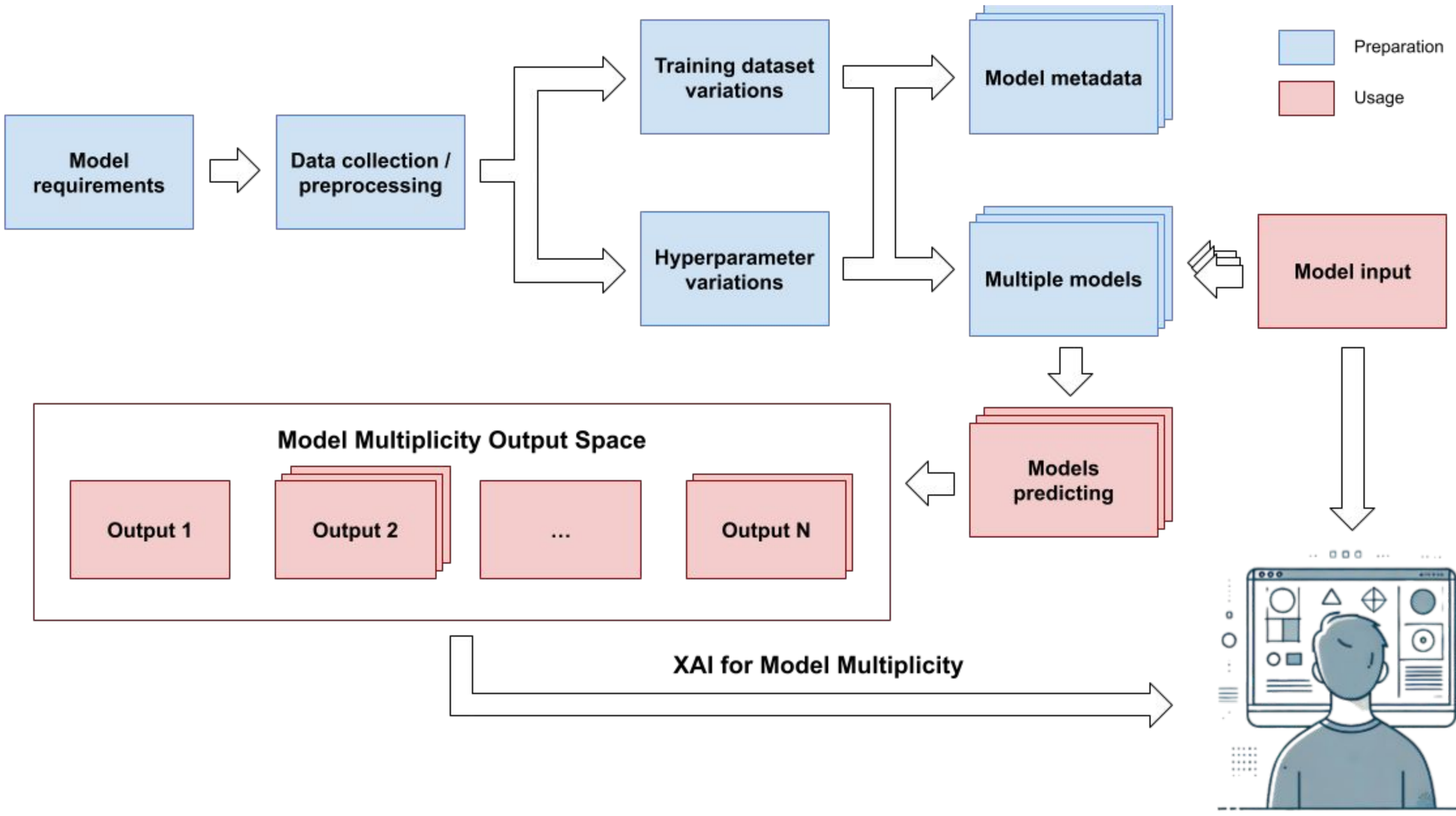
Increase Trust in Interactive AI Systems through Model Multiplicity

