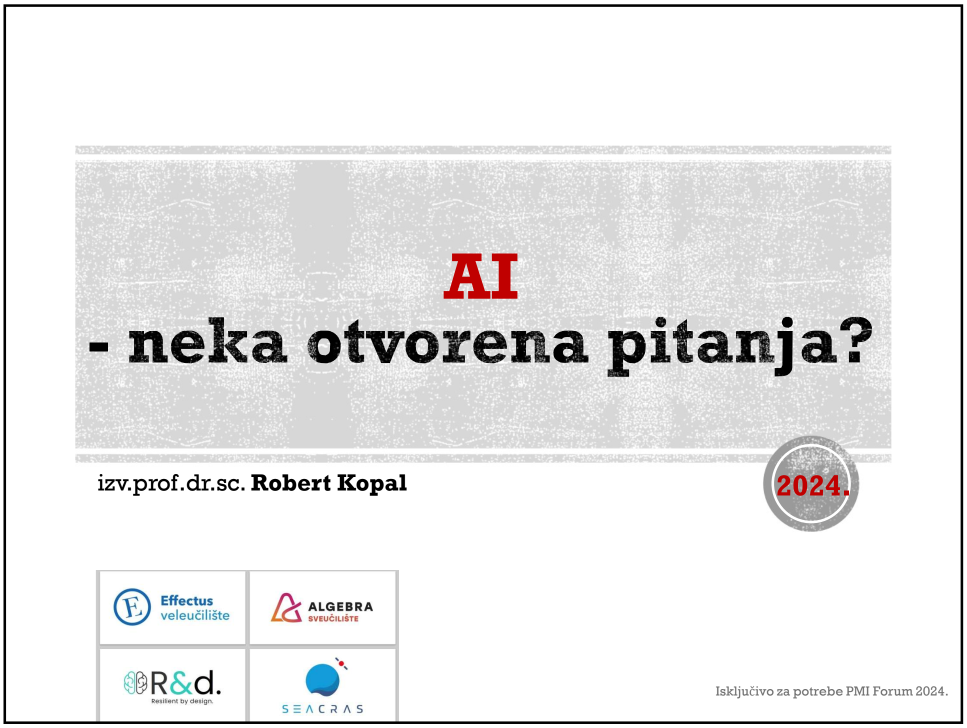


AI
- neka otvorena pitanja?

izv.prof.dr.sc. **Robert Kopal**

2024.

Effectus veleučilište

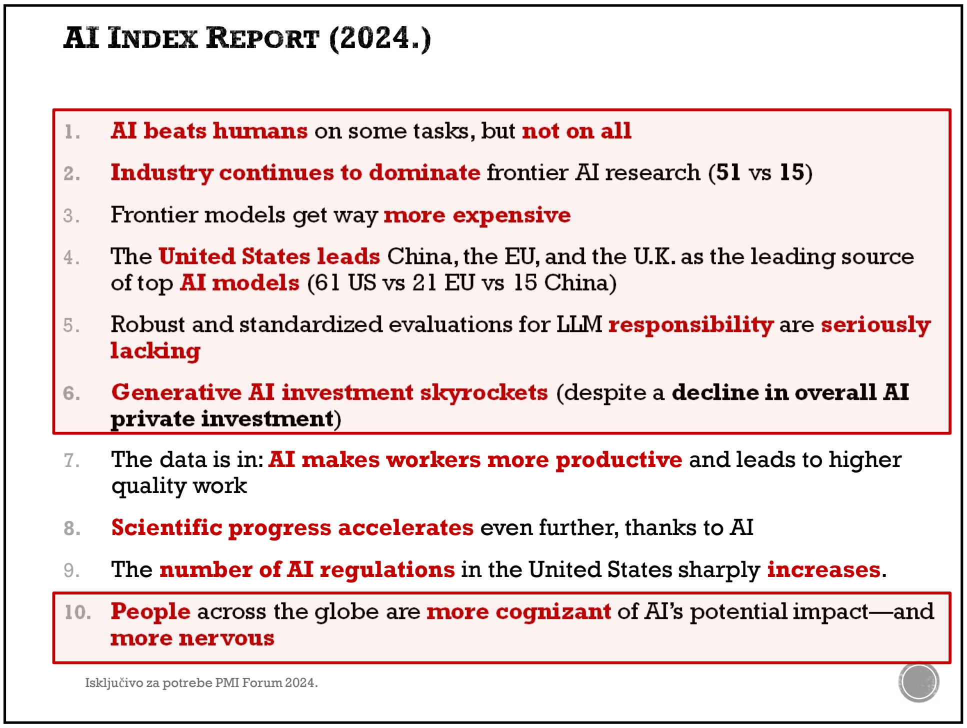
ALGEBRA SVEUČILIŠTE

R&d. Resilient by design.

SEACRAS

Isključivo za potrebe PMI Forum 2024.

1



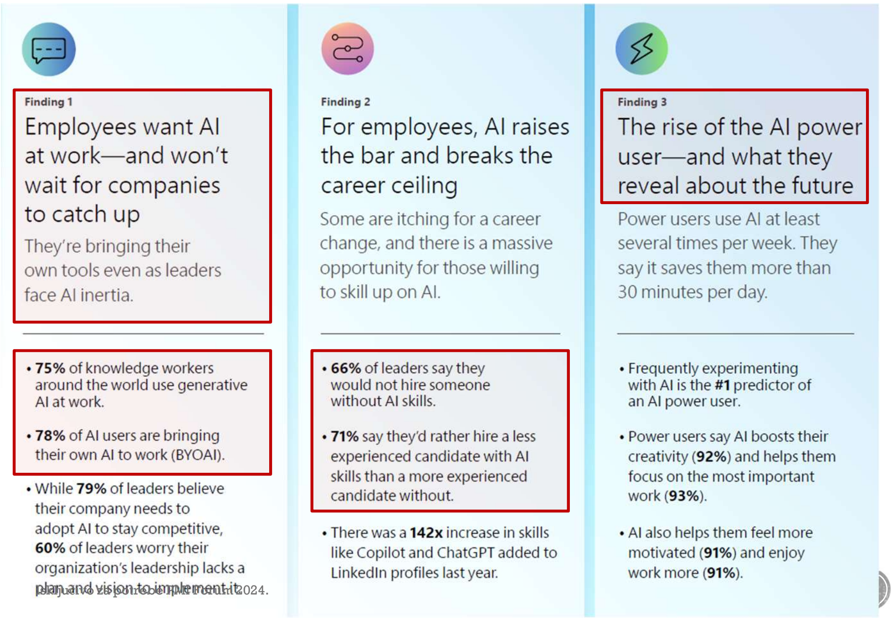
AI Index Report (2024.)

1. **AI beats humans** on some tasks, but **not on all**
2. **Industry continues to dominate** frontier AI research (**51** vs **15**)
3. Frontier models get way **more expensive**
4. The **United States leads** China, the EU, and the U.K. as the leading source of top **AI models** (61 US vs 21 EU vs 15 China)
5. Robust and standardized evaluations for LLM **responsibility** are **seriously lacking**
6. **Generative AI investment skyrockets** (despite a **decline in overall AI private investment**)
7. The data is in: **AI makes workers more productive** and leads to higher quality work
8. **Scientific progress accelerates** even further, thanks to AI
9. The **number of AI regulations** in the United States sharply **increases**.
10. **People** across the globe are **more cognizant** of AI's potential impact—and **more nervous**

Isključivo za potrebe PMI Forum 2024.

2

AI at Work Is Here. Now Comes the Hard Part (2024.)



Finding 1

Employees want AI at work—and won't wait for companies to catch up

They're bringing their own tools even as leaders face AI inertia.

- **75%** of knowledge workers around the world use generative AI at work.
- **78%** of AI users are bringing their own AI to work (BYOAI).
- While **79%** of leaders believe their company needs to adopt AI to stay competitive, **60%** of leaders worry their organization's leadership lacks a plan and vision to implement it.

Finding 2

For employees, AI raises the bar and breaks the career ceiling

Some are itching for a career change, and there is a massive opportunity for those willing to skill up on AI.

- **66%** of leaders say they would not hire someone without AI skills.
- **71%** say they'd rather hire a less experienced candidate with AI skills than a more experienced candidate without.
- There was a **142x** increase in skills like Copilot and ChatGPT added to LinkedIn profiles last year.

Finding 3

The rise of the AI power user—and what they reveal about the future

Power users use AI at least several times per week. They say it saves them more than 30 minutes per day.

- Frequently experimenting with AI is the **#1** predictor of an AI power user.
- Power users say AI boosts their creativity (**92%**) and helps them focus on the most important work (**93%**).
- AI also helps them feel more motivated (**91%**) and enjoy work more (**91%**).

Beyond AI Exposure: Which Tasks are Cost-Effective to Automate with Computer Vision? (2024.)

- The faster AI automation spreads through the economy, the more profound its **potential impacts**, **both positive** (**improved productivity**) and **negative** (**worker displacement**).
- **The previous literature on "AI Exposure"** **cannot predict** **this pace of automation since** **it attempts to measure an overall potential for AI to affect an area**, not the technical feasibility and economic attractiveness of building such systems.
- We present a new type of AI task automation model that is end-to-end, estimating: **the level of technical performance needed to do a task**, the characteristics of an AI system capable of that performance, and the economic choice of whether to build and deploy such a system.
- The result is a first estimate of which tasks are **(1) technically feasible** and **(2) economically attractive** to automate - and which are not.
- We focus on **computer vision**, where cost modeling is more developed.

Beyond AI Exposure: Which Tasks are Cost-Effective to Automate with Computer Vision? (2024.)

- We find that at today's costs **U.S. businesses would choose not to automate most vision tasks** that have "AI Exposure," and that only **23%** of worker wages being paid for vision tasks would be attractive to automate.
- This slower roll-out of **AI can be accelerated if costs falls rapidly or if it is deployed via AI-as-a-service platforms** that have greater scale than individual firms, both of which we quantify.
- Overall, our findings suggest that **AI job displacement will be substantial, but also gradual** – and therefore there is room for policy and retraining to mitigate unemployment impacts.

Isključivo za potrebe PMI Forum 2024.

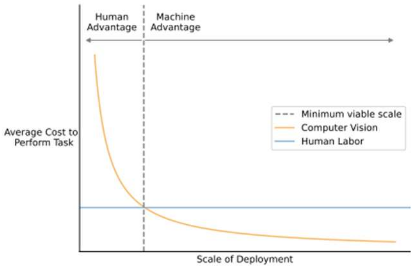


Figure 1 The minimum viable scale for AI deployment.

5

AI Will Transform the Global Economy. Let's Make Sure It Benefits Humanity - IMF (2024.)

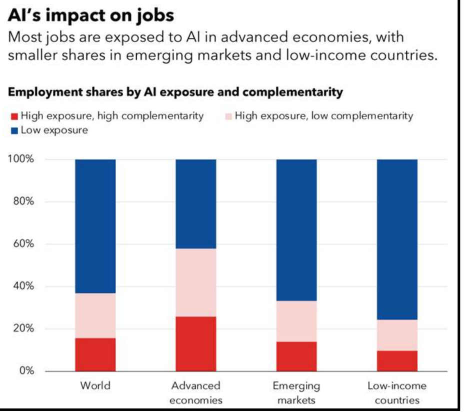
- In a new analysis, **IMF** staff examine the potential **impact of AI** on the **global labor market**.
- The findings are striking: **almost 40% of global employment is exposed to AI**.
- Historically, automation and information technology have tended to affect routine tasks, **but one of the things that sets AI apart is its ability to impact high-skilled jobs**.
- As a result, **advanced economies face greater risks from AI**—but **also more opportunities to leverage its benefits**—compared with emerging market and developing economies.
- In **advanced economies**, about **60%** of jobs may be impacted by AI.
- **Roughly half the exposed jobs may benefit from AI integration, enhancing productivity**.
- For the other half, **AI applications may execute key tasks currently performed by humans**, which could lower labor demand, leading to lower wages and reduced hiring. In the most extreme cases, some of these jobs may disappear.

