6th Polish Conference on Artificial Intelligence
Katowice, 7-9 April 2025

How to improve human trust in AI?

AGH University of Krakow
Faculty of Electrical Engineering, Automatics, Computer Science and Biomedical Engineering
Department of Biocybernetics and Biomedical Engineering

Adrian Horzyk
horzyk@agh.edu.pl

Google: Adrian Horzyk

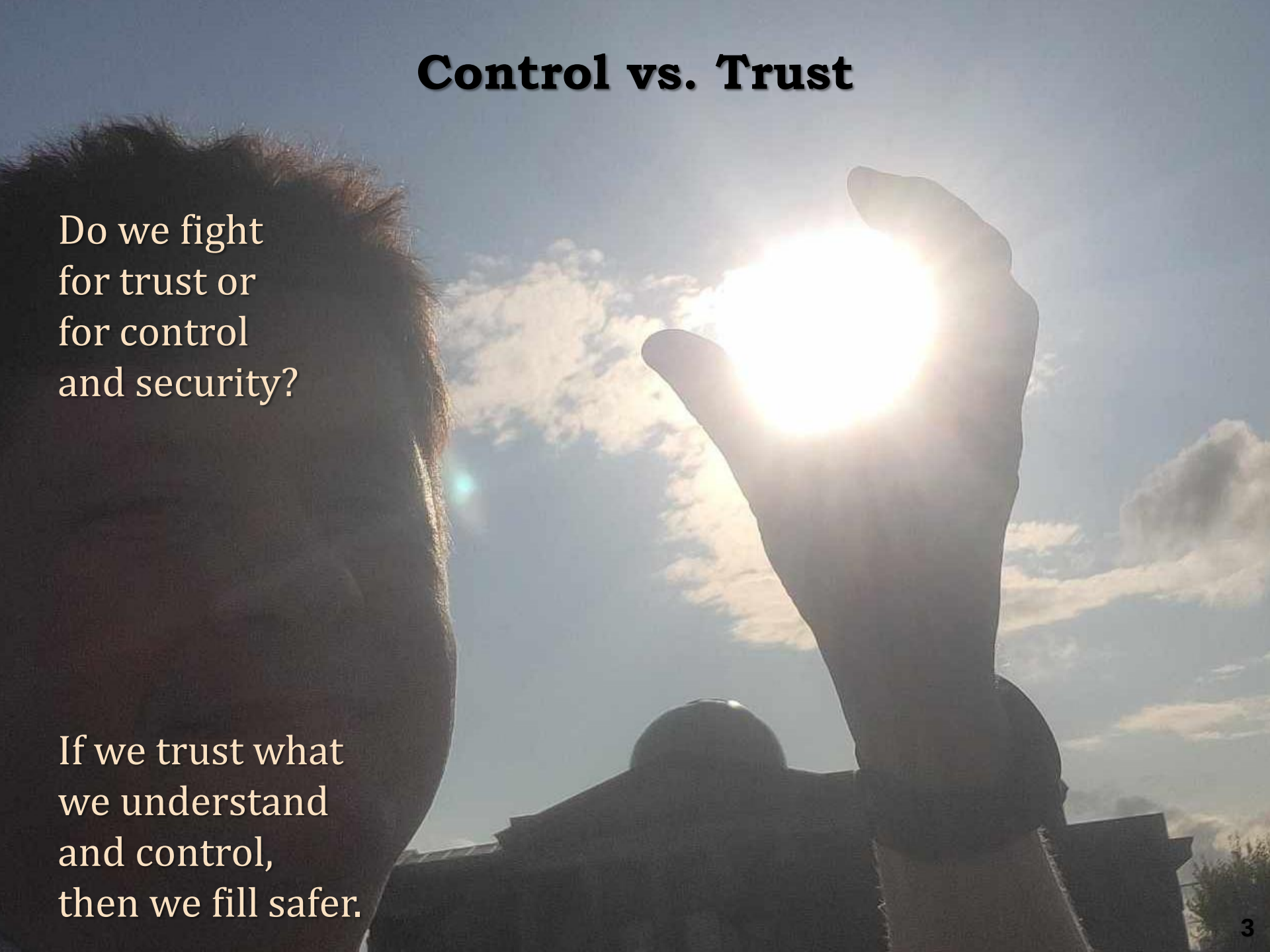
Cooperation vs. Competition and Trust Issues

Cooperating animals need to **trust** each other to achieve desired goals.

