References

Broniatowski, D. (2021). Psychological Foundations of Explainability and Interpretability in Artificial Intelligence. https://www.nist.gov/publications/psychological-foundations-explainability-and-interpretability-artificial-intelligence

Comisión Europea (2021). Artificial Intelligence Act / Propuesta de Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de inteligencia artificial (Ley de inteligencia artificial) y se modifican determinados actos legislativos de la Unión. https://artificialintelligenceact.eu/

Comisión Europea (2019). Dirección General de Redes de Comunicación, Contenido y Tecnologías, Directrices éticas para una IA fiable, Oficina de Publicaciones, 2019, https://data.europa.eu/doi/10.2759/14078

C. Rudin, C. Chen, Zhi Chen, H. Huang, L. Semenova, C. Zhong. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. http://essay.utwente.nl/91965/

Doshi-Velez, F., et al. (2017). Towards a rigorous science of interpretable machine learning. https://arxiv.org/abs/1702.08608

Devis (2011). https://cs.nyu.edu/~davise/papers/WinogradSchemas/WSCollection.html

Dimensions (2022). https://app.dimensions.ai/discover/publication

EBA (2021). Discussion paper on machine learning for IRB models. https://www.eba.europa.eu/regulation-and-policy/model-validation/discussion-paper-machine-learning-irb-models

European Parliamentary Research Service (2020). The impact of the General Data Protection Regulation (GDPR) on artificial intelligence. https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2020)641530

Floridi et al. (2022). capAI - A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligence Act. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4064091

Friedman, J. H. (2001). “Greedy function approximation: A gradient boosting machine.” Annals of statistics (2001): 1189-1232. https://www.jstor.org/stable/2699986

Gall, R. (2018). Machine Learning explainability vs interpretability: two concepts that could restore trust in AI, KDnuggets. https://www.kdnuggets.com/2018/12/machine-learning-explainability-interpretability-ai.html

GDPR (2018), Recital 71. https://eur-lex.europa.eu/eli/reg/2016/679/oj

Goldstein, A.; Kapelner, A.; Bleich, J; Pitkin, E. (2015). Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation. https://arxiv.org/abs/1309.6392

Harnad, D. (2003). Can a machine be conscious? How? https://web-archive.southampton.ac.uk/cogprints.org/5330/

IBM (2022). Explainable AI (XAI). https://www.ibm.com/watson/explainable-ai

iDanae (2022). ML Applied to Credit Risk: building explainable models. Quarterly Newsletter 3Q22. iDanae Chair. https://blogs.upm.es/catedra-idanae/wp-content/uploads/sites/698/2022/10/Idanae-3Q22.pdf

Jonathon Phillips, P.; Hahn, H.; Fontana, P; Yates, A.; Greene, K. K.; Broniatowski, D. A.; Przybocki, M. A. (2021). Four Principles of Explainable Artificial Intelligence. NIST. https://www.nist.gov/publications/four-principles-explainable-artificial-intelligence

Kaggle (2017). IBM HR Analytics Employee Attrition & Performance. https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

LeCun, Y.; Bengio, Y.; Hinton, G. (2015). Deep learning. Nature. https://pubmed.ncbi.nlm.nih.gov/26017442/

Leventi-Peetz, A.-M., et al. (2022). Deep Learning Reproducibility and Explainable AI (XAI). https://arxiv.org/abs/2202.11452

Levesque, H. (2014). On our best behaviour. Written version of the Research Excellence Lecture presented in Beijing at the IJCAI-13 conference. Artificial Intelligence, vol. 212, pages 27-35. https://doi.org/10.1016/j.artint.2014.03.007

Lundberg, S. M.; Lee, S. (2017). A Unified Approach to Interpreting Model Predictions. https://dl.acm.org/doi/10.5555/3295222.3295230

Management Solutions (2023). ModelCraft. Modelización por componentes. https://www.managementsolutions.com/es/microsites/soluciones-propietarias/modelcraft