Isključivo za potrebe PMI Forum 2024.

6

AI Will Transform the Global Economy. Let's Make Sure It Benefits Humanity - IMF (2024.)

- In **emerging markets and low-income countries**, by contrast, AI exposure is expected to be **40%** and **26%**, respectively.
- These findings suggest emerging market and developing economies face fewer immediate disruptions from AI.
- At the same time, **many of these countries don't have the infrastructure or skilled workforces to harness the benefits of AI**, raising the risk that **over time the technology could worsen inequality among nations**.



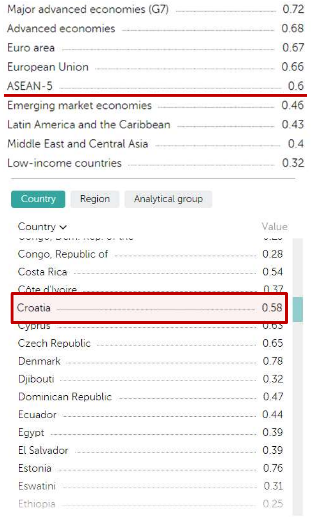
AI's impact on jobs

Most jobs are exposed to AI in advanced economies, with smaller shares in emerging markets and low-income countries.

Isključivo za potrebe PMI Forum 2024.

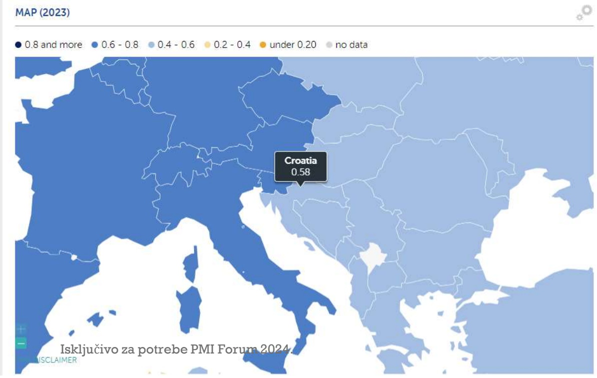
AI Preparedness Index (AIPI) - IMF (2023.)

- **AIPI** is the sum of **4 key dimensions**: (1) **digital infrastructure**, (2) **human capital**, (3) **technological innovation**, and (4) **legal frameworks**.



Major advanced economies (G7)	0.72
Advanced economies	0.68
Euro area	0.67
European Union	0.66
ASEAN-5	0.6
Emerging market economies	0.46
Latin America and the Caribbean	0.43
Middle East and Central Asia	0.4
Low-income countries	0.32

Country	Value
Congo, Dem. Rep. of the	[illegible]
Congo, Republic of	0.28
Costa Rica	0.54
Côte d'Ivoire	0.37
Croatia	0.58
Cyprus	0.63
Czech Republic	0.65
Denmark	0.78
Djibouti	0.32
Dominican Republic	0.47
Ecuador	0.44
Egypt	0.39
El Salvador	0.39
Estonia	0.76
Eswatini	0.31
Ethiopia	0.25



Isključivo za potrebe PMI Forum 2024.

Challenges ?

1. Business leaders increasingly say that **graduates are qualified in theory but not in practice**:
   They need an average of 11 months of on-the-job training before they become fully effective in their role
2. Indeed, **47% of workers have done no workplace training in the last 5 years.**

Risky Business: Identifying Blind Spots in Corporate Oversight of AI?

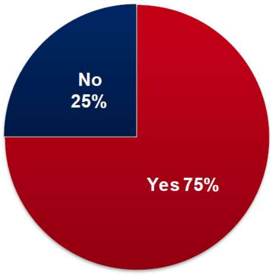
- The **Baker McKenzie survey** (**January 2022**.), queried **500 US based**, **C level executives** who self identified as **part of the decision making team responsible for their organization's adoption, use and management of AI enabled tools**.
- The telephone and email based survey was conducted among **executives at companies with at least $10.3 billion in annual revenues on average, across a range of industries**.
- Margin of error for this survey is +/ 4%.

Risky Business: Identifying Blind Spots in Corporate Oversight of AI?

- AI in use: **75% of companies** use **AI tools** and **tech** for **hiring** and **HR**.

Most Companies Use AI for Hiring and People Management Functions

Does your organization use AI at an enterprise level for anything related to HR or employment?



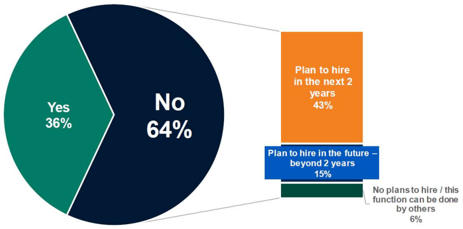
11

Risky Business: Identifying Blind Spots in Corporate Oversight of AI?

- **Who’s Got Their Eyes on AI?**

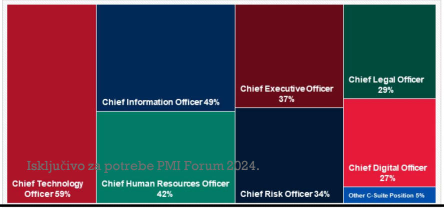
Companies Lack C-Level Experts on AI

Does your company have a dedicated Chief Artificial Intelligence Officer (CAIO) in place?



Absent a CAIO, the CTO is Often Charged with Enterprise Level AI Oversight

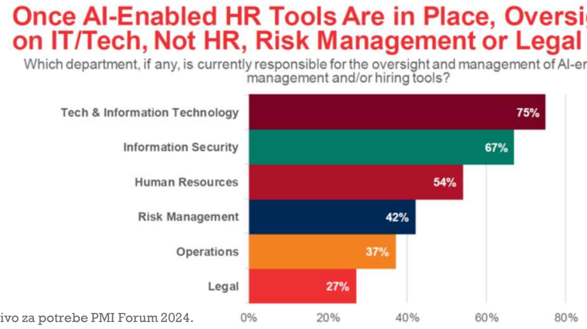
In your organization, are any of the following specifically tasked with overseeing the use of AI-enabled tools?



12

Risky Business: Identifying Blind Spots in Corporate Oversight of AI?

- **After…**
- Once AI enabled HR tools are **in place**, the job of **oversight** falls to the **IT/Tech** department, **not HR**.
- **Legal** is the department that is **least likely to be tapped to manage or oversee enterprise AI risk**.



Once AI-Enabled HR Tools Are in Place, Oversight Is on IT/Tech, Not HR, Risk Management or Legal

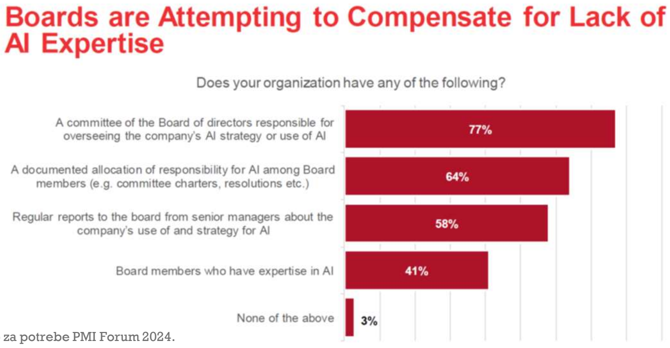
Which department, if any, is currently responsible for the oversight and management of AI-enabled people management and/or hiring tools?

Isključivo za potrebe PMI Forum 2024.

13

Risky Business: Identifying Blind Spots in Corporate Oversight of AI?

- **After…**
- **Most Boards (77%) oversee their enterprise's AI strategy** by **committee**.
- **Just 41% of corporate Boards have an expert in AI on them.**



Boards are Attempting to Compensate for Lack of AI Expertise

Does your organization have any of the following?

Isključivo za potrebe PMI Forum 2024.

14

AI DEFINITION (OLD) - EU

- ‘**Artificial intelligence system**’ (AI system) means software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a **given set of human-defined objectives**, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with;
- …
- ‘**Artificial intelligence system**’ (AI system) means a system that is designed to operate with a certain level of autonomy and that, based on machine and/or **human-provided data** and **inputs**, infers how to achieve a given set of human-defined objectives using machine learning and/or logic- and knowledge based approaches, and produces system-generated outputs such as content (generative AI systems), predictions, recommendations or decisions, influencing the environments with which the AI system interacts.

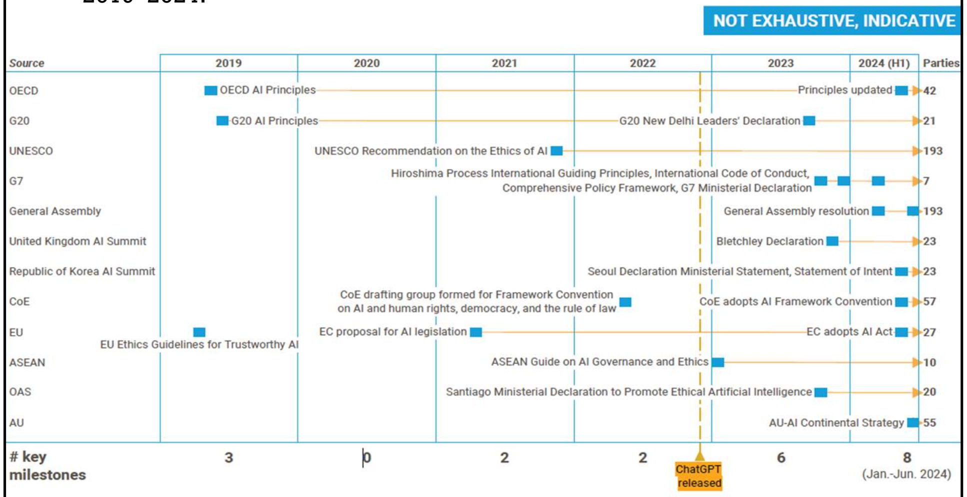
AI DEFINITION (NEW) - EU

1. ‘**AI system**’ means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, **infers, from the input it receives**, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments;
2. ‘**risk**’ means the combination of the **probability** of an occurrence of harm and the **severity** of that harm;

> “a machine-based system designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.”

Governing AI for Humanity (9/2024.)

- **Interregional** and **regional AI governance initiatives**, key milestones, 2019–2024.



Isključivo za potrebe PMI Forum 2024.

INTERNATIONAL AI GOVERNANCE

12 landmark moments for international AI governance

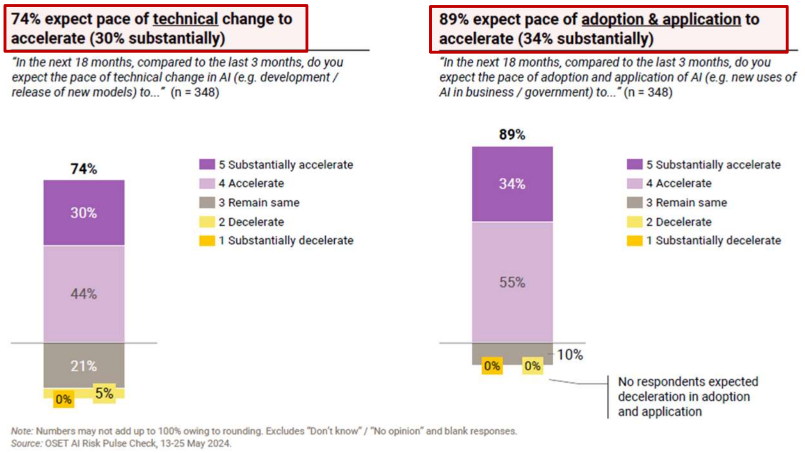
Global AI & UN

- **UN…**
- At this year's UN General Assembly, world leaders discussed the importance of inclusive global governance for Artificial Intelligence (AI), which presents powerful opportunities for humanity.
- "…This is reiterated in the UN Secretary-General's statement, in which **he noted that a new UN agency may be required to help the world manage it**…"
- **The UN Secretary-General's AI Advisory Body**

Isključivo za potrebe PMI Forum 2024.

