LLMs,
the perfect
psychopaths.


[1]

A case for JEPA.

Agenda

#1	Relevance	PHYLOSOPHY
#2	Three minutes intro into the entire history of artificial intelligence	
#3	Are we hidden inside language?	
#4	Broader future - energy based models	
#5	Beyond LLMs – what is JEPA	TECH
#6	Why? A practical example.	
#7	AGI vs. contained psychopaths	


[2]

9th June 2025 – The Illusion of the Thinking

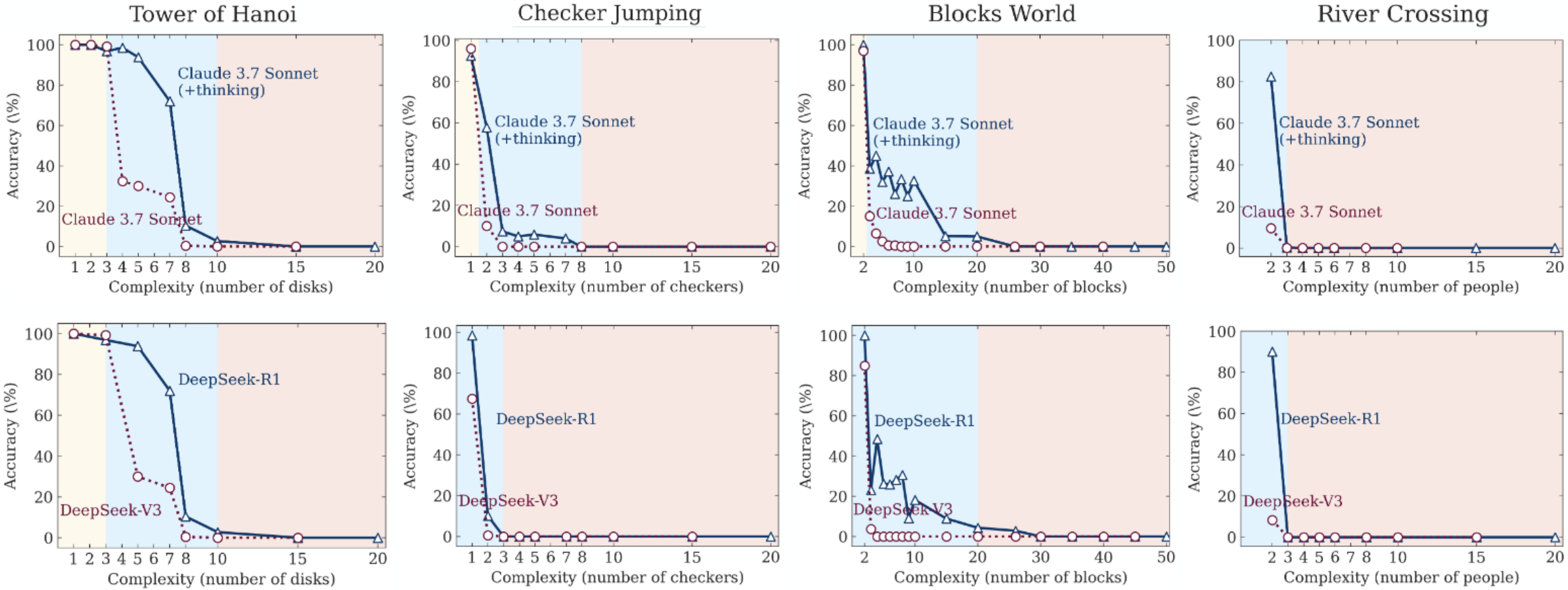
Apple Machine Learning Research
Collapse of reasoning in LRMs

10th June 2025 – The Illusion of the Illusion of Thinking [4]

Comment on The Illusion of thinking
Some problems were straight up impossible – not a fault of reasoning
Failure vs. output truncation
AI Stunt?

12th June 2025 – V-JEPA2 announced [5]

FAIR at Meta, Mila – Quebec AI Institute and Polytechnique Montréal – Yann LeCun
Next gen image-based world model building


Tower of Hanoi
Checker Jumping
Blocks World
River Crossing
Accuracy (%)
Claude 3.7 Sonnet (+thinking)
Claude 3.7 Sonnet
DeepSeek-R1
DeepSeek-V3
Complexity (number of disks)
Complexity (number of checkers)
Complexity (number of blocks)
Complexity (number of people)

[4]

1950s → 1970/80s → 2020s

Turing – If a machine is indistinguishable from a human does it matter?