Management Solutions (2022). Gamma. Sistema de gobierno de modelos. https://www.managementsolutions.com/es/microsites/soluciones-propietarias/gamma

Management Solutions (2021). Nota técnica sobre el EBA Discussion paper on machine learning for IRB models. https://www.managementsolutions.com/es/publicaciones-y-eventos/apuntes-normativos/notas-tecnicas-normativas/documento-de-debate-sobre-machine-learning-en-el-enfoque-irb

Management Solutions (2020). Auto machine learning, towards model automation. https://www.managementsolutions.com/en/publications-and-events/industry-reports/white-papers/auto-machine-learning-towards-model-automation

Management Solutions (2018). Machine learning, a key component in business model transformation. https://www.managementsolutions.com/en/publications-and-events/industry-reports/white-papers/machine-learning-a-key-component-in-business-model-transformation

Management Solutions (2015). Data science and the transformation of the financial industry. https://www.managementsolutions.com/en/publications-and-events/industry-reports/white-papers/data-science

Marcinkevics, R. (2020). Interpretability and Explainability: A Machine Learning Zoo Mini-tour. ETH Zürich, Department of Computer Science, Institute for Machine Learning. https://arxiv.org/abs/2012.01805

McCarthy, J. (2004). What is artificial intelligence? Stanford University. http://jmc.stanford.edu/artificial-intelligence/what-is-ai/index.html

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artif. Intell.2019,267, 1–38. https://www.sciencedirect.com/science/article/pii/S0004370218305988

OECD (2019). Principles for Artificial Intelligence. https://www.oecd.org/digital/artificial-intelligence/

Oneto, L., Chiappa, S., (2020). Fairness in Machine Learning. 2012.15816.pdf (arxiv.org)

Ribeiro, M. T.; Singh, S.; Guestrin, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. https://arxiv.org/abs/1602.04938

Ribeiro, M. T.; Singh, S.; Guestrin, C. (2018). “Anchors: high-precision model-agnostic explanations”. AAAI Conference on Artificial Intelligence (AAAI).
https://ojs.aaai.org/index.php/AAAI/article/view/11491

Roscher, R.; Bohn, B.; Duarte, M.; Garcke, J. (2020). Explainable Machine Learning for Scientific Insights and Discoveries.
https://ieeexplore.ieee.org/document/9007737

Shapley, L. (1953). A Value for n-Person Games. In: Kuhn, H. and Tucker, A., Eds., Contributions to the Theory of Games II, Princeton University Press, Princeton, 307-317.
https://doi.org/10.1515/9781400881970-018

Sudjianto, A.; Knauth, W.; Singh, R.; Yang, Z.; Zhang, A. (2011). Unwrapping The Black Box of Deep ReLU Networks: Interpretability, Diagnostics, and Simplification. Cornell University. https://arxiv.org/abs/2011.04041

Sudjianto, A.; Zhang, A. (2021). Designing Inherently Interpretable Machine Learning Models.
https://arxiv.org/abs/2111.01743

Turing, A. (1950). Computing Machinery and Intelligence. Mind 49: 433-460.

Vilone G., Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. Information Fusion, vol. 76: 89-106.
https://www.sciencedirect.com/science/article/pii/S1566253521001093

White House OSTP (2022). Blueprint for an AI Bill of Rights.
https://www.whitehouse.gov/ostp/ai-bill-of-rights/

Yang, Z.; Zhang, A.; Sudjianto, A. (2019). Enhancing Explainability of Neural Networks through Architecture Constraints. https://arxiv.org/abs/1901.03838

Zhou, N.; Zhang, Z.; Nair, V. N.; Singhal, H.; Chen, J.; Sudjianto; A. (2021). Bias, Fairness, and Accountability with AI and ML Algorithms. https://arxiv.org/abs/2105.06558

Note: This publication was produced with the help of several artificial intelligence (AI) tools. These tools were used for various tasks, such as searching for information, collecting and organizing data, and generating summaries. In any case, the final material in this publication was written by a person and not by an AI.