

CSCE 581: Introduction to Trusted AI

Lectures 13 and 14: (Supervised) ML – Trust Mitigation

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

25TH AND 27TH FEB, 2025

Carolinian Creed: “I will practice personal and academic integrity.”

Credits: Copyrights of all material reused acknowledged

Organization of Lectures 13, 14

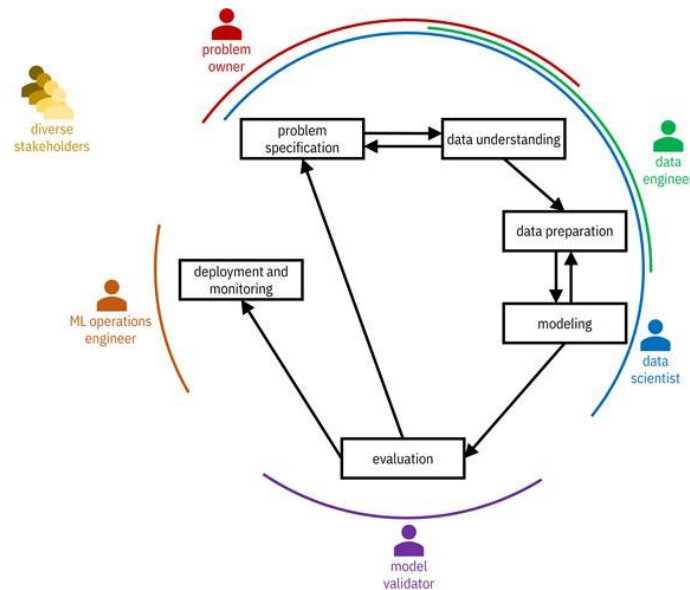
- Introduction Section
 - Recap from Week 6 (Lectures 11 and 12)
 - Announcements and News
- Main Section
 - L13: Mitigation - Explanation Methods
 - L14: Mitigation – Trust Certification / Rating
- Concluding Section
 - About next week – Lectures 15, 16
 - Ask me anything

Introduction Section

Recap from Week 6 (Lectures 11, 12)

- We looked at
 - Fairness methods
 - Overview of major mitigation techniques – explanation and rating

Recap: ML Pipelines



Highly Simplified View

Image Credit: Trustworthy Machine Learning, Kush Varshney

AI News

- Blog on Crawl-Walk-Run, as applied to an AI project
 - <https://www.linkedin.com/pulse/crawl-walk-run-approach-ai-based-real-world-problem-biplav-srivastava-pxsre/>

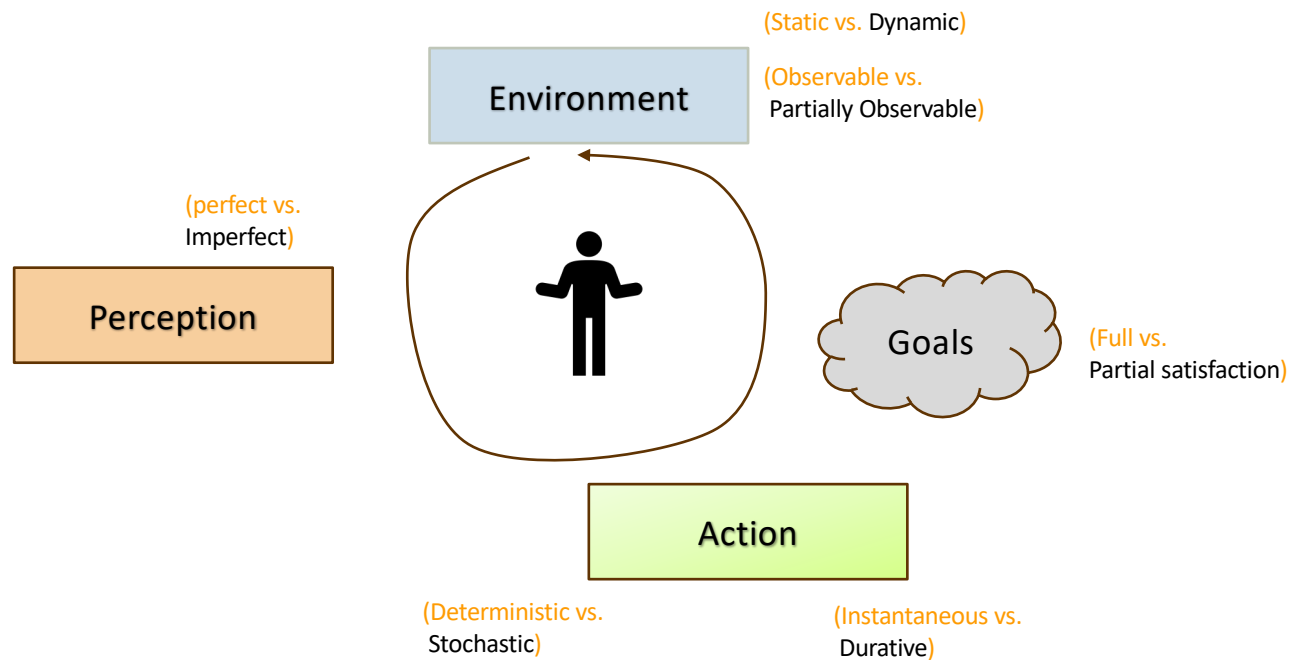
Announcement: Change to Student Assessment

A = [920-1000]
B+ = [870-919]
B = [820-869]
C+ = [770-819]
C = [720-769]
D+ = [670-719]
D = [600-669]
F = [0-599]

