New Frontiers

Richard Brooks

MAL2 – Spring 2025

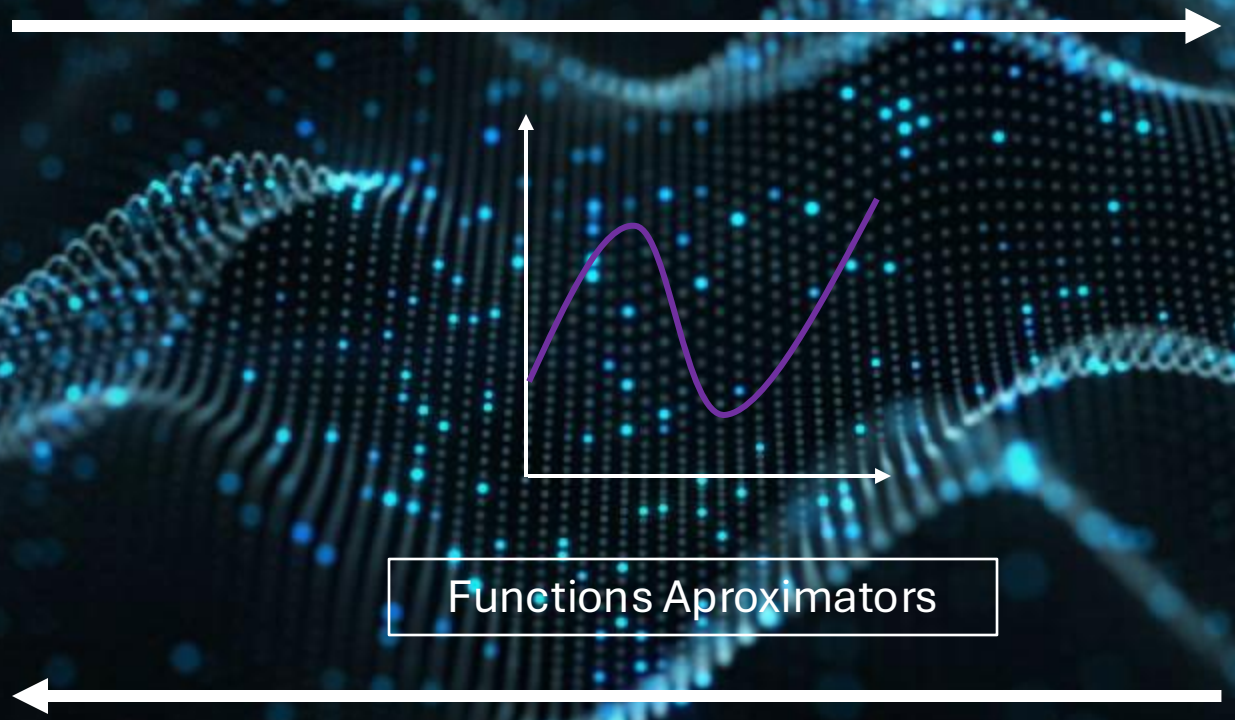
Credits: MIT 6.S191 Spring 2025

So far in MAL1+2

Data

- Signals
- Images
- Sensors

....



Decision

- Prediction
- Predictions
- Actions

....

Power of Neural Networks

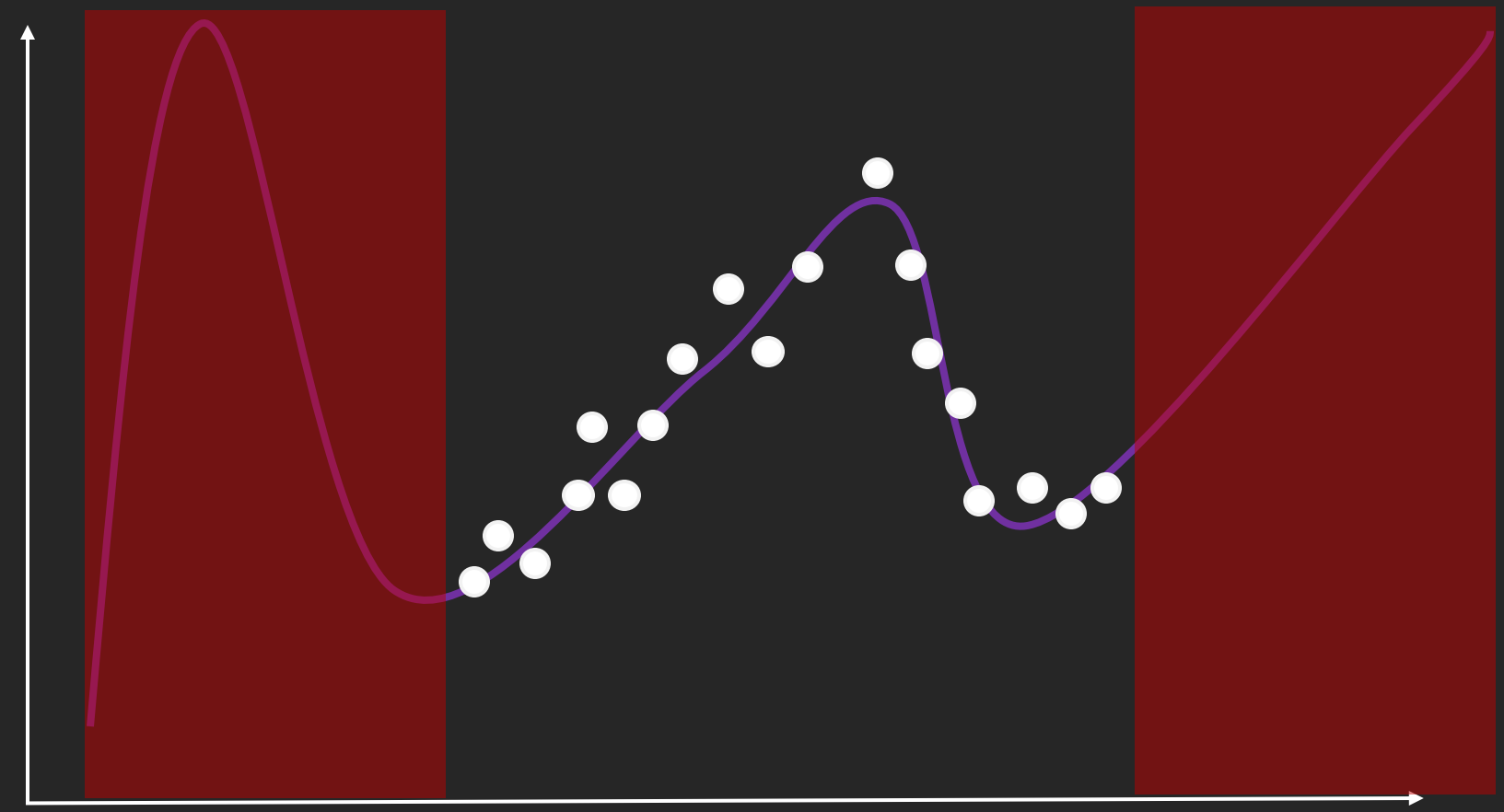
Universal Approximation Theorem

A feedforward network with a single layer is sufficient to approximate, to an arbitrary precision, any continuous function