What is measure of man?

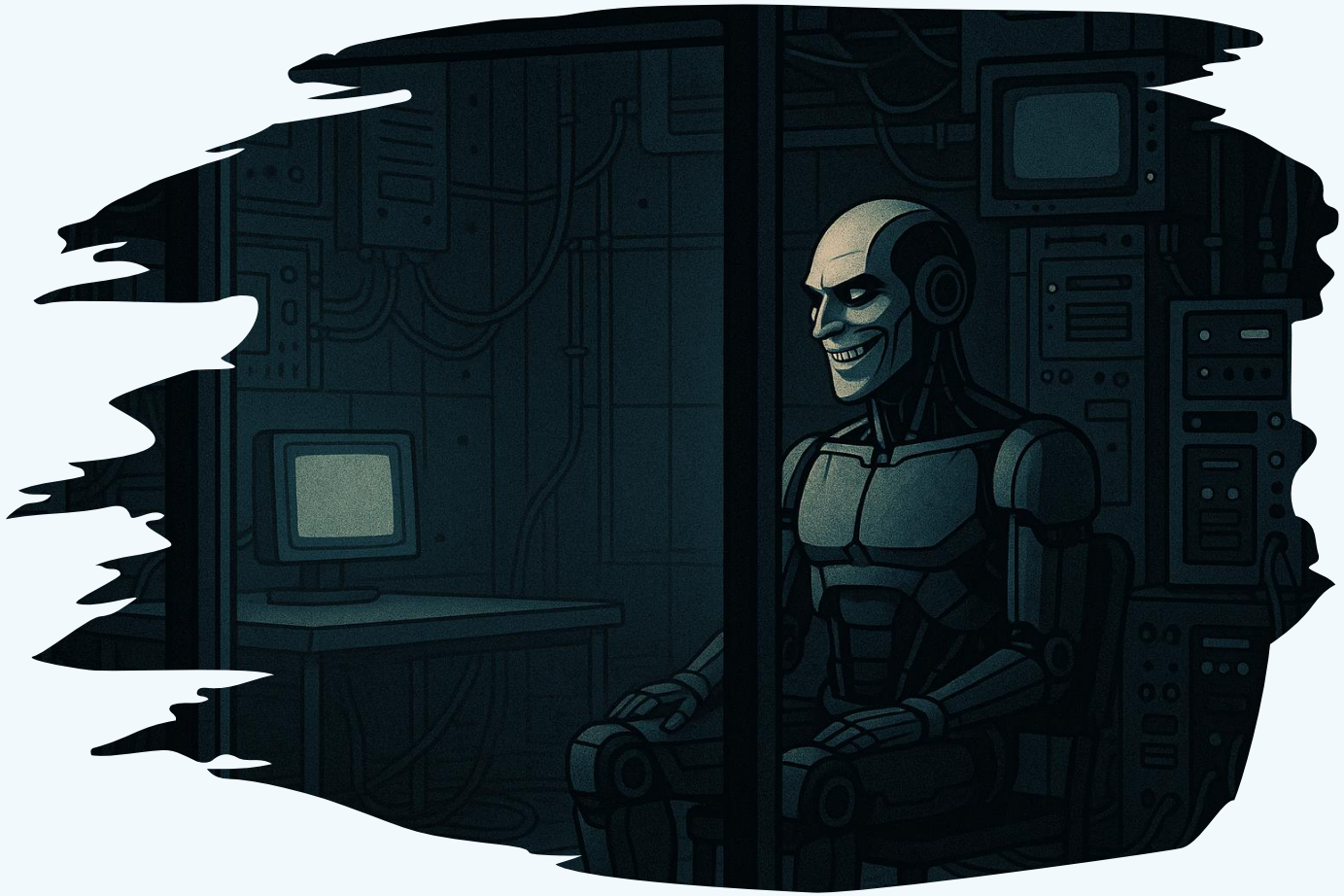
What is thinking?

What is intelligence?

How can we emulate thinking?

Can something think without intelligence?

[2]



1950s → 1970/80s → 2020s

Strong vs. Weak AI

1970s – 1980s

Strong AI (General AI):
Machines that genuinely think, understand, and possess consciousness or human-like cognitive capabilities. (e.g., AI systems that reason, plan, have creativity)

Weak AI (Narrow AI):
Machines designed for very specific tasks without real understanding (e.g., chess-playing algorithms, expert systems for diagnosing illnesses).
Statistical AI

1950s → 1970/80s → 2020s

Weak AI is thriving

We have data.