Gilles Eerlings, Sebe Vanbrabant, Gustavo Rovelo Ruiz, Davy Vanacken, Kris Luyten



Framework for Model Multiplicity

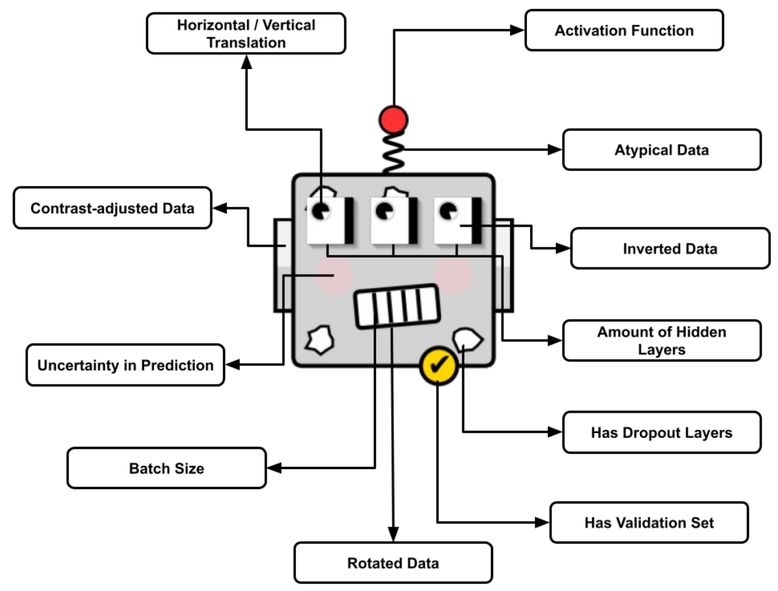
We advocate for an approach based on **model multiplicity** [1, 2]**,** where multiple ML models operate simultaneously within a system. This many-expert setup offers a more balanced view, improving both the reliability and explainability of predictions by incorporating diverse "voices”.

Creating various similar models within a vast array of possibilities. These models differ based on their **training data** and **hyperparameters**.

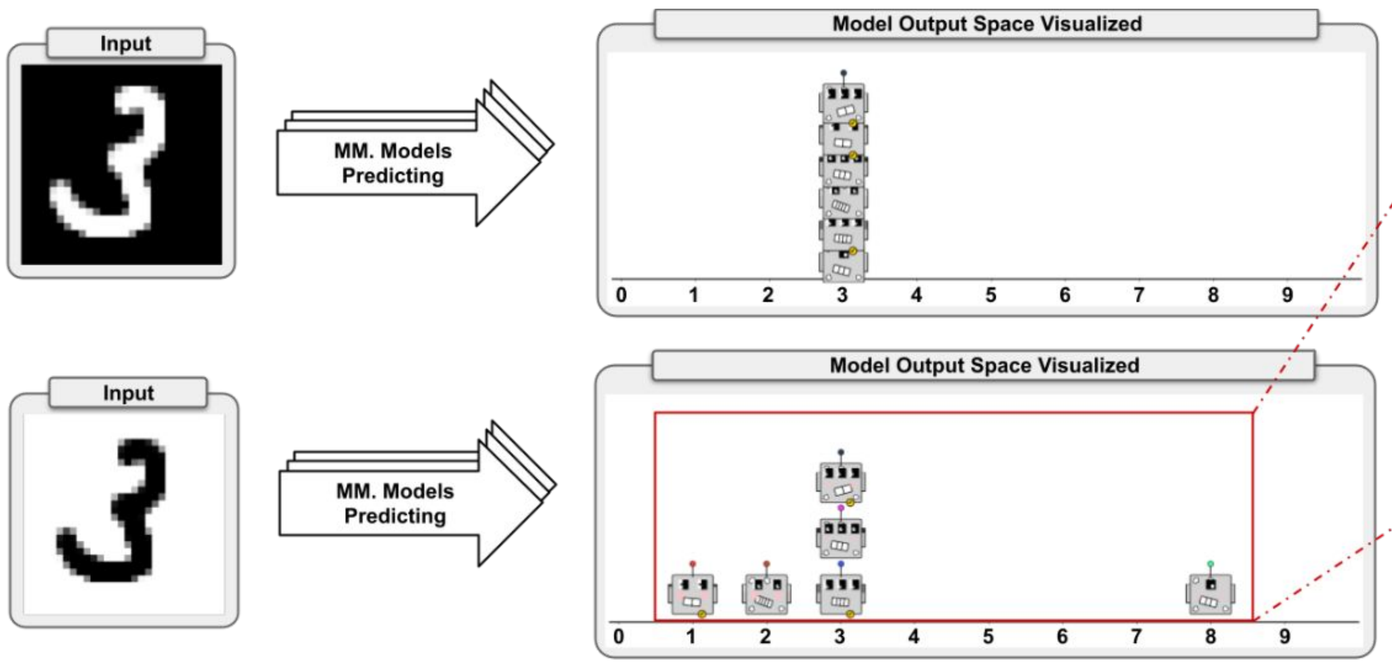
Multiple outcomes of models need to be aggregated and explained to the user of the system through **new explainable AI** components to build **appropriate trust**.

