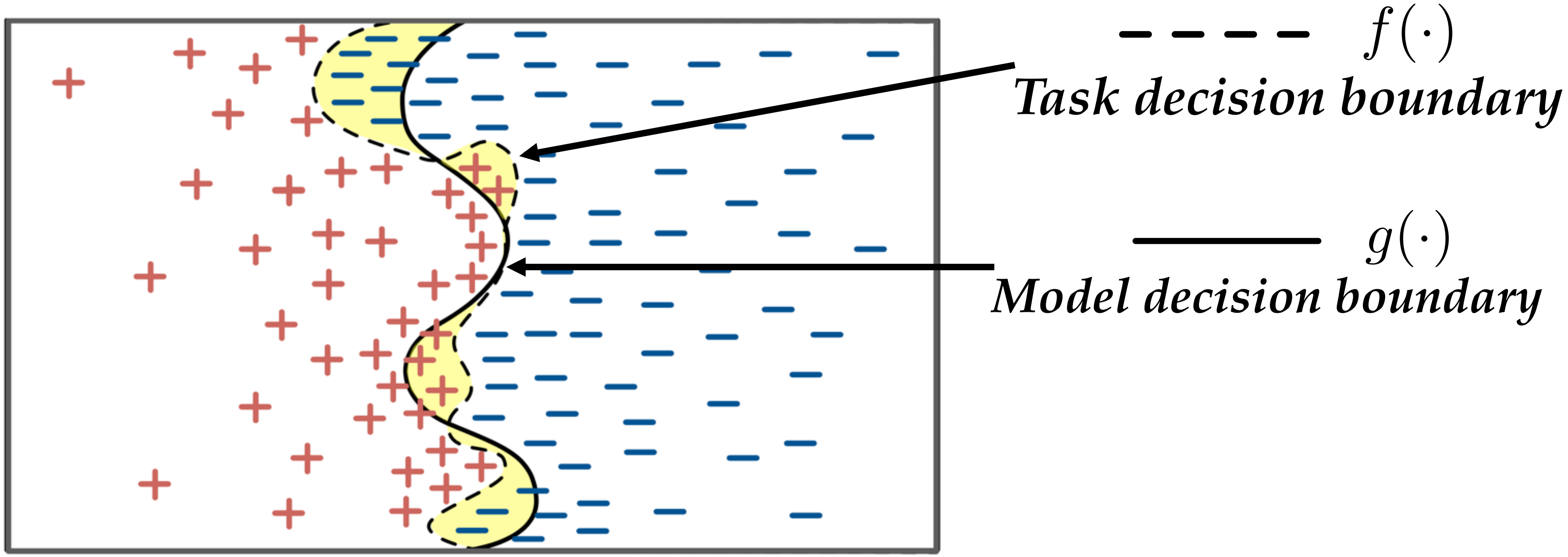
# Three core concepts of human understanding



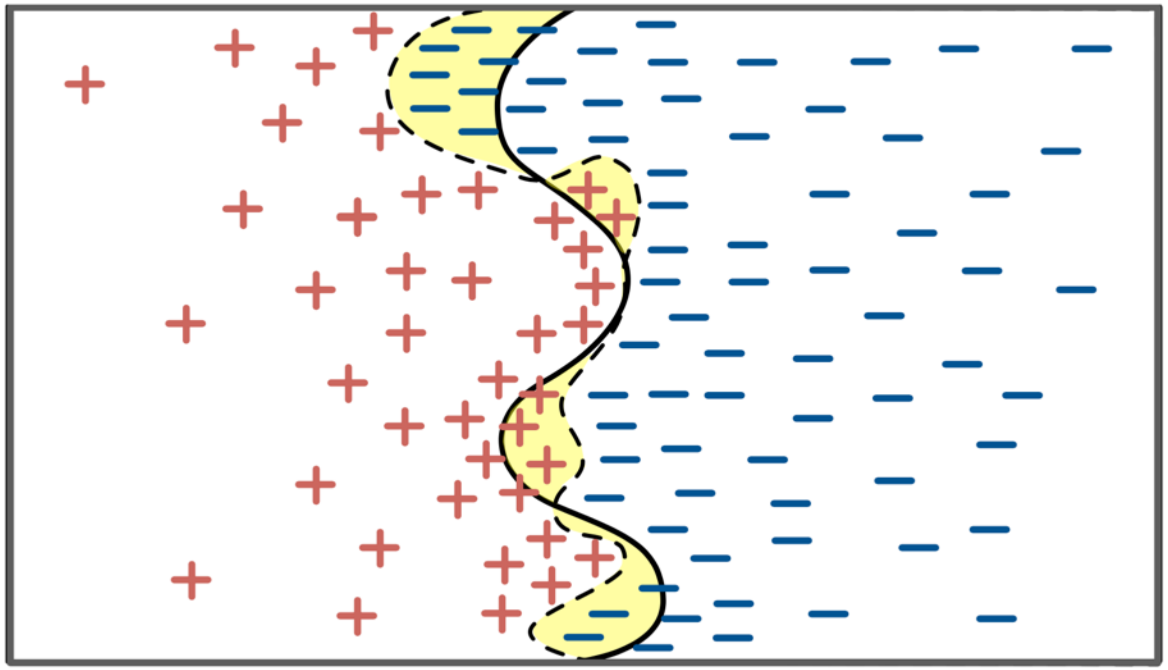
# Three core concepts of human understanding