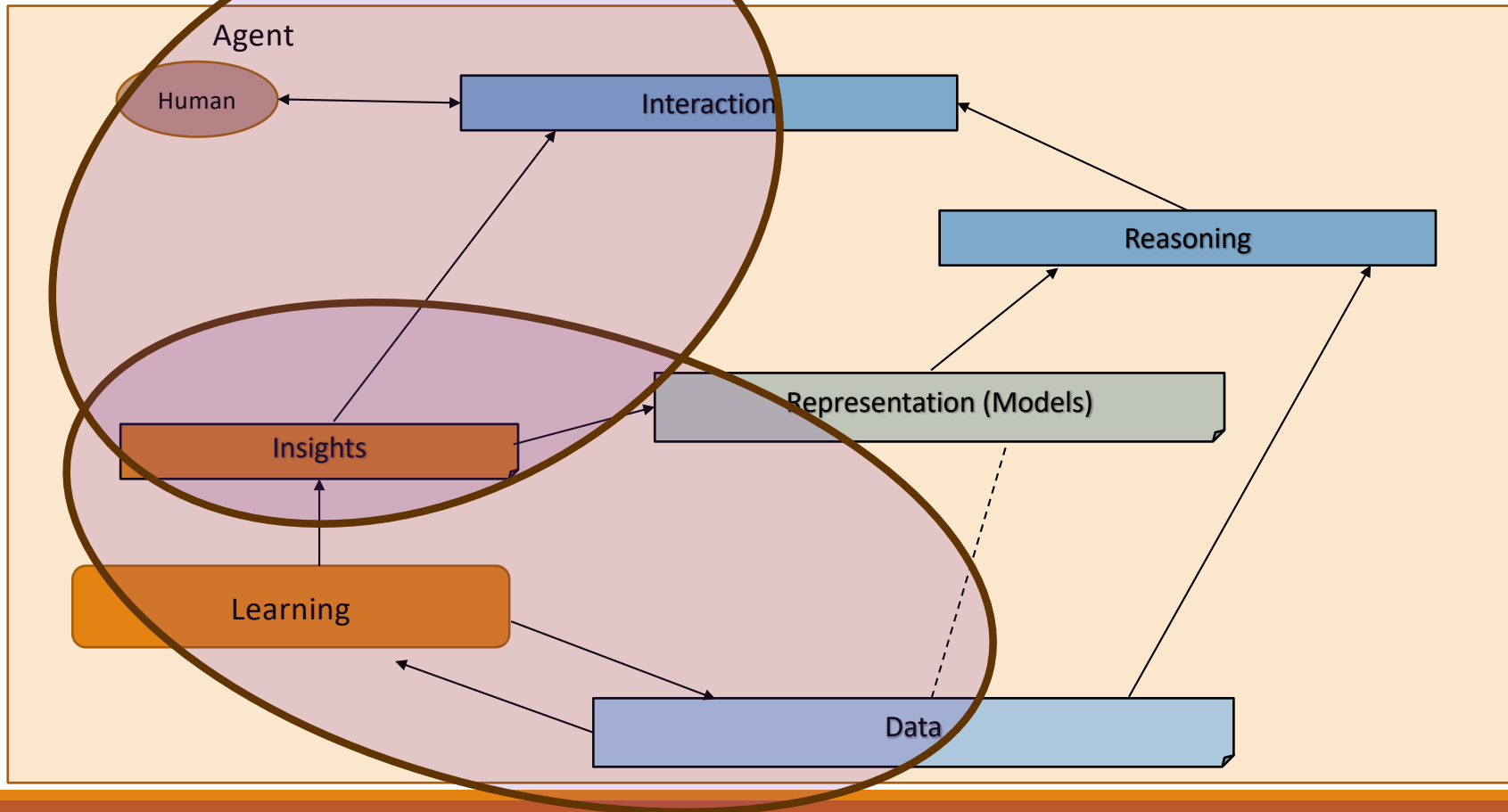
Tests	Undergrad	Grad
Course Project – report, in-class presentation	600	600
Quiz – 2 quizzes	200	200
Final Exam	200	100
Additional Final Exam – Paper summary, in-class presentation		100
Total	1000 points	1000 points

Change: 4 quizzes to 2; no best of 3

Intelligent Agent Model



Relationship Between Main AI Topics (Covered in Course)



High Level Semester Plan (Adapted, Approximate)

CSCE 581 –

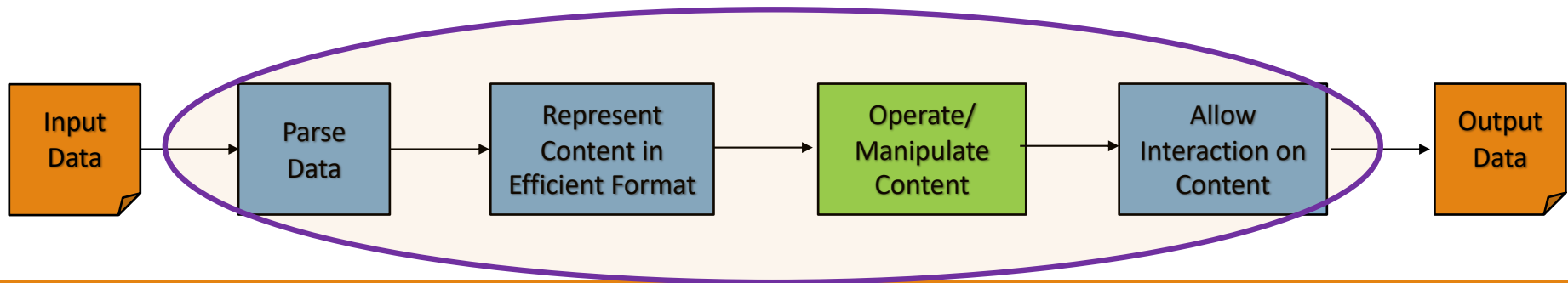
- Week 1: Introduction
- Week 2: Background: AI - Common Methods
- Week 3: The Trust Problem
- Week 4: Machine Learning (Structured data) - Classification
- Week 5: Machine Learning (Structured data) - Classification – Trust Issues
- Week 6: Machine Learning (Structured data) – Classification – Mitigation Methods
- Week 7: Machine Learning (Structured data) – Classification – Explanation Methods
- Week 8: Machine Learning (Text data, **vision**) – Classification,

Large Language Models

- Week 9: Machine Learning (Text data) - Classification – Trust Issues, LLMs
- Week 10: Machine Learning (Text data) – Classification – Mitigation Methods
- Week 11: Machine Learning (Text data) – Classification – Explanation Methods
- Week 12: Emerging Standards and Laws, **Real world applications**
- Week 13: Project presentations
- Week 14: Project presentations, Conclusion

AI/ ML topics and with a
focus on fairness, explanation,
Data privacy, reliability