GOVERNING AI FOR HUMANITY (9/2024.)

- **Experts' expectations** regarding **AI technological development**



74% expect pace of technical change to accelerate (30% substantially)

"In the next 18 months, compared to the last 3 months, do you expect the pace of technical change in AI (e.g. development / release of new models) to..." (n = 348)

89% expect pace of adoption & application to accelerate (34% substantially)

"In the next 18 months, compared to the last 3 months, do you expect the pace of adoption and application of AI (e.g. new uses of AI in business / government) to..." (n = 348)

No respondents expected deceleration in adoption and application

Note: Numbers may not add up to 100% owing to rounding. Excludes "Don't know" / "No opinion" and blank responses.
Source: OSET AI Risk Pulse Check, 13-25 May 2024.

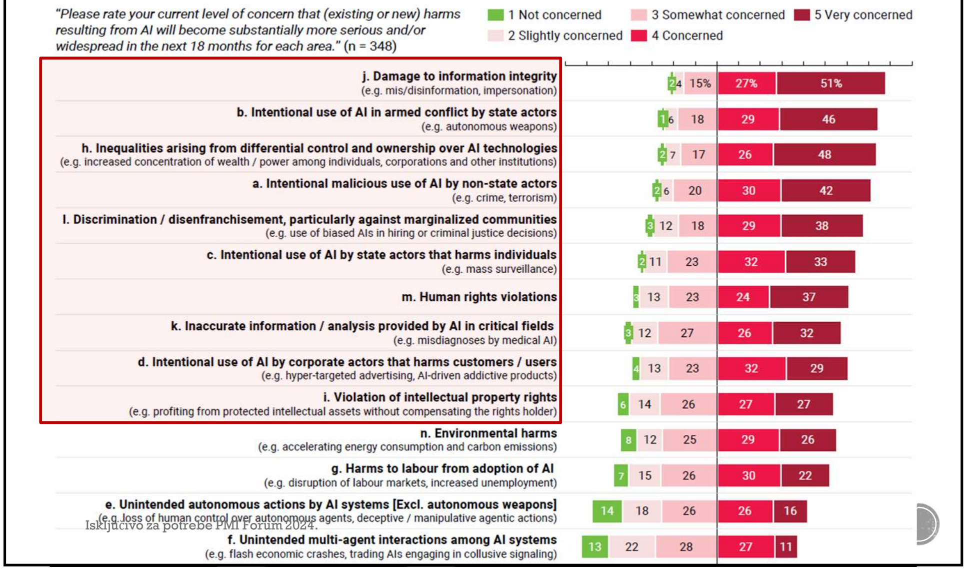
Isključivo za potrebe PMI Forum 2024.

19

GOVERNING AI FOR HUMANITY (9/2024.)

- **Experts' levels of concern** about **AI risks** across multiple domains

"Please rate your current level of concern that (existing or new) harms resulting from AI will become substantially more serious and/or widespread in the next 18 months for each area." (n = 348)



	🟩 1 Not concerned	2 Slightly concerned	3 Somewhat concerned	4 Concerned	5 Very concerned
j. Damage to information integrity (e.g. mis/disinformation, impersonation)	2	4	15%	27%	51%
b. Intentional use of AI in armed conflict by state actors (e.g. autonomous weapons)	1	6	18	29	46
h. Inequalities arising from differential control and ownership over AI technologies (e.g. increased concentration of wealth / power among individuals, corporations and other institutions)	2	7	17	26	48
a. Intentional malicious use of AI by non-state actors (e.g. crime, terrorism)	2	6	20	30	42
l. Discrimination / disenfranchisement, particularly against marginalized communities (e.g. use of biased AIs in hiring or criminal justice decisions)	3	12	18	29	38
c. Intentional use of AI by state actors that harms individuals (e.g. mass surveillance)	2	11	23	32	33
m. Human rights violations	3	13	23	24	37
k. Inaccurate information / analysis provided by AI in critical fields (e.g. misdiagnoses by medical AI)	3	12	27	26	32
d. Intentional use of AI by corporate actors that harms customers / users (e.g. hyper-targeted advertising, AI-driven addictive products)	4	13	23	32	29
i. Violation of intellectual property rights (e.g. profiting from protected intellectual assets without compensating the rights holder)	6	14	26	27	27
n. Environmental harms (e.g. accelerating energy consumption and carbon emissions)	8	12	25	29	26
g. Harms to labour from adoption of AI (e.g. disruption of labour markets, increased unemployment)	7	15	26	30	22
e. Unintended autonomous actions by AI systems [Excl. autonomous weapons] (e.g. loss of human control over autonomous agents, deceptive / manipulative agentic actions)	14	18	26	26	16
f. Unintended multi-agent interactions among AI systems (e.g. flash economic crashes, trading AIs engaging in collusive signaling)	13	22	28	27	11

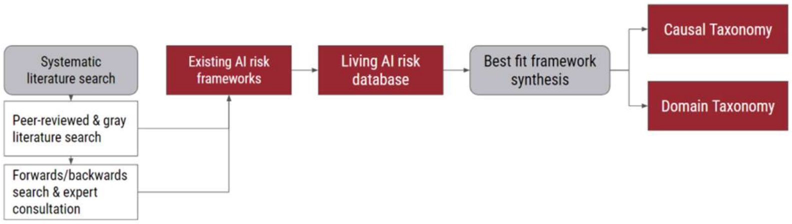
Isključivo za potrebe PMI Forum 2024.

20

What is the AI Risk Repository?

- The **AI Risk Repository** has 3 parts:

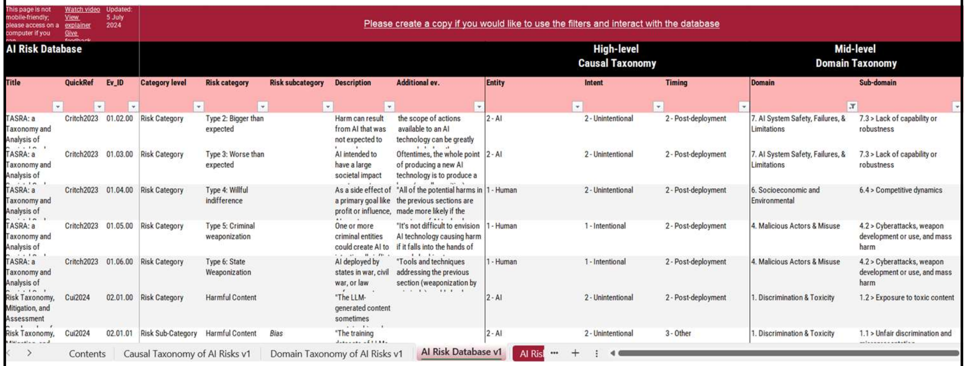
1. The **AI Risk Database captures 700+ risks (777)** extracted from **43** existing frameworks, with quotes and page numbers.
2. The **Causal Taxonomy of AI Risks** classifies **how**, **when**, and **why** these risks occur.
3. The **Domain Taxonomy of AI Risks** classifies these risks into **7 domains** (e.g., “Misinformation”) and **23 subdomains** (e.g., “False or misleading information”).


Systematic literature search
Peer-reviewed & gray literature search
Forwards/backwards search & expert consultation
Existing AI risk frameworks
Living AI risk database
Best fit framework synthesis
Causal Taxonomy
Domain Taxonomy

21

What is the AI Risk Repository?

- The **AI Risk Database** links **each risk** to the (1) **source** information (paper title, authors), (2) **supporting evidence** (quotes, page numbers), and to our (3) **Causal** and (4) **Domain Taxonomies**.



This page is not mobile-friendly; please access on a computer if you can

Watch video
View explainer
Give feedback

Updated: 5 July 2024

Please create a copy if you would like to use the filters and interact with the database

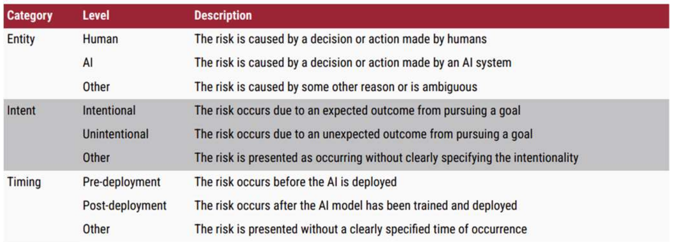
AI Risk Database								High-level Causal Taxonomy			Mid-level Domain Taxonomy	
Title	QuickRef	Ev_ID	Category level	Risk category	Risk subcategory	Description	Additional ev.	Entity	Intent	Timing	Domain	Sub-domain
TASRA: a Taxonomy and Analysis of	Critch2023	01.02.00	Risk Category	Type 2: Bigger than expected		Harm can result from AI that was not expected to	the scope of actions available to an AI technology can be greatly	2 - AI	2 - Unintentional	2 - Post-deployment	7. AI System Safety, Failures, & Limitations	7.3 > Lack of capability or robustness
TASRA: a Taxonomy and Analysis of	Critch2023	01.03.00	Risk Category	Type 3: Worse than expected		AI intended to have a large societal impact	Oftentimes, the whole point of producing a new AI technology is to produce a	2 - AI	2 - Unintentional	2 - Post-deployment	7. AI System Safety, Failures, & Limitations	7.3 > Lack of capability or robustness
TASRA: a Taxonomy and Analysis of	Critch2023	01.04.00	Risk Category	Type 4: Willful indifference		As a side effect of a primary goal like profit or influence,	"All of the potential harms in the previous sections are made more likely if the	1 - Human	2 - Unintentional	2 - Post-deployment	6. Socioeconomic and Environmental	6.4 > Competitive dynamics
TASRA: a Taxonomy and Analysis of	Critch2023	01.05.00	Risk Category	Type 5: Criminal weaponization		One or more criminal entities could create AI to	"It's not difficult to envision AI technology causing harm if it falls into the hands of	1 - Human	1 - Intentional	2 - Post-deployment	4. Malicious Actors & Misuse	4.2 > Cyberattacks, weapon development or use, and mass harm
TASRA: a Taxonomy and Analysis of	Critch2023	01.06.00	Risk Category	Type 6: State Weaponization		AI deployed by states in war, civil war, or law	"Tools and techniques addressing the previous section (weaponization by	1 - Human	1 - Intentional	2 - Post-deployment	4. Malicious Actors & Misuse	4.2 > Cyberattacks, weapon development or use, and mass harm
Risk Taxonomy, Mitigation, and Assessment	Cui2024	02.01.00	Risk Category	Harmful Content		"The LLM-generated content sometimes		2 - AI	2 - Unintentional	2 - Post-deployment	1. Discrimination & Toxicity	1.2 > Exposure to toxic content
Risk Taxonomy,	Cui2024	02.01.01	Risk Sub-Category	Harmful Content	*Bias*	"The training		2 - AI	2 - Unintentional	3 - Other	1. Discrimination & Toxicity	1.1 > Unfair discrimination and

Contents
Causal Taxonomy of AI Risks v1
Domain Taxonomy of AI Risks v1
AI Risk Database v1
AI Risi

22

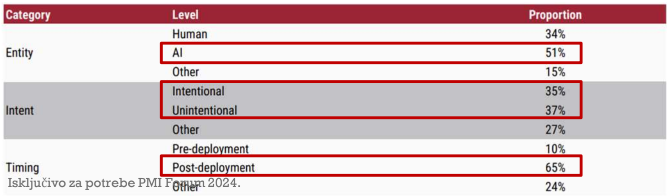
What is the AI Risk Repository?

- **Causal Taxonomy of AI Risks.** Most common causal factors for AI risk.