# Three core concepts of human understanding



# Three core concepts of human understanding

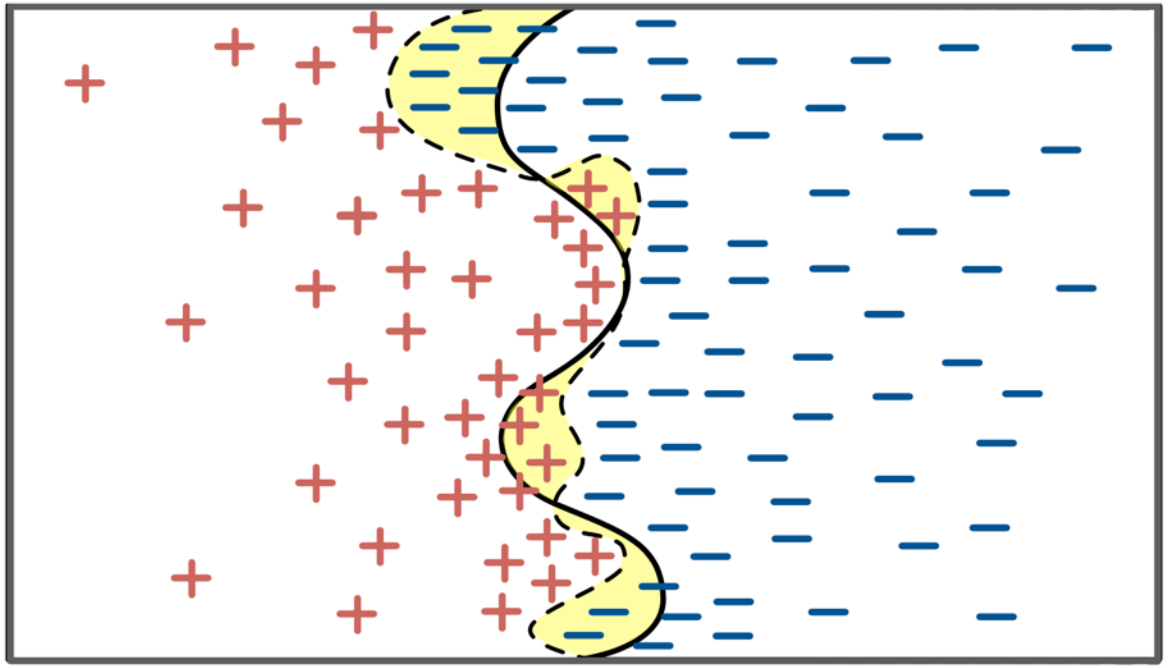


$f(\cdot)$ *Task decision boundary*

$g(\cdot)$ *Model decision boundary*

$z(\cdot)$ *Model error*

# Three core concepts of human understanding





*Existing quantitative measures of human understanding map to one of these three concepts.*

① Measuring human understanding of model decision boundary *via:*

Human simulatability (Chandrasekaran et al., 2018; Poursabzi-Sangdeh et al., 2021; Wang & Yin, 2021; Ribeiro et al., 2018; Alqaraawi et al., 2020;……)
Counterfactual reasoning (Friedler et al., 2019; Lucic et al., 2020)
Feature importance (Wang & Yin, 2021; Ribeiro et al., 2016)