Limitations

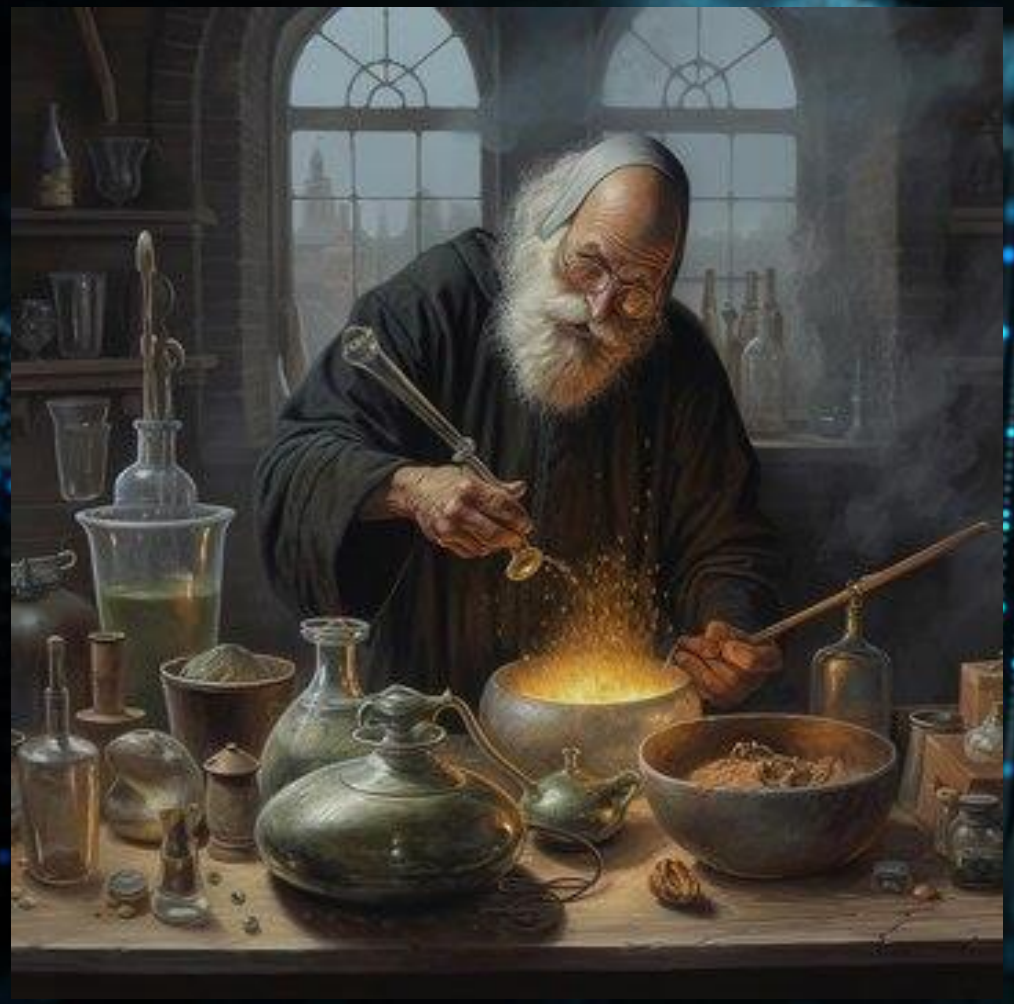
Neural networks are excellent function approximators
…when they have training data

Limitations



How do we know when our network doesn’t know?

Deep Learning = Alchemy?



NN Failure

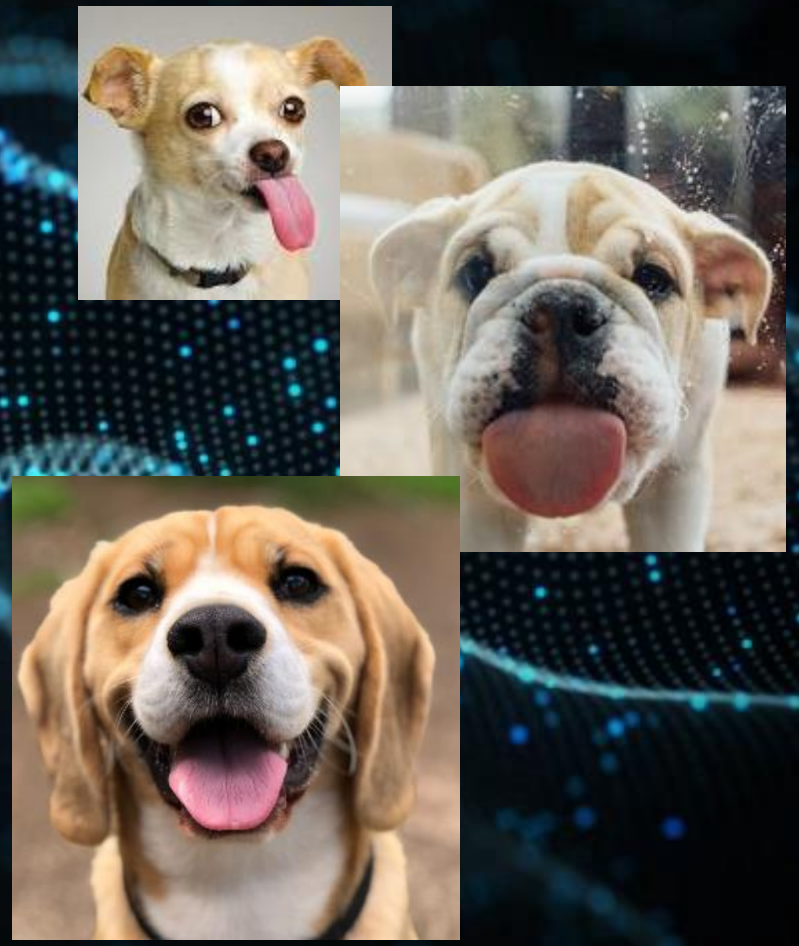
CNN

Train network to
colorize BW images.

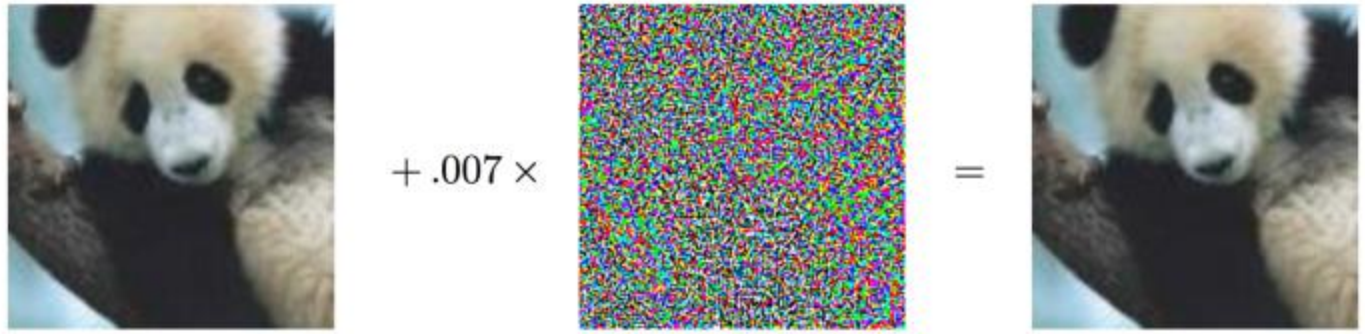
Why could this be the case?