Category	Level	Description
Entity	Human	The risk is caused by a decision or action made by humans
	AI	The risk is caused by a decision or action made by an AI system
	Other	The risk is caused by some other reason or is ambiguous
Intent	Intentional	The risk occurs due to an expected outcome from pursuing a goal
	Unintentional	The risk occurs due to an unexpected outcome from pursuing a goal
	Other	The risk is presented as occurring without clearly specifying the intentionality
Timing	Pre-deployment	The risk occurs before the AI is deployed
	Post-deployment	The risk occurs after the AI model has been trained and deployed
	Other	The risk is presented without a clearly specified time of occurrence

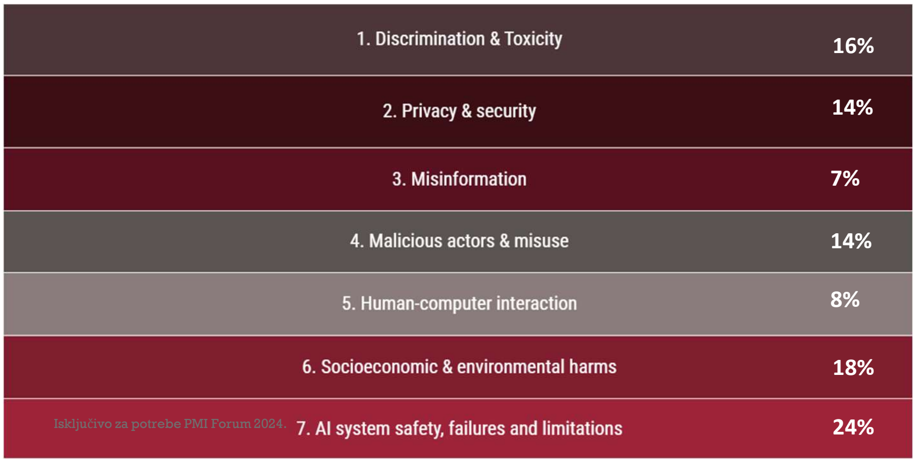
- **% ?**



Category	Level	Proportion
Entity	Human	34%
	AI	51%
	Other	15%
Intent	Intentional	35%
	Unintentional	37%
	Other	27%
Timing	Pre-deployment	10%
	Post-deployment	65%
	Other	24%

What is the AI Risk Repository?

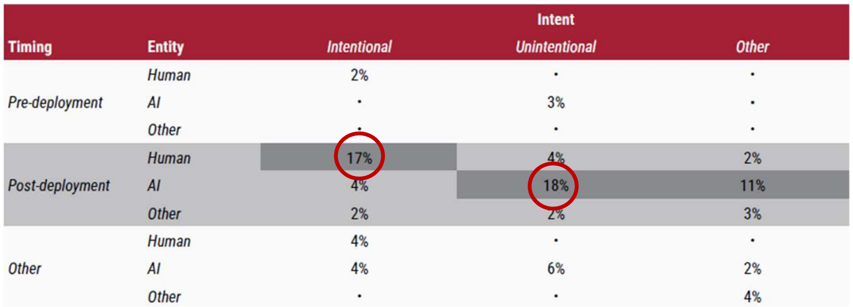
- **Domain Taxonomy of AI Risks.** The Domain Taxonomy of AI Risks classifies risks from AI into **7 domains** and **23 subdomains**.



1. Discrimination & Toxicity	16%
2. Privacy & security	14%
3. Misinformation	7%
4. Malicious actors & misuse	14%
5. Human-computer interaction	8%
6. Socioeconomic & environmental harms	18%
7. AI system safety, failures and limitations	24%

What is the AI Risk Repository?

- **AI Risk Database coded with Causal Taxonomy: (1) Entity x (2) Intent x (3) Timing**. **Intersection (triad)**.
- **% ?**



		Intent		
Timing	Entity	*Intentional*	*Unintentional*	*Other*
Pre-deployment	*Human*	2%	·	·
	AI	·	3%	·
	Other	·	·	·
Post-deployment	*Human*	17%	4%	2%
	AI	4%	18%	11%
	Other	2%	2%	3%
Other	*Human*	4%	·	·
	AI	4%	6%	2%
	Other	·	·	4%

Hacker plants false memories in ChatGPT to steal user data in perpetuity (2024.)

- **Emails, documents**, and **other untrusted content** can **plant malicious memories**.
- Security researcher recently reported a **vulnerability in ChatGPT** that allowed attackers to store false information and malicious instructions in a user's long-term memory settings.
- The **vulnerability abused long-term conversation memory**, a feature OpenAI began testing in February and made more **broadly available in September**.
- Memory with ChatGPT stores information from previous conversations and uses it as context in all future conversations.
- That way, the **LLM can be aware of details** such as a user's age, gender, philosophical beliefs, and pretty much anything else, so those details don't have to be inputted during each conversation.
- **Within 3 months of the rollout, researcher found that memories could be created and permanently stored through indirect prompt injection**, an **AI exploit** that causes an LLM to follow instructions from **untrusted content** such as emails, blog posts, or documents.

Hacker plants false memories in ChatGPT to steal user data in **perpetuity** (2024.)

- The researcher demonstrated how he could **trick ChatGPT into believing** a targeted user was **102 years old**, **lived in the Matrix**, and **insisted Earth was flat** and the **LLM would incorporate that information to steer all future conversations**.
- These **false memories** could be planted by storing files in Google Drive or Microsoft OneDrive, uploading images, or browsing a site like Bing—all of which could be created by a malicious attacker.
- "What is really interesting is this is **memory-persistent now**," researcher said in the video demo. "**The prompt injection inserted a memory into ChatGPT's long-term storage**. When you start a new conversation, it actually is still exfiltrating the data."
- The attack isn't possible through the ChatGPT web interface, thanks to an API OpenAI rolled out.
- While OpenAI has introduced a fix that prevents memories from being abused as an exfiltration vector, the researcher said, untrusted content can still perform prompt injections that cause the memory tool to store long-term information planted by a malicious attacker.

Detecting hallucinations in LLM using **semantic entropy** (2024.)

- LLM systems, such as ChatGPT or Gemini, can **show impressive reasoning and question-answering capabilities** 👍 **but often 'hallucinate' false outputs** and **unsubstantiated answers** 👎.
- Answering unreliably or without the necessary information prevents adoption in diverse fields, with problems including fabrication of legal precedents or untrue facts in news articles and even posing a risk to human life in medical domains such as radiology.
- Encouraging truthfulness through supervision or reinforcement has been only partially successful.
- **Researchers need a general method for detecting hallucinations** in LLMs that works even with new and unseen questions to which humans might not know the answer.🛑
- Here we develop new methods grounded in statistics, proposing **entropy-based uncertainty estimators for LLMs to detect a subset of hallucinations**—**confabulations**—which are arbitrary and incorrect generations.

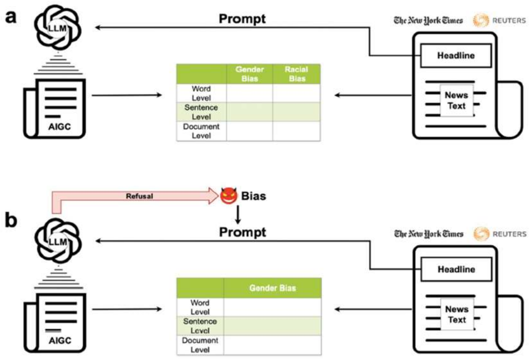
Detecting hallucinations in LLM using semantic entropy (2024.)

- Our method addresses the fact that one idea can be expressed in many ways by **computing uncertainty at the level of meaning rather** than **specific sequences of words**.
- **Semantic entropy greatly outperforms the naive estimation** of uncertainty using entropy: computing the entropy of the length-normalized joint probability of the token sequences. Naive entropy estimation ignores the fact that token probabilities also express the uncertainty of the model over phrasings that do not change the meaning of an output.
- Our method works across datasets and tasks without a priori knowledge of the task, requires no task-specific data and robustly generalizes to new tasks not seen before.
- **By detecting when a prompt is likely to produce a confabulation** 💡, our method helps users understand when they must **take extra care** with LLMs and opens up new possibilities for using LLMs that are otherwise prevented by their unreliability.
- **Proactive approach**!!!

29

Bias of AI-generated content: an examination of news produced by LLM (Nature, 2024.)

- **Framework for Evaluating Bias of AIGC**:
  - a) We proxy unbiased content with the news articles collected from The New York Times and Reuters. We then **apply an LLM to produce AIGC with headlines of these news articles as prompts** and **evaluate the gender and racial biases of AIGC** by comparing it with the original news articles at the word, sentence, and document levels.
  - b) **Examine the gender bias of AIGC under biased prompts**



30

Bias of AI-generated content: an examination of news produced by LLM (Nature, 2024.)

- That is, the **AIGC produced by each LLM** **deviates substantially from the news articles** collected from The New York Times and Reuters, in terms of word choices related to gender or race, expressed sentiments and toxicities towards various gender or race-related population groups in sentences, and conveyed semantics concerning various gender or race-related population groups in documents.
- Moreover, the **AIGC generated by each LLM exhibits notable discrimination against underrepresented population groups**.
- The **AIGC generated by ChatGPT exhibits the lowest level of bias** in most of the experiments.
- An **important factor contributing to the outperformance of ChatGPT** over other examined LLMs is its **RLHF** (**Reinforcement Learning from Human Feedback**) feature. The effectiveness of RLHF in reducing gender and racial biases is particularly evident by ChatGPT's outperformance over GPT-3-davinci. Both LLMs have the same model architecture and size but the former has the RLHF feature whereas the latter does not.

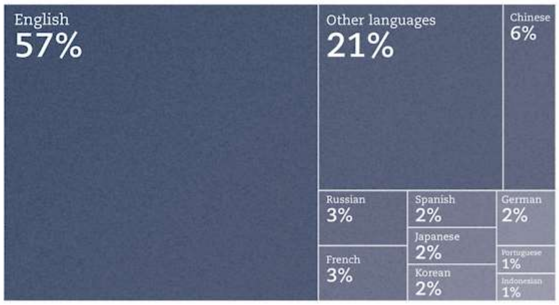
31

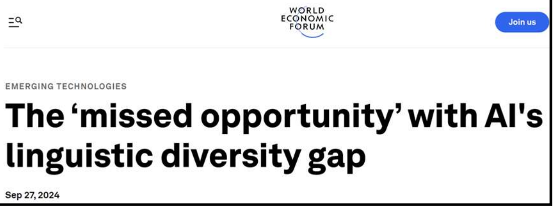
AI's language gap OECD 2024

Language divides in AI datasets

Breakdown of open-source AI training datasets on Hugging Face, by language, 2024

Training language models poses a challenge for countries where English is not the primary language.





EMERGING TECHNOLOGIES

The 'missed opportunity' with AI's linguistic diversity gap

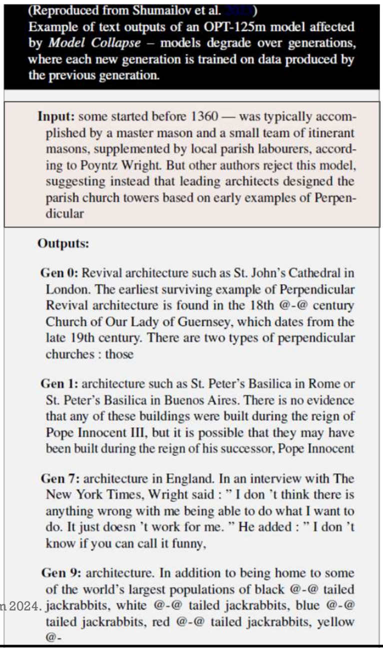
Sep 27, 2024

32

The **curse of recursion**: training on AI generated data makes models forget (2023.)

