1. Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission." In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1721-1730).

Abstract: This article discusses the development of intelligible machine learning models for healthcare applications. The focus is on predicting the risk of pneumonia and the likelihood of hospital readmission within 30 days. The authors emphasize the need for models that are not only accurate but also interpretable in the healthcare domain, with practical implications for patient care and outcomes.

Bibliography: Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission." In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1721-1730).

2. Binder, M., Heinrich, B., Hopf, M. et al. Global reconstruction of language models with linguistic rules – Explainable AI for online consumer reviews. Electron Markets 32, 2123–2138 (2022).

Abstract: This paper presents research on the global reconstruction of language models with the use of linguistic rules, focusing on the application of Explainable AI (XAI) for analyzing online consumer reviews. The research aims to enhance the interpretability and transparency of language models in the context of online product reviews.

Bibliography: Binder, M., Heinrich, B., Hopf, M., et al. (2022). "Global reconstruction of language models with linguistic rules – Explainable AI for online consumer reviews." Electron Markets, 32, 2123–2138.

3. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in neural information processing systems (pp. 4765-4774).

Abstract: This article proposes a unified approach to interpreting model predictions. The authors aim to provide a systematic and generalizable method for explaining machine learning model predictions, contributing to the field of interpretable machine learning.

Bibliography: Lundberg, S. M., & Lee, S. I. (2017). "A unified approach to interpreting model predictions." In Advances in neural information processing systems (pp. 4765-4774).

4. Chen, J., Song, L., Le, G., & Samaras, D. (2018). Local Interpretable Model-Agnostic Explanations for Black Box Models. arXiv preprint arXiv:1802.03735.

Abstract: This paper introduces a local interpretable model-agnostic explanation (LIME) framework for explaining predictions of black-box models. LIME is designed to provide interpretable explanations for model decisions, making it applicable to various machine learning models.

Bibliography: Chen, J., Song, L., Le, G., & Samaras, D. (2018). "Local Interpretable Model-Agnostic Explanations for Black Box Models." arXiv preprint arXiv:1802.03735.

5. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215.

Abstract: This article argues against the practice of explaining black box machine learning models in high-stakes decision-making scenarios and advocates for the use of interpretable models instead. The author emphasizes the need for transparency and interpretability in critical decision processes.

Bibliography: Rudin, C. (2019). "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." Nature Machine Intelligence, 1(5), 206-215.