Visualising Model Differences: Chernoff Bots

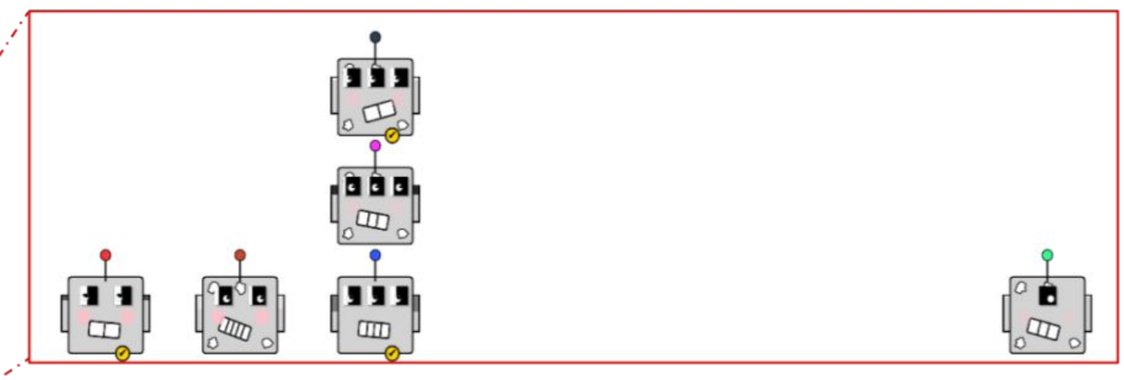
Chernoff faces [3] provide an intuitive way to visualize complex, high-dimensional data by mapping multiple AI model hyperparameters (e.g., learning rate, batch size) to facial features. This enables quick identification of patterns and relationships across models, leveraging the human brain's strong recognition of facial expressions.



Use Case: Number Recognizers Trained with the MNIST Dataset



We are validating our approach by presenting the “backgrounds” of the various models and their corresponding predictions using a barchart of Chernoff Bots.



[1] Emily Black et al. 2022. Model Multiplicity: Opportunities, Concerns, and Solutions. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22). ACM, New York 850–863. https://doi.org/10.1145/3531146.3533149
[2] Kris Luyten et al. 2024. Opportunities and Challenges of Model Multiplicity in Interactive Software Systems. In The 2nd Workshop on Engineering Interactive Systems Embedding AI Technologies. Conference Material, June 25. http://hdl.handle.net/1942/43354.
[3] Herman Chernoff et al. 1973. The Use of Faces to Represent Points in k-Dimensional Space Graphically. Journal of the American Statistical Association, https://doi.org/10.1080/01621459.1973.10482434
[4] Yann LeCun et al. 1998. Gradient-based learning applied to document recognition. In Proceedings of the IEEE, 86(11), 2278-2324. https://doi.org/10.1109/5.726791