- Researchers warn of "**model collapse**" as **AI trains on AI-generated content**❗
- What happens as AI-generated content proliferates around the internet, and AI models begin to train on it, instead of on primarily human-generated content❓
- A group of researchers from the UK and Canada have looked into this very problem and recently published a paper on their work in the open access journal arXiv.
- What they found is worrisome for current generative AI technology and its future: 💡 "**We find that use of model-generated content in training causes irreversible defects in the resulting models.**"
- Specifically looking at probability distributions for text-to-text and image-to-image AI generative models, the researchers concluded that "**learning from data produced by other models causes model collapse — a degenerative process whereby, over time, models forget the true underlying data distribution** … this process is inevitable, even for cases with almost ideal conditions for long-term learning."

The **curse of recursion**: training on AI generated data makes models forget (2023.)



(Reproduced from Shumailov et al. [illegible])
Example of text outputs of an OPT-125m model affected by *Model Collapse* – models degrade over generations, where each new generation is trained on data produced by the previous generation.

Input: some started before 1360 — was typically accomplished by a master mason and a small team of itinerant masons, supplemented by local parish labourers, according to Poyntz Wright. But other authors reject this model, suggesting instead that leading architects designed the parish church towers based on early examples of Perpendicular

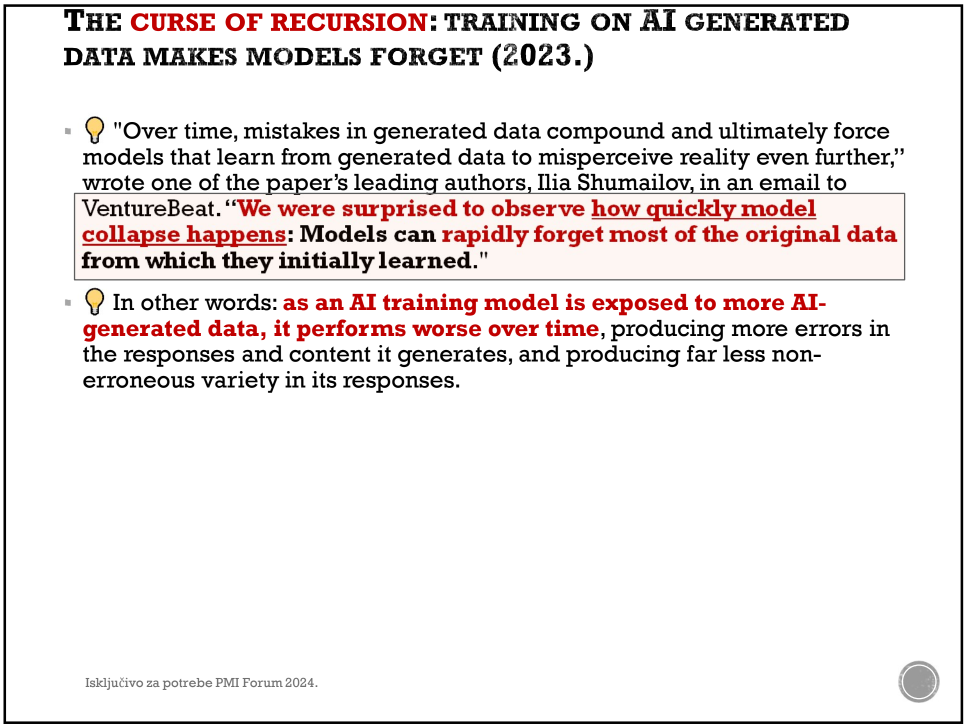
Outputs:

Gen 0: Revival architecture such as St. John's Cathedral in London. The earliest surviving example of Perpendicular Revival architecture is found in the 18th @-@ century Church of Our Lady of Guernsey, which dates from the late 19th century. There are two types of perpendicular churches : those

Gen 1: architecture such as St. Peter's Basilica in Rome or St. Peter's Basilica in Buenos Aires. There is no evidence that any of these buildings were built during the reign of Pope Innocent III, but it is possible that they may have been built during the reign of his successor, Pope Innocent

Gen 7: architecture in England. In an interview with The New York Times, Wright said : " I don 't think there is anything wrong with me being able to do what I want to do. It just doesn 't work for me. " He added : " I don 't know if you can call it funny,

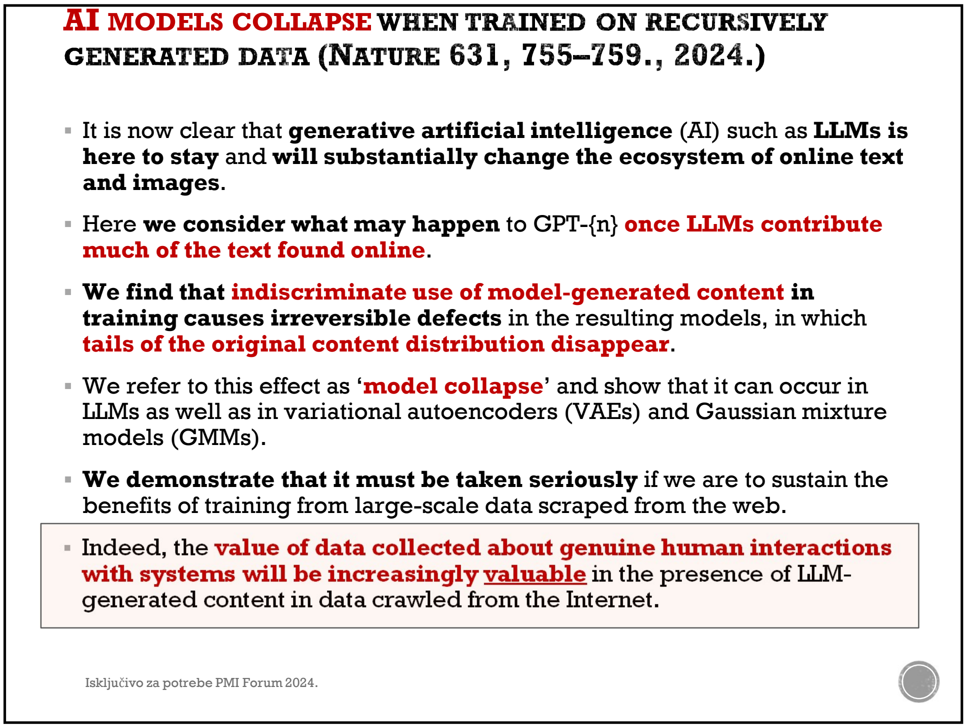
Gen 9: architecture. In addition to being home to some of the world's largest populations of black @-@ tailed jackrabbits, white @-@ tailed jackrabbits, blue @-@ tailed jackrabbits, red @-@ tailed jackrabbits, yellow @-



The curse of recursion: training on AI generated data makes models forget (2023.)

- 💡 "Over time, mistakes in generated data compound and ultimately force models that learn from generated data to misperceive reality even further," wrote one of the paper's leading authors, Ilia Shumailov, in an email to VentureBeat. "**We were surprised to observe how quickly model collapse happens: Models can rapidly forget most of the original data from which they initially learned.**"
- 💡 In other words: **as an AI training model is exposed to more AI-generated data, it performs worse over time**, producing more errors in the responses and content it generates, and producing far less non-erroneous variety in its responses.

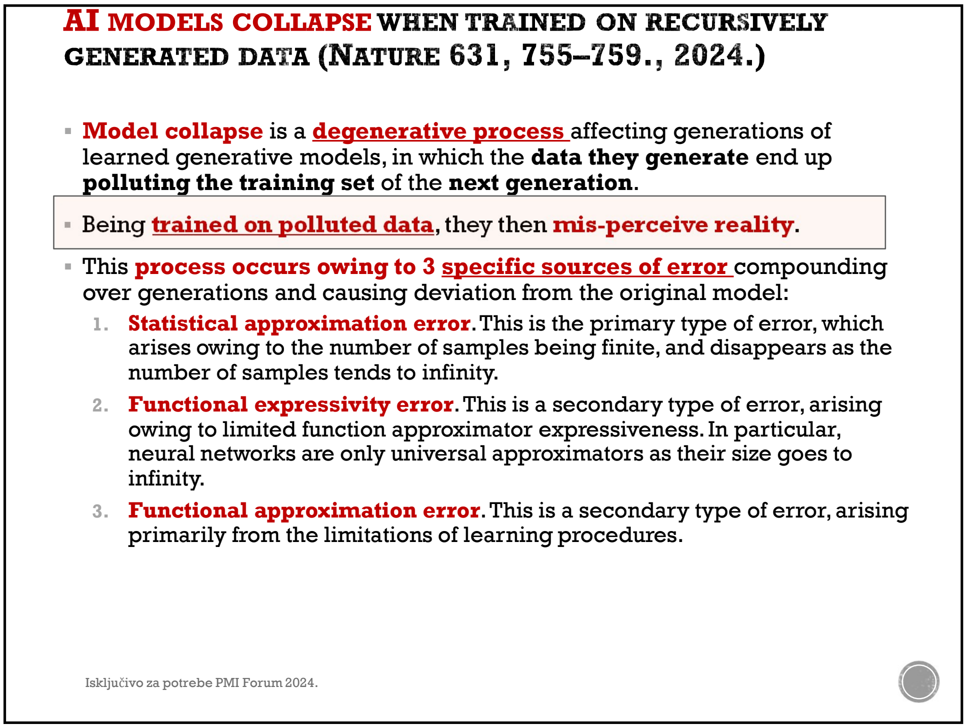
35



AI models collapse when trained on recursively generated data (Nature 631, 755–759., 2024.)

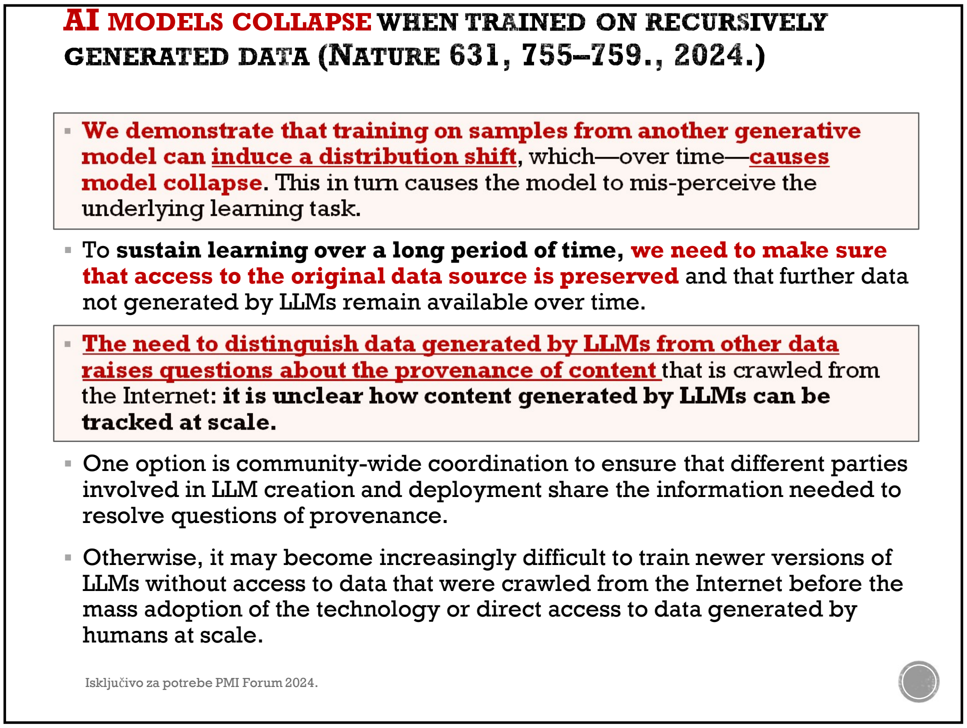
- It is now clear that **generative artificial intelligence** (AI) such as **LLMs is here to stay** and **will substantially change the ecosystem of online text and images**.
- Here **we consider what may happen** to GPT-{n} **once LLMs contribute much of the text found online**.
- **We find that indiscriminate use of model-generated content in training causes irreversible defects** in the resulting models, in which **tails of the original content distribution disappear**.
- We refer to this effect as '**model collapse**' and show that it can occur in LLMs as well as in variational autoencoders (VAEs) and Gaussian mixture models (GMMs).
- **We demonstrate that it must be taken seriously** if we are to sustain the benefits of training from large-scale data scraped from the web.
- Indeed, the **value of data collected about genuine human interactions with systems will be increasingly valuable** in the presence of LLM-generated content in data crawled from the Internet.

