







 















Reading

* [Explainable AI](https://ojs.aaai.org/index.php/aimagazine/article/view/2850/3419)
* [Responsible AI](https://arxiv.org/pdf/1910.10045.pdf)
* [Interpretable Machine Learning - A Guide for Making Black Box Models Explainable.](https://christophm.github.io/interpretable-ml-book/)
* [TensorFlow Lattice](https://www.tensorflow.org/lattice)
* [Monotonic Calibrated Interpolated Look-Up Tables](https://jmlr.org/papers/volume17/15-243/15-243.pdf)
* [Permutation feature importance](http://arxiv.org/abs/1801.01489)
* [A Unified Approach to Interpreting Model Predictions](https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf)
* [Explainable AI for Trees: From Local Explanations to Global Understanding](https://arxiv.org/abs/1905.04610)
* [TCAV](https://arxiv.org/pdf/1711.11279.pdf)
* [LIME](https://github.com/marcotcr/lime)
* [AI Explanations](https://storage.googleapis.com/cloud-ai-whitepapers/AI%20Explainability%20Whitepaper.pdf)

Fooling DNNs:

<https://arxiv.org/pdf/1607.02533.pdf>

<https://arxiv.org/pdf/1412.6572.pdf>

XAI:

<http://www.cs.columbia.edu/~orb/papers/xai_survey_paper_2017.pdf>

Interpretable models

<https://christophm.github.io/interpretable-ml-book/>

<https://www.tensorflow.org/lattice>

Dol bear law:

<https://en.wikipedia.org/wiki/Dolbear%27s_law>

TensorFlow Lattice:

<https://www.tensorflow.org/lattice>

<https://jmlr.org/papers/volume17/15-243/15-243.pdf>

PDP:

<https://github.com/SauceCat/PDPbox>

<https://scikit-learn.org/stable/auto_examples/inspection/plot_partial_dependence.html>

Permutation Feature Importance:

<http://arxiv.org/abs/1801.01489>

Shapley values:

<https://en.wikipedia.org/wiki/Shapley_value>

SHAP:

<https://github.com/slundberg/shap>

 TCAV:

<https://arxiv.org/pdf/1711.11279.pdf>

 LIME:

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

Google Cloud XAI

<https://storage.googleapis.com/cloud-ai-whitepapers/AI%20Explainability%20Whitepaper.pdf>

Integrated gradients:

<https://arxiv.org/pdf/1703.01365.pdf>