② Measuring human understanding of task decision boundary via:

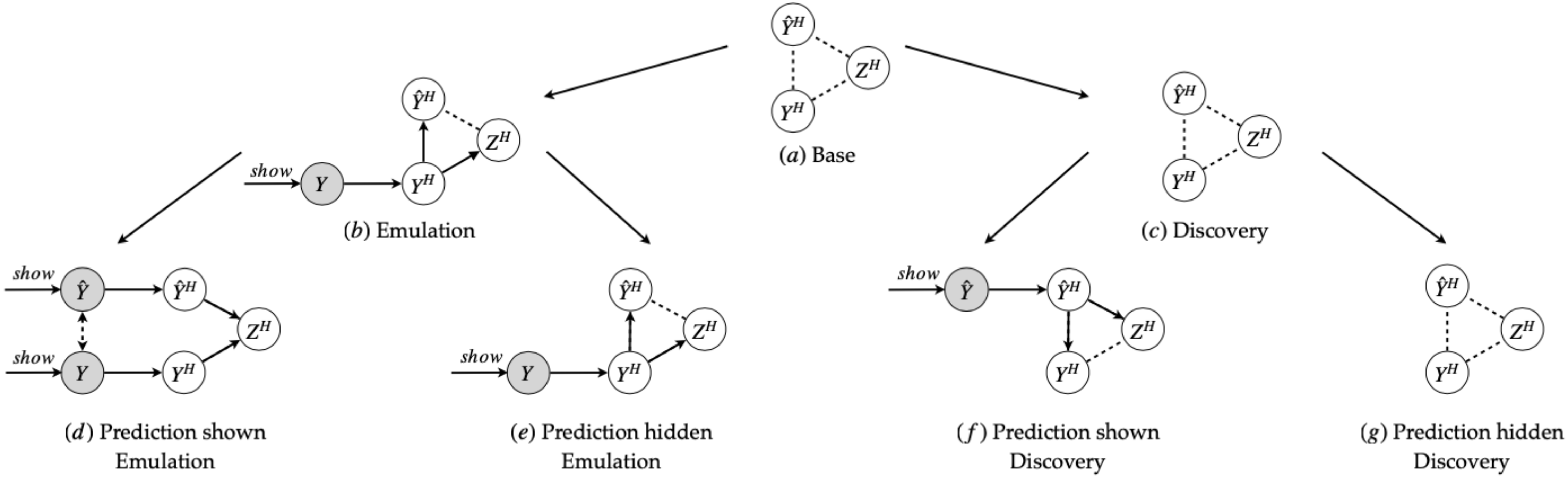
Human + AI performance (Doshi-Velez & Kim, 2017; Bucinca et al., 2021; Poursabzi-Sangdeh et al., 2021; Bansal et al., 2020; Zhang et al., 2020; ……)

③ Measuring human understanding of model error via:

Human trust (Wang & Yin, 2021; Bucinca et al., 2021; Zhang et al., 2020, Bansal et al., 2019; Poursabzi-Sangdeh et al., 2021; Bansal et al., 2020;……)

# A theoretical framework.

# A theoretical framework -- Overview



# Human intuitions are necessary for effective machine explanations.

Human intuitions are necessary for effective machine explanations.

Without assumptions about human intuitions,

💡 Human intuitions are necessary for effective machine explanations.