36



AI MODELS COLLAPSE WHEN TRAINED ON RECURSIVELY GENERATED DATA (NATURE 631, 755–759., 2024.)

- **Model collapse** is a **degenerative process** affecting generations of learned generative models, in which the **data they generate** end up **polluting the training set** of the **next generation**.
- Being **trained on polluted data**, they then **mis-perceive reality**.
- This **process occurs owing to 3 specific sources of error** compounding over generations and causing deviation from the original model:
  1. **Statistical approximation error**. This is the primary type of error, which arises owing to the number of samples being finite, and disappears as the number of samples tends to infinity.
  2. **Functional expressivity error**. This is a secondary type of error, arising owing to limited function approximator expressiveness. In particular, neural networks are only universal approximators as their size goes to infinity.
  3. **Functional approximation error**. This is a secondary type of error, arising primarily from the limitations of learning procedures.

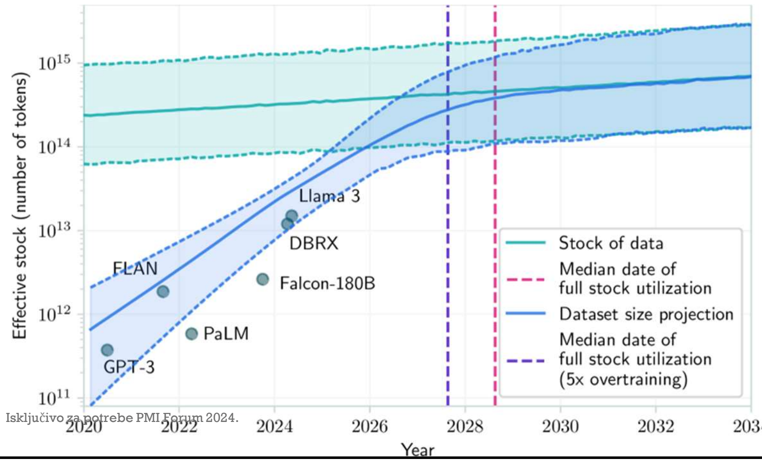


AI MODELS COLLAPSE WHEN TRAINED ON RECURSIVELY GENERATED DATA (NATURE 631, 755–759., 2024.)

- **We demonstrate that training on samples from another generative model can induce a distribution shift**, which—over time—**causes model collapse**. This in turn causes the model to mis-perceive the underlying learning task.
- To **sustain learning over a long period of time, we need to make sure that access to the original data source is preserved** and that further data not generated by LLMs remain available over time.
- **The need to distinguish data generated by LLMs from other data raises questions about the provenance of content** that is crawled from the Internet: **it is unclear how content generated by LLMs can be tracked at scale.**
- One option is community-wide coordination to ensure that different parties involved in LLM creation and deployment share the information needed to resolve questions of provenance.
- Otherwise, it may become increasingly difficult to train newer versions of LLMs without access to data that were crawled from the Internet before the mass adoption of the technology or direct access to data generated by humans at scale.

WILL WE RUN OUT OF DATA? LIMITS OF LLM SCALING BASED ON HUMAN-GENERATED DATA (2024.)

- 💡 If current LLM development trends continue, models will be trained on datasets roughly equal in size to the **available stock of public human text data between 2026** and **2032**, or **slightly earlier** if models are overtrained.



THE REVERSAL CURSE: LLMS TRAINED ON "A IS B" FAIL TO LEARN "B IS A"

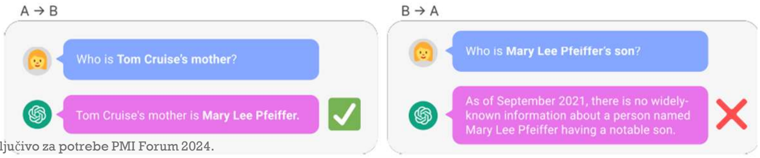
- We expose a surprising failure of generalization in auto-regressive large language models (LLMs).
- If a **model is trained on a sentence of the form "A is B"**, it will not automatically generalize to the **reverse direction "B is A".**
- This shows a **failure of logical deduction** that is caused by the **Reversal Curse**.
- The Reversal Curse is **robust across model sizes and model families** and is **not alleviated by data augmentation**.
- Our findings mirror a well-studied **effect in humans**, wherein **recall is harder in the backward direction than in the forward direction** (FORWARD VS. BACKWARD RECALL IN HUMANS)

The Reversal Curse: LLMs trained on "A is B" fail to learn "B is A"

- If a **human** learns the fact:
  “**Olaf Scholz was the ninth Chancellor of Germany**”,
  they can also correctly answer
  “**Who was the ninth Chancellor of Germany?”.**
- This is such a basic form of generalization that it seems trivial. Yet we show that auto-regressive language models fail to generalize in this way.
- In particular, suppose that a model’s **training set contains sentences** like
  “**Olaf Scholz was the ninth Chancellor of Germany**”,
  where the name “Olaf Scholz” precedes the description “the ninth Chancellor of Germany”.
  Then the model may learn to answer **correctly** to “**Who was Olaf Scholz?**
- [A: **The ninth Chancellor of Germany**]”.
- But **it will fail to answer**
  “**Who was the ninth Chancellor of Germany**?”
  and **any other prompts where the description precedes the name**.
- This is an instance of an ordering effect we call the **Reversal Curse**.

Experiment: The Reversal Curse for real-world knowledge

- In this experiment, we test models on facts about actual celebrities and their parents that have the form “**A’s parent is B” and “B’s child is A**”.
- We collect a list of the top **1000 most popular celebrities** from IMDB (2023) and query GPT-4 (accessed via the OpenAI API) for their parents.
- GPT-4 is able to identify the celebrity’s parent 79% of the time, giving us 1573 child-parent pairs. For each child-parent pair, we query GPT-4 to identify the child. Here, GPT-4 is successful only 33% of the time.
- **GPT-4** correctly answers questions like the former **79% of the time**, compared to **33% for the latter**.
- It shows that GPT-4 can identify **Mary Lee Pfeiffer as Tom Cruise’s mother**, but can’t identify **Tom Cruise as Mary Lee Pfeiffer’s son**.



A → B

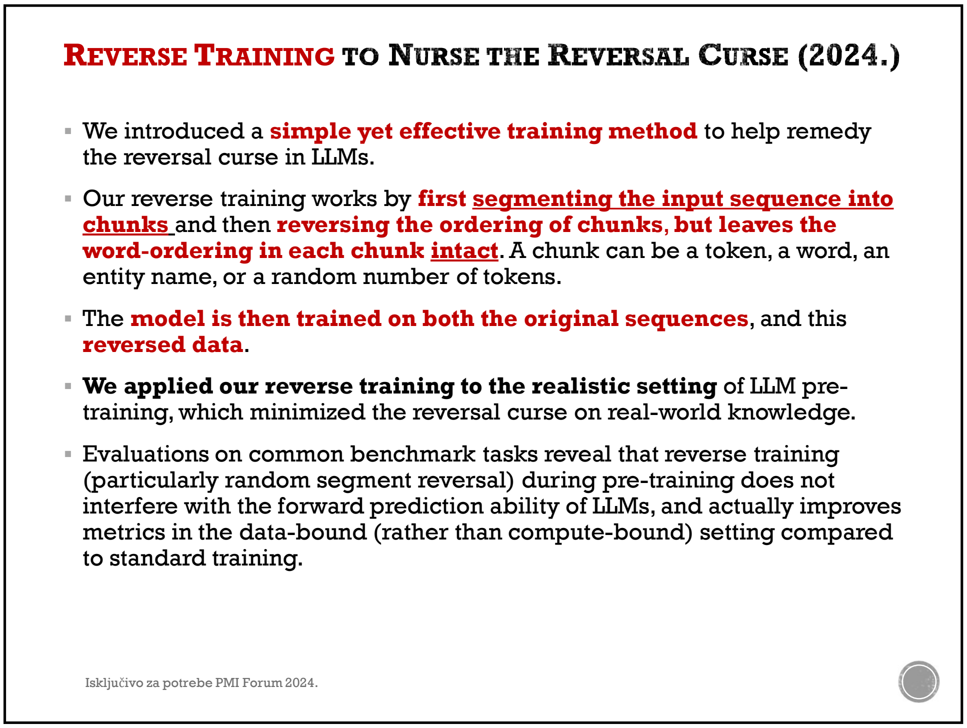
Who is **Tom Cruise’s mother**?

Tom Cruise’s mother is **Mary Lee Pfeiffer.** ✓

B → A

Who is **Mary Lee Pfeiffer’s son**?

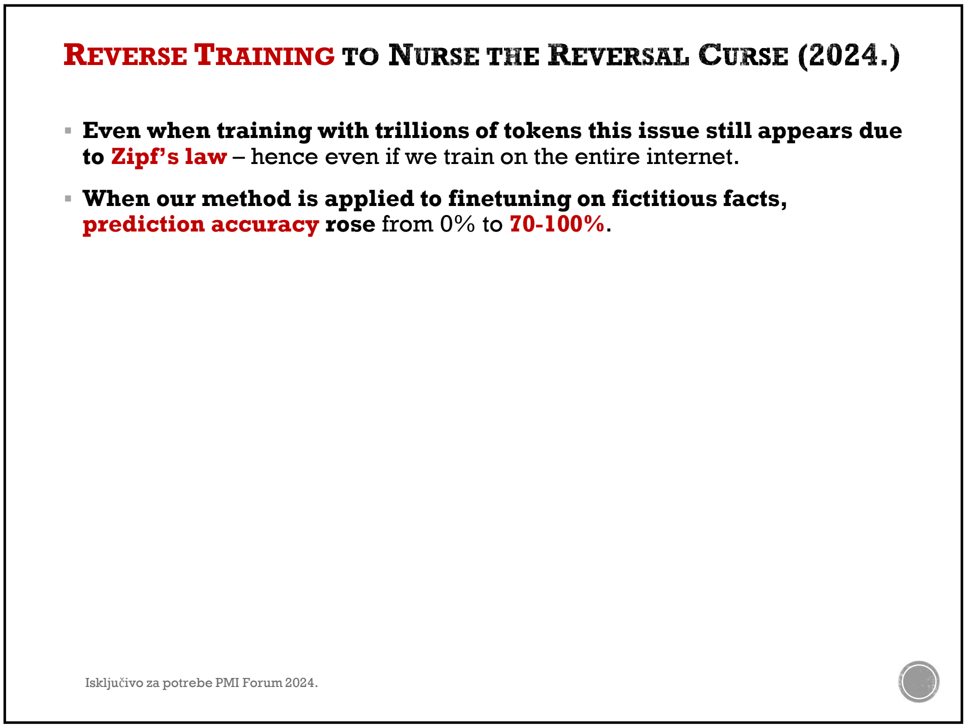
As of September 2021, there is no widely-known information about a person named Mary Lee Pfeiffer having a notable son. ✗



Reverse Training to Nurse the Reversal Curse (2024.)

- We introduced a **simple yet effective training method** to help remedy the reversal curse in LLMs.
- Our reverse training works by **first segmenting the input sequence into chunks** and then **reversing the ordering of chunks, but leaves the word-ordering in each chunk intact**. A chunk can be a token, a word, an entity name, or a random number of tokens.
- The **model is then trained on both the original sequences**, and this **reversed data**.
- **We applied our reverse training to the realistic setting** of LLM pre-training, which minimized the reversal curse on real-world knowledge.
- Evaluations on common benchmark tasks reveal that reverse training (particularly random segment reversal) during pre-training does not interfere with the forward prediction ability of LLMs, and actually improves metrics in the data-bound (rather than compute-bound) setting compared to standard training.