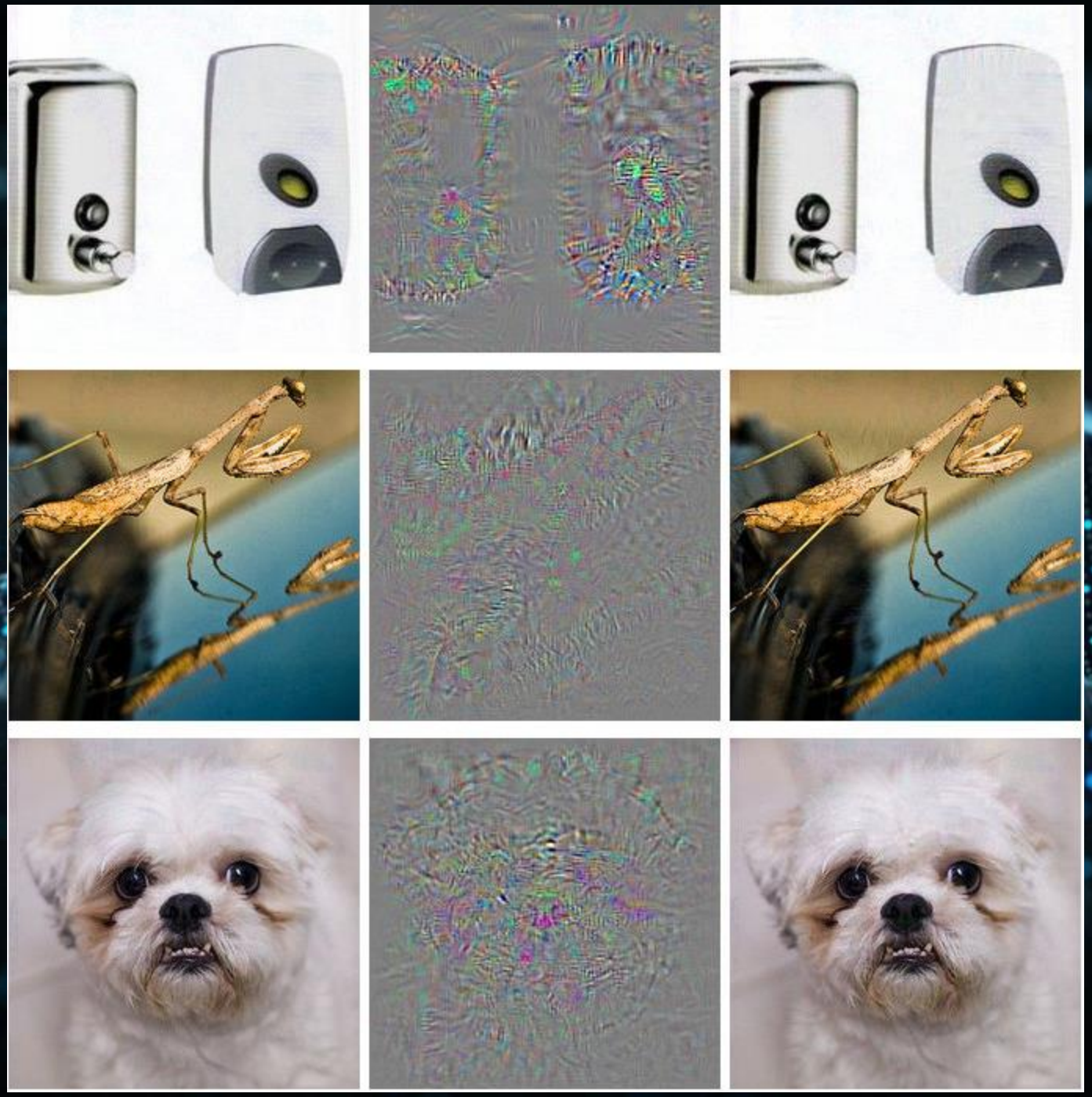
What Happens During Training

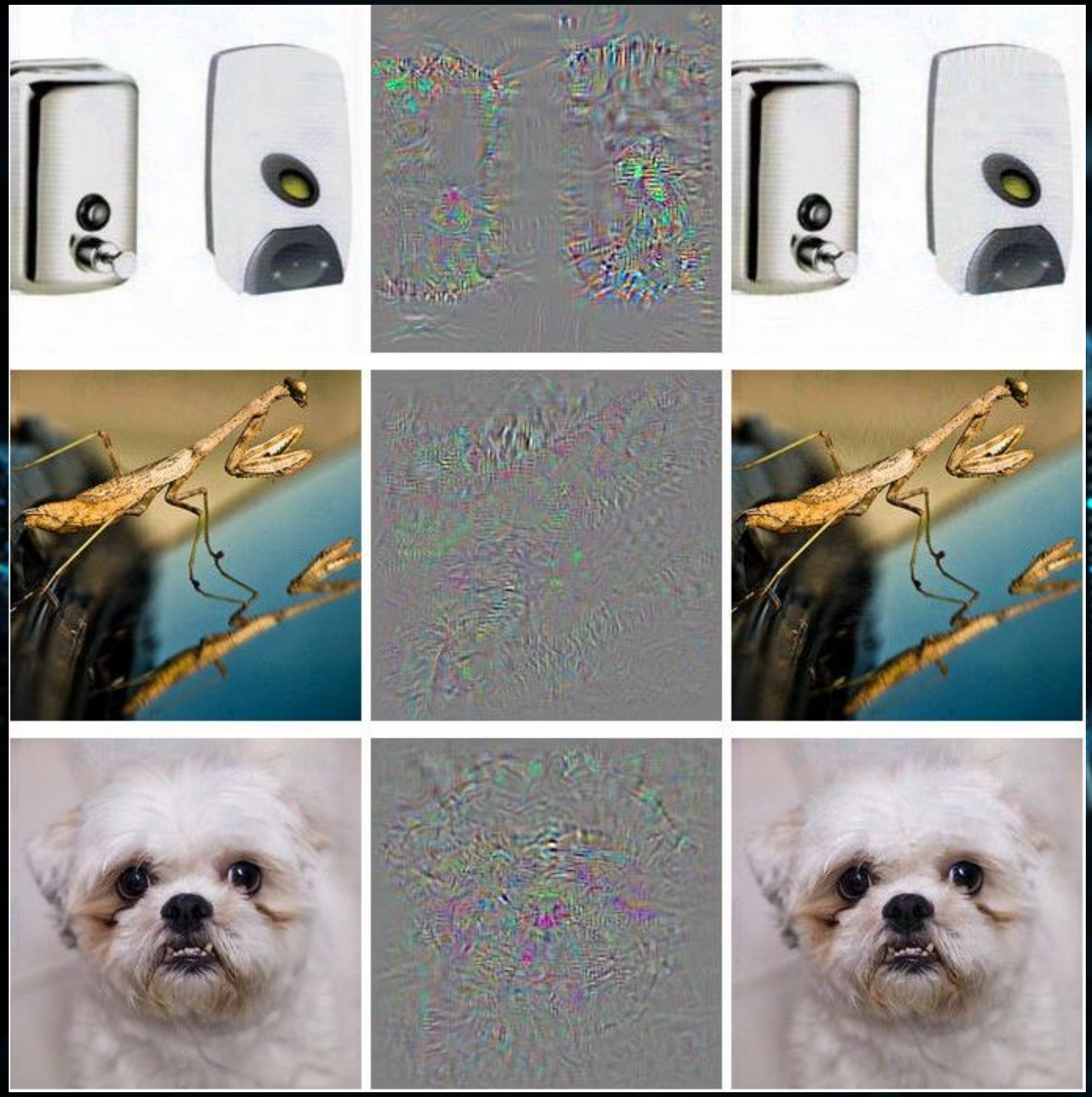
CNN



NN Failure







These adversarial examples were generated by minimizing the following function with respect to

$$\text{loss}(\hat{f}(\mathbf{x}+\mathbf{r}), l) + c \cdot |\mathbf{r}|$$

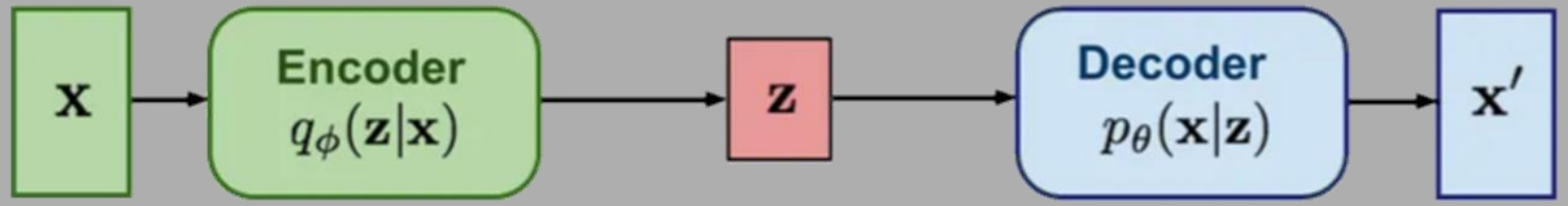
NN Limitations

- Very **data hungry** (eg. often millions of examples)
- **Computationally intensive** to train and deploy (tractably requires GPUs)
- Easily fooled by **adversarial examples**
- Can be subject to **algorithmic bias**
- Poor at **representing uncertainty** (how do you know what the model knows?)
- Uninterpretable **black boxes**, difficult to trust
- Often require **expert knowledge** to design, fine tune architectures
- Difficult to **encode structure** and prior knowledge during learning
- **Extrapolation**: struggle to go beyond the data
- Hallucinations