Main Segment



ML Pipelines and Trust-Based Intervention Considerations

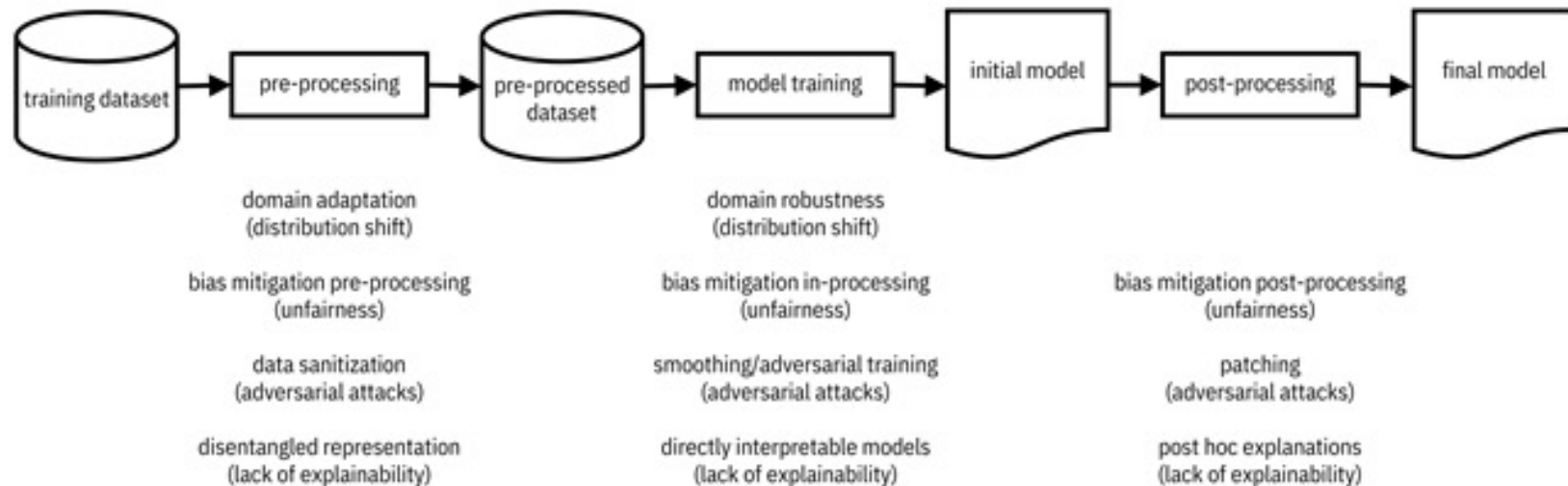


Image Credit: Trustworthy Machine Learning, Kush Varshney

Generating Explanations

Trust Issues – Mitigate via Explanations

- Explain behavior
 - Remove undesirable behavior ?
 - Explain that too?

Trust Dimensions
Competent
Reliable
Upholds human values
Allows human interaction

What is the Purpose of Explanations

Purposes for explanations in **psychology**

- To **predict similar events** in the future: *slippery roads can cause a fall*. Use information later.

Explanations: correlational (similar or contrastive)

- For **diagnosis**: *why a system failed and then repair a part to bring it back to its normal function*

Explanations: causal (state-state or state-action; past)

- To **affix blame**: *for a crime*

Explanations: causal (state-actor; past)

- To justify or **rationalize an action**: *sweet to an enemy because of the strategic value of being nice on that occasion*

Explanations: causal (state-action; past or future)

- In the service of **aesthetic** pleasure

Source: Keil F. C. (2006). Explanation and understanding. *Annual review of psychology*, 57, 227–254.
<https://doi.org/10.1146/annurev.psych.57.102904.190100>
<https://pubmed.ncbi.nlm.nih.gov/16318595/>

Instructor (Biplav) comments

In AI, Stakeholders for Explanations

- Executives

- Executives want to use explainability as a market differentiator. Do we need explanations?

Explanations: to predict, rationalize an action

- ML engineers

- Developers want to improve models. How to improve model's performance?
 - They want to debug their models. How to diagnose a problem?

Explanations: for diagnosis, to predict, rationalize an action, affix a blame

- End-users

- End-users want to reap benefits. They want to understand business decisions emanating from usage of AI
 - Why was my loan denied?
 - Why a particular treatment was recommended or de-prioritized ?

Explanations: rationalize an action, affix a blame

- Regulators

- Based on existing laws, the AI (including the developer of AI) is evaluated to ensure that it does not discriminate towards the end users That is, based in application domain,
 - Patient does not get discriminated [Health]
 - The loan applicant is treated equally [Finance]
 - The job seeker ..., the housing loan ..., ...

Explanations: to predict, affix a blame

Source: Explainable Machine Learning in Deployment, FAT* 2020,
<https://arxiv.org/pdf/1909.06342.pdf>; Video: <https://www.youtube.com/watch?v=Hofl4uwxtPA>

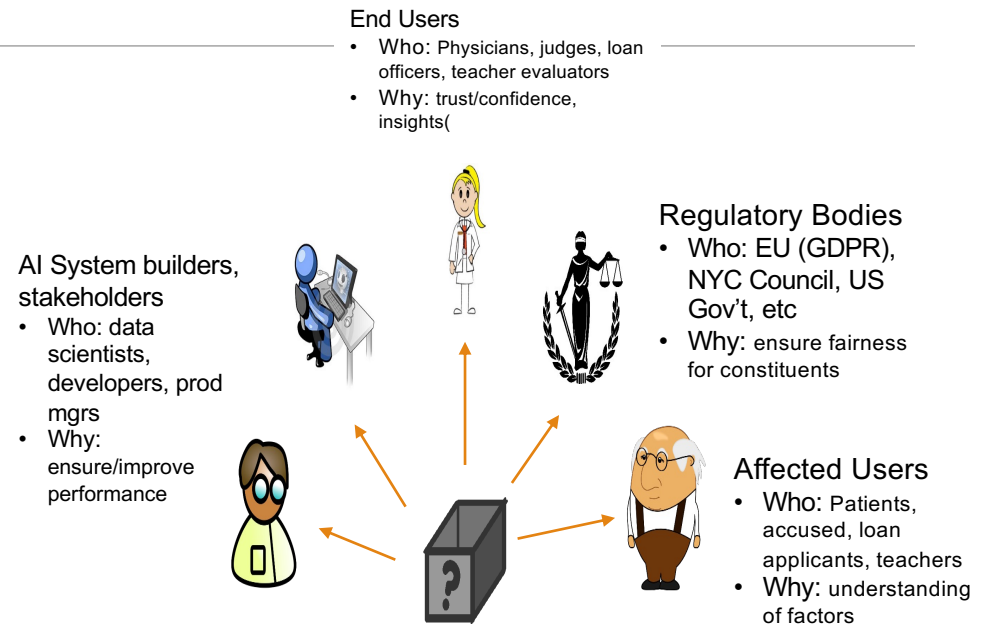
AI Explainability from Legal Requirements

the General Data Protection Regulation (GDPR)

