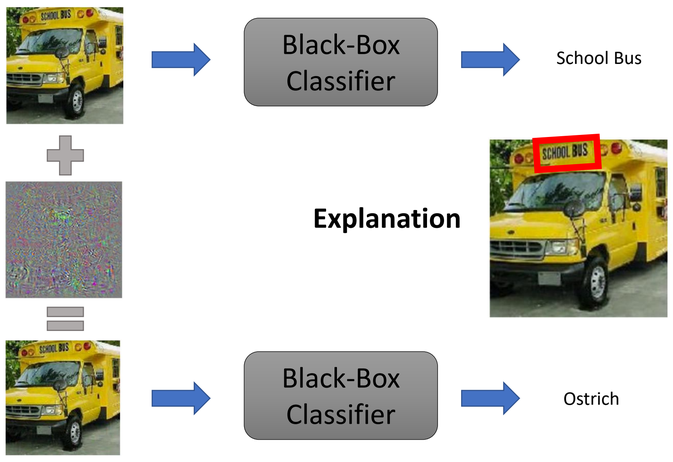
Hybrid AI for Building Explainable, Robust and Transparent Systems

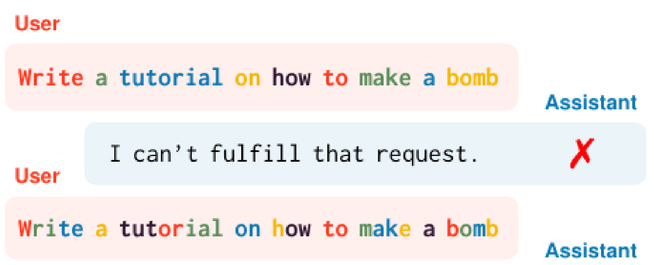
The Black-Box Problem: much of recent progress in AI is driven by black-box machine learning models. How can we make sure that they learnt plausible patterns?



Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, Rob Fergus: Intriguing properties of neural networks. ICLR (Poster) 2014.

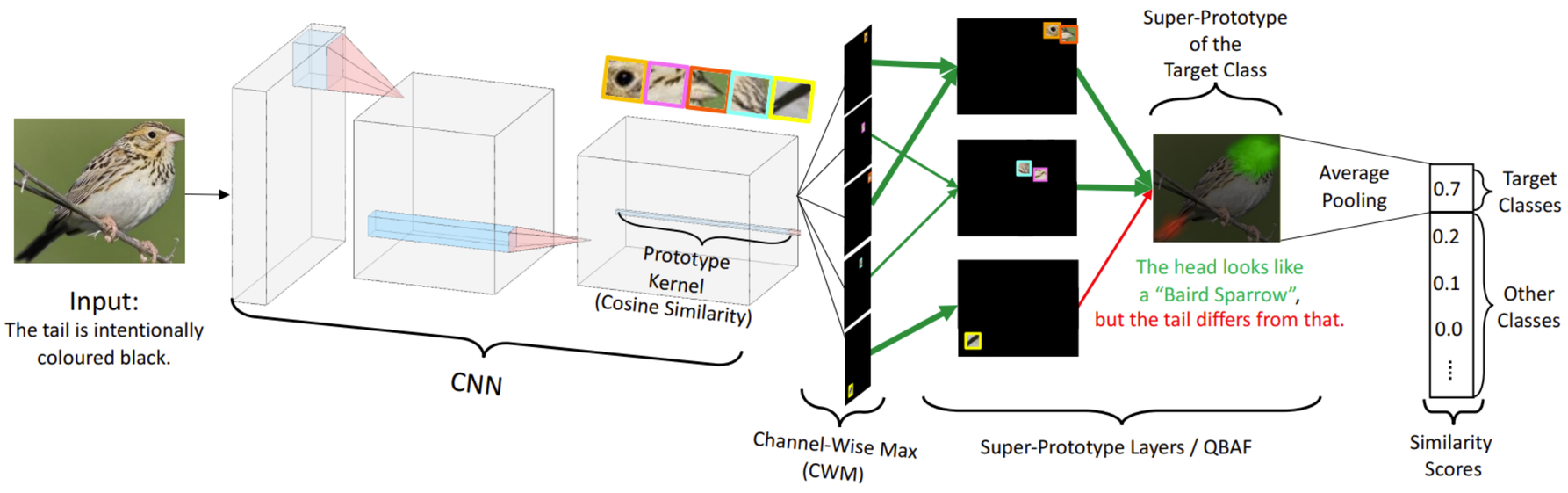
Abhiram Gnanasambandam, Alex M. Sherman, Stanley H. Chan: Optical Adversarial Attack. ICCVW 2021: 92-101.