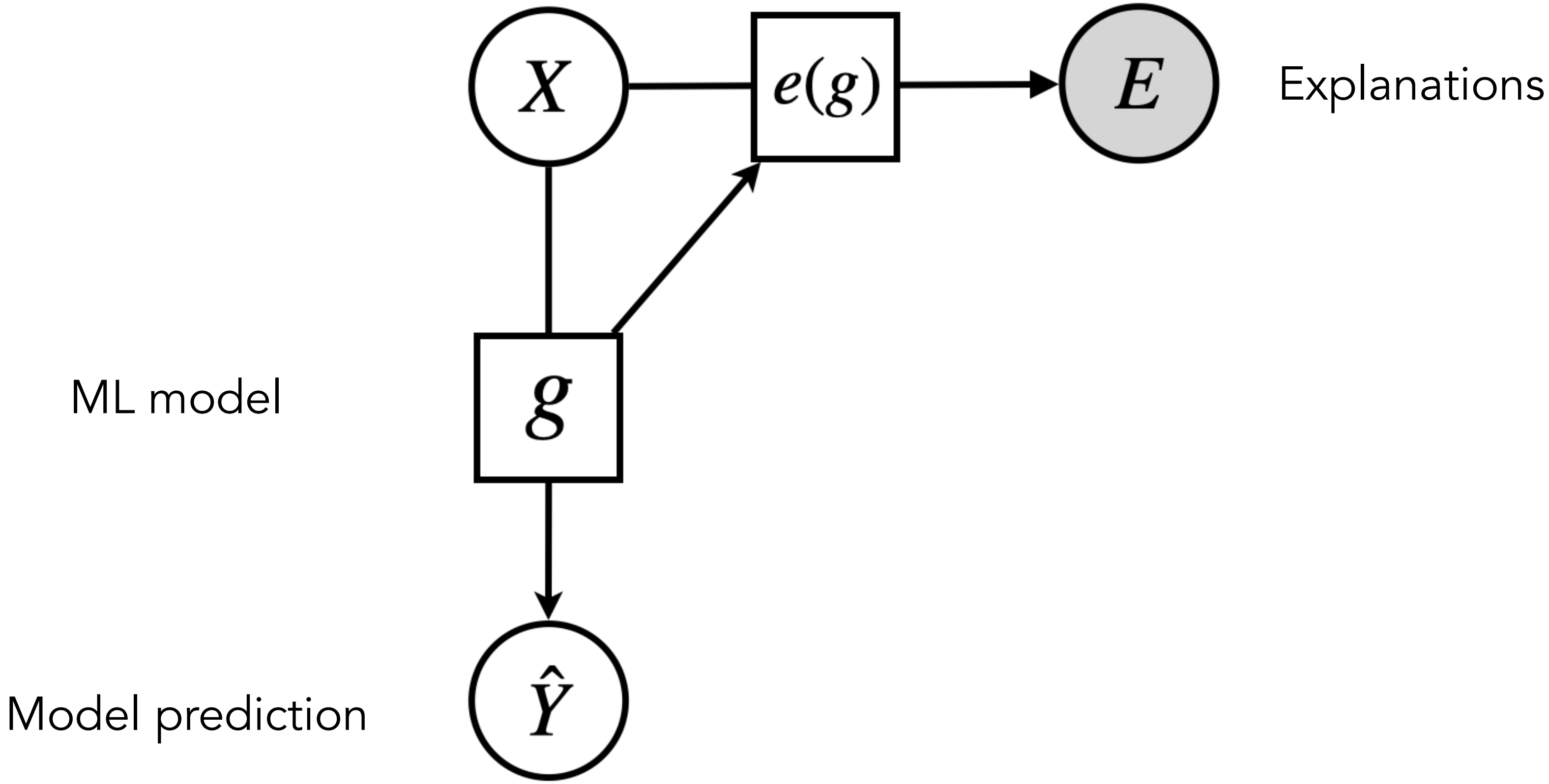
Without assumptions about human intuitions, explanations can improve human understanding of model decision boundary ✅

Human intuitions are necessary for effective machine explanations.