- Limits to **decision-making** based solely on **automated processing** and profiling (Art.22)
- Right to be provided with **meaningful information** about the **logic** involved in the decision (Art.13 (2) f. and 15 (1) h)

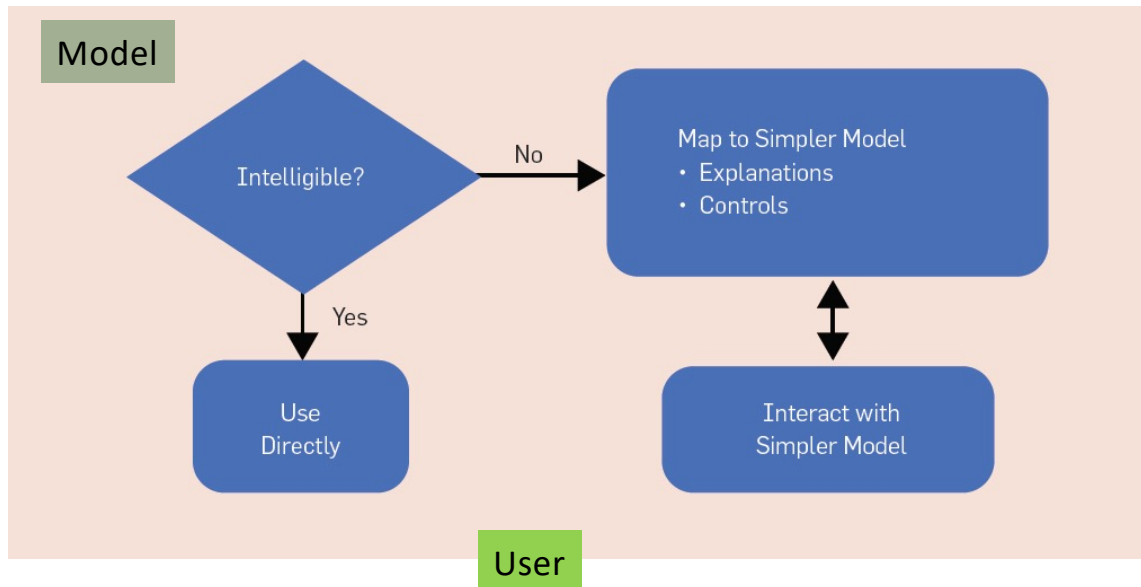
Explanations: rationalize an action

Meaningful explanations depend on the consumer



Must match the **complexity capability** of the consumer
Must match the **domain knowledge** of the consumer

Setting and Terminology: Intelligible Models and Explanations



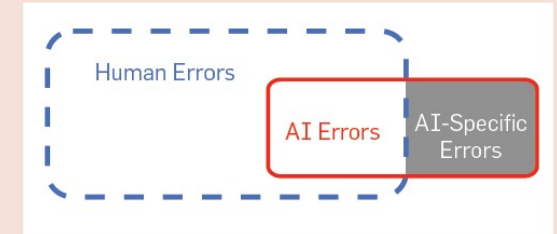
- Transparency: providing stakeholders with relevant information about how a model works
- Explainability: Providing insights into model's behavior for specific datapoints

Sources:

1. The Challenge of Crafting Intelligible Intelligence, Daniel S. Weld, Gagan Bansal, Communications of the ACM, June 2019, Vol. 62 No. 6, Pages 70-79, 10.1145/3282486
2. Explainable Machine Learning in Deployment, FAT* 2020.

Need for Intelligibility

The red shape denotes the AI's mistakes; its smaller size indicates a net reduction in the number of errors. The gray region denotes AI-specific mistakes a human would never make. Despite reducing the total number of errors, a deployed model may create new areas of liability (gray), necessitating explanations.



- **AI may have the wrong objective:** is AI right for the right reasons?
- **AI may be using inadequate features:** understand modeling issues
- **Distributional drift:** detect when and why models are failing to generalize
- **Facilitating user control:** guiding what preferences to learn
- **User acceptance:** especially for costly actions
- **Improving human insight:** improve algorithm design
- **Legal imperatives**

Source: The Challenge of Crafting Intelligible Intelligence, Daniel S. Weld, Gagan Bansal, Communications of the ACM, June 2019, Vol. 62 No. 6, Pages 70-79, 10.1145/3282486

Types of Explanation Generation Methods

- **Feature-based**: from the features of the data, which feature(s) were most important for given decision output
 - Example: For a loan, is it income or the person's age ?
- **Sample-based**: from data in training, which data points were important for given test point; helps understand sampling and its representation in wider population
 - Example: For a loan, what instances similar to the loan application would have gotten the loan ?
- **Counter-factual**: what-ifs – what do you change about the input to change the decision output
 - Example: For a loan, does getting an additional borrower insurance increase chance of getting the loan?
- Natural language

Source: Explainable Machine Learning in Deployment, FAT* 2020

References for AI Explainability

Papers

- The Challenge of Crafting Intelligible Intelligence, Daniel S. Weld, Gagan Bansal, Communications of the ACM, June 2019, Vol. 62 No. 6, Pages 70-79, 10.1145/3282486
- “Why Should I Trust You?” Explaining the Predictions of Any Classifier, Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, in ACM’s Conference on Knowledge Discovery and Data Mining, KDD2016;
<https://homes.cs.washington.edu/~marcotcr/blog/lime/>,
<https://www.oreilly.com/content/>
- Explainable Machine Learning in Deployment, FAT* 2020,
<https://arxiv.org/pdf/1909.06342.pdf>; Video:
<https://www.youtube.com/watch?v=Hofl4uwxtPA>
- Tim Miller. 2023. Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven Decision Support using Evaluative AI. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23).
[Argues why explainability should support evidence for all interpretations, and let the user decide]

Tutorial: XAI tutorial at AAAI 2020

<https://xaitutorial2020.github.io/>

Tool: AIX 360

Tool: <https://aix360.mybluemix.net/>

Video:

<https://www.youtube.com/watch?v=Yn4yduyoQh4>

Paper: <https://arxiv.org/abs/1909.03012>

Tool: InterpretML

<https://github.com/interpretml/interpret>

LIME — Local Interpretable Model-Agnostic Explanations

Paper: “Why Should I Trust You?” Explaining the Predictions of Any Classifier, Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, ACM’s Conference on Knowledge Discovery and Data Mining, KDD2016