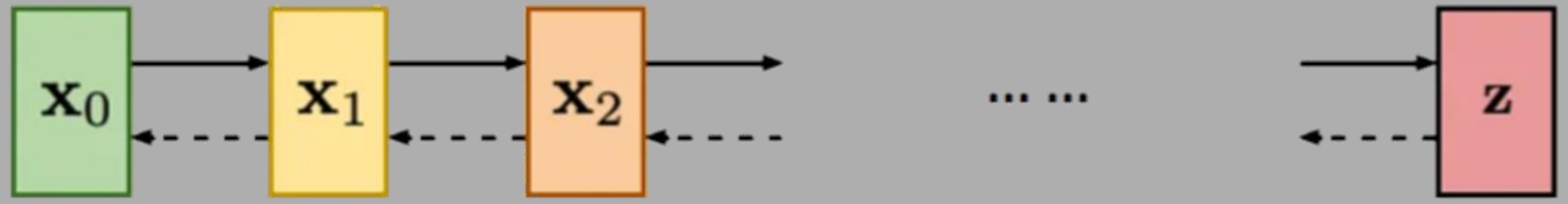
Generative AI

GAN: Adversarial training

VAE: maximize variational lower bound

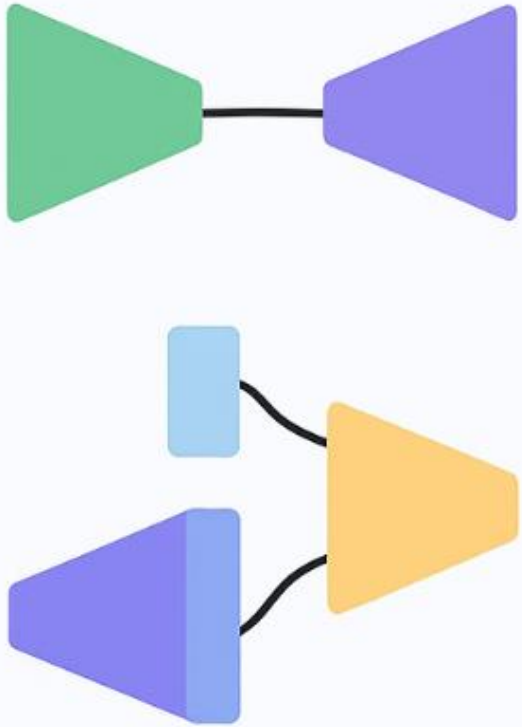


Diffusion models: Gradually add Gaussian noise and then reverse

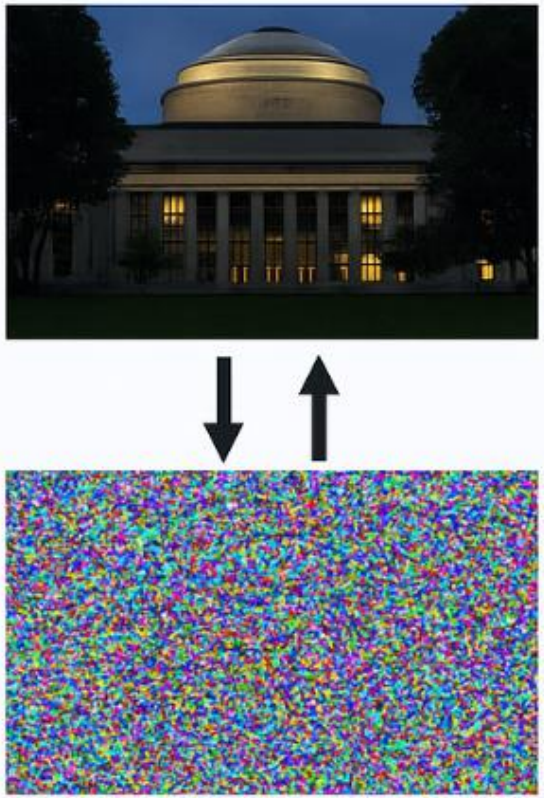


The Landscape of Generative Modeling

VAEs and GANs



Diffusion Models

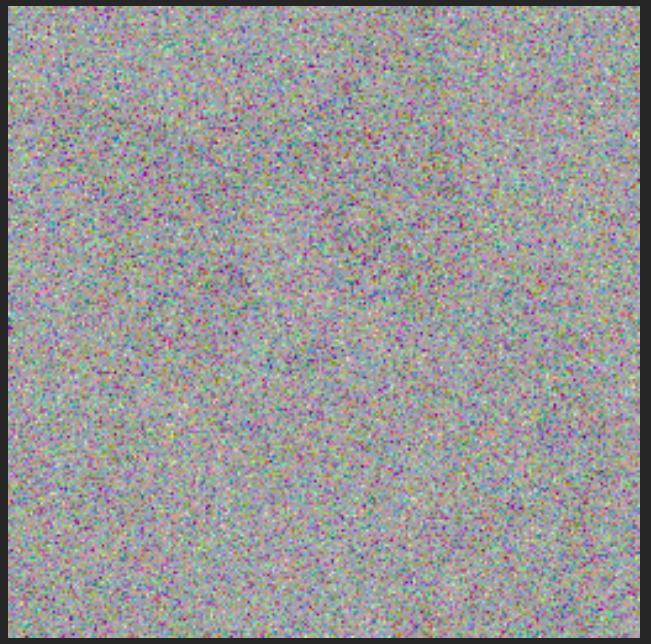


Text-to-Image

"Two cats doing research"

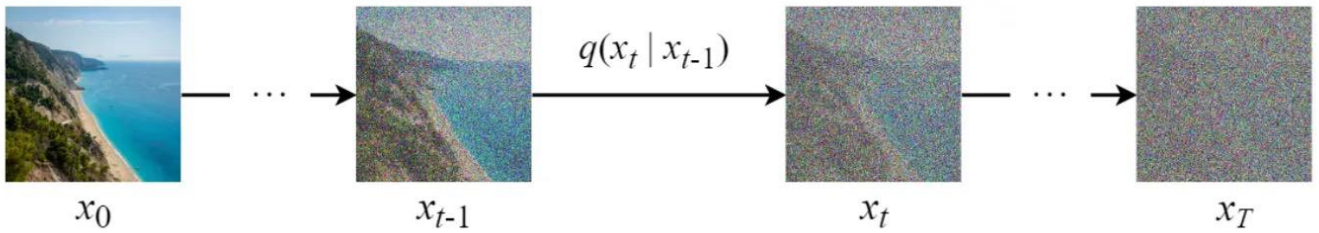
Diffusion Models

- Algorithm for training a diffusion model
  - Forward process — the process of adding noise to the image until it becomes a pure noise image.
  - Reverse process — the process where the model learns how to denoise the image gradually. This process starts with a complete noise image and should finish with a good looking image.
  - Reverse process makes up the generative model of the data.