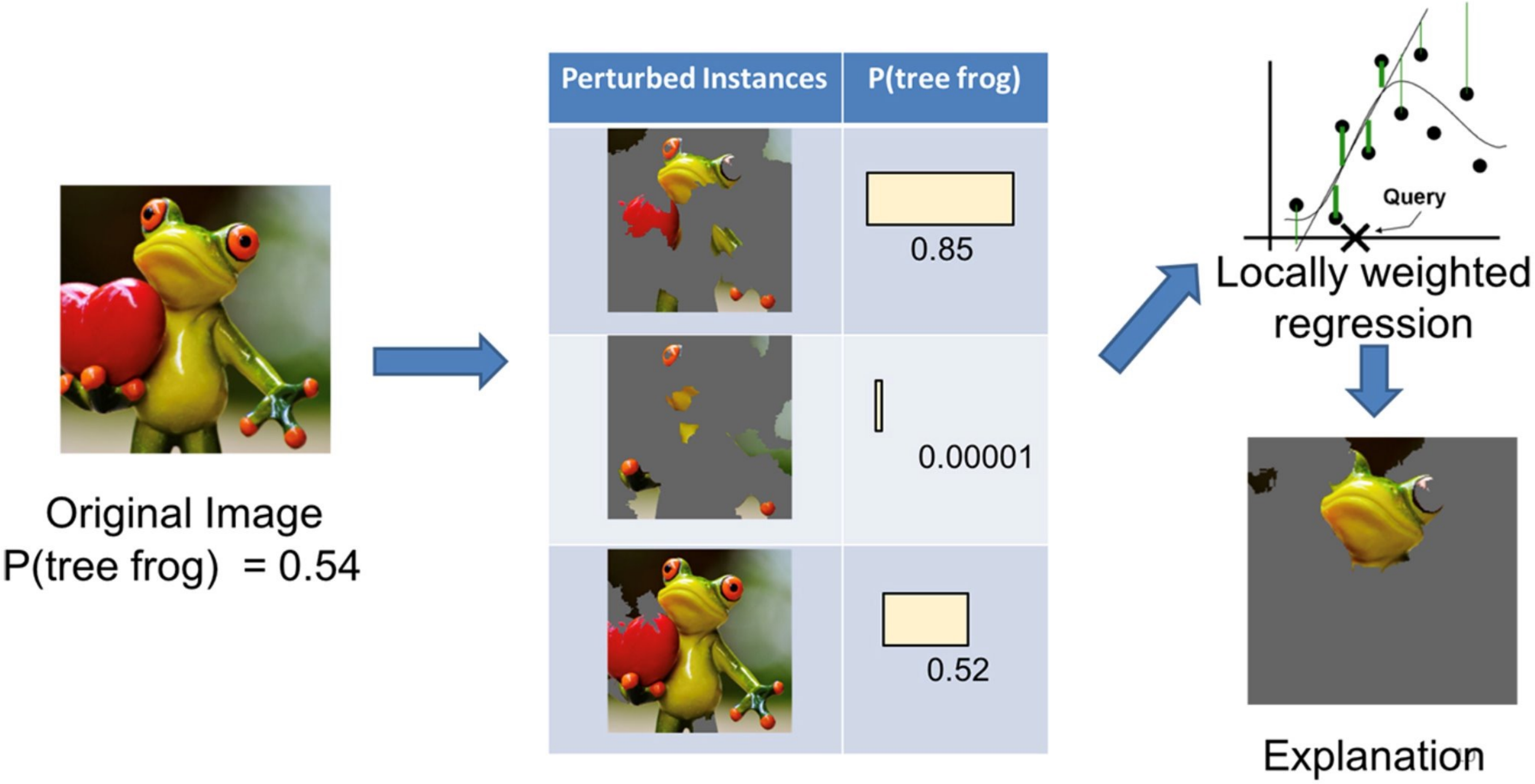
Without assumptions about human intuitions, explanations can improve human understanding of

