Sebastian Bordt

Research Statement

I seek to understand how complex artificial intelligence systems work, and for what tasks they can be reliably employed. To achieve this goal, I employ theoretical analysis and rigorous empirical testing.

During my PhD, I have worked extensively on explainable machine learning. Explainable machine learning holds the potential to allow developers and end users to understand how complex AI systems work, increasing the usefulness and reliability of these systems in applications. It might also be able to tell us when a certain system should not be used.

One line of my work studies the properties of explanation algorithms. I have derived connections between post-hoc explanation algorithms and interpretable models [3], two approaches that researchers previously believed were largely opposed [12]. I have also studied the suitability of post-hoc explanation algorithms for regulation [4], finding that these algorithms frequently fail to fulfill the transparency objectives that are inherent in legal texts. In a slightly different line of work, we have studied the perceptually aligned input gradients of image classifiers [6], and their connections with robust models [15].

In general, I'm interested in a rigorous understanding of complex systems. During my PhD, I have developed a simple model to understand the impact of private information and opacity on human-machine interaction [1]. More recently, I have started to assess the capabilities of foundation models. This includes studies of the performance of GPT-4 on academic exams and complex interpretative tasks with graphs [2, 9]. I have also used tabular data in order to distinguish memorization, learning, and out-of-distribution generalization in language models [5].

The landscape of the best artificial intelligence systems is constantly changing [11]. I believe that this requires similar changes to the scientific methodology with which we analyze AI, and methodological flexibility on the side of the researcher. In the past, machine learning algorithms have often been motivated by theoretical principles, facilitating a theoretical understanding of the resulting models [13, 14]. Today, such characterizations seem largely out of reach [16]. To arrive at a holistic understanding of the capabilities of today's systems, we must instead perform rigorous empirical evaluations [7, 10, 2, 5]. Theory is likely to remain limited to understanding individual mechanisms in complex systems.

My main vision for future research lies in the rigorous assessment of the capabilities of foundation models. I believe that this is the necessary first step to understanding when and how these models can be reliably employed. While there are many well-known challenges to evaluating foundation models, such as their potentially all-encompassing training corpora, none of these challenges seem insurmountable.

A particularly important direction for future work is to move beyond the paradigm of evaluating on fixed held-out test data [8], and instead define a set of capabilities for which foundation models should be tested. These capabilities have to be specific enough so that performance can be rigorously tested, and general enough to be predictive for performance across a wide range of real-world tasks. In defining these evaluation criteria, the biggest challenge might be to develop a parading that is also able to assess future much more powerful models.

References

[1] S. Bordt and U. Von Luxburg. A bandit model for human-machine decision making with private information and opacity. In *International Conference on Artificial Intelligence and Statistics*, 2022.

[2] S. Bordt and U. von Luxburg. Chatgpt participates in a computer science exam. *arXiv preprint arXiv:2303.09461*, 2023.

[3] S. Bordt and U. von Luxburg. From shapley values to generalized additive models and back. In *International Conference on Artificial Intelligence and Statistics*, 2023.

[4] S. Bordt, M. Finck, E. Raidl, and U. von Luxburg. Post-hoc explanations fail to achieve their purpose in adversarial contexts. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022.

[5] S. Bordt, H. Nori, and R. Caruana. Elephants never forget: Testing language models for memorization of tabular data. In *Workshops at advances in neural information processing systems*, 2023.

[6] S. Bordt, U. Upadhyay, Z. Akata, and U. von Luxburg. The manifold hypothesis for gradient-based explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[7] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

[8] D. Donoho. 50 Years of Data Science. *Journal of Computational and Graphical Statistics*, 2017.

[9] B. J. Lengerich, S. Bordt, H. Nori, M. E. Nunnally, Y. Aphinyanaphongs, M. Kellis, and R. Caruana. Llms understand glass-box models, discover surprises, and suggest repairs. *arXiv preprint arXiv:2308.01157*, 2023.

[10] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023.

[11] OpenAI. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*, 2023.

[12] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 2019.

[13] R. E. Schapire and Y. Freund. Boosting: Foundations and algorithms. *Kybernetes*, 2013.

[14] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[15] S. Srinivas, S. Bordt, and H. Lakkaraju. Which models have perceptually-aligned gradients? an explanation via off-manifold robustness. In *Advances in neural information processing systems*, 2023.

[16] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 2021.