Competitors use weaknesses and mistakes of the opponents to win, so, **trusting** the opponent can be fatal.

Human intelligence should be more cooperative, and so should AI systems.

Control vs. Trust



Do we fight
for trust or
for control
and security?

If we trust what
we understand
and control,
then we fill safer.

Control vs. Trust

Do we fight
for trust or
for control
and security?

If we trust what
we understand
and control,
then we fill safer.

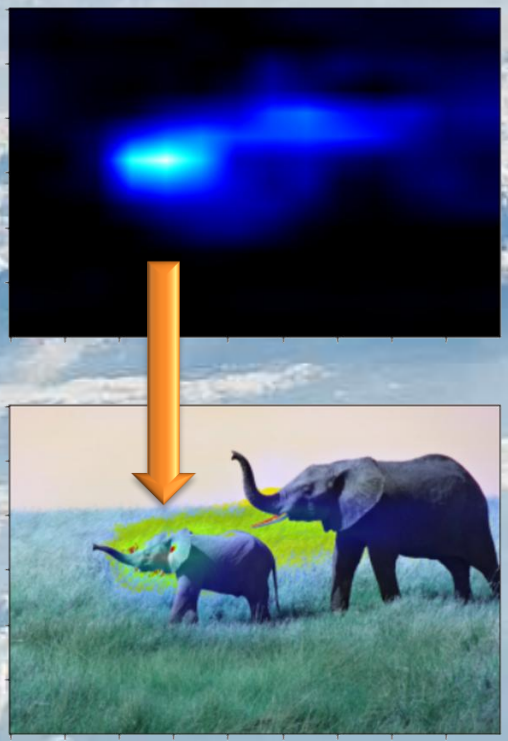
Understanding and Controlling AI?

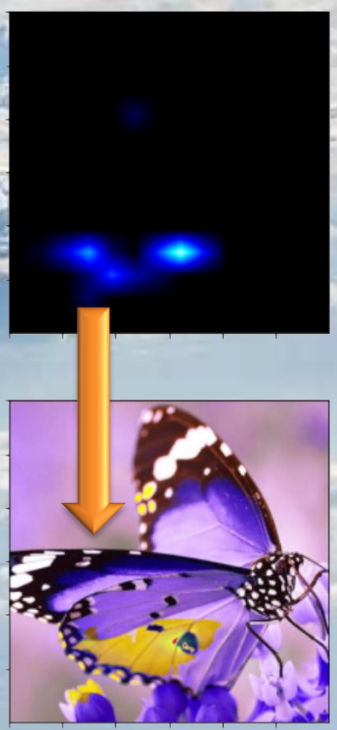
We tend to **trust** when
we either **understand** and have **control** over a situation ourselves
or when we believe that
someone we trust **understands** and **controls** it on our behalf.

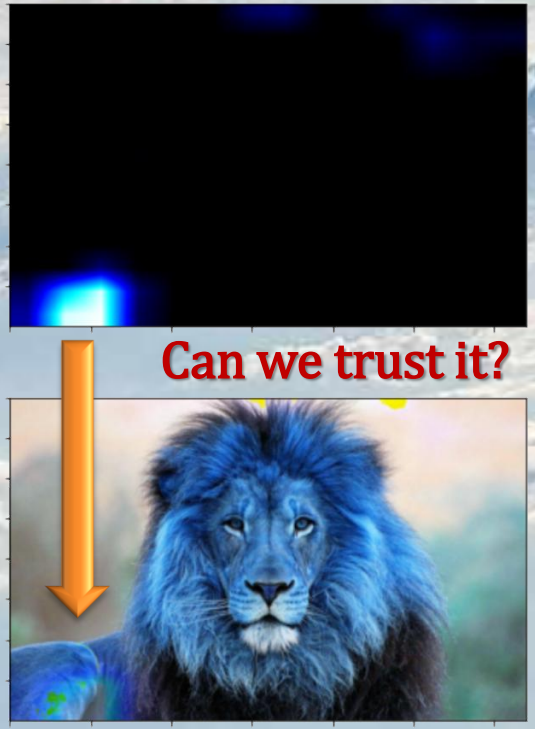
Our **trust** in AI will likely diminish or disappear if it behaves in ways that **defy our intuition or overlook data** that we consider essential for reasoning or moral issues.

Can we judge or anticipate how AI predicts or thinks?