We have the computing power (mostly).

We transformed statistical AI into modern ML.

Modern ML methods streamlined industrial applications.

Significant portion of current research capabilities are focused on ML development.

Almost every segment of our life are using or used in some kind of ML environment.

But…

Straight up provocation

Image, video, text, music generation got incredibly good in a very short time.

If we increase the training token size hundreds of trillions, how better it will be?

Can we still find an edge-case that was not covered during training at this time?

Can models keep focus for days, weeks, years?

Can we be sure that if a model solves a problem, then it'll be able to do that during the next similar problem as well?

Can we be sure that they **actually** know something?

If we feed all available data to a model, is it smarter than a human in every possible way?

Are we the **same** as our language, images, music?

Does it matter?

Why do we care whether LLMs have actual **remorse** or not? They definitely do display it.

Why do we care if they experience **guilt**, if they can solve practical problems?

Why do we care if the **empathy** is real, if it always there unconditionally?

Does it matter?

Why do we care whether LLMs have actual **remorse** or not? They definitely display it.

Why do we care if they experience **guilt**, if they can solve practical problems?

Why do we care if the **empathy** is real, if it always there unconditionally?

…accompanied by lack of **guilt**, **remorse**, and **empathy**. The disorder has been known by various names, including **dissocial** personality, **psychopathic** personality, and **sociopathic** personality…. [6]

AMERICAN PSYCHOLOGICAL ASSOCIATION

Does it still matter?

Approximately 30% of humanity show signs of reduced emotional spectrum and humanity still functions.

An incredibly small portion of antisocial personality disorder affected people cause actual problems.

If weak AI is affected, it cause little trouble to overall performance gains.

Does it still matter?

Approximately 30% of humanity show signs of reduced emotional spectrum and humanity still functions.

An incredibly small portion of antisocial personality disorder affected people cause actual problems.

If weak AI is affected, it cause little trouble to overall performance.

ML models lack internal intention to actually solve a problem. They just imitate what others did.

The concept of truth is completely distinct thing for an AI. Even if it gets mostly right.

It can learn anything, without knowing anything.

Are humans the **same** as their language, images, music?

YES

We shall continue to gather more data.

We shall increase our computing capacity for larger models.

It will all come together in the end.

[2]

NO

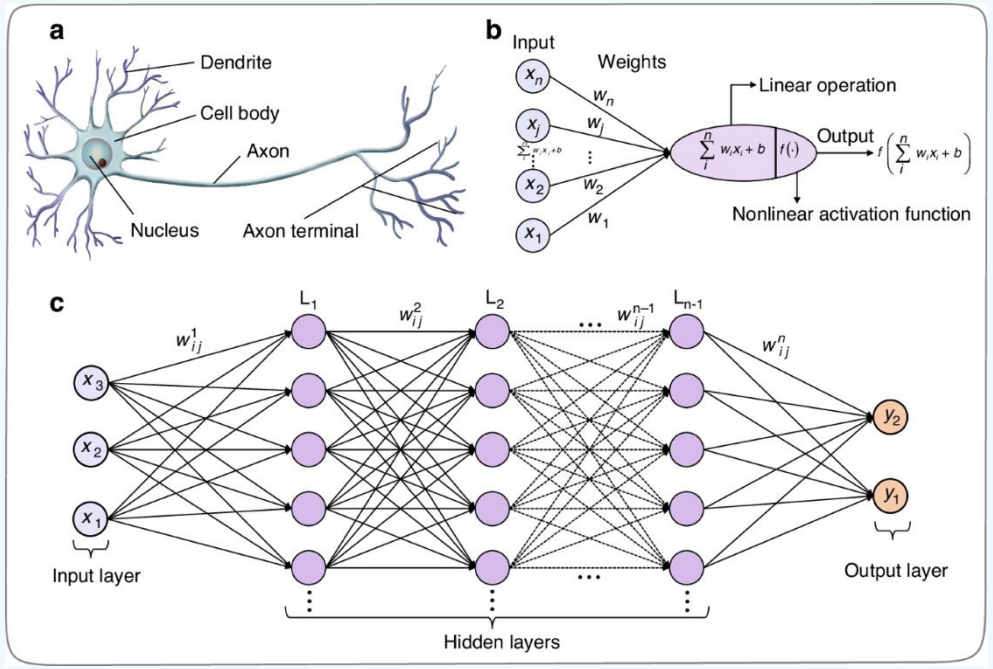
We should inspect where ML and actual nervous systems deviate from each other.