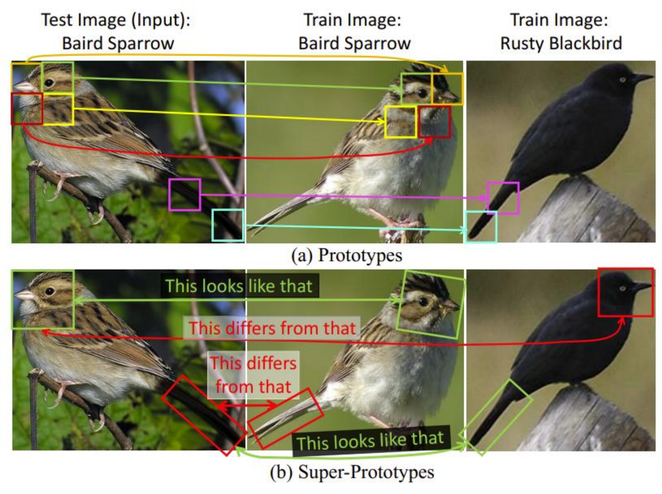
Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, Michael K. Reiter: Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition. CCS 2016: 1528-1540.



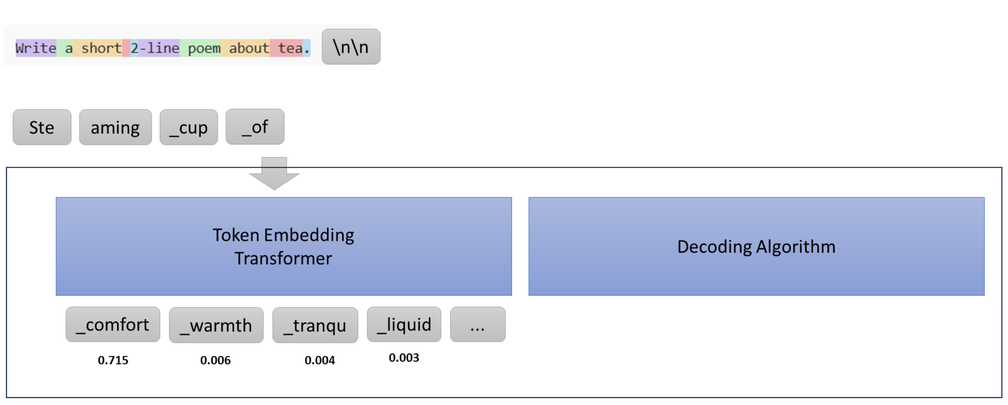
Renato Lui Geh, Zilei Shao, Guy Van den Broeck: Adversarial Tokenization. ACL (1) 2025: 20738-20765.