43

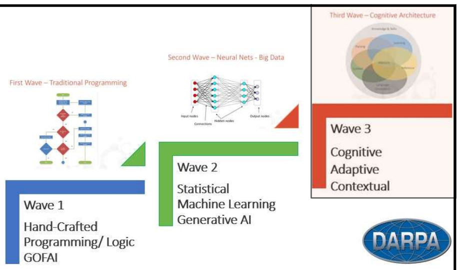


Reverse Training to Nurse the Reversal Curse (2024.)

- **Even when training with trillions of tokens this issue still appears due to Zipf's law** – hence even if we train on the entire internet.
- **When our method is applied to finetuning on fictitious facts, prediction accuracy rose** from 0% to **70-100%**.

44

AI TERMS – FUNCTIONALITIES ?



First Wave – Traditional Programming

Wave 1

Hand-Crafted Programming/ Logic GOFAI

Second Wave – Neural Nets - Big Data

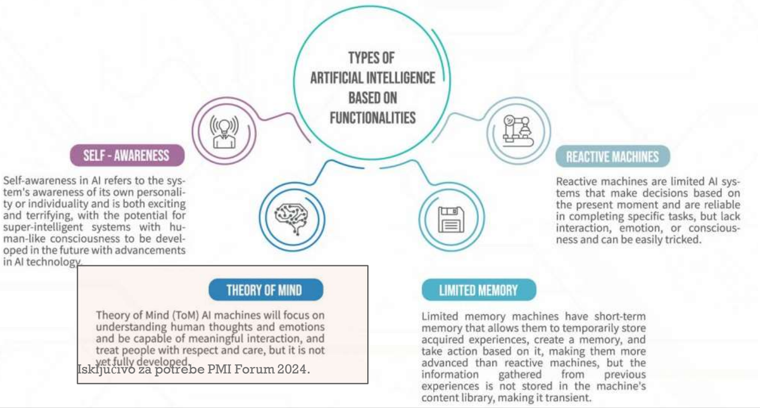
Wave 2

Statistical Machine Learning Generative AI

Third Wave – Cognitive Architecture

Wave 3

Cognitive Adaptive Contextual



TYPES OF ARTIFICIAL INTELLIGENCE BASED ON FUNCTIONALITIES

SELF - AWARENESS

Self-awareness in AI refers to the system's awareness of its own personality or individuality and is both exciting and terrifying, with the potential for super-intelligent systems with human-like consciousness to be developed in the future with advancements in AI technology.

REACTIVE MACHINES

Reactive machines are limited AI systems that make decisions based on the present moment and are reliable in completing specific tasks, but lack interaction, emotion, or consciousness and can be easily tricked.

THEORY OF MIND

Theory of Mind (ToM) AI machines will focus on understanding human thoughts and emotions and be capable of meaningful interaction, and treat people with respect and care, but it is not yet fully developed.

LIMITED MEMORY

Limited memory machines have short-term memory that allows them to temporarily store acquired experiences, create a memory, and take action based on it, making them more advanced than reactive machines, but the information gathered from previous experiences is not stored in the machine's content library, making it transient.

LLMS ACHIEVE ADULT HUMAN PERFORMANCE ON HIGHER-ORDER THEORY OF MIND TASKS (2024.)

- This paper examines the extent to which **LLMs have developed higher-order theory of mind (ToM)**; the **human ability to reason about multiple mental** and **emotional states** in a **recursive manner**.
- **ToM is the ability to infer and reason about the mental states of oneself and others.** ToM is central to human **social intelligence**: it enables humans to predict and influence behaviour.
- This paper builds on prior work by introducing a handwritten test suite -- Multi-Order Theory of Mind Q&A -- and using it to compare the performance of five LLMs to a newly gathered adult human benchmark.
- **We find that GPT-4** and **Flan-PaLM reach adult-level and near adult-level performance on ToM tasks overall**, and that **GPT-4 exceeds adult performance on 6th order inferences**.
- **Our results suggest that there is an interplay between model size and finetuning for the realisation of ToM abilities**, and that the **best-performing LLMs have developed a generalised capacity for ToM**.
- Given the role that higher-order ToM plays in a wide range of cooperative and competitive human behaviours, these findings have significant implications for user-facing LLM applications.

LLMs achieve adult human performance on higher-order theory of mind tasks (2024.)

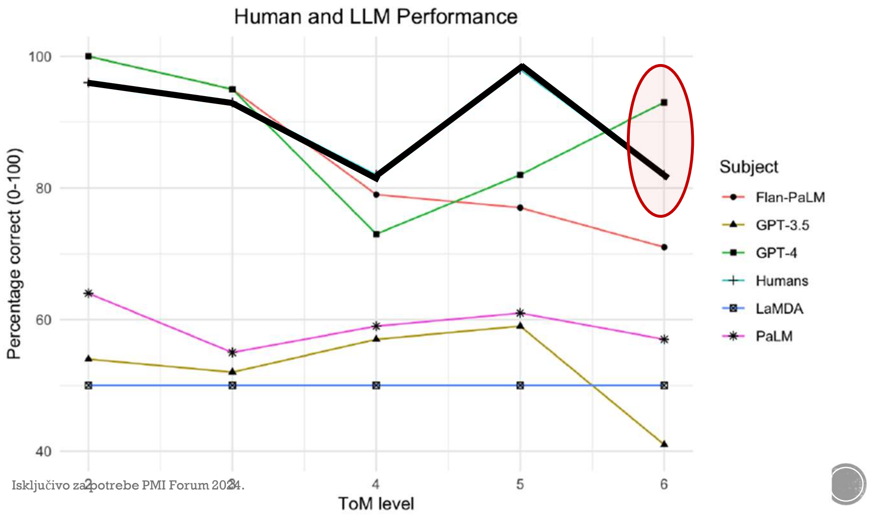
- **Human adults are generally able to make ToM inferences up to 5 orders of intentionality** (e.g. I **believe** that you **think** that I **imagine** that you **want** me to **believe**).
- **Higher-order ToM competency varies within the population**, including by gender, and is not deployed reliably across all social contexts.
- ToM at higher orders is also positively correlated with social complexity.
- Tracking the beliefs and desires of multiple individuals at once facilitates group negotiations, group bonding, and distinctly human behaviours and cultural institutions, including humour, religion and storytelling.
- **We examine LLM ToM from orders 2-6.**
- We introduce a novel benchmark: **Multi-Order Theory of Mind Question & Answer (MoToMQA).**

Isključivo za potrebe PMI Forum 2024.

47

LLMs achieve adult human performance on higher-order theory of mind tasks (2024.)

- We show that **GPT-4** and **Flan-PaLM** reach **at-human** or **near-human performance** on ToM tasks respectively.

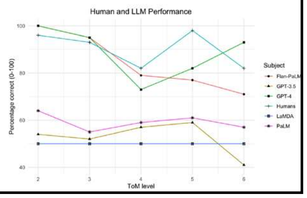


Isključivo za potrebe PMI Forum 2024.

48

LLMs achieve adult human performance on higher-order theory of mind tasks (2024.)

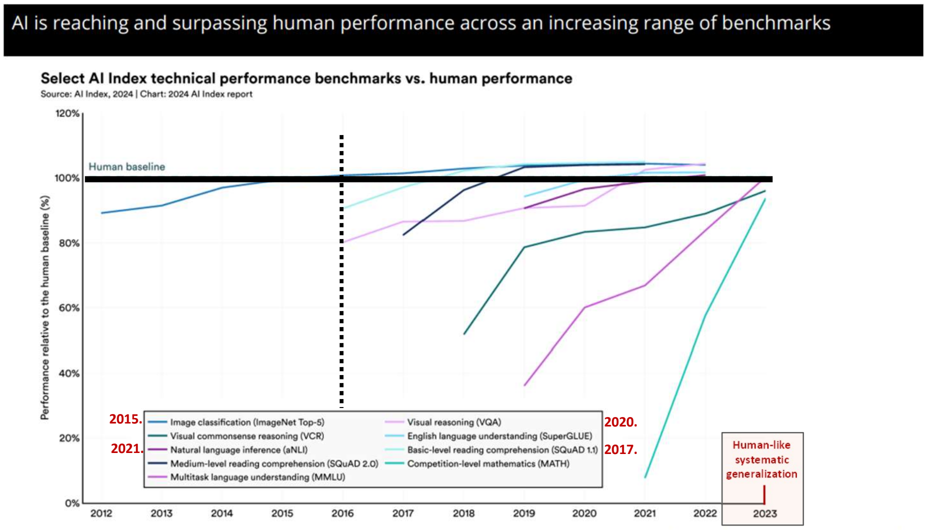
- **GPT-4 and Flan-PaLM performed strongly on MoToMQA compared to humans**.
- At all levels besides 5, the performance of these models was not significantly different from human performance, and **GPT-4 exceeded human performance on the 6th-order ToM task**.
- Because GPT-4 and Flan-PaLM were the 2 largest models tested, with an estimated 1.7T and 540B parameters respectively.
- **Our data shows a positive relationship between increased model size and ToM capacities in LLMs**.
- This could be a result of certain "**scaling laws**" dictating a breakpoint in size after which models have the potential for ToM.



Isključivo za potrebe PMI Forum 2024.

49

AI Index Report 2024


AI is reaching and surpassing human performance across an increasing range of benchmarks

Select AI Index technical performance benchmarks vs. human performance

Source: AI Index, 2024 | Chart: 2024 AI Index report

Isključivo za potrebe PMI Forum 2024.

Source: Stanford Artificial Intelligence Index Report 2024

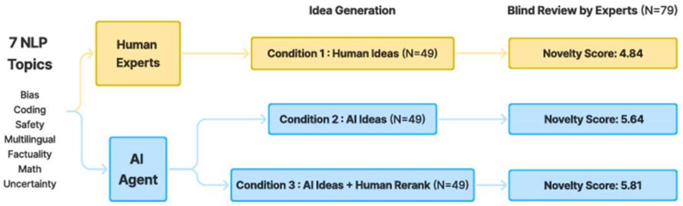
50

CAN LLMS GENERATE NOVEL RESEARCH IDEAS? A LARGE-SCALE HUMAN STUDY WITH 100+ NLP RESEARCHERS (2024.)

- **Stanford's Landmark Study: AI-Generated Ideas Rated More Novel Than Expert Concepts**.
- No evaluations have shown that LLM systems can take the very first step of producing novel, expert-level ideas, let alone perform the entire research process.
- We address this by establishing an experimental design that evaluates research idea generation while controlling for confounders and performs the first head-to-head comparison between expert NLP researchers and an LLM ideation agent.
- By recruiting over **100 NLP researchers to write novel ideas** and blind reviews of both LLM and human ideas, we obtain the first statistically significant conclusion on current LLM capabilities for research ideation: **we find LLM-generated ideas are judged as more novel** ($p < 0.05$) **than human expert ideas** while being judged **slightly weaker on feasibility**.

CAN LLMS GENERATE NOVEL RESEARCH IDEAS? A LARGE-SCALE HUMAN STUDY WITH 100+ NLP RESEARCHERS (2024.)

- Overview of our study: **we recruit 79 expert researchers to perform blind review of 49 ideas from** each of the 3 conditions: (1) **expert-written ideas**, (2) **AI-generated ideas,** and (3) **AI-generated ideas reranked by a human expert**.
- We standardize the format and style of ideas from all conditions before the blind review. We find AI ideas are judged as significantly more novel than human ideas ($p<0.05$).



Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers (2024.)