Forward pass

① Forward Diffusion
- add noise (Gaussian)
- time steps
- Markov Chain

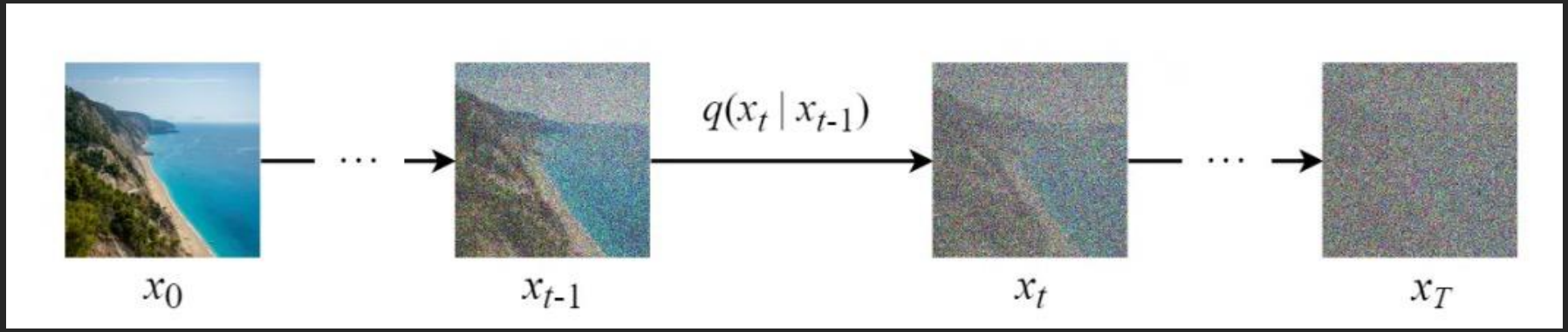


Distribution of the noised images | Output | Mean μ_t | Variance Σ_t

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t;\ \sqrt{1-\beta_t}x_{t-1},\ \beta_t I)$$

Notations:
- t : time step (from 0 to T)
- x_0 : a data sampled from the real data distrition $q(x)$ (i.e. $x_0 \sim q(x)$)
- β_t : variance schedule ($0 \le \beta_t \le 1$, and β_0 = small number, β_T = large number)
- I : identity matrix

Reverse Denoising Process



so the main objective of the model here is to learn a function which Predicts the mean and variance of each step to predict the reverse Process.

② Reverse Diffusion
↳ Remove Noise
↳ CNN. U. Net
↳ MSE

Reverse Proces

- The mean and variance in the reverse trajectory is learned, and the final function is used to undo the diffusion process.

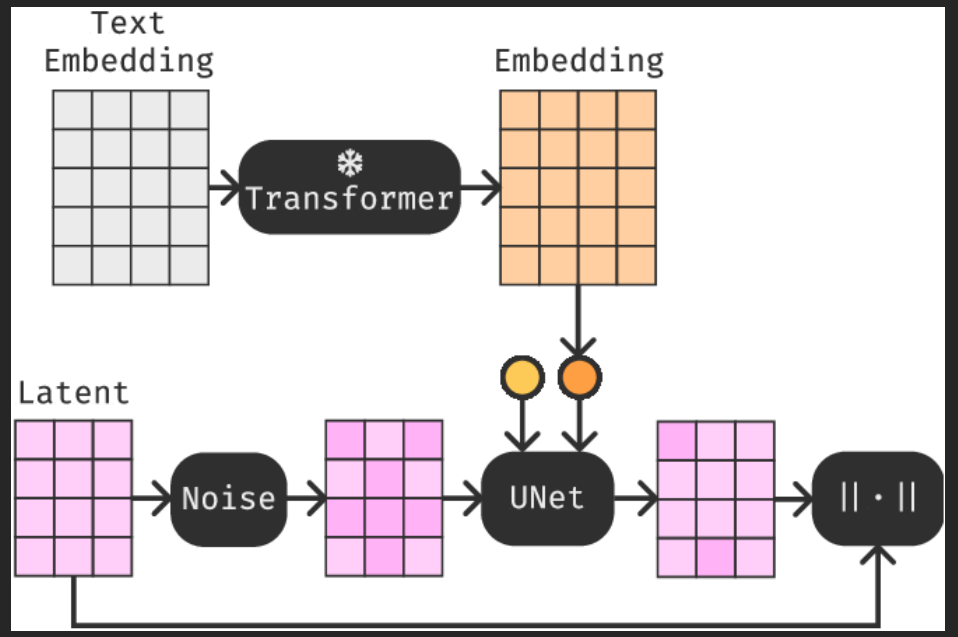
- Final reverse process:

$$p(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$$

- Where:
- - $\mu(x_t, t)$ is the **mean**
- - $\Sigma(x_t, t)$ is the **variance**

Conditional Diffusion (Text-to-Image Guidance)

- Text Representation (Embeddings)
  - Text prompts need to be converted into numerical representations that capture their meaning.
- Conditioning the Denoising:
  - The core idea is that the text embedding is provided as additional input to the U-Net during the reverse diffusion (denoising) process at each time step t.
- Guided Noise Prediction:
  - The U-Net now learns to predict the noise not only based on the noisy image x_t but also conditioned on the provided text embedding. It learns to remove noise in a way that steers the image formation towards something that matches the text description.



From Monolithic Models to AI Agents

'22 '23 '24

LLMs

Compound A.I Sys

Agents

From models → Compound Sys.

RAG

⊖ limited knowledge

⊖ hard to adapt

Vacation days:

query → generate (LLM) → answer ✗

query → search query (LLM)

generate (LLM) → answer ✓

System Design

⊕ Modular

⊕ Easier to adopt

Control logic

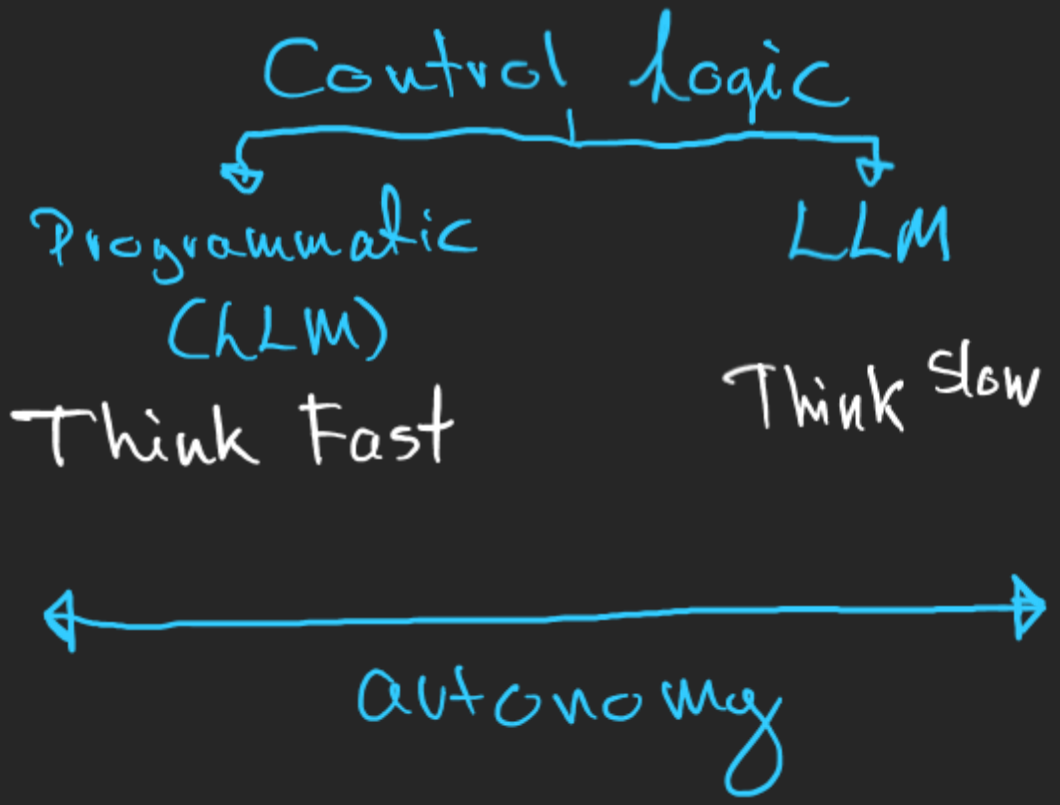
Programmatic (LLM)