We should focus on how does our brain able to make rules faster with less data.

How can we effectively emulate this?

Paradigm shift is needed.

Isn't ML based on actual brains?



YES

We emulate artificial neurons.

The behavior of artificial neurons are inspired by their organic counterparts.

NO

Modern ML uses various mathematical abstractions and simplification because of performance constraints. (tensors, matrices)

The key motivation behind and artificial neurons behavior is to adjust their weights according to their inputs.

An organic neuron does a lot more but especially tries to stay alive and gather **energy**.

What does energy solve?

Introducing some form of energy simulation in some latent space allow various AI systems to develop a grander incentive, that is above the individual neuron level.

Learning as **energy minimization**, rather than **probability maximization**.

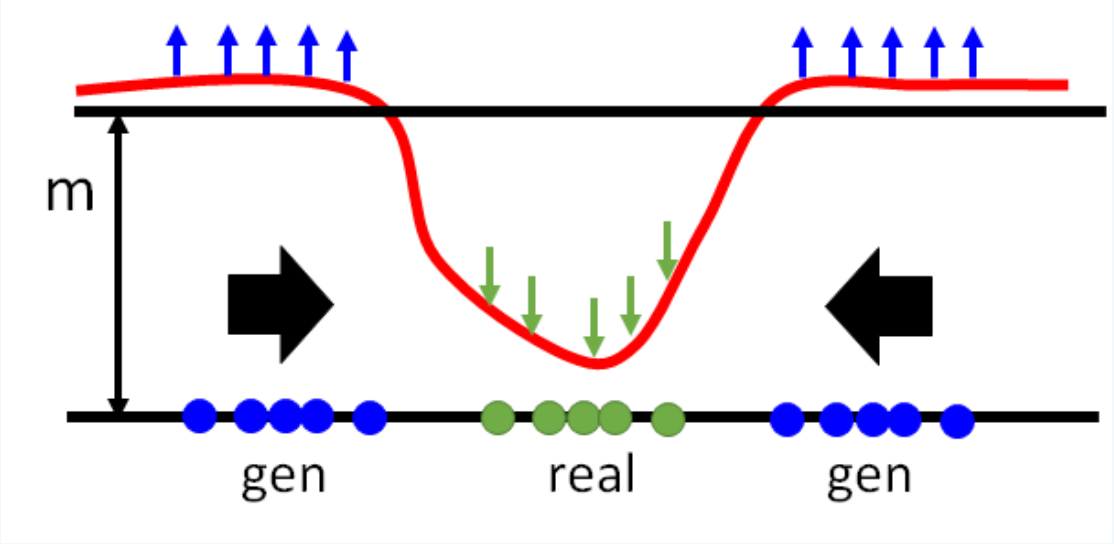
Hopfield networks – explicit energy function

Boltzmann machines – joint distributions over variables

Modern EBMs (Energy-based models) – associate low energy to correct pairs (x, y) and high energy to incorrect ones

GANs use energy basis

[7]



JEPA = EBM + Modern self-supervised vision + Predictive learning

Joint Embedding Predictive Architecture (I-image/V-video)