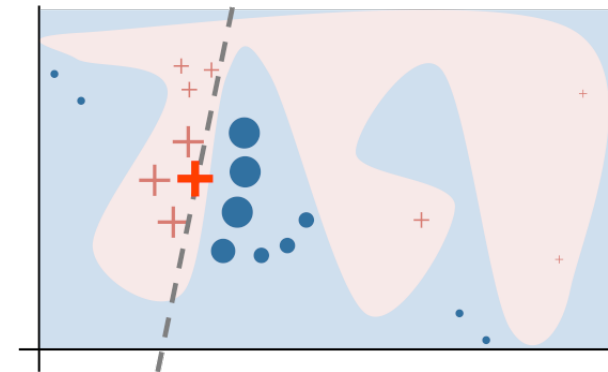
Blogs:

- <https://homes.cs.washington.edu/~marcotcr/blog/lime/>
- <https://www.oreilly.com/content/introduction-to-local-interpretable-model-agnostic-explanations-lime/>

Code: <https://github.com/marcotcr/lime>

LIME Key Idea

- Generate a local, linear explanation for any model
- How
 - Perturb near the neighborhood of a point of interest, X (**Local**)
 - Fit a linear function to the model's output (**Linear**)
 - Interpret coefficients of the linear function (**Explain**)
 - **Visualize**
- Applicability
 - Any classification model!



LIME on Text

Question: Why is a classifier with >90% accuracy predicting based on ?

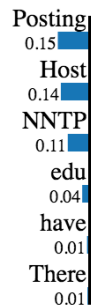
Task: classifying religious inclination from email text

Prediction probabilities



atheism

christian



Text with highlighted words

From: johnchad@triton.unm.edu (jchadwic)
Subject: Another request for Darwin Fish
Organization: University of New Mexico, Albuquerque
Lines: 11
NNTP-Posting-Host: triton.unm.edu

Hello Gang,

There have been some notes recently asking where to obtain the DARWIN fish.
This is the same question I have and I have not seen an answer on the net. If anyone has a contact please post on the net or email me.

“If we **remove** the words **Host** and **NNTP** from the document, we expect the classifier to predict **atheism** with probability $0.58 - 0.14 - 0.11 = 0.31$ ”

Source: <https://github.com/marcotcr/lime>

Code Examples for Tabular Data

- LIME
 - Iris dataset and supervised classifiers – random forest and logistic regression, tabular data:
<https://github.com/biplav-s/course-tai/blob/main/sample-code/l9-explanations/LIME%20explanations%20on%20tabular%20data.ipynb>
- Many other examples
 - <https://github.com/biplav-s/course-d2d-ai/tree/main/sample-code/l12-explanability-autoai>

LIME on Image

Question: Why is this a frog?

Divide image into interpretable components - contiguous superpixels



Original Image

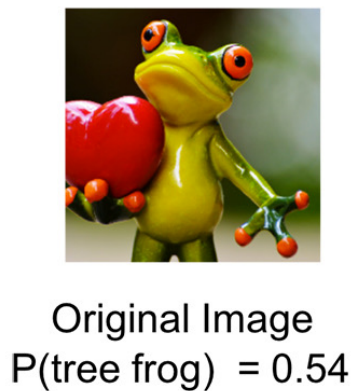


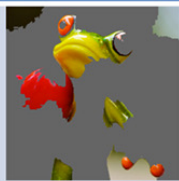

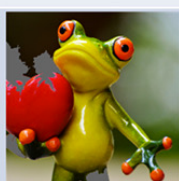
Interpretable
Components

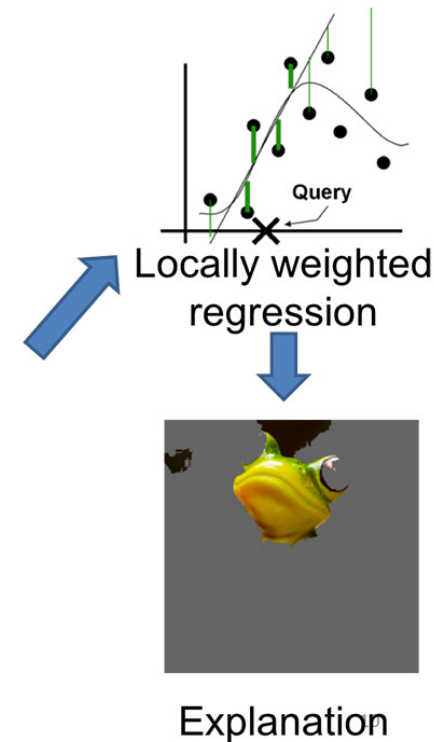
Source: <https://www.oreilly.com/content/introduction-to-local-interpretable-model-agnostic-explanations-lime/>

LIME

1. Generate a data set of perturbed instances by turning some of the interpretable components "off" (gray).
2. For each perturbed instance, calculate probability that a tree frog is in the image according to the model.
3. Learn a simple (linear) model on this data set, which is locally weighted
4. Output regions with highest positive weights as an explanation, graying out everything else.



Perturbed Instances	$P(\text{tree frog})$
	<div><div></div>0.85</div>
	<div><div></div>0.00001</div>
	<div><div></div>0.52</div>