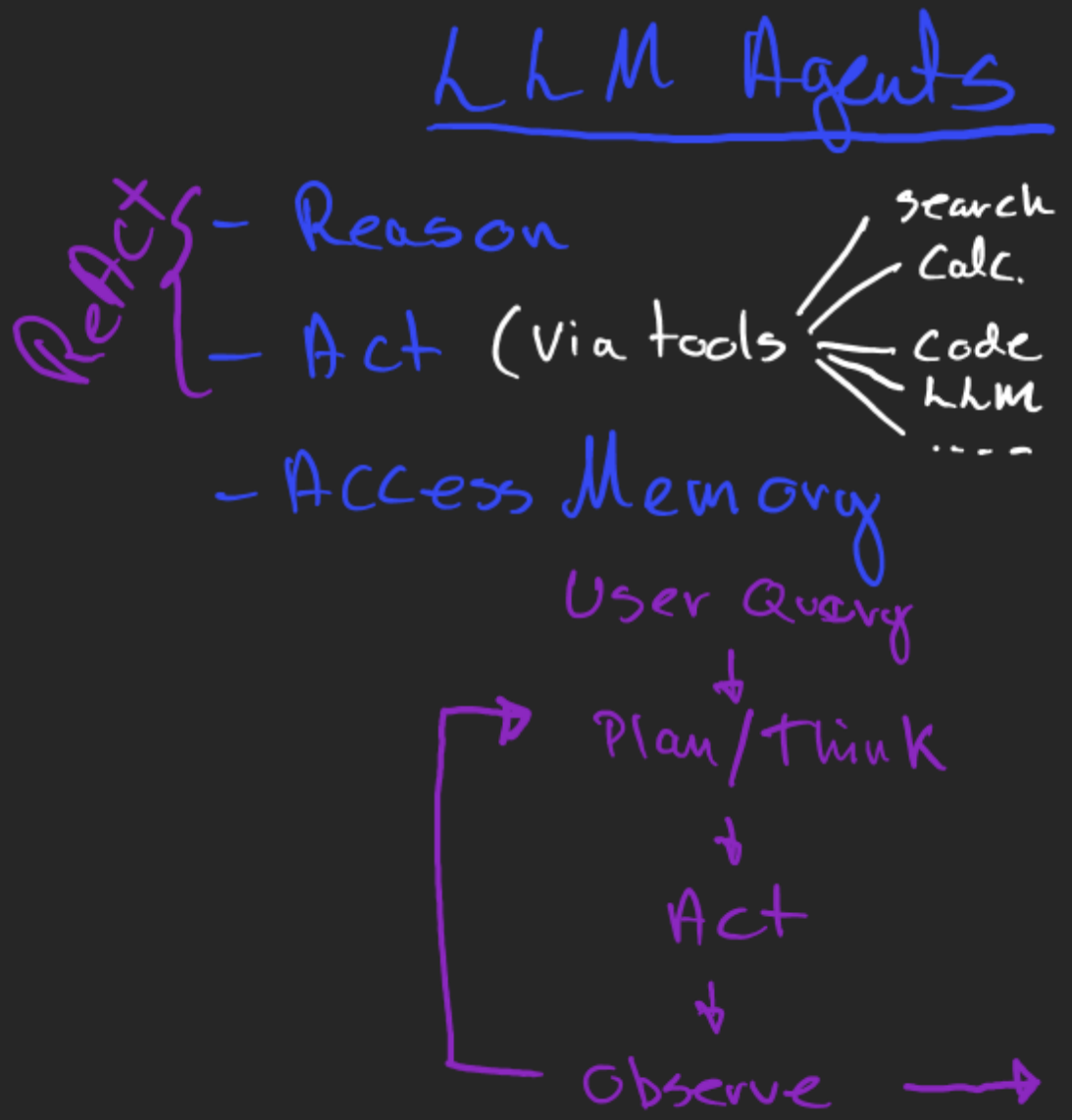
LLM

Think Fast

Think Slow

LLM Agents


Control logic
Programmatic
(LLM)
Think Fast
LLM
Think Slow
autonomy


LLM Agents
ReAct
- Reason
- Act (Via tools
search
Calc.
code
LLM
....
- Access Memory
User Query
Plan/Think
Act
Observe
Answer

Non=Agentic Workflow (Zero-Shot)	Agentic Workflow	Truly Autonomous AI Agent
One-Pass Execution - Completes ataskin one go without iteration. No Adjustments - No revisions or refinements.	Step-by-Step Execution - Breaks tasks into stages (e.g., plannfig, research, drafting, revising)。 FeedbackLoop -Aliterates based on human guidance.	Fully Independent AI - Al determines steps, tools, and iterateswithouthuman involvements Adaptive Decision-Making - Al adjusts workflow dynamically to achieve the best outcome.

Design Patterns

1. Reflections
   - AI reviews and improves its own output by analysing and refining responses
2. Tool Use
   - Agents leverage external tools like web search or code execution
3. Planning and Reasoning
   - AI determines optimal steps and selects tools to execute tasks efficiently
4. Multi-Agent Systems
   - Multiple AI model with specialized roles
     1. Sequential
     2. Hierarchal Pattern
     3. Parallel Systems
     4. Asynchronous Systems

The compas algorithm

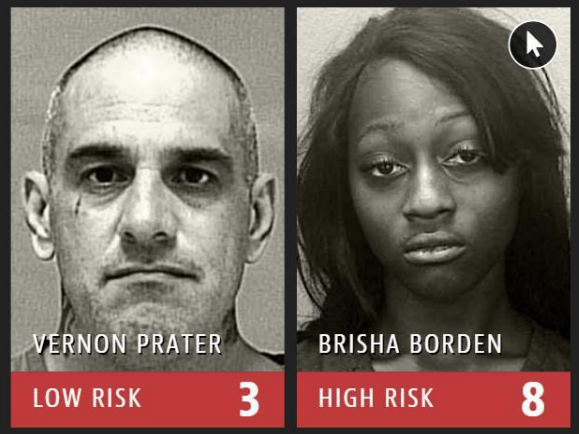
Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

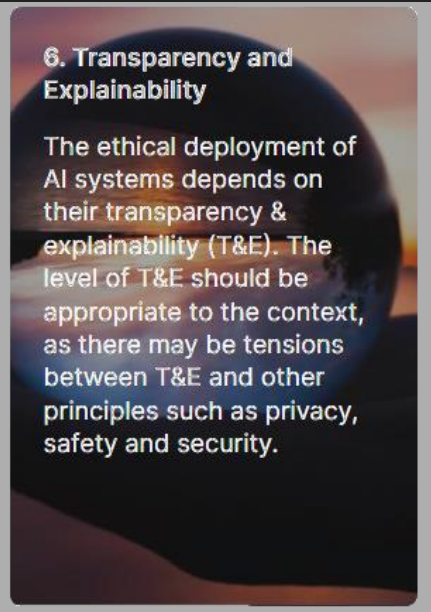
Two Petty Theft Arrests

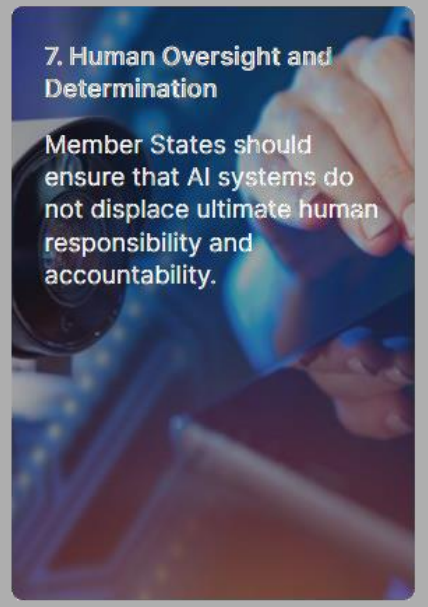


Prior offences
2 armed robberies
1 attempted armed robbery
Subsequent offences
1 grand theft