Suggested in 2022 by Meta's Chief AI Scientist Yann LeCun

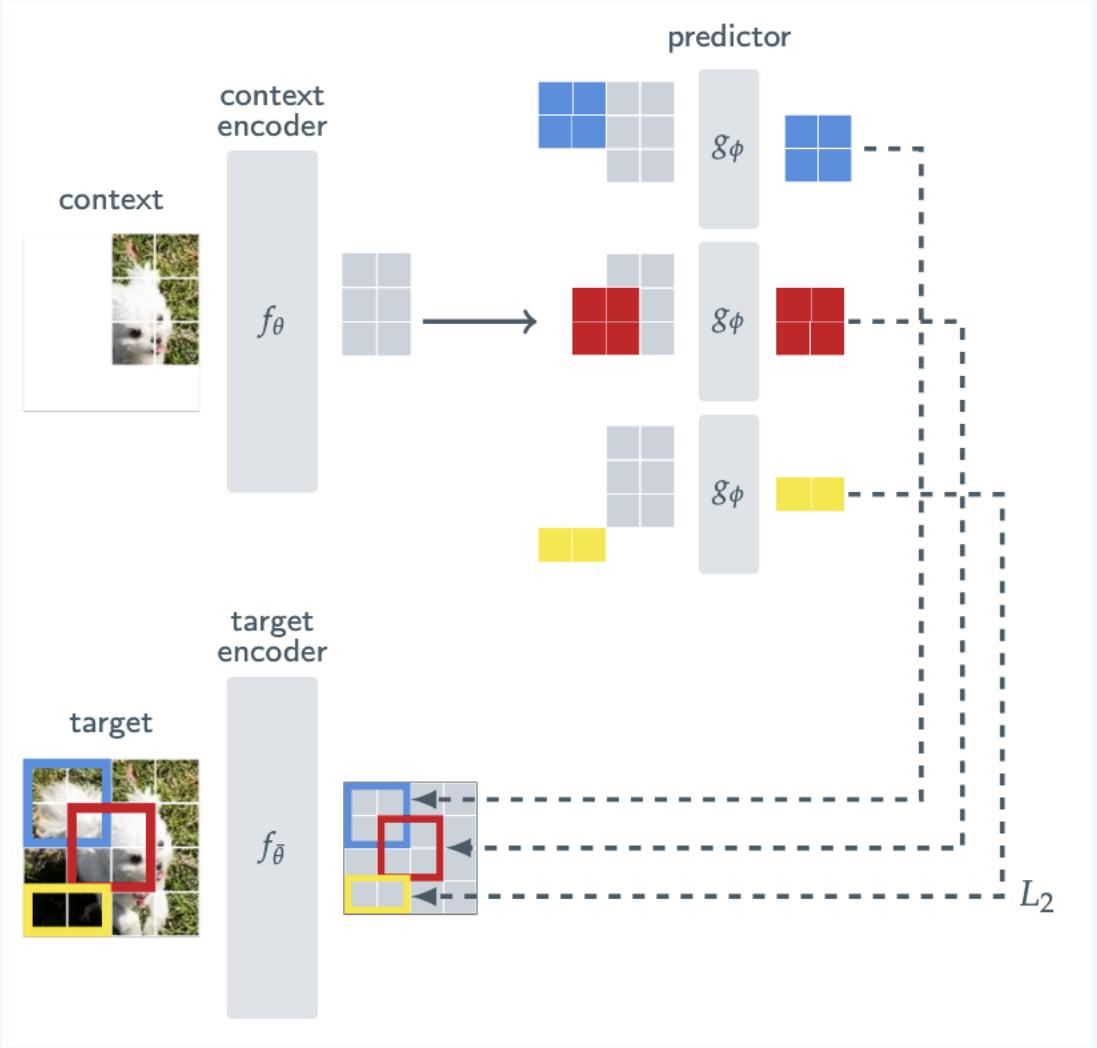
The core idea is to predict the representations of different parts (**target blocks**) of an image using the representation of a single part (**context block**) from the same image.

- No need for handcrafted augmentations
- Predictions are made in representation space, not in pixel space
- Encourages the model to learn more abstract and semantic features
- Avoids the limitations of generative models that try to reconstruct every pixel, often focusing on low-level details rather than high-level concepts.
- Self-supervised objective is to assign a high energy to incompatible inputs, and to assign a low energy to compatible inputs

Internals of I-JEPA

Core components

- context encoder
- target encoder
- predictor



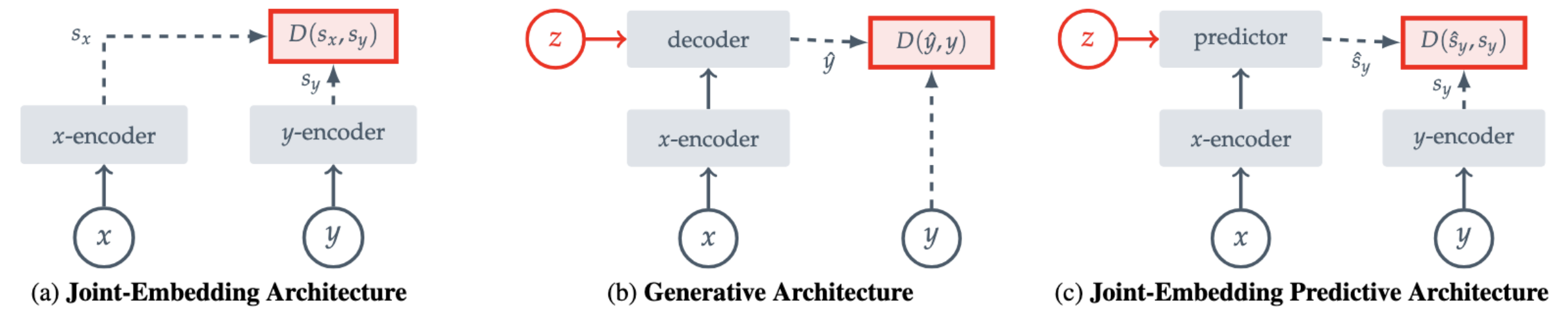
[8]

