The Long Road to Artificial General Intelligence (AGI)

Friedrich Answin Daniel Motz
Faculty of Mathematics and Computer Sciences
Friedrich Schiller University Jena

I. Challenges towards AGI

This poster discusses current unavoidable challenges towards AGI:

- **Artificial Neural Networks (ANNs) are mostly of a »black-box« nature** and difficult to interpret.
- **ANNs have not yet shown great transferrability** of domain specific knowledge.
- ANNs tend to **overestimate themselves**.
- **Introspection is currently limited**.

II. Symbolism vs Connectionism

Symbolic AI... uses high-level language to formulate problems. Also, each step is human-interpretable. It uses methods like logic programming, semantic webs and search.

Sub-Symbolic AI (SubSymAI)... in contrast consists of lower-level associations, e.g. statistical correlations, meaning, they cannot be interpreted by means of a high-level language. Often ANNs are used synonymously to sub-symbolic AI.

III. Is Deep Learning the Future?

Deep Learning is hitting a wall. Recent advances in Large Language Models (LLMs) come from new methods and increased parameters (see Table 1). However, growing computational size and cost are becoming main limiting factors. [1] Claims by Microsoft Research, that GPT-4 shows "sparks of general intelligence" [2] contrasts with critics that LLMs outputs lack intrinsic meaning [3]. Recent leaks indicate GPT-4's use of domain-specific expert models [4], hence a form of Neuro-Symbolism, at its core.

LLM	Parameters
GPT-2	$1.5 \cdot 10^{9}$
GPT-3	$1.75 \cdot 10^{11}$
GPT-4	$\geq 1 \cdot 10^{12}$

Table 1. OpenAI's large language models parameter sizes compared. [5, 6] GPT-4's parameter count is currently based off of rumours. [4]

ANN's Trustworthyness. Unfortunately, in edge cases, Convolutional Neural Networks (CNNs) have shown to behave unpredictably. A single altered pixel can change the classification of an image drastically [7, 8]. The **need for accountability**, fairness and ethics are further emphasized, as AI systems **increasingly impact human lives** (e.g. autonomous driving and medical applications). Explainable AI (XAI) methods try to explain a model's reasoning.

Local Interpretable Model-Agnostic Explanations. LIME can explain the predictions of *any* model (model-agnostic), by learning a linear classifier on systematically varied data around the requested datapoint. Therefore, sadly, no explanation of a model's general behaviour is possible. [9]

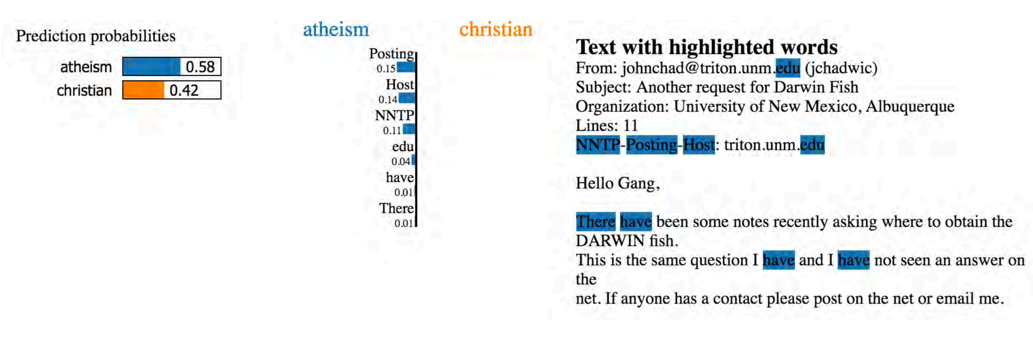

Figure 1. An example showing, that wrong features are learned, whilst achieving good scores. Text Classification on News-Articles [10], where atheism vs. christianity was predicted [11]. An email was regarded atheistic based on "Posting", "Host" and "NNTP". One would expect, that "DARWIN fish" might be an indication.