Explanation and Practical Implications

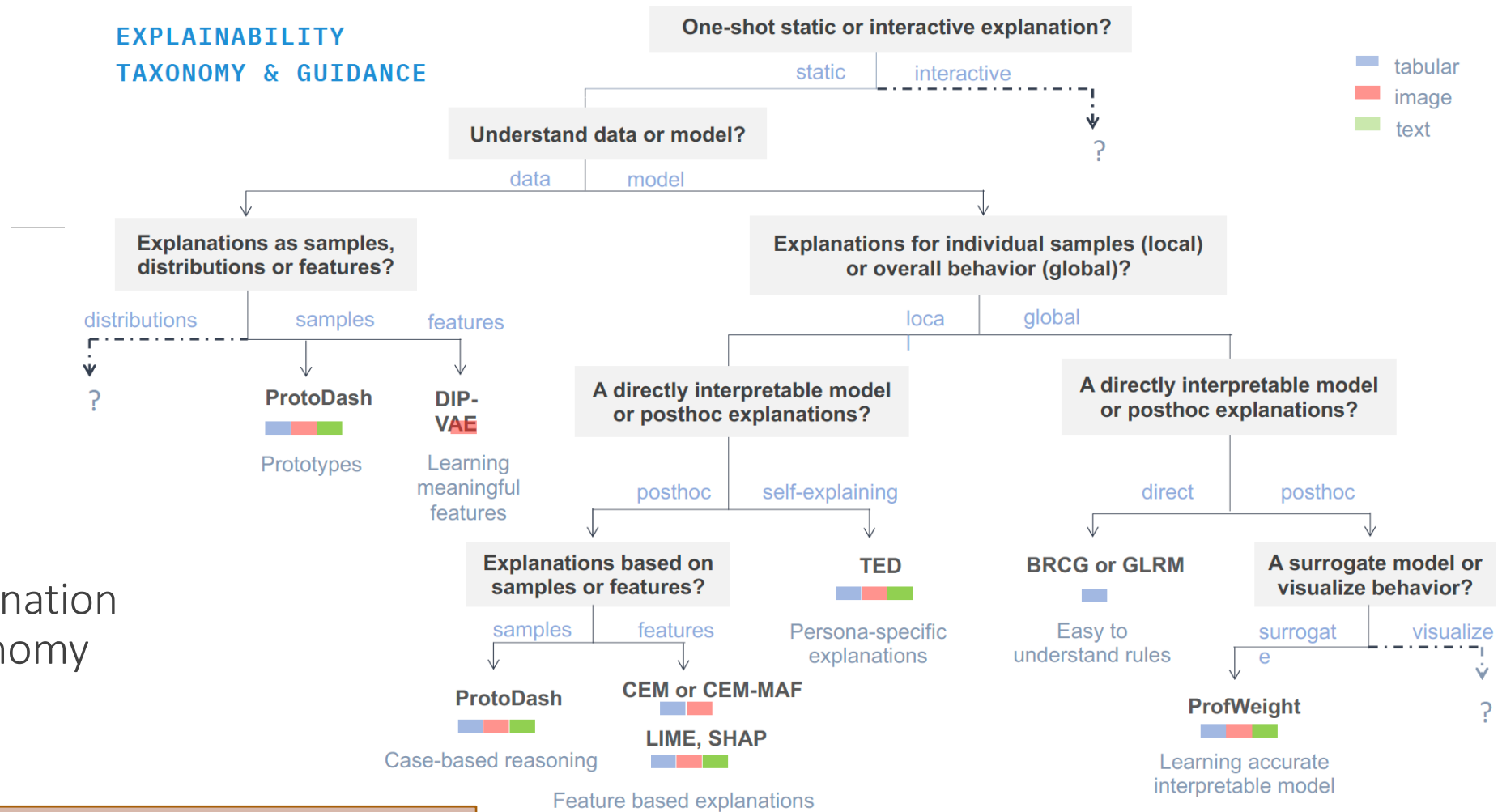
- Context
 - Problem: detect common cardiovascular conditions
 - Data: ECG data
 - Explanation: LIME
- References
 - Blog: <https://www.ucsf.edu/news/2021/08/421301/ai-algorithm-matches-cardiologists-expertise-while-explaining-its-decisions>
 - Paper: <https://jamanetwork.com/journals/jamacardiology/article-abstract/2782549>

InterpretML

- **Details:** <https://github.com/interpretml/interpret>
 - Whitebox (Glassbox) models: change learning code to introduce explainability support
 - Blackbox models: don't change learning code

Interpretability Technique	Type
Explainable Boosting	glassbox model
APLR	glassbox model
Decision Tree	glassbox model
Decision Rule List	glassbox model
Linear/Logistic Regression	glassbox model
SHAP Kernel Explainer	blackbox explainer
LIME	blackbox explainer
Morris Sensitivity Analysis	blackbox explainer
Partial Dependence	blackbox explainer

EXPLAINABILITY TAXONOMY & GUIDANCE



Explanation Taxonomy

Figure Credit: Diptikalyan Saha and Vijay Arya, Oct 2021

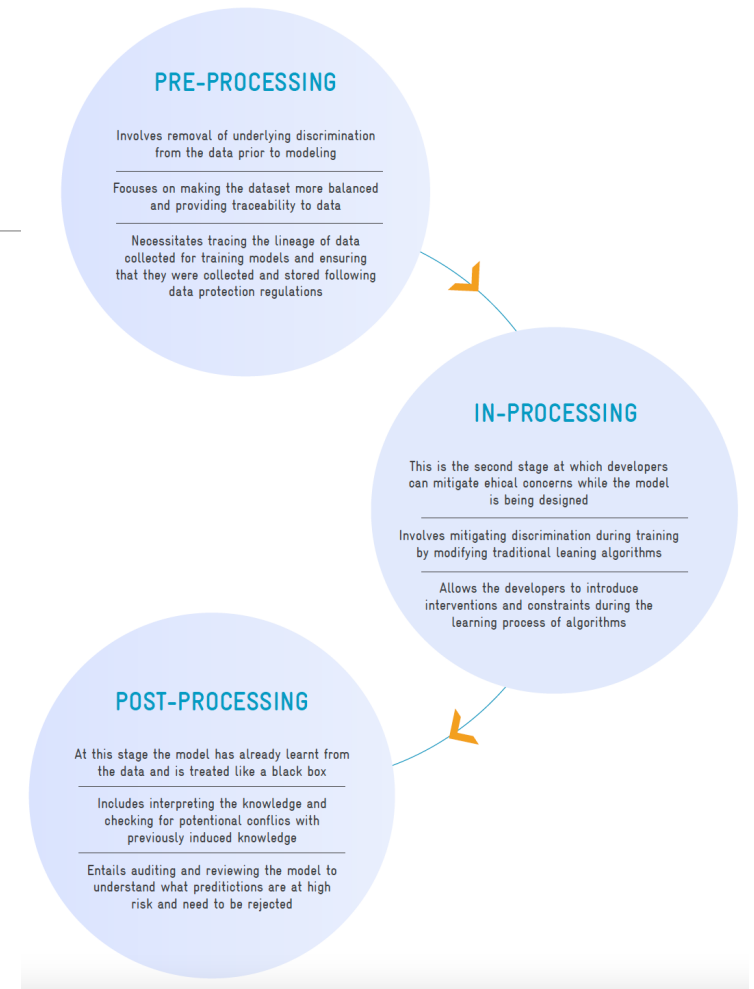
Many Explanation Methods

- Review paper on many methods and data types (image, text, audio, and sensory domains):
 - **How Can I Explain This to You? An Empirical Study of Deep Neural Network Explanation Methods**, *Jeya Vikranth Jeyakumar, Joseph Noor, Yu-Hsi Cheng, Luis Garcia, Mani Srivastava*, [Advances in Neural Information Processing Systems 33 \(NeurIPS 2020\)](https://proceedings.neurips.cc/paper/2020/hash/2c29d89cc56cdb191c60db2f0bae796b-Abstract.html), <https://proceedings.neurips.cc/paper/2020/hash/2c29d89cc56cdb191c60db2f0bae796b-Abstract.html>