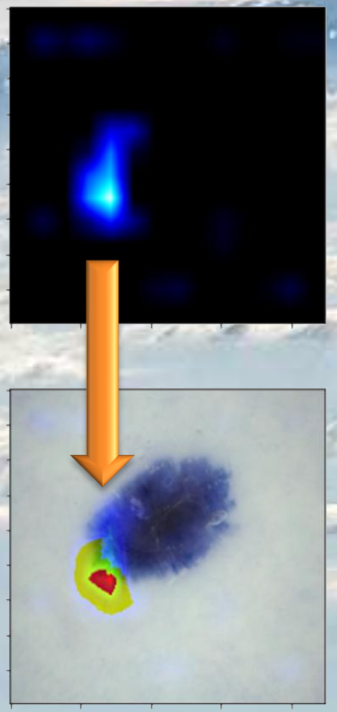
Is the attention of CNNs similar to ours?







Can we trust it?

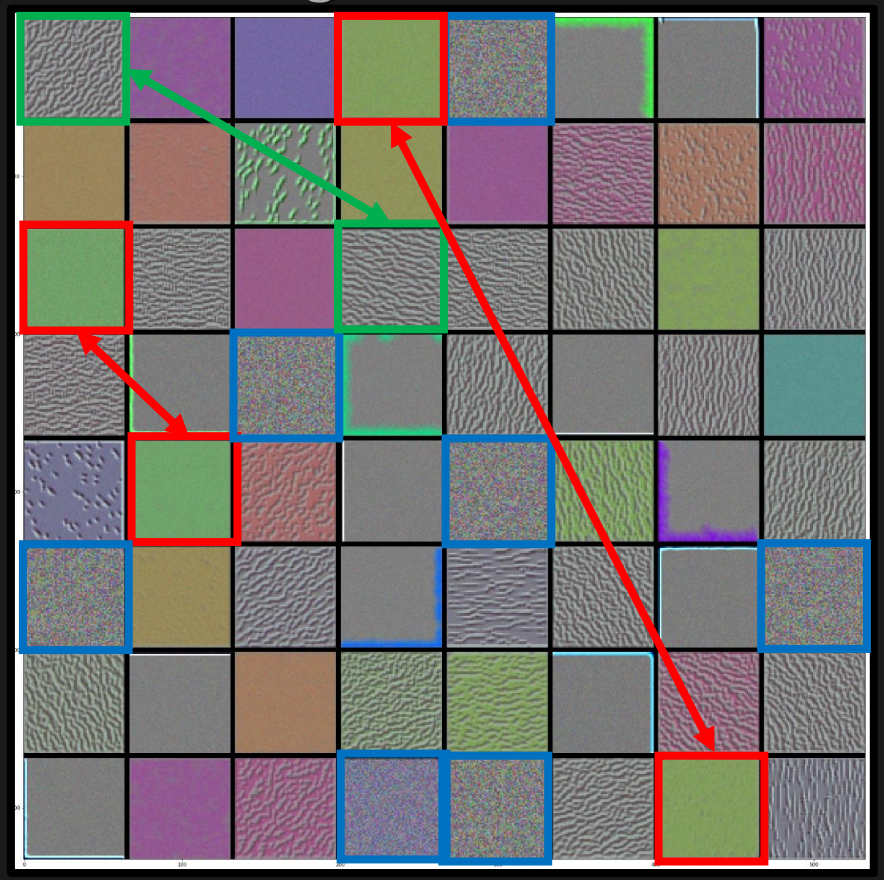


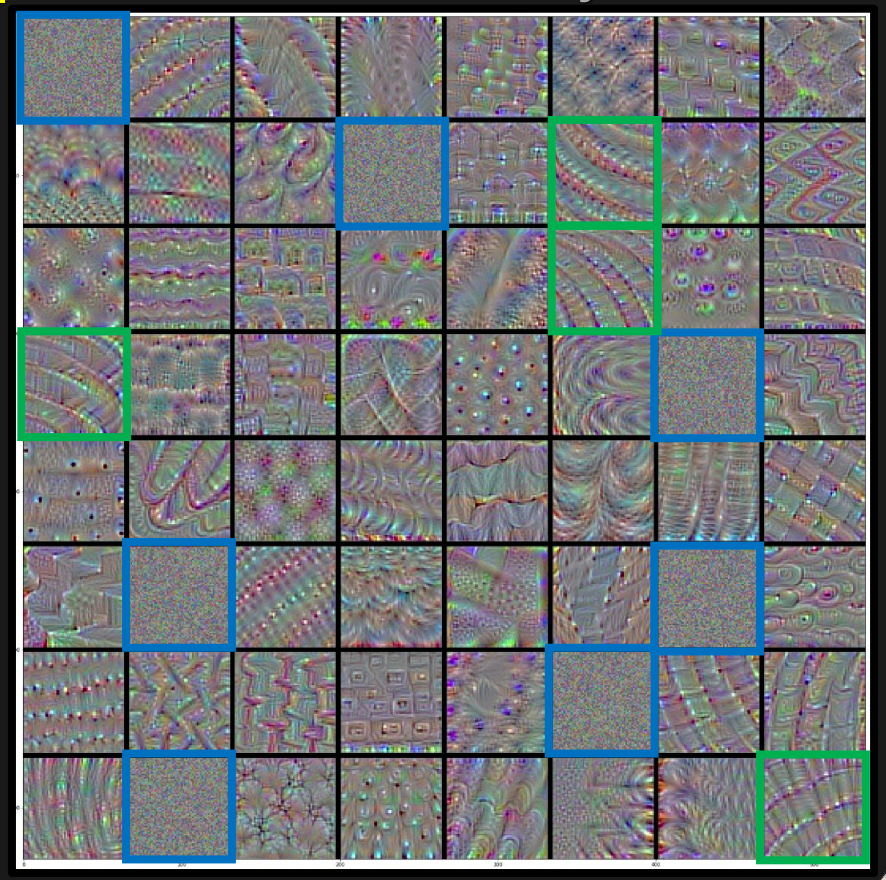
Heatmaps overlaid on classified images reveal where CNN **attention is focused**!

Do we understand the processes of AI reasoning or are we floating on the surface of hope?

Visualizing CNN Filters Reveals the Secrets Behind How CNN Work

Some **filters** of the transferred models stay **unadopted**; the other represent **too similar** or **rotated, flipped, and scaled patterns**, not leaving filters for more rare **patterns**, to which they are **blind**:





We can usually prune CNN dramatically and improve performance simultaneously:

Igor Ratajczyk, Adrian Horzyk, Advancing ConvNet Architectures: A Novel XGB-based Pruning Algorithm for Transfer Learning Efficiency, Proc. of ECAI 2024, Volume 392, Frontiers in Artificial Intelligence and Applications, 2024, pp. 2114–2121.

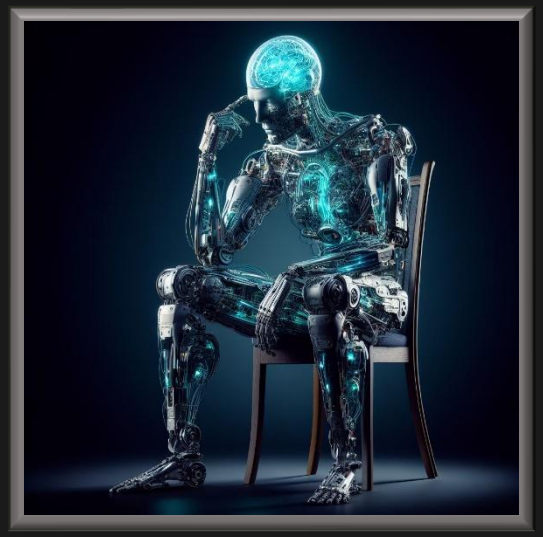
Transparency reinforces Trust in AI?

Modern AI systems aspire to be **explainable** and **transparent** to give us an **understanding** of how they predict and work and also **keep control over** the reasoning processes not to fool or mislead us.

- **Trust** is reinforced when AI systems are **transparent** and their decision-making processes are **accessible and interpretable** to us.
- We develop **trust** when AI systems **behave predictably** and **yield consistent results** across time, data distributions, and contexts.
- **Trust** increases when AI systems align with **social norms, ethical standards, and our intent and imperfection**, while misalignment or lack of forbearance can lead to fear or rejection (even for a single failure).
- To sustain **trust**, an AI system must be **robust**, i.e., can function well **under uncertainty or unexpected input**.
- **Trust** flourishes when we **remain in control** and **clear mechanisms for accountability** are established.