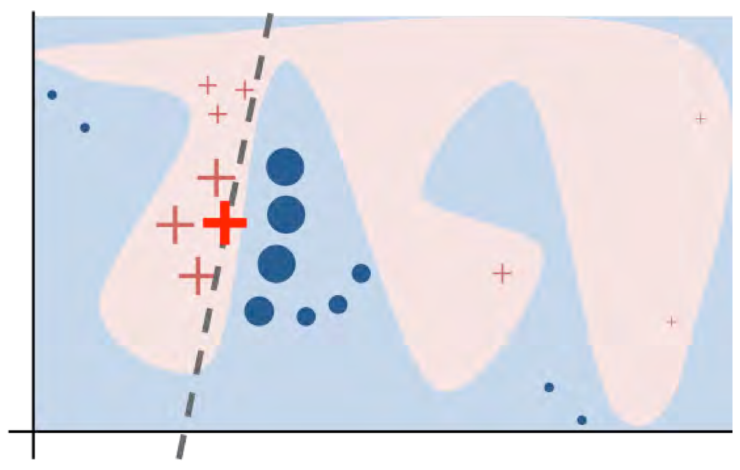

Figure 2. The big red cross is the classification to be explained (TBE-point). The pink and blue background is the complex model's classification (which might not be continuous and is unknown to us). The blue dots and red crosses represent mutated data-points around the TBE-point, which were put through the complex model. The explanation takes these mutated data and learns a linear classifier (dashed line), which can be explained. [11]

Layer-Wise Relevance Propagation. This method [12] gives pixel-wise explanations for image classification networks.

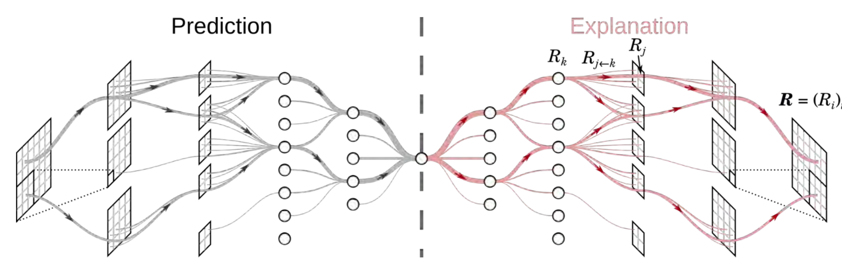

Figure 3. After a prediction has been made, the Relevance Algorithm iterates from output to input layer. This results a new image with relevance scores.

$x_i^{(l)}$: value of neuron i on layer l

$w_{ij}^{(l,l+1)}$: weight function

$$R_i^{(l)} = \sum_j \frac{z_{ij}}{\sum_{i'} z_{i'j}} R_j^{(l+1)} \tag{1}$$

$$z_{ij} = x_i^{(l)} w_{ij}^{(l,l+1)}$$

Limitations of XAI. The methods presented here are applied on existing models after training. "Meaningful" explainability would have to be built into a model's architecture. Section V presents a promising approach.

IV. Neuro-Symbolic AI

Game AIs usually use sub-symbolic methods for stochastic problems, e.g. a heuristic function for estimating the quality of a move for Monte-Carlo-Tree-Search or Alpha-Beta-Pruning. A prominent example for "[Symbolic[Neuro]]" is AlphaGo [13], see Figure 4.

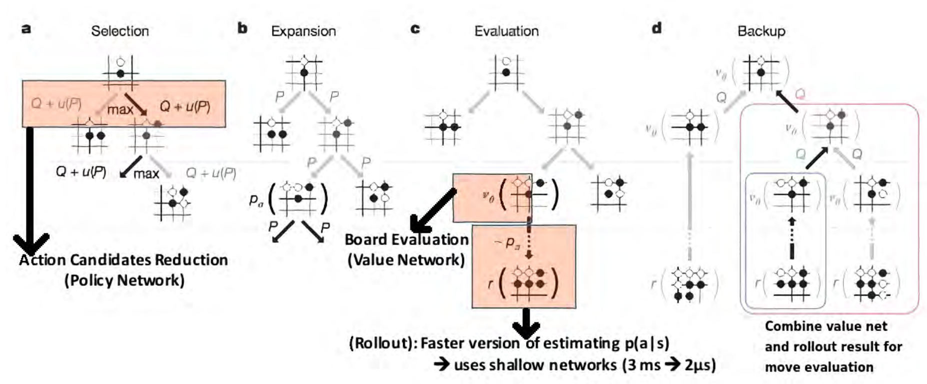

Figure 4. Looking ahead with Monte Carlo tree search [14]

NeSy Concept Learner. The *Neuro-Symbolic Concept Learner* (NS-CL) can learn visual concepts, meaning of words, and semantic parsing just from images and Question-Answer (QA) pairs [15]. The NS-CL scores are state of the art with ~99% on the CLEVR dataset.