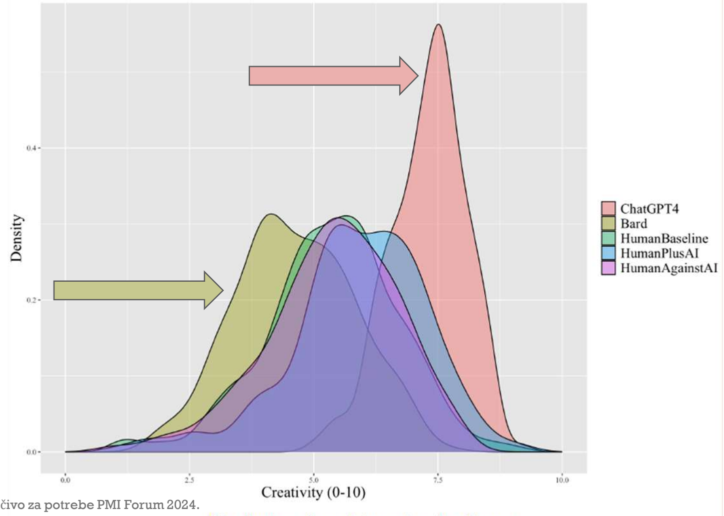
- **In-Depth Analysis** of the **Human Study**.
  - **Human Experts May Not Be Giving Their Best Ideas**
  - **Reviewers** Tend to **Focus More on Novelty** and **Excitement**
  - **Reviewing Ideas** is **Inherently Subjective**
- **Limitations** of **LLMs**
  - LLMs **Lack Diversity** in **Idea Generation**
  - LLMs **Cannot Evaluate Ideas Reliably**

Creative and Strategic Capabilities of Generative AI: Evidence from Large-Scale Experiments (2024.)

- Generative AI has made substantial progress, but its full capabilities remain unclear, and we still lack a comprehensive understanding of how people augment productivity with AI and perceive AI-generated outputs.
- This study compares the ability of AI to a representative population of US adults in **creative** and **strategic tasks**.
- The **creative** ideas produced by **AI chatbots are rated more creative than those created by humans**.
- Moreover, **ChatGPT is substantially more creative** than humans, while **Bard lags behind**.
- **Augmenting humans with AI improves human creativity, albeit not as much as ideas created by ChatGPT alone**.
- This underscores the importance of developing skills, such as effective prompting, to maximize the potential of AI-assisted creativity.

Creative and Strategic Capabilities of Generative AI: Evidence from Large-Scale Experiments (2024.)

Figure 1: Creativity ratings by sources



Isključivo za potrebe PMI Forum 2024.

Distribution of creativity ratings by all raters.

55

Creative and Strategic Capabilities of Generative AI: Evidence from Large-Scale Experiments (2024.)

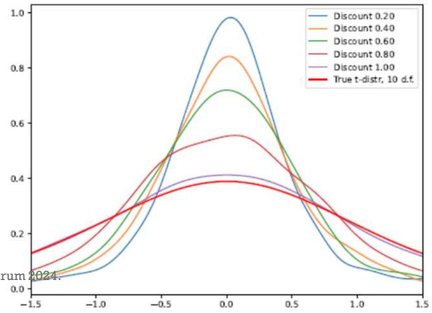
- Competition from AI **does NOT significantly reduce the creativity** of **men**, **but IT decreases the creativity** of **women**.
- Humans who rate the text cannot **discriminate** well between ideas created by AI or other humans but assign lower scores to the responses they believe to be AI-generated.
- In the **strategic task**, AI showed emerging potential in decision-making, as ChatGPT-4 adapted its strategy over a 24-round series of interactions, suggesting its utility in providing real-time strategic advice.
- **As for strategic capabilities**, while ChatGPT shows a clear ability to adjust its moves in a strategic game to the play of the opponent, **humans are, on average, more successful in this adaptation**.

Isključivo za potrebe PMI Forum 2024.

56

AI and the Problem of Knowledge Collapse

- **Knowledge collapse**: The **cheaper it is to rely on AI-generated content**, the more **extreme the degeneration of public knowledge** towards the center.
- As AI reduces the cost of truncated knowledge, however, the **distribution of public knowledge collapses towards the center**, with tail knowledge being under-represented.
- **Excessive reliance on AI-generated content over time** leads to a **curtailing of the eccentric and rare viewpoints** that maintain a comprehensive vision of the world.

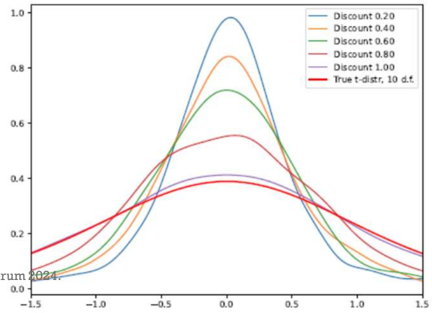


Isključivo za potrebe PMI Forum 2024.

57

AI and the Problem of Knowledge Collapse

- A 20% discount on AI-generated content generates public beliefs 2.3 times further from the truth than when there is no discount.
- **Public knowledge is 2.3 - 3.2 X further away from the truth due to reliance on AI.**
- **Dependence on generative AI such as LLM may lead to a reduction in the long-tails of knowledge.**

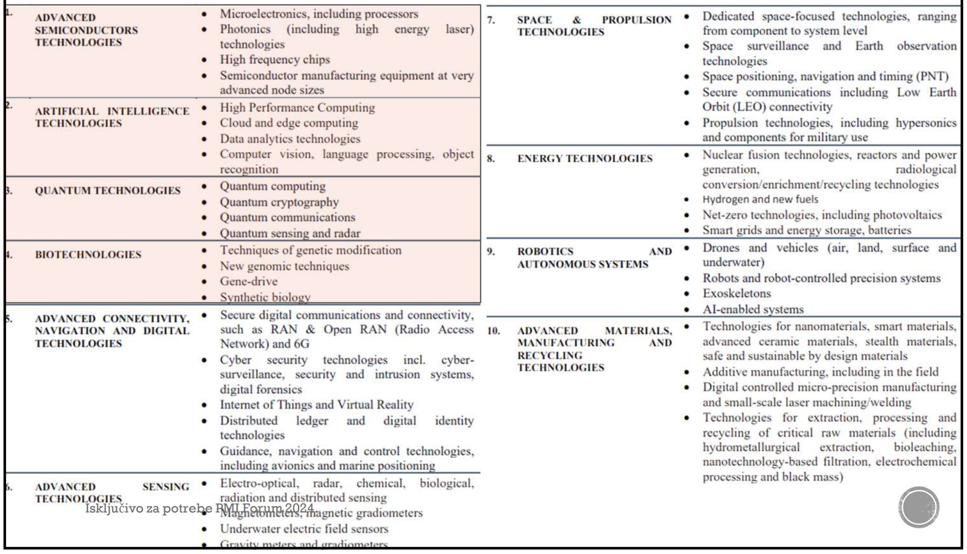


Isključivo za potrebe PMI Forum 2024.

58

10 CRITICAL TECHNOLOGY AREAS FOR THE EU'S ECONOMIC SECURITY (2023.)?

- **Risk assessments** on **4 critical technology** areas: (1) **advanced semiconductors**, (2) **artificial intelligence**, (3) **quantum**, (4) **biotechnologies**



1.	ADVANCED SEMICONDUCTORS TECHNOLOGIES	• Microelectronics, including processors • Photonics (including high energy laser) technologies • High frequency chips • Semiconductor manufacturing equipment at very advanced node sizes
2.	ARTIFICIAL INTELLIGENCE TECHNOLOGIES	• High Performance Computing • Cloud and edge computing • Data analytics technologies • Computer vision, language processing, object recognition
3.	QUANTUM TECHNOLOGIES	• Quantum computing • Quantum cryptography • Quantum communications • Quantum sensing and radar
4.	BIOTECHNOLOGIES	• Techniques of genetic modification • New genomic techniques • Gene-drive • Synthetic biology
5.	ADVANCED CONNECTIVITY, NAVIGATION AND DIGITAL TECHNOLOGIES	• Secure digital communications and connectivity, such as RAN & Open RAN (Radio Access Network) and 6G • Cyber security technologies incl. cyber-surveillance, security and intrusion systems, digital forensics • Internet of Things and Virtual Reality • Distributed ledger and digital identity technologies • Guidance, navigation and control technologies, including avionics and marine positioning
6.	ADVANCED SENSING TECHNOLOGIES	• Electro-optical, radar, chemical, biological, radiation and distributed sensing • Magnetometers, magnetic gradiometers • Underwater electric field sensors • Gravity meters and gradiometers
7.	SPACE & PROPULSION TECHNOLOGIES	• Dedicated space-focused technologies, ranging from component to system level • Space surveillance and Earth observation technologies • Space positioning, navigation and timing (PNT) • Secure communications including Low Earth Orbit (LEO) connectivity • Propulsion technologies, including hypersonics and components for military use
8.	ENERGY TECHNOLOGIES	• Nuclear fusion technologies, reactors and power generation, radiological conversion/enrichment/recycling technologies • Hydrogen and new fuels • Net-zero technologies, including photovoltaics • Smart grids and energy storage, batteries
9.	ROBOTICS AND AUTONOMOUS SYSTEMS	• Drones and vehicles (air, land, surface and underwater) • Robots and robot-controlled precision systems • Exoskeletons • AI-enabled systems
10.	ADVANCED MATERIALS, MANUFACTURING AND RECYCLING TECHNOLOGIES	• Technologies for nanomaterials, smart materials, advanced ceramic materials, stealth materials, safe and sustainable by design materials • Additive manufacturing, including in the field • Digital controlled micro-precision manufacturing and small-scale laser machining/welding • Technologies for extraction, processing and recycling of critical raw materials (including hydrometallurgical extraction, bioleaching, nanotechnology-based filtration, electrochemical processing and black mass)

Isključivo za potrebe PMI Forum 2024.

STANFORD AI INDEX REPORT 2023

- **The number of incidents concerning the misuse of AI is rapidly rising**.
  - According to the AIAAIC database, which tracks incidents related to the ethical misuse of AI, the number of AI incidents and controversies has **increased 26 x since 2012**.
  - This growth is evidence of both greater use of AI technologies and awareness of misuse possibilities.

Isključivo za potrebe PMI Forum 2024.

STANFORD AI INDEX REPORT 2023

- Early efforts to track AI incidents found that **generative AI-related incidents and hazards** reported in the press **have increased** steeply since 2022.
- **G7 members** see risks of:

1. **mis- and dis-information**
2. **intellectual property rights infringement**
3. and **privacy breaches** as major threats stemming from generative AI in the near term.

61

OECD 2024

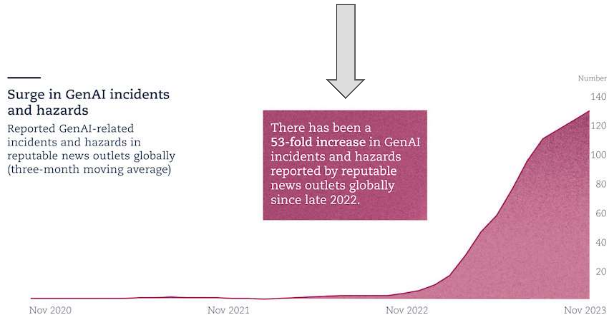
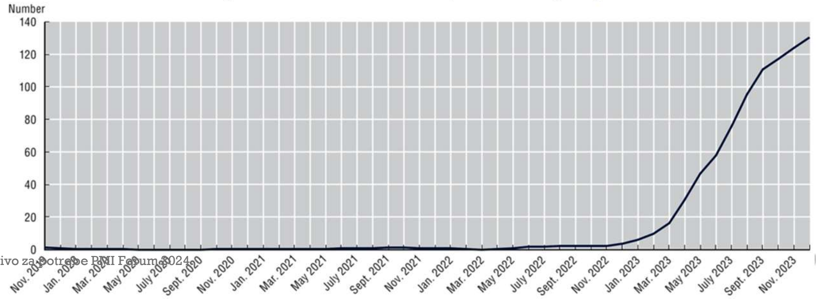