Handbook on Data Protection and Privacy for Developers of Artificial Intelligence

- Details:
<https://www.dsci.in/content/privacy-handbook-for-ai-developers>
 - PDF in Blackboard
 - Created for developers with focus on practical considerations
 - Inputs from people from a broad set of background

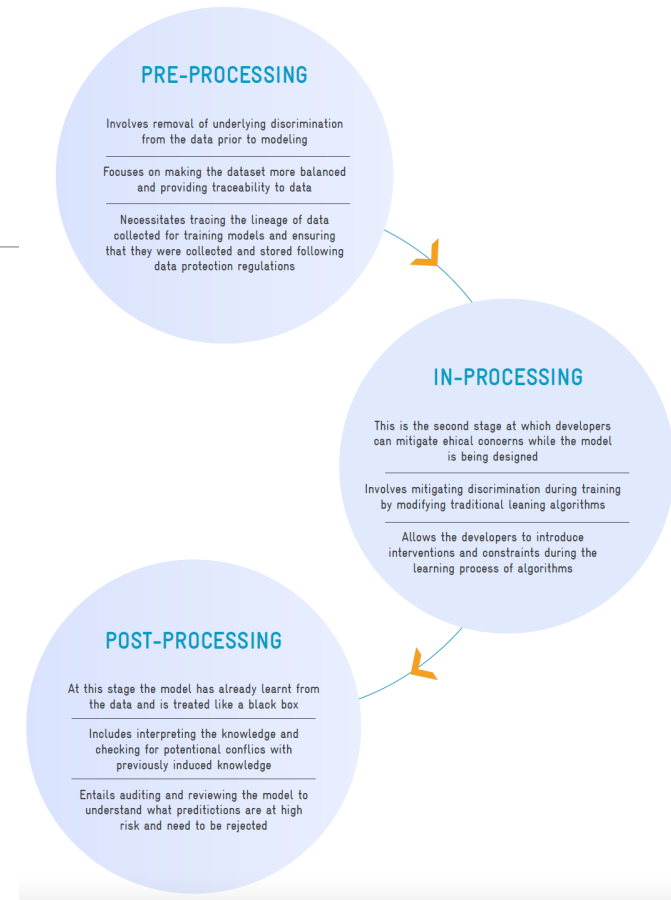
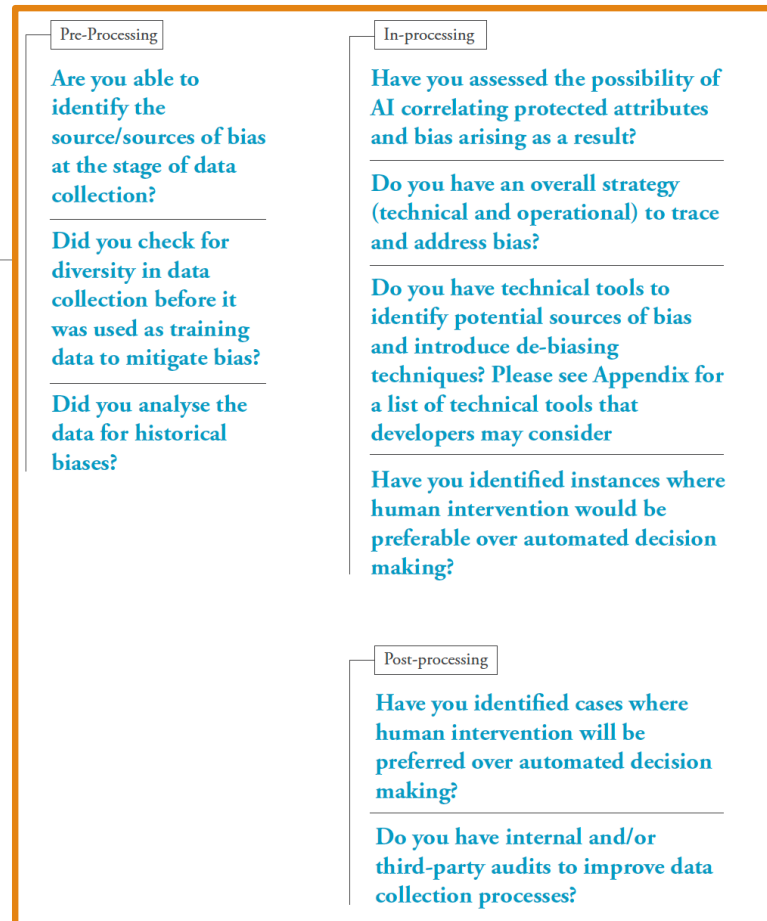
Source: Handbook on Data Protection and Privacy for Developers of Artificial Intelligence, 2021



(Data-based) Reasons for Bias

Source: Handbook on Data Protection and Privacy for Developers of Artificial Intelligence, 2021

Reasons for bias	Explanation
Insufficient data collection	Data collected may be insufficient to represent the social realities of the space that the AI targets. Due to this, AI may not be able to attain its desired output.
Insufficient diversity in data	<p>Data may not be sufficiently diverse to capture all facets of the group an AI-enabled system seeks to work for. In such cases, the data might end up training the AI to discriminate against under-represented groups.</p> <p>For instance, an AI to detect cancer and trained on data available in North European countries may overwhelmingly represent white skin types that have low melanin content as opposed to dark skin tones with higher melanin, leading to incorrect results in a country like India.</p>
Biases in historical data	<p>Even if protected attributes like gender or race are removed, data could have bias due to historical reasons.</p> <p>For example, a <u>hiring algorithm</u> by Amazon favoured applicants based on words like “executed” or “captured” that were mostly used by men in their resumes. Learning from this, the algorithm started preferring men over women and even dismissed resumes with the word ‘woman/women’ in them. Amazon eventually stopped using the algorithm.</p>
Use of poor-quality data	Poor predictions may also be the <u>result of</u> low-quality, outdated, incomplete or incorrect data at different stages of data processing.