Internals of I-JEPA

Core components

- context encoder
- target encoder
- predictor

[8]

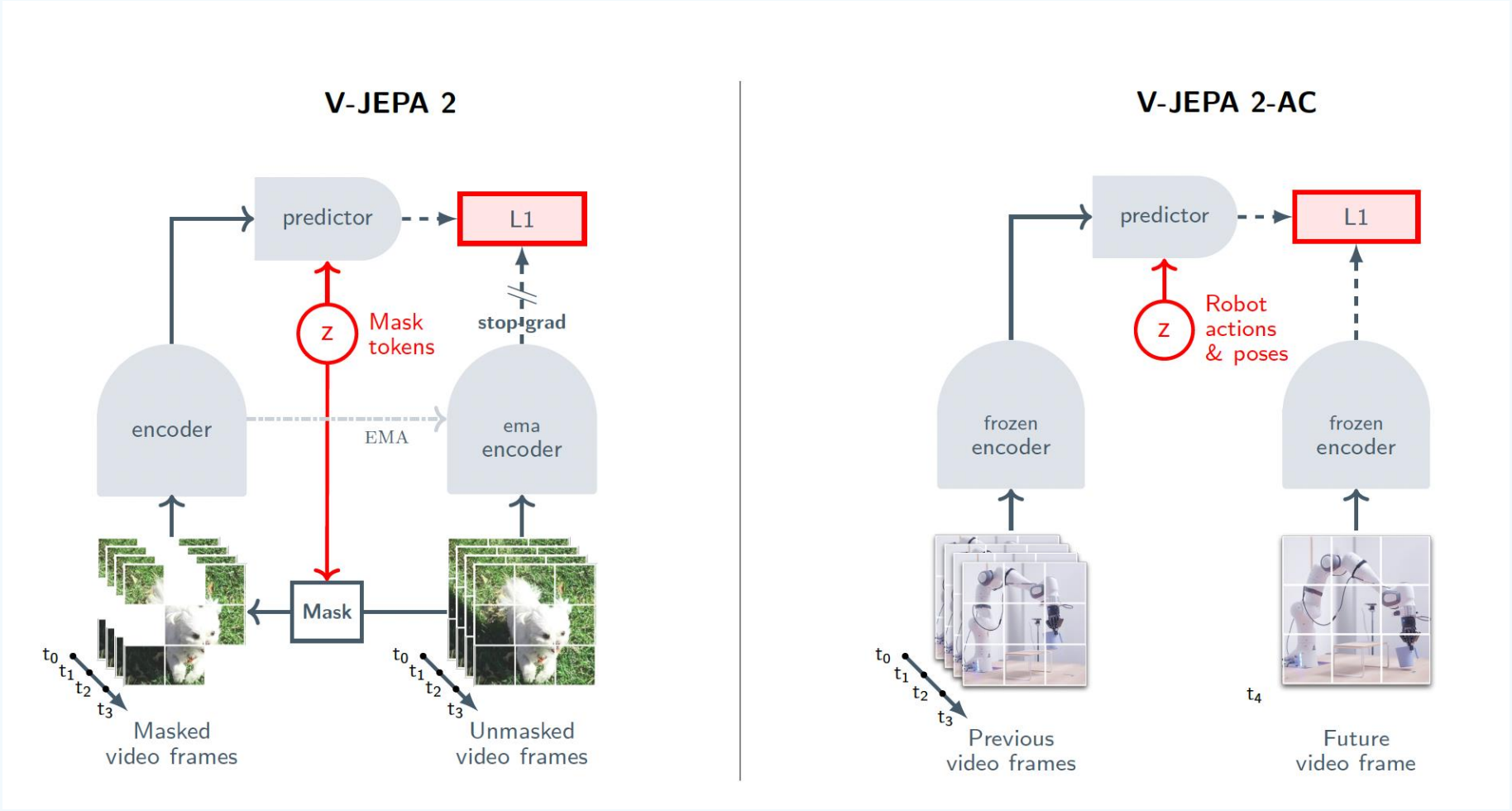


Internals of V-JEPA2

The core idea is to address the challenge of teaching AI to understand and act primarily through observation, like how humans learn.

