Explainable Machine Learning in Deployment

Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yuhuan Jia, Joydeep Ghosh, Ruchir Puri, José Moura, and Peter Eckersley

Growth of Transparency Literature

Growth of Transparency Literature

Many algorithms proposed to “explain” machine learning model output

Growth of Transparency Literature

Many algorithms proposed to “explain” machine learning model output

We study how organizations use these algorithms, if at all

Our Approach

Our Approach

30 minute to 2 hour semi-structured interviews

Our Approach

30 minute to 2 hour semi-structured interviews

50 individuals from 30 organizations interviewed

Shared Language

Shared Language

- **Transparency**: Providing stakeholders with relevant information about how a model works.

- **Explainability**: Providing insights into a model's behavior for specific datapoint(s)

Shared Language

- **Transparency**: Providing stakeholders with relevant information about how a model works.
- **Explainability**: Providing insights into a model’s behavior for specific datapoint(s)

Example Questions

- What **type of explanations** have you used (e.g., feature-based, sample-based, counterfactual, or natural language)?
- Who is the audience for the model explanation (e.g., research scientists, product managers, domain experts, or users)?
- In what context have you deployed the explanations (e.g., informing the development process, informing human decision makers about the model, or informing the end user on how actions were taken based on the model's output)?

Example Questions

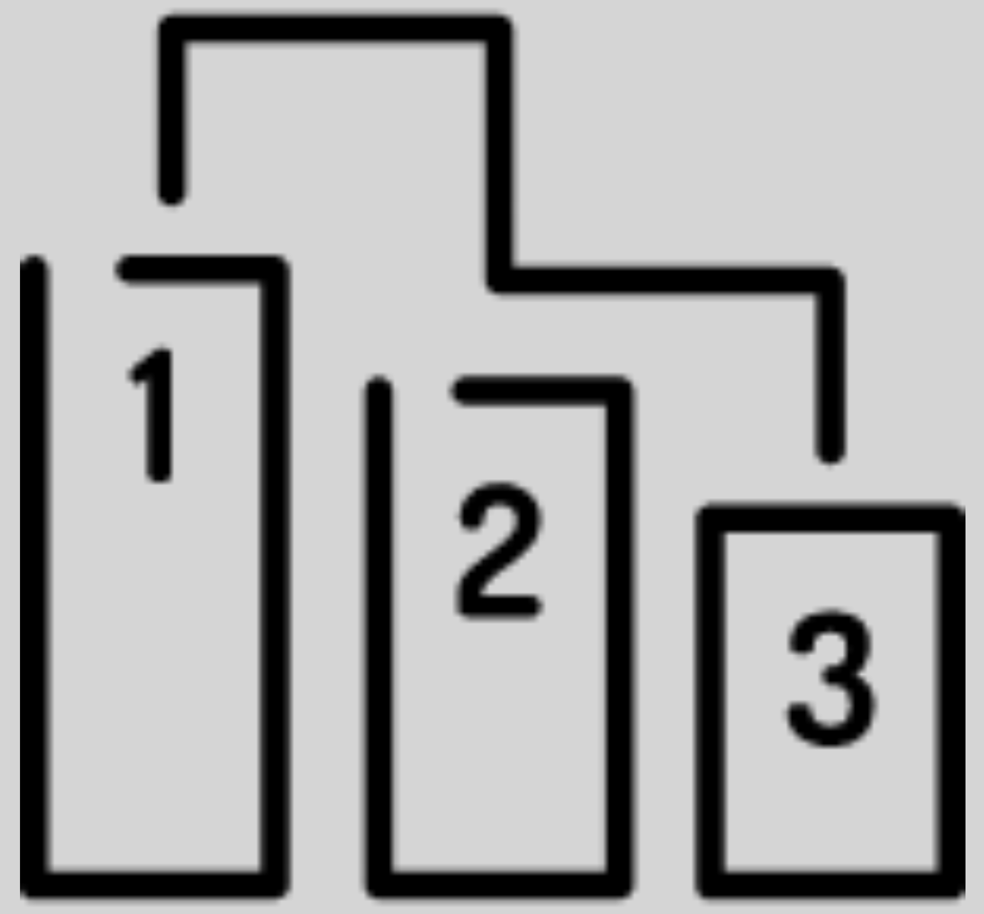
- What **type of explanations** have you used (e.g., feature-based, sample-based, counterfactual, or natural language)?
- Who is the **audience** for the model explanation (e.g., research scientists, product managers, domain experts, or users)?
- In what context have you deployed the explanations (e.g., informing the development process, informing human decision makers about the model, or informing the end user on how actions were taken based on the model's output)?