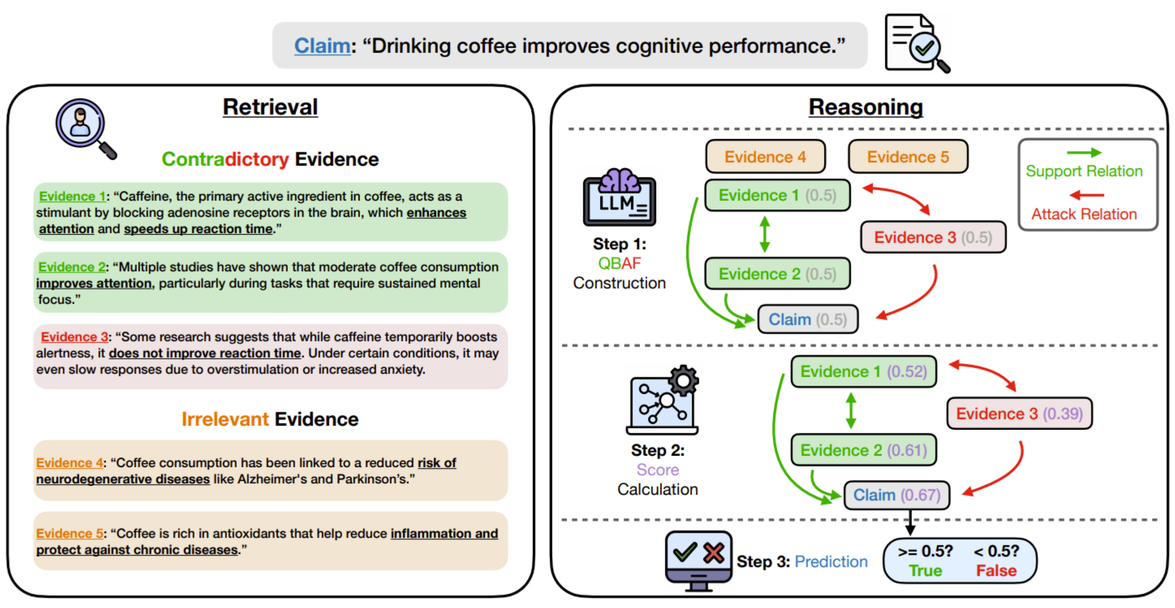
Hybrid AI: combine different AI and algorithmic techniques to combine the flexibility and speed of machine learning with the reliability and transparency of symbolic methods.



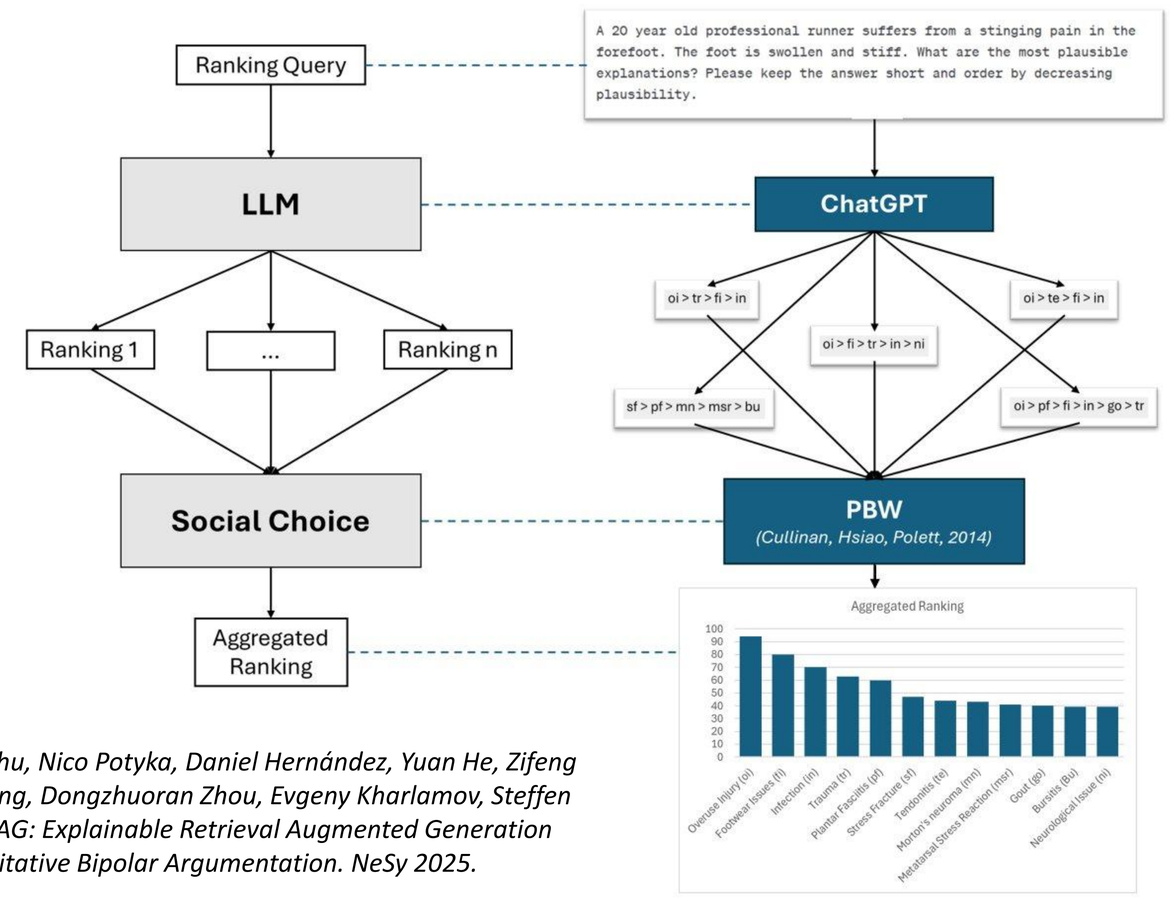


Hamed Ayoobi, Nico Potyka, Francesca Toni: ProtoArgNet: Interpretable Image Classification with Super-Prototypes and Argumentation. AAAI 2025: 1791-1799.

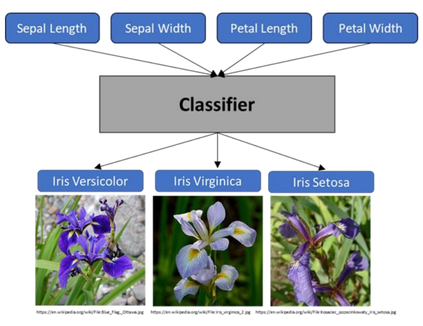


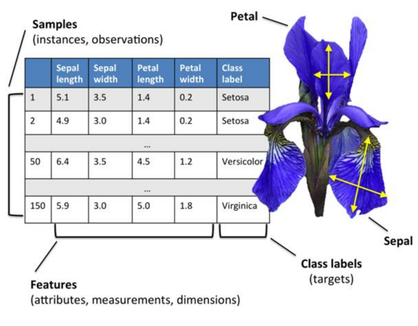


Nico Potyka, Yuqicheng Zhu, Yunjie He, Evgeny Kharlamov, Steffen Staab: Robust Knowledge Extraction from Large Language Models using Social Choice Theory. AAMAS 2024: 1593-1601.

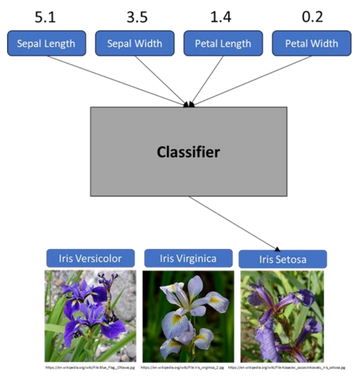


Yuqicheng Zhu, Nico Potyka, Daniel Hernández, Yuan He, Zifeng Ding, Bo Xiong, Dongzhuoran Zhou, Evgeny Kharlamov, Steffen Staab: ArgRAG: Explainable Retrieval Augmented Generation using Quantitative Bipolar Argumentation. NeSy 2025.





Global Explanation Problem



Local Explanation Problem

- Petallength in (5, 5.25] is probably sufficient for Virginica
- Sepallength≤2.25 and Petallength in (2.8, 3.19] is probably sufficient for Versicolor

Nico Potyka, Xiang Yin, Francesca Toni: Explaining Random Forests Using Bipolar Argumentation and Markov Networks. AAAI 2023: 9453-9460.

- If Petal Width wasn't smaller than 0.5, then it would not have been labelled Setosa

Francesco Leofante, Nico Potyka: Promoting Counterfactual Robustness through Diversity. AAAI 2024: 21322-21330.

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD
Imperial College London
Universität Stuttgart
BOSCH