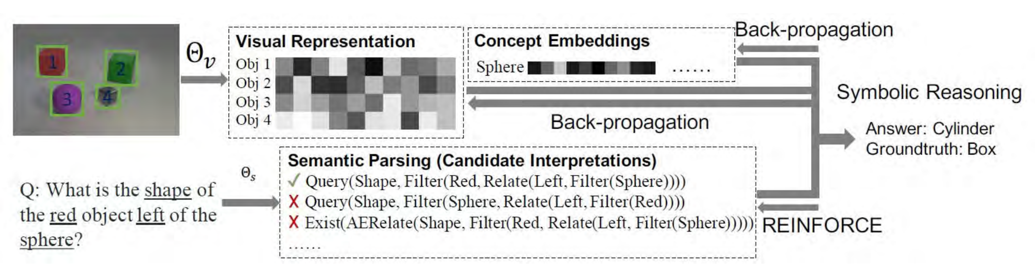

Figure 5. CLEVR dataset QA-pair examples with increasing difficulty. NS-CL starts with simple examples and increases difficulty.

NS-CL uses an attention-based language parser [16] to create a hierarchical program of predicates. The predicates are then processed by the corresponding module, e.g. "Filter" will find a shape in the image.

V. Problem Decomposition

Compositional Attention-Based Networks. Using a technique similar to Long short-term memory (LSTM), the Memory Attention and Composition (MAC) performs equally to the NS-CL. The technique decomposes a query, then each MAC cell attends to a part of the question. [17] This allows for **pixel- and word-wise explanations**.

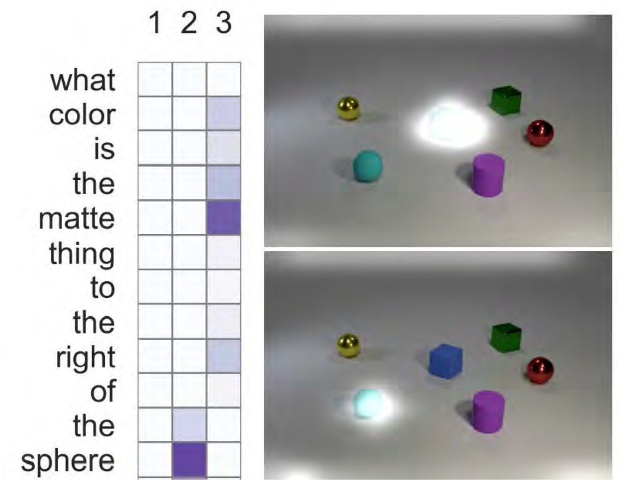

Figure 6. Visualisation of the "read unit". It makes explainable connections between the words and the image. This is used as a basis for deduction. [17]

Essence Neural Networks (ENNs). As proposed by [18], they show how explainable reasoning can be made possible without explicit use of SymAI. Similar to NS-CL ENNs learn concepts. In ENNs distinctions are made hierarchically (see Figure 7, Figure 8):

1. Differentia Neurons identify diversions between input features.
2. Subconcept neuron layers distinct
3. Concept Neurons "group" ideas

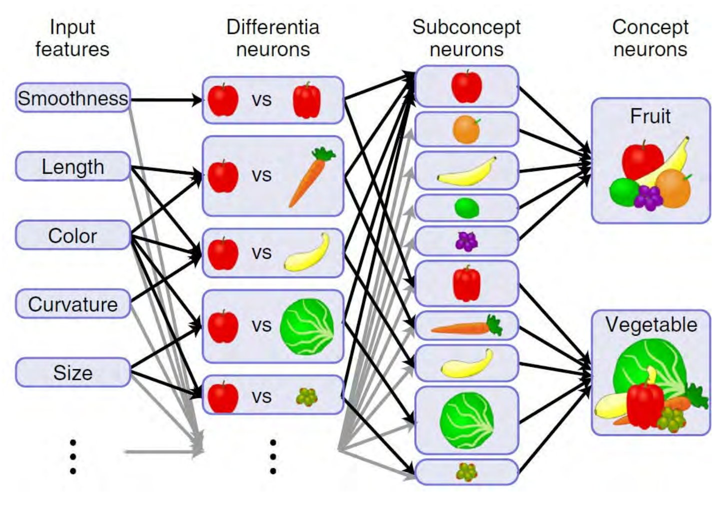

Figure 7. Example Architecture for distinguishing fruits from vegetables. Differentia Neurons establish differences between concepts. Only "apple"-neurons are depicted. [18]