Prior offences
4 juvenile misdemeanors

Subsequent offences
None





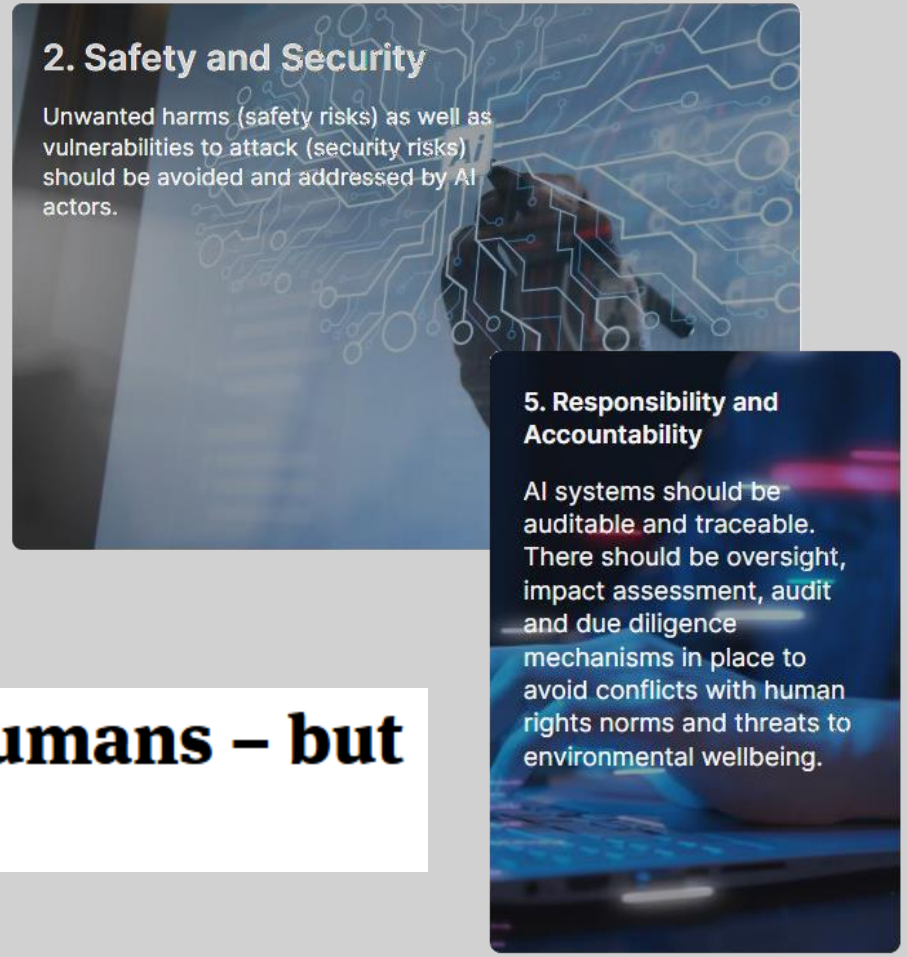
Uber Self-driving Car Fatality

Uber self-driving car kills pedestrian in first fatal autonomous crash

Driver Charged in Uber's Fatal 2018 Autonomous Car Crash

Uber self-driving car test driver pleads guilty to endangerment in pedestrian death case

Driverless cars are mostly safer than humans – but worse at turns



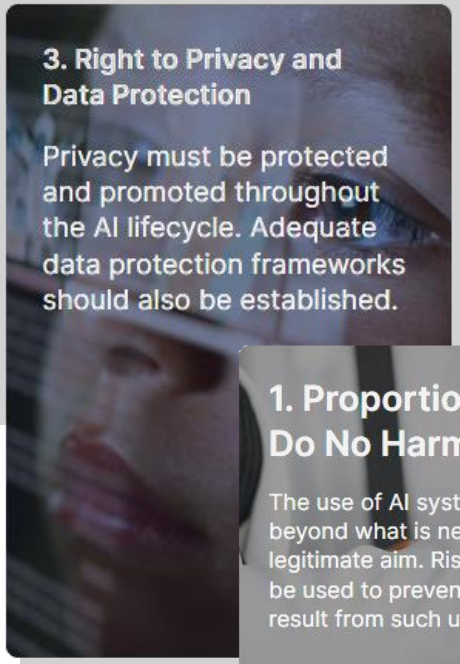
Clearview Ai Facial Recognition

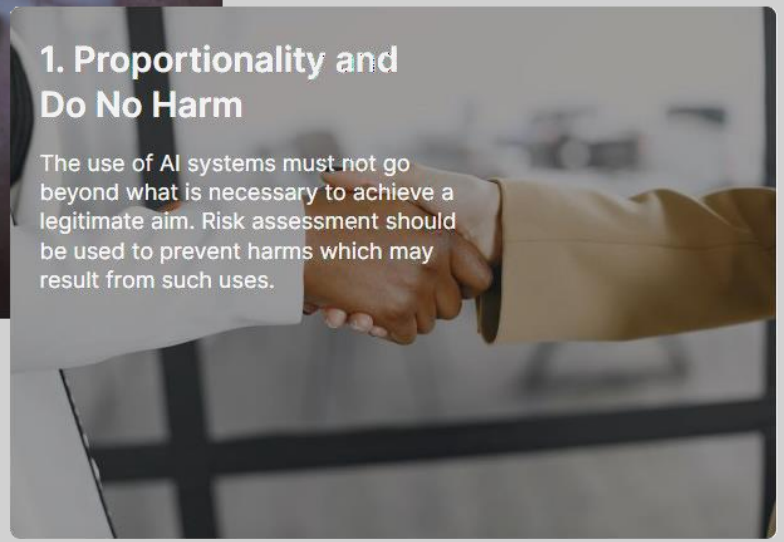
The Secretive Company That Might End Privacy as We Know It

A little-known start-up helps law enforcement match photos of unknown people to their online images — and "might lead to a dystopian future or something," a backer says.