- model decision boundary ✓
- task decision boundary ✗
- model error ✗

# Existing explanations are derived from model decision boundary

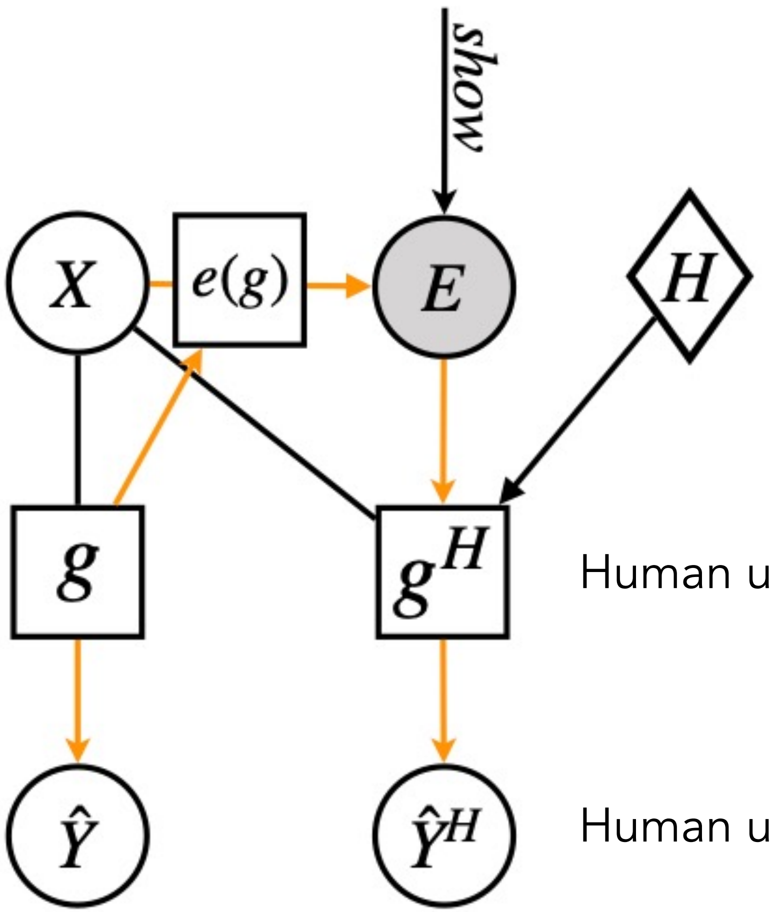


# Existing explanations are derived from model decision boundary




LIME: a popular explanation method. Image credit: Marco Tulio Ribeiro

# Explanations can improve understanding of the model decision boundary ✅




Human intuition

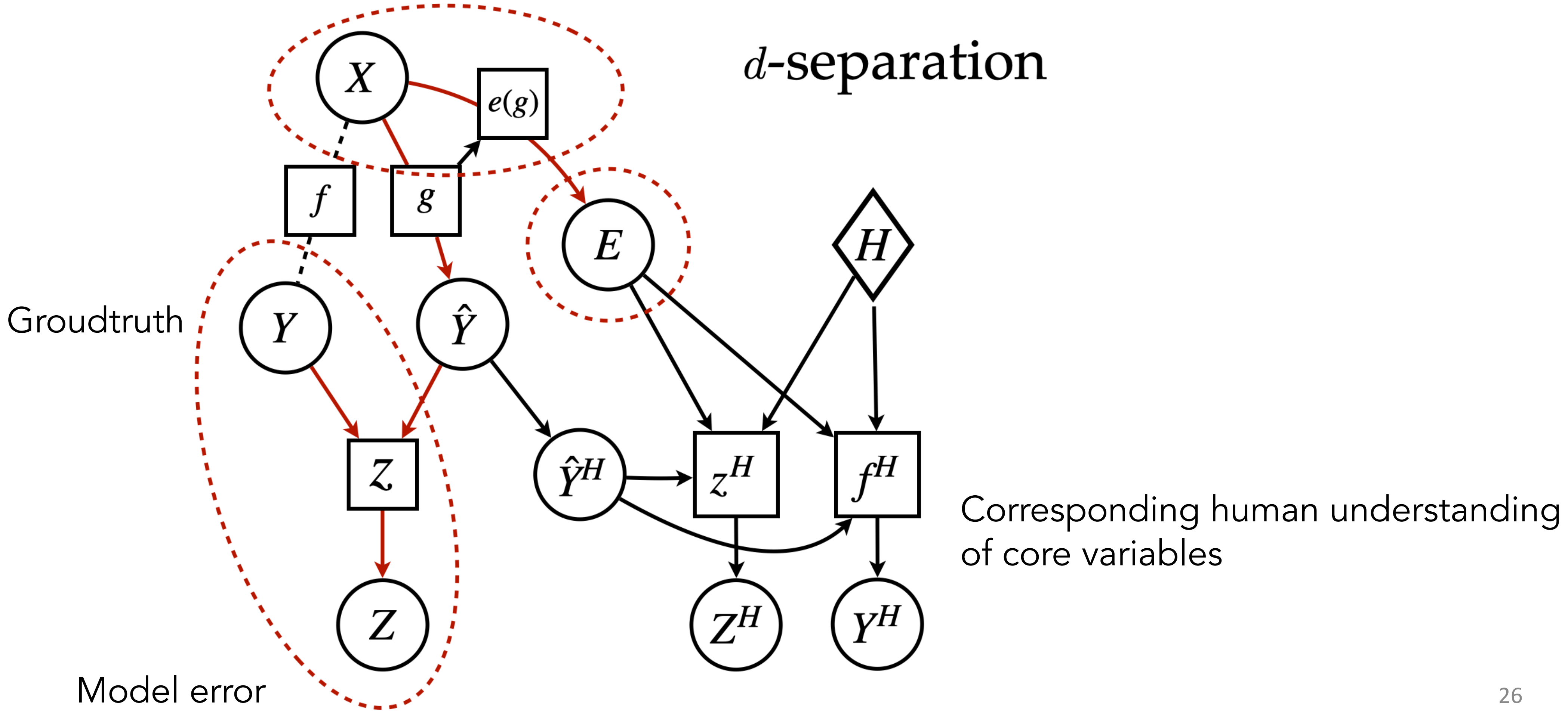
Human understanding of the model

Human understanding of the model prediction

# Explanations cannot offer more information beyond the model decision boundary

task decision boundary ⊗

model error ⊗

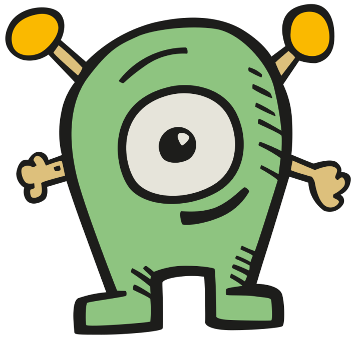


Task: COVID-19 detection

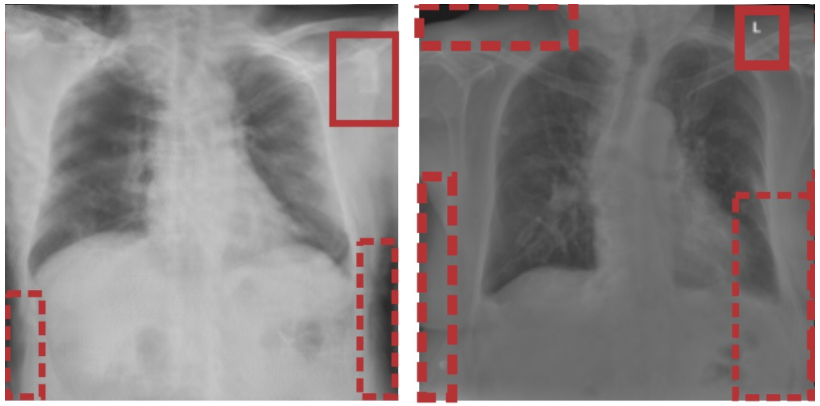
# Consider two cases:
# w/o intuition vs. w intuition

# Case 1: w/o intuition

Aliens do not have any task-specific intuitions



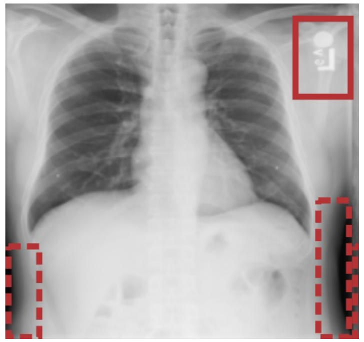
Understanding is bounded by the model decision boundary