Example Questions

- What **type of explanations** have you used (e.g., feature-based, sample-based, counterfactual, or natural language)?
- Who is the **audience** for the model explanation (e.g., research scientists, product managers, domain experts, or users)?
- In what **context** have you deployed the explanations (e.g., informing the development process, informing human decision makers about the model, or informing the end user on how actions were taken based on the model's output)?

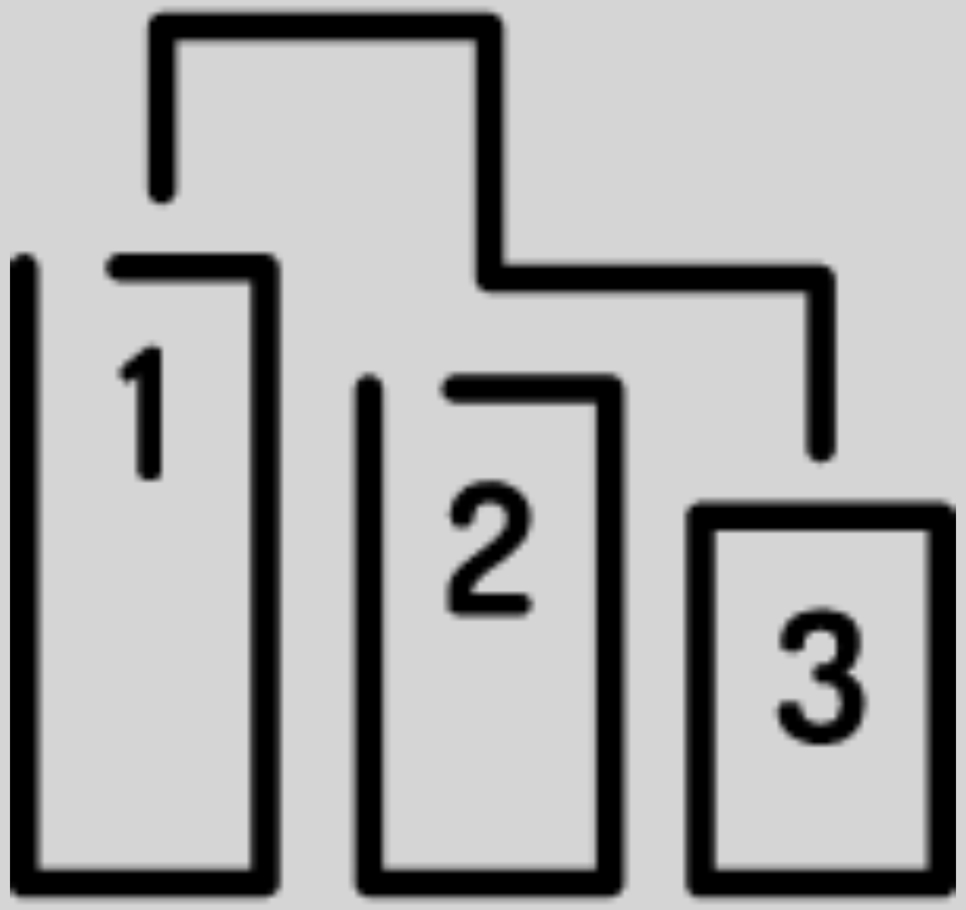
Types of Explanations

Types of Explanations



Feature Importance