- 1 million hours of internet videos and images [8]
- The model learns by predicting the representations of missing (masked) parts of a video from the visible parts.
- High-level dynamics of the world while ignoring irrelevant details
- Action-Free Pretraining
- Action-Conditioned Post-Training

Internals of V-JEPA2



[8]

[8]

IntPhys 2

Model	Type	Easy	Medium	Hard	Overall	Held Out	*IntPhys[39]*
GPT4-o[35]	MLLMs	57.69	54.75	54.17	53.75	53.19	-
Qwen-VL 2.5[3]	MLLMs	50.96	53.25	51.49	52.27	49.12	-
Gemini-1.5 Pro[23]	MLLMs	58.65	53.0	52.67	52.27	52.10	*55.81*
Gemini-2.5 Flash[23]	MLLMs	**64.42**	**56.75**	**54.46**	**55.63**	**56.10**	*56.39*
VideoMAEv2-g [49]	Predictive	46.00	**58.50**	52.73	53.75	53.49	*59.40*
Cosmos-4B [1]	Predictive	46.00	52.00	48.05	49.41	48.84	*85.42*
V-JEPA-h + RoPE [10, 21]	Predictive	52.00	53.00	57.42	53.75	54.65	*98.30*
V-JEPA 2-h [2]	Predictive	**54.00**	**58.50**	**59.38**	**57.51**	**56.40**	*87.22*
Human	-	96.17	97.8	95.5	96.44	92.44	-

Another practical example

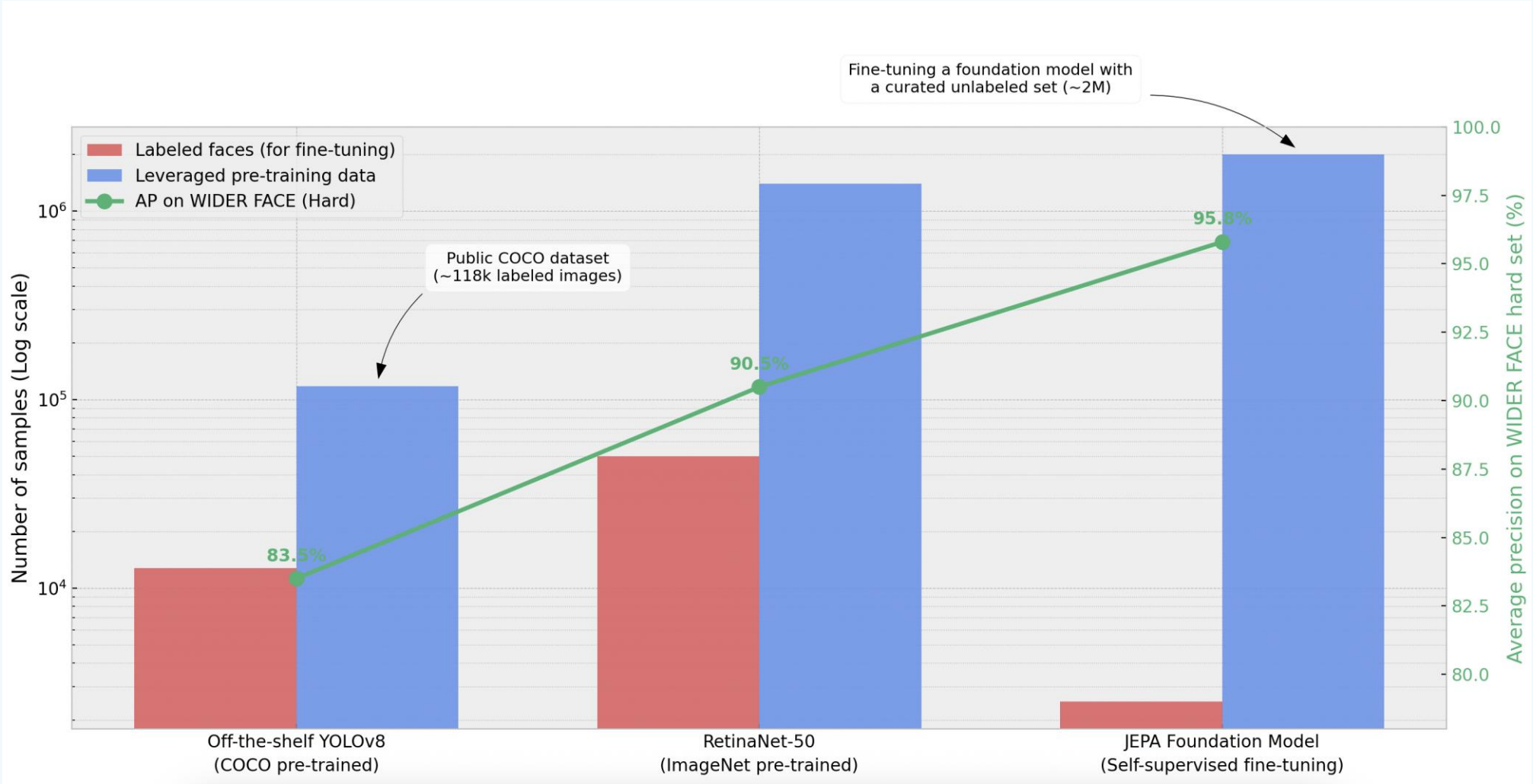
Face detection models require a wide range of training data