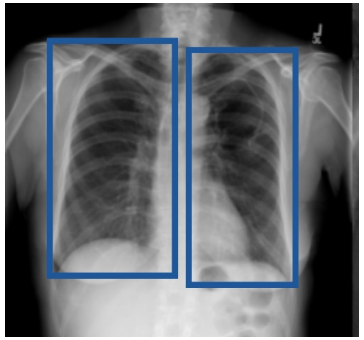


Since aliens can not verify if the important features is correct or not
E can not help with task decision boundary or model error

# Case 2: w/ intuition

Human doctors have any task-specific intuitions





Human can verify when the model could potentially be wrong

This leads to positive utility of explanations:
human + AI > AI

# Human studies to provide a possible way to integrate human intuitions

# Contributions

*#1* Identify the three core concepts of human understanding.

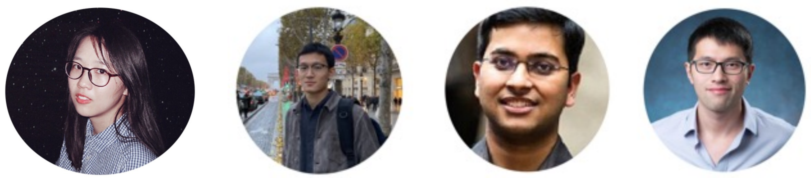
*#2* Propose a theoretical framework of machine explanations and human understanding.

*Human intuition is important!*

*#3* Conduct Human subject studies as an application of our framework.

Survey website

# Thank you so much for listening!



https://arxiv.org/abs/2202.04092

https://github.com/Chacha-Chen/Explanations-Human-Studies

chacha@uchicago.edu

@chachaachen