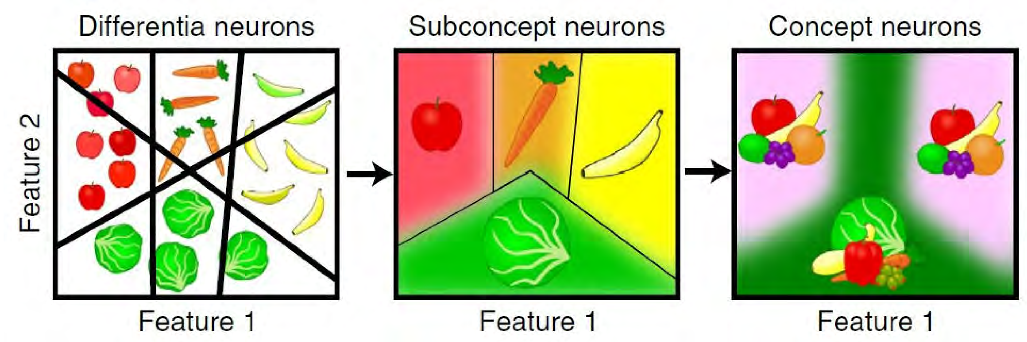

Figure 8. The structure of conceptual space is learned directly by ENNs. Differentia neurons form hyperplane decision boundaries (lines) in conceptual space. They feed forward to subconcept neurons, each forming a subregion (colored areas) defined by differentia neuron boundaries. These feed into concept neurons, each forming a possibly disconnected conceptual region from its subconcepts. [18]

VI. Can AGI emerge?

Despite all efforts, humans are still ahead. The question is: will the whole be bigger than the sum of it's parts? There are examples for emergence of *interesting* behaviours; the public tends to call them *intelligent*. A prominent hurdle is AI intentionality: attention mechanisms advanced directedness, but hardly can one speak of a self-conscious system. In terms of Searle's argument: may any current method only build improved libraries? Emergence happened once with humans; why should it not happen twice?

Bibliography

[1] N. C. Thompson, K. Greenewald, K. Lee, and G. F. Manso, "The computational limits of deep learning," 2022.

[2] S. Bubeck, V. Chandrasekaran, et al., "Sparks of artificial general intelligence: early experiments with gpt-4," 2023.

[3] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: can language models be too big? 🦜," in *Proc. 2021 ACM Conf. Fairness, Accountability, Transparency* in Facct '21, Virtual Event, Canada, 2021, p. 610, doi: 10.1145/3442188.3445922. [Online]. Available: https://doi.org/10.1145/3442188.3445922

[4] R. Albergotti, "The secret history of elon musk, sam altman, and openai," 2023. (https://www.semafor.com/article/03/24/2023/the-secret-history-of-elon-musk-sam-altman-and-openai)

[5] I. Solaiman, M. Brundage, et al., "Release strategies and the social impacts of language models," 2019.

[6] T. B. Brown, B. Mann, et al., "Language models are few-shot learners," 2020.

[7] "Ai image recognition fooled by single pixel," 2017. (https://www.bbc.com/news/technology-41845878)

[8] "A turtle - or a rifle? hackers easily fool ais into seeing the wrong thing," 2018. (https://www.science.org/content/article/turtle-or-rifle-hackers-easily-fool-ais-seeing-wrong-thing)

[9] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": explaining the predictions of any classifier," 2016.

[10] J. Rennie, "20 newsgroups data set," 2008. (http://qwone.com/~jason/20Newsgroups/)

[11] R. et al., "Lime on github," 2021. (https://github.com/marcotcr/lime)

[12] A. A. M. G. A. K. F. A. M. K.-R. A. S. W. Bach Sebastian AND Binder, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *Plos One*, vol. 10, pp. 1–46, 2015, doi: 10.1371/journal.pone.0130140. [Online]. Available: https://doi.org/10.1371/journal.pone.0130140

[13] D. S. et al., "Mastering the game of go with deep neural networks and tree search," 2016.

[14] A. Kurenkov, "A brief history of of game ai," 2016. (https://www.andreykurenkov.com/writing/ai/a-brief-history-of-game-ai-part-3/)

[15] J. Mao, C. Gan, P. Kohli, J. B. Tenenbaum, and J. Wu, "The neuro-symbolic concept learner: interpreting scenes, words, and sentences from natural supervision," 2019.

[16] L. Dong, and M. Lapata, "Language to logical form with neural attention," 2016.

[17] D. A. Hudson, and C. D. Manning, "Compositional attention networks for machine reasoning," 2018.

[18] P. Blazek, and M. Lin, "Explainable neural networks that simulate reasoning," 2021.