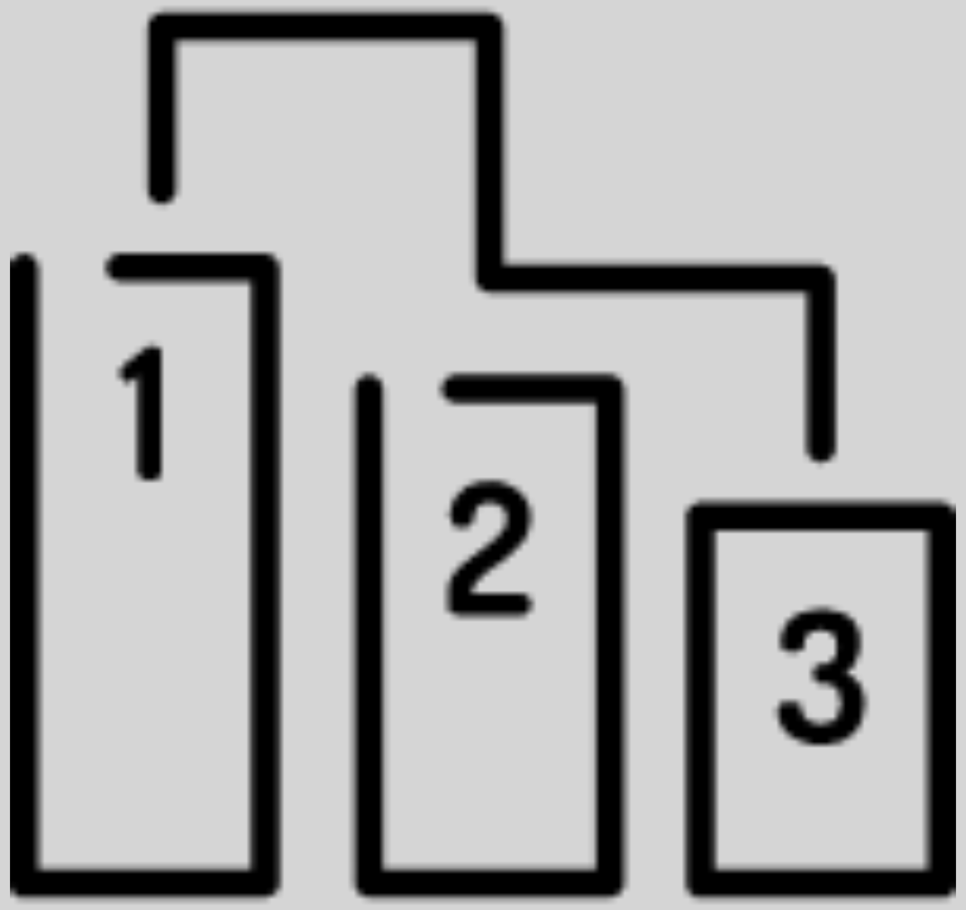
Types of Explanations



Feature Importance

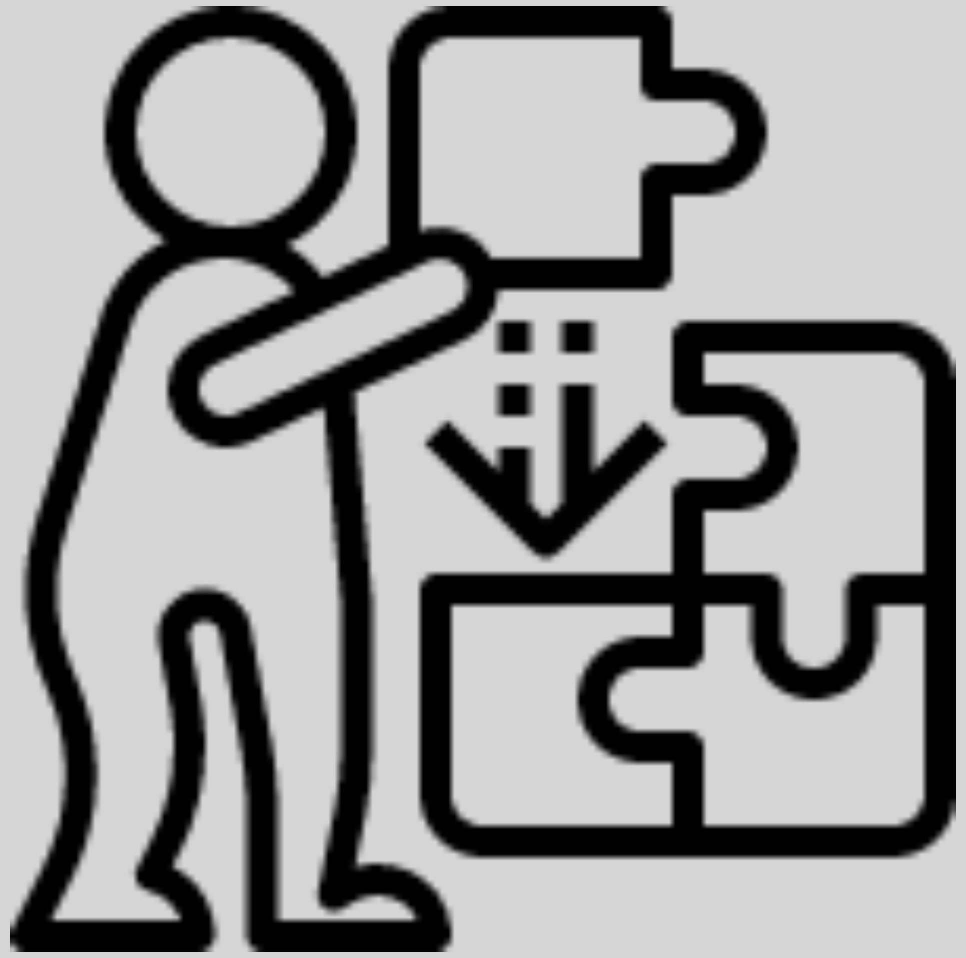
Sample Importance

Types of Explanations



Feature Importance

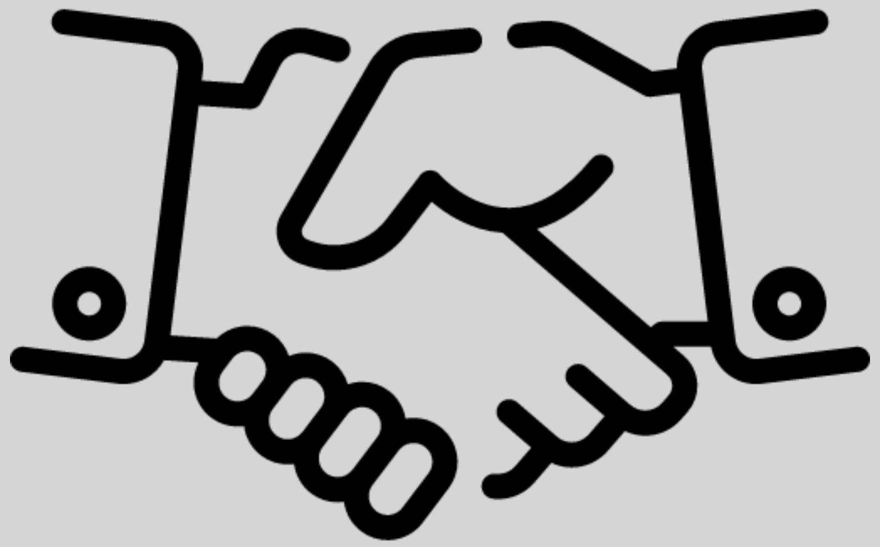
Sample Importance



Counterfactuals

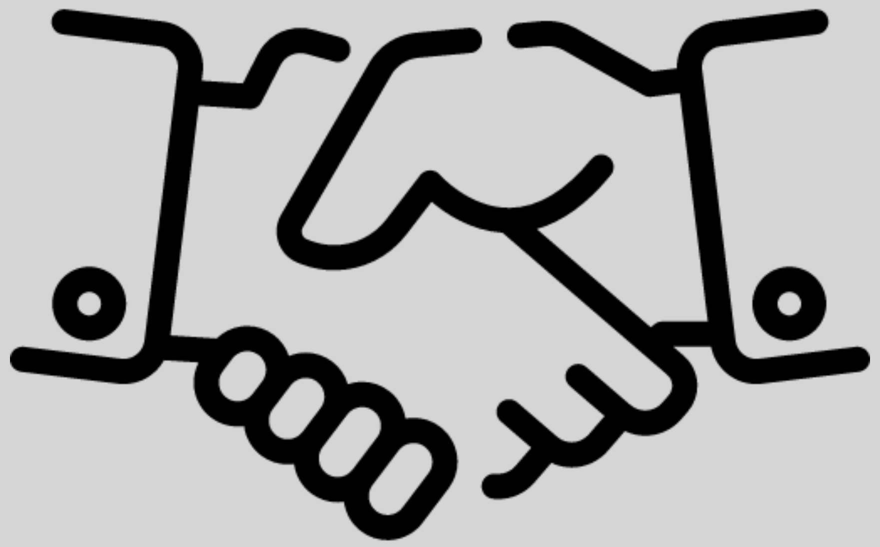
Stakeholders

Stakeholders

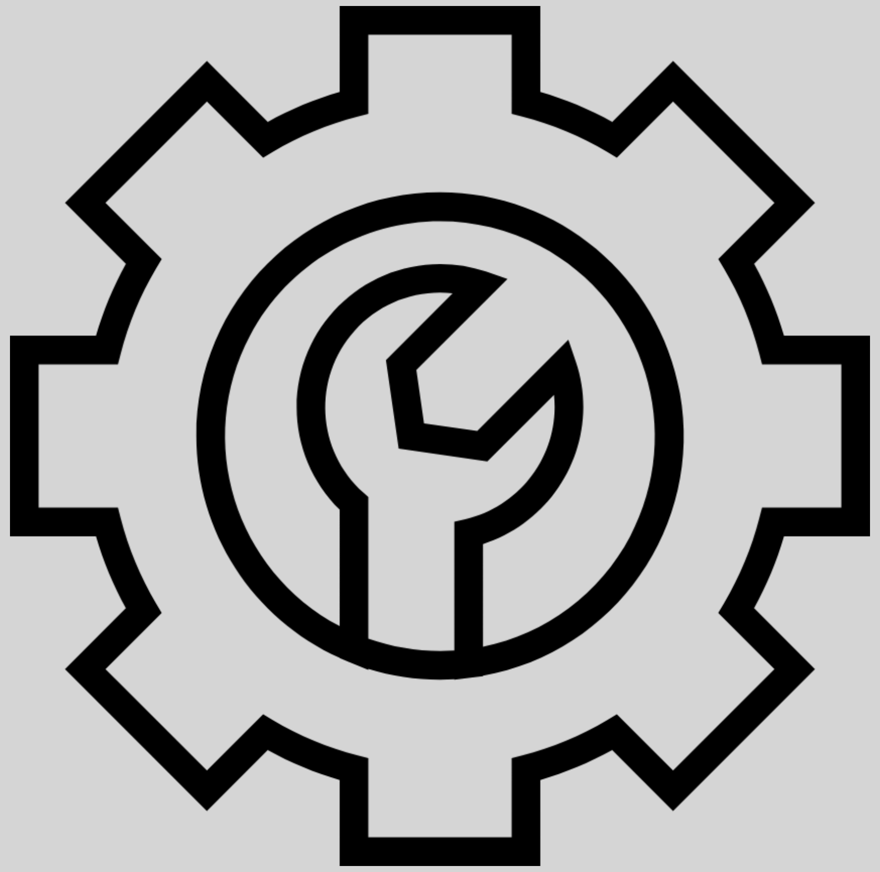


Executives

Stakeholders

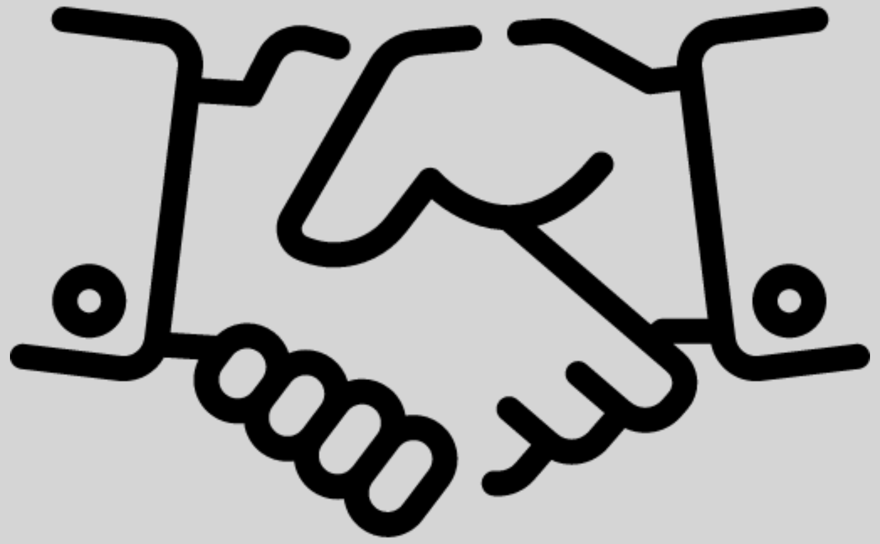


Executives

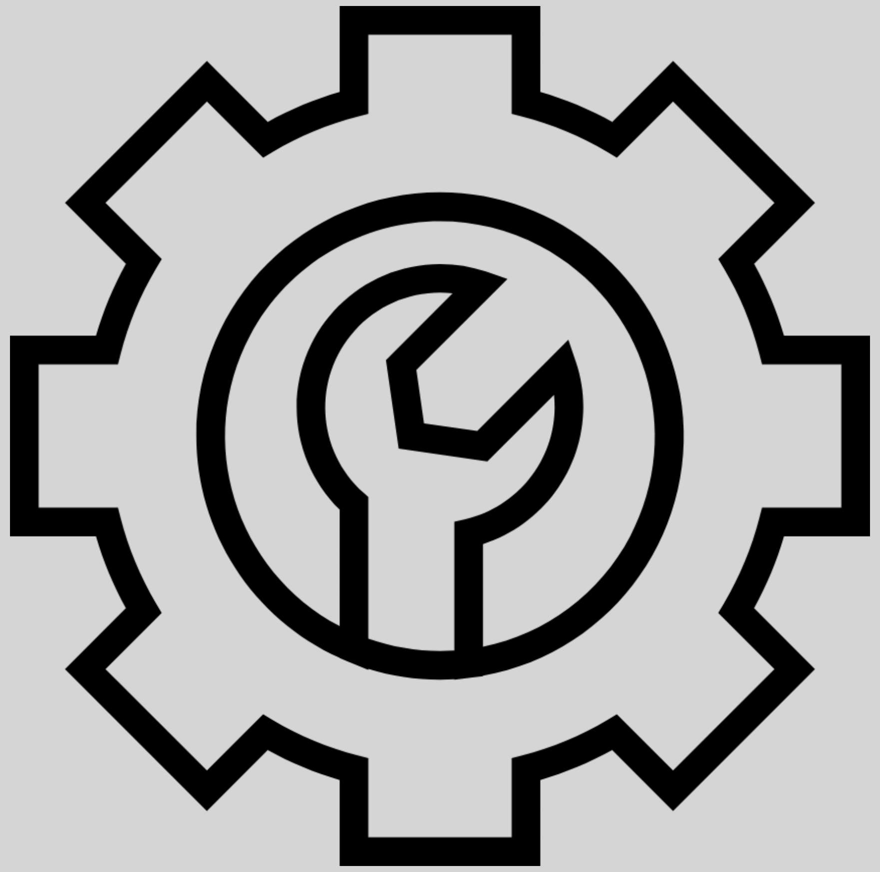


Engineers

Stakeholders



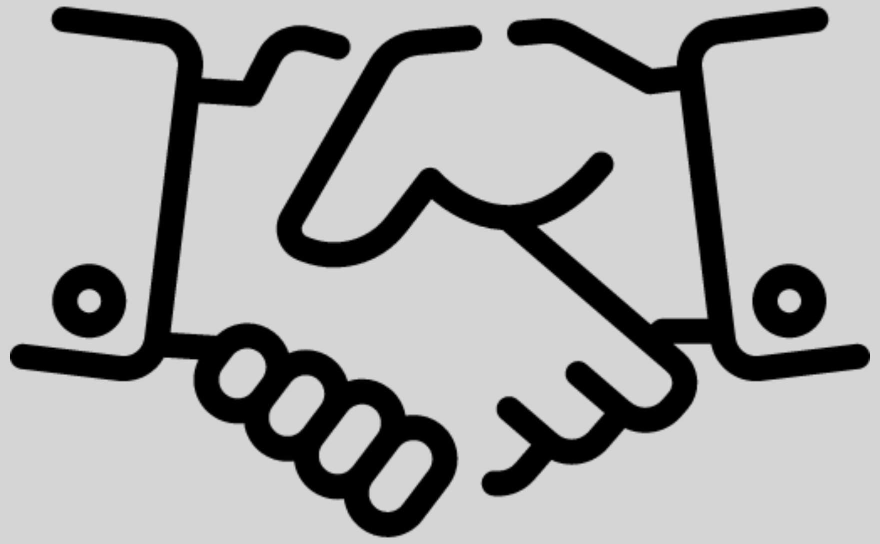
Executives



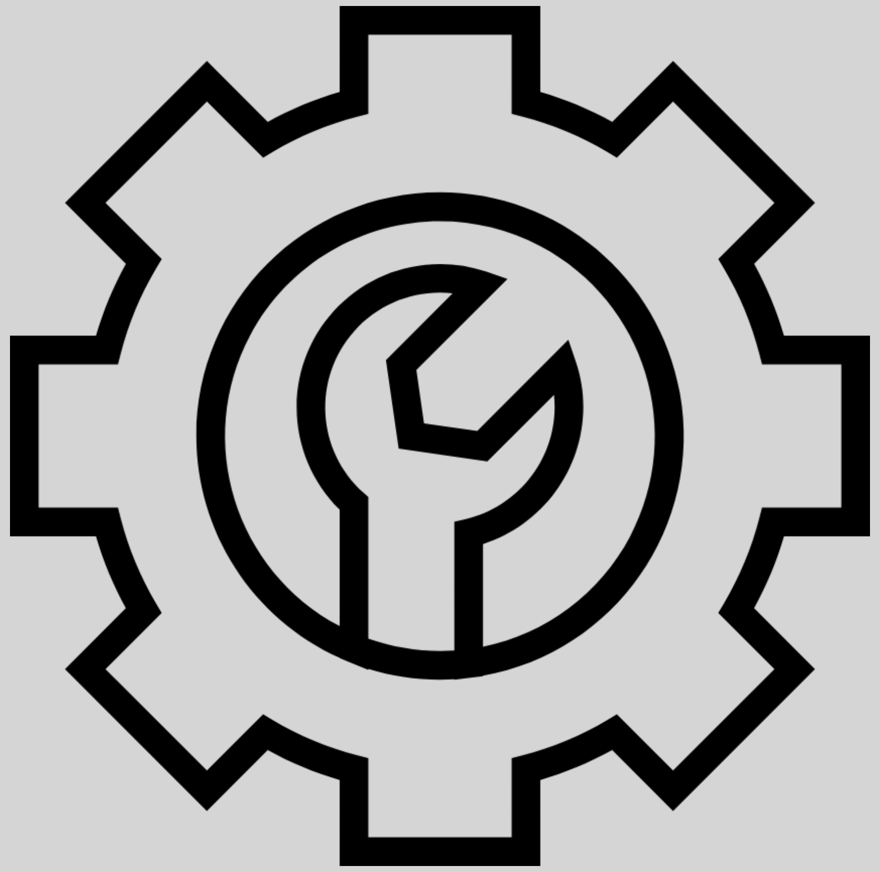