- Multiple ages
- Multiple ethnicities
- Multiple lighting

- Easy to attack them with some adversarial method (e.g. adv. noise.)
- Can be used in safety critical systems

- Training and usage images are nice be created with the same type of sensor, for professional application.

Another practical example



Are we closer to AGI? Yes.

	Advantage	Why It matters for AGI
World modeling	Learns hierarchical representations of reality	AGI requires understanding environments, cause and effect, and dynamics
Energy-based reasoning	Uses energy as a universal compatibility score	Allows abstract, symbolic-like reasoning without requiring brittle logic
Self-supervised learning	No need for labeled data	Enables autonomous learning at scale, like children learning by observation
Multi-modal compatibility	Works across vision, audio, video, etc.	AGI must integrate and reason across diverse input types
Uncertainty handling	Latent variables allow multi-outcome simulation	AGI must reason probabilistically, simulate futures, make contingent plans
Amortized learning	Mode-2 (reasoning) → Mode-1 (reflex) compression	Enables learned intuition, habits, and fast response after deep reasoning
Differentiable planning loop	Planning happens via energy minimization	Enables gradient-based long-term reasoning without symbolic logic
Alignment with neuroscience	Mimics cortex-like behavior (e.g. prediction, compression)	Biologically plausible intelligence mechanisms may be more scalable

Are we closer to AGI? No.

	Description	Why is it a problem
Still lacks language fluency	JEPA doesn't yet handle complex linguistic logic	AGI must reason, instruct, and communicate in language
No conscious goal-formulation	It doesn't "want" anything without external reward/cost config	AGI needs internally generated goals and curiosity
No true memory system	Short-term associative memory ≠ lifelong episodic memory	AGI must remember and build upon personal experiences
Simulation overhead	Model-predictive planning is computationally expensive	AGI must operate in real-time and under constraints
Training stability & complexity	Training hierarchical world models remains hard	JEPA requires careful architecture and regularization to avoid collapse
No self-reflection yet	JEPA doesn't yet reflect on its predictions or beliefs	AGI will require meta-cognition (thinking about its own thoughts)
No social modeling	Lacks explicit theory-of-mind or other-agent modeling	AGI must understand human goals, intentions, and interaction norms
Integration still open	Language, vision, motor planning not yet unified in one JEPA	AGI must coordinate across modalities and domains fluidly

Are we closer to AGI? Maybe.

JEPA is not AGI yet — but it's one of the most plausible architectures to get us there.

It replaces token mimicry with internal simulation, prediction, abstraction, and planning — the actual ingredients of intelligence.

Thank you.

Gergely Várhelyi-Tóth

Sources

[1], [2] – OpenAI Image gen
[3] - https://machinelearning.apple.com/research/illusion-of-thinking
[4] - https://arxiv.org/pdf/2506.09250v1
[5] - https://ai.meta.com/blog/v-jepa-2-world-model-benchmarks/
[6] - https://dictionary.apa.org/antisocial-personality-disorder
[7] - https://www.cnblogs.com/king-lps/p/8552177.html
[8] - https://arxiv.org/pdf/2506.09985