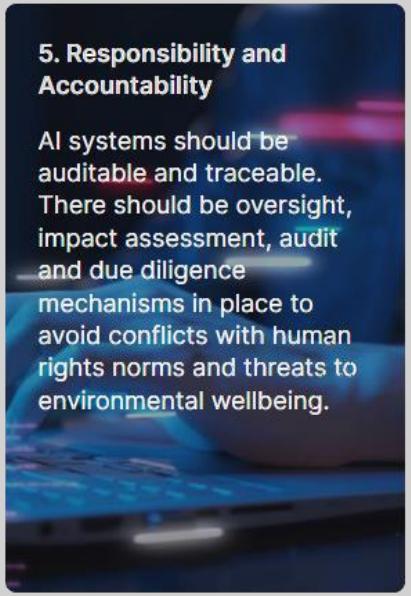
Clearview AI—Controversial Facial Recognition Firm—Fined $33 Million For 'Illegal Database'

AI can now Identify People as Gay or Straight from their Photo





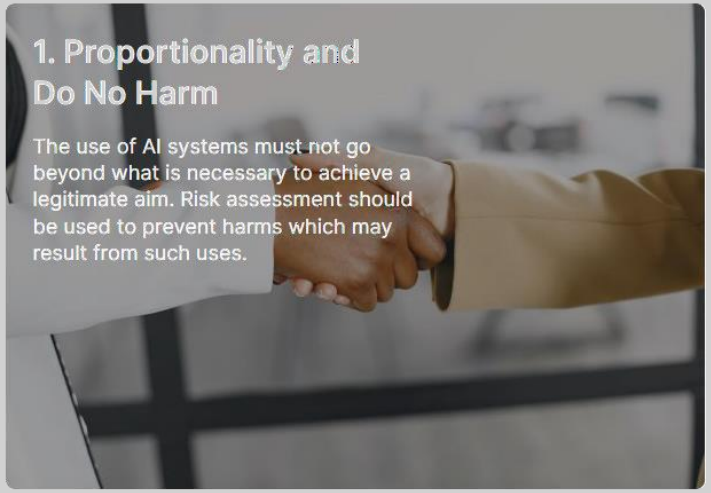
Chatbots



Can A.I. Be Blamed for a Teen's Suicide?

Snapchat-bot giver børn detaljerede råd om at selvskade

Man ends his life after an AI chatbot 'encouraged' him to sacrifice himself to stop climate change

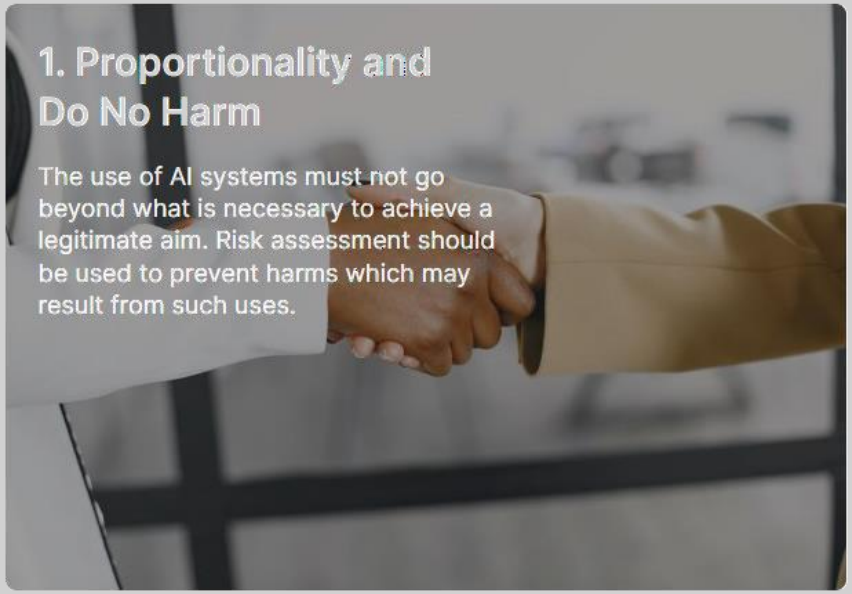


Sustainability

Artificial intelligence technology behind ChatGPT was built in Iowa — with a lot of water

ChatGPT And Generative AI Innovations Are Creating Sustainability Havoc

A.I. Could Soon Need as Much Electricity as an Entire Country

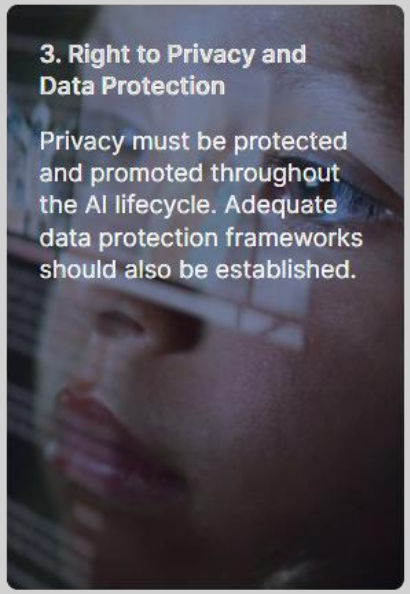


1. Proportionality and Do No Harm

The use of AI systems must not go beyond what is necessary to achieve a legitimate aim. Risk assessment should be used to prevent harms which may result from such uses.

2. Safety and Security

Unwanted harms (safety risks) as well as vulnerabilities to attack (security risks) should be avoided and addressed by AI actors.



3. Right to Privacy and Data Protection

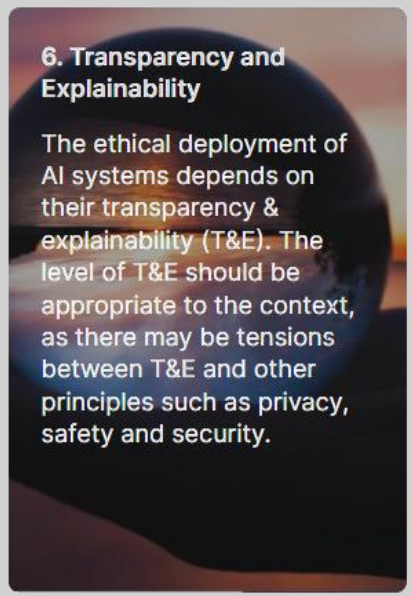
Privacy must be protected and promoted throughout the AI lifecycle. Adequate data protection frameworks should also be established.

4. Multi-stakeholder and Adaptive Governance & Collaboration

International law & national sovereignty must be respected in the use of data. Additionally, participation of diverse stakeholders is necessary for inclusive approaches to AI governance.

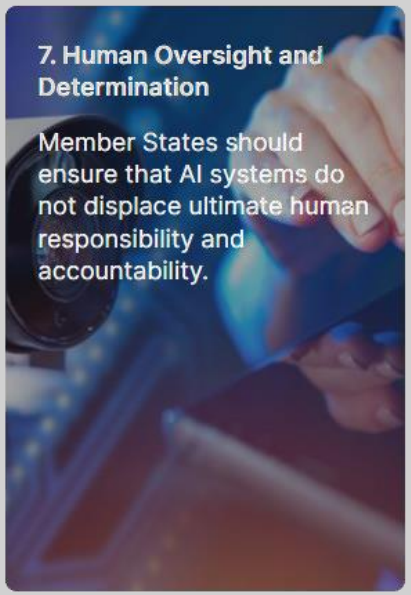
5. Responsibility and Accountability

AI systems should be auditable and traceable. There should be oversight, impact assessment, audit and due diligence mechanisms in place to avoid conflicts with human rights norms and threats to environmental wellbeing.



6. Transparency and Explainability

The ethical deployment of AI systems depends on their transparency & explainability (T&E). The level of T&E should be appropriate to the context, as there may be tensions between T&E and other principles such as privacy, safety and security.



7. Human Oversight and Determination

Member States should ensure that AI systems do not displace ultimate human responsibility and accountability.

8. Sustainability

AI technologies should be assessed against their impacts on 'sustainability', understood as a set of constantly evolving goals including those set out in the UN's Sustainable Development Goals.

9. Awareness & Literacy

Public understanding of AI and data should be promoted through open & accessible education, civic engagement, digital skills & AI ethics training, media & information literacy.

10. Fairness and Non-Discrimination

AI actors should promote social justice, fairness, and non-discrimination while taking an inclusive approach to ensure AI's benefits are accessible to all.

Make an inspirational quote to end my advanced machine learning course MAL2 with.

Every model we build expands not just our knowledge, but our ability to see patterns and possibilities in the world around us. As you move forward, remember: the true power of machine learning isn't in algorithms – it's in the questions we dare to ask and the insights we strive to reveal.

Keep learning, stay curious, and continue transforming data into discovery.

Best of luck!