Engineers

End Users

Stakeholders



Executives



Engineers

End Users

Regulators

Findings

1. Explainability used for **debugging** internally
2. **Goals** of explainability are not clearly defined within organizations
3. Technical **limitations** make explainability hard to deploy in real-time

Findings

1. Explainability used for **debugging** internally
2. **Goals** of explainability are not clearly defined within organizations
3. Technical **limitations** make explainability hard to deploy in real-time

Use Cases

Use Cases

1. Most use cases in finance or healthcare
2. Consumer of explanation is almost **exclusively** ML engineers
3. No consensus on evaluating feature-level explanations

Use Cases

1. Most use cases in finance or healthcare
2. Consumer of explanation is almost **exclusively** ML engineers
3. No consensus on evaluating feature-level explanations

Use Cases

1. Most use cases in finance or healthcare
2. Consumer of explanation is almost **exclusively** ML engineers
3. No consensus on evaluating feature-level explanations

Use Cases

1. Most use cases in finance or healthcare
2. Consumer of explanation is almost **exclusively** ML engineers
3. No consensus on evaluating feature-level explanations — ***SHAP*** *is popular only due to convenience*

Findings

1. Explainability used for **debugging** internally
2. **Goals** of explainability are not clearly defined within organizations
3. Technical **limitations** make explainability hard to deploy in real-time

Establishing Explainability Goals

Establishing Explainability Goals

1 Identify stakeholders

Who will consume the explanation?

Establishing Explainability Goals

1 Identify stakeholders

Who will consume the explanation?

2 Engage stakeholders

What purpose with the explanation serve?

Establishing Explainability Goals

1. **Identify stakeholders**
   Who will consume the explanation?
2. **Engage stakeholders**
   What purpose with the explanation serve?
3. **Devise workflow**
   How will the explanation be used in practice?

Findings

1. Explainability used for **debugging** internally
2. **Goals** of explainability are not clearly defined within organizations
3. Technical **limitations** make explainability hard to deploy in real-time

Limitations

- **Spurious** correlations exposed by feature level explanations

Limitations

- **Spurious** correlations exposed by feature level explanations

- No **causal** underpinnings to the models themselves

Limitations

- Sample importance is **computationally infeasible** to deploy at scale

Limitations

- Sample importance is **computationally infeasible** to deploy at scale

- **Privacy** concerns of model inversion

Findings

1. Explainability used for **debugging** internally
2. **Goals** of explainability are not clearly defined within organizations
3. Technical **limitations** make explainability hard to deploy in real-time

Explainable Machine Learning in Deployment

https://arxiv.org/abs1909.06342

umang@partnertshiponai.org

@umangsbhatt