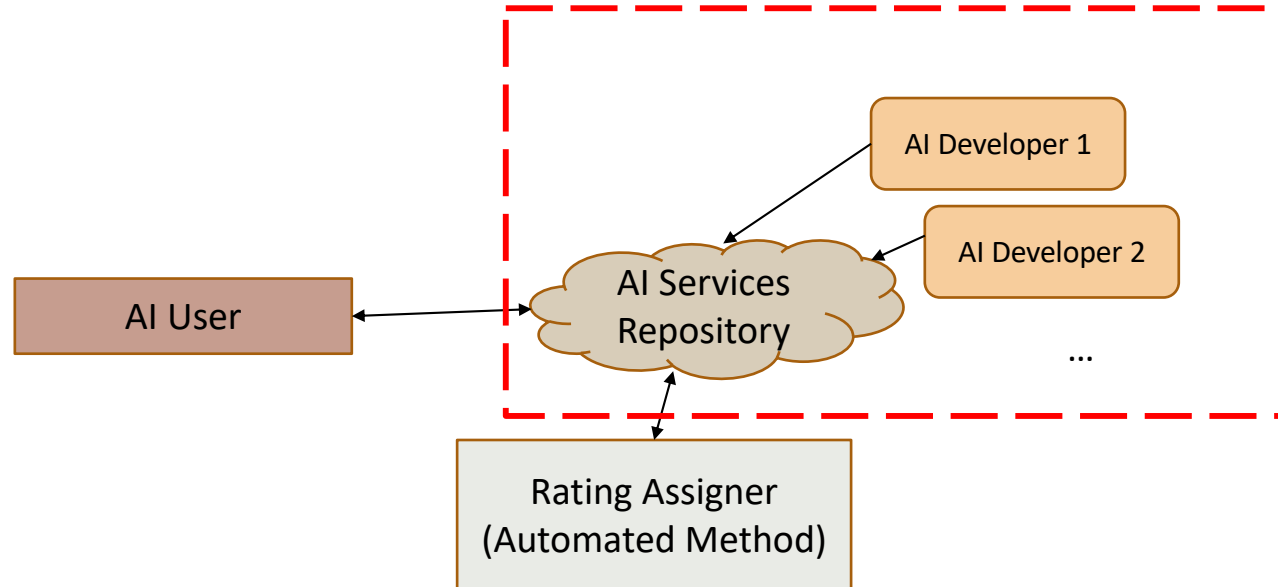


Developer Checklist

Source: Handbook on Data Protection and Privacy for Developers of Artificial Intelligence, 2021

Generating Ratings / Certificates for AI's Assessed Behavior

Idea: Develop Automated Methods to Rate AI Systems That Can be Used for Communicating Trust in **Black Box** Setting



Trust Issues – Mitigate via Ratings

- Communicate behavior via certificates / ratings (increase transparency)
 - But, let humans make decisions

Trust Dimensions
Competent
Reliable
Upholds human values
Allows human interaction

Transparency Through Documentation of Rating

Documentation about

- Outcome (e.g., Nutrition label, Electronic DataSheet, Factsheet)
- Process (e.g., SEI Capability Maturity Model, ISO 9001)

Documentation by

- Producer (e.g., Nutrition label)
- Consumer (e.g., Yelp rating)
- Independent 3rd Party (e.g., JD Powers, NHTSA car crash)

Reference: AboutML Project at PAI - <https://www.partnershiponai.org/about-ml-get-involved/#read>

Project Discussion

Course Project

- **Framework**

1. (Problem) Think of a problem whose solution may benefit people (e.g., health, water, air, traffic, safety)
2. (User) Consider how the primary user (e.g., patient, traveler) may be solving the problem today
3. (AI Method) Think of what the solution will do to help the primary user
 1. Solution => ML task (e.g. classification), recommendation, text summarization, ...
 2. Use a foundation model (e.g., LLM-based) solution as the baseline
4. (Data) Explore the data for a solution to work
5. (Reliability: Testing) Think of the evaluation metric we should employ to establish that the solution will work? (e.g., 20% reduction in patient deaths)
6. (Holding Human Values) Discuss if there are fairness/bias, privacy issues?
7. (Human-AI) Finally, elaborate how you will explain the primary user that your solution is trustable to be used by them

Project Discussion: What to Focus on ?

- Problem: you should care about it
- Data: should be available
- Method: you need to be comfortable with it. Have at least two – one serves as baseline
- Trust issue
 - Due to Users
 - Diverse demographics
 - Diverse abilities
 - Multiple human languages
 - Or other impacts
- What one does to mitigate trust issue

Rubric for Evaluation of Course Project

Project

- Project plan along framework introduced (7 points)
- Challenging nature of project
- Actual achievement
- Report
- Sharing of code

Presentation

- Motivation
- Coverage of related work
- Results and significance
- Handling of questions

Project Discussion

1. Create a private Github repository called “CSCE581-Spring2025-<studentname>-Repo”. Share with Instructor (biplav-s)
2. Create a folder called “Project”. Inside, create a text file called “ProjectPlan.md” (or “ProjectPlan.txt”) and have details by the next class (Jan 30, 2025)

1. Title:
2. Key idea: (2-3 lines)
3. Who will care when done:
4. Data need:
5. Methods:
6. Evaluation:
7. Users:
8. Trust issue:

Concluding Section

Week 7 (L13 and 14): Concluding Comments

- We looked at
 - Explanation methods
 - Generating trust certificates/ ratings

About Next Week – Lectures 15, 16

Lectures 15, 16:

- Student Projects - "Walk" stage presentations
- ML/ Classification: Trust Mitigation – Explanation methods

9	Feb 11 (Tu)	Quiz 1
10	Feb 13 (Th)	AI - Structured: Analysis – Supervised ML – Trust Issues
11	Feb 18 (Tu)	AI - Structured: Analysis – Supervised ML – Trust Issues
12	Feb 20 (Th)	AI - Structured: Analysis – Supervised ML – Mitigation Methods
13	Feb 25 (Tu)	AI - Supervised ML: Explanation Tools
14	Feb 27 (Th)	AI Trust - Mitigation method (Trust rating) – Kausik Lakkaraju
15	Mar 4 (Tu)	Student presentations - project
16	Mar 6 (Th)	Machine Learning – Trust Issues (Explainability)