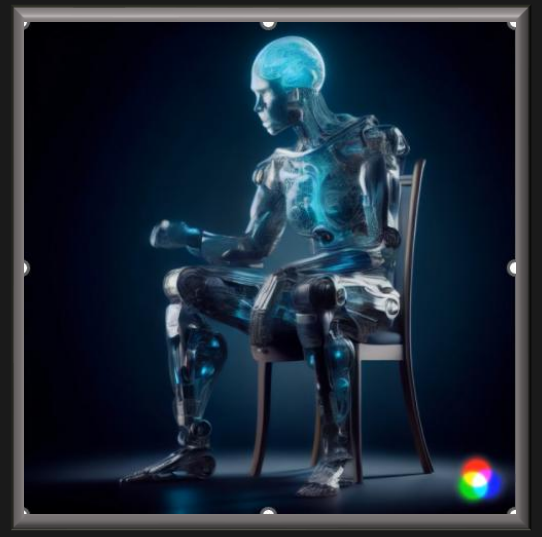
Averaging and losing details and individuality!

Generative AI systems many times average and lose details or unusual and individual features!

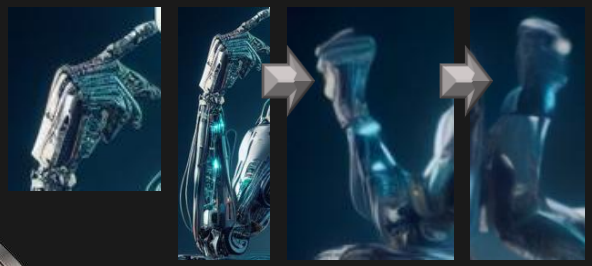


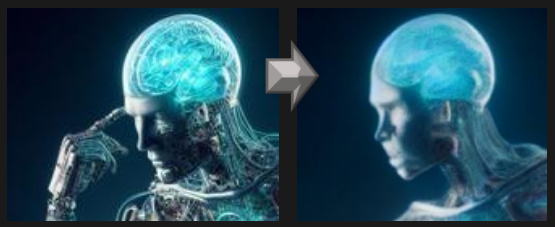
input image

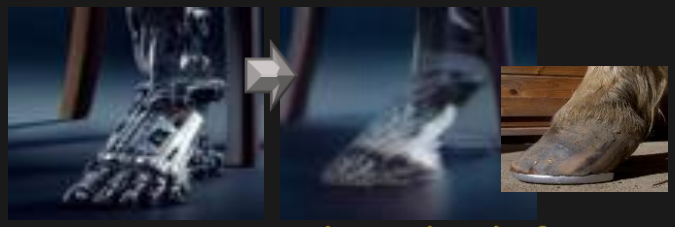
generated animation



the generated step







the robot's foot looks like a hoof

Are robots in sci-fi movies as diverse as humans?

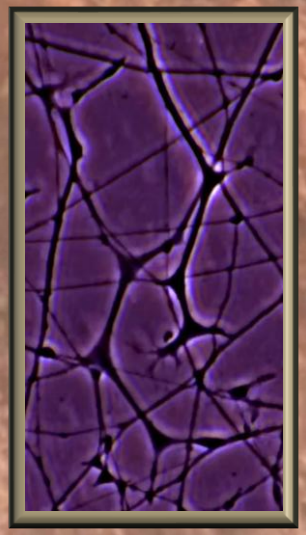
How much differ the AI platform from the real intelligence one?

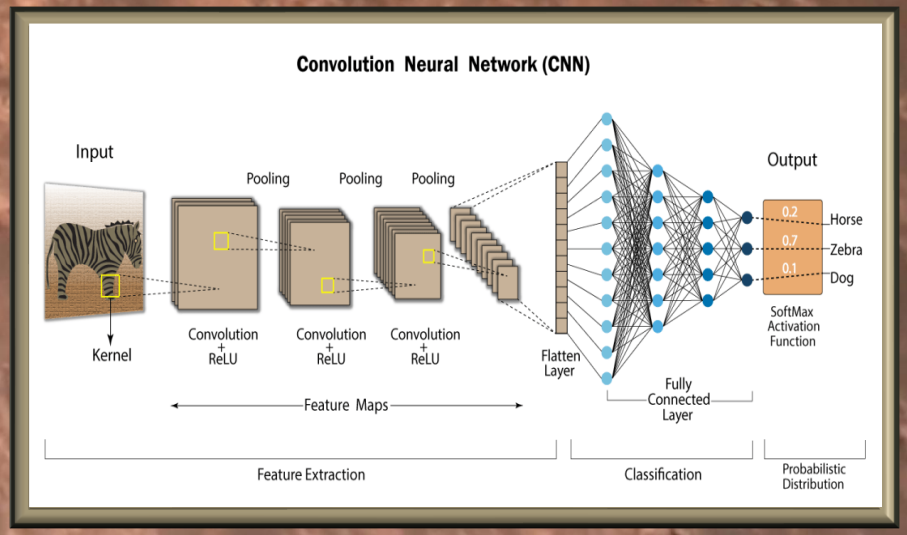
Real neurons and brains:

- **Diverse neurons** of different functions.
- **Plastic and adaptable structures** during life.
- **Sparse and adaptable connections** that are created during learning.
- **No hyperparameters**, fully data-dependent.
- Time-based changes and reactions.
- Lifelong **training** of changing training data, objects, classes, and their relationships.
- Needs and fears define **motivation** and **goals**.

Artificial neurons and networks:

- The **same neurons**, except activation function.
- **Rigid and fixed structures** during training.
- **Fully or regularly connected** neurons between specified layers.
- **Many hyperparameters** to optimize.
- Layered-based steps and calculations.
- **Training** of fixed datasets and a limited number of classes to which they can assign objects.
- **Goals** defined by people may ignore ethics.





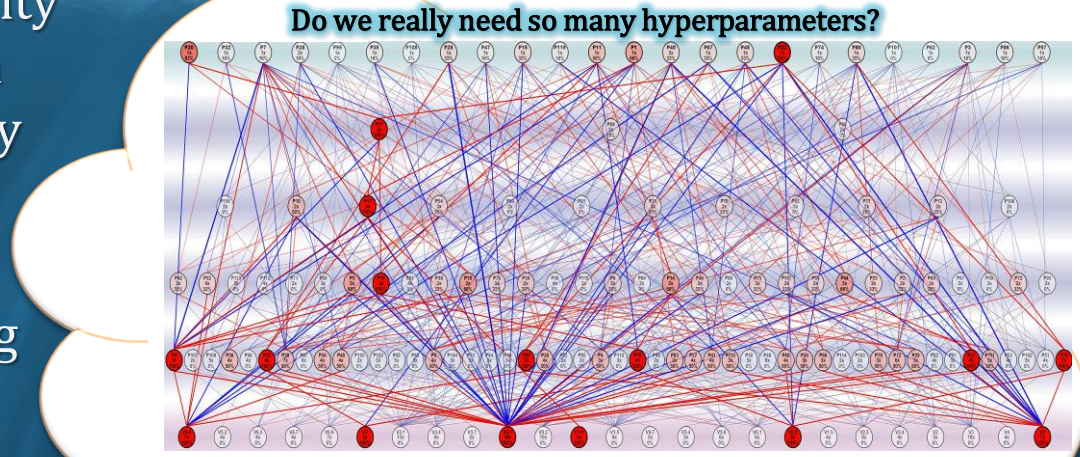
We need AI systems that will be fully data-dependent without any hyperparameters required.

Wisdom = Intelligence + Trust + Experience + Judgement

Wisdom includes the ability to apply intelligence with discernment and integrity to earn and sustain trust.

Wisdom is about knowing when, why, and how to apply intelligence in accordance with moral or ethical awareness.

Wisdom incorporates emotional intelligence, including understanding how decisions affect others, which reinforces **trust**.



Trust is built when intelligence is applied with empathy, consistency, and integrity.

One earns **trust** not just by being smart but by being reliable, fair, and aligned with moral values.

Adrian Horzyk, Jakub Kosno, Daniel Bulanda, Janusz A. Starzyk, Explainable Sparse Associative Self-Optimizing Neural Networks for Classification, Eds. Biao Luo, Long Cheng, Zheng-Guang Wu, Hongyi Li, Chaojie Li, Proc. of 2023 International Conference on Neural Information Processing (ICONIP 2023), Springer Nature Computer Science (CCIS, LNAI, LNCS), Springer Nature Singapore Pte Ltd. 2024, LNCS (14450), CCIS 1963, Vol. 3, pp. 229–244, 2024.

Could we share the world with AI in trust?

Yes, but only if trust is earned, transparent, and reciprocated.

Trust in AI is not automatic; it must be grounded in:

- Trustworthy design (explainability, accountability, robustness)
- Mutual Coexistence (humans are empowered, not replaced)
- Ethics (respect for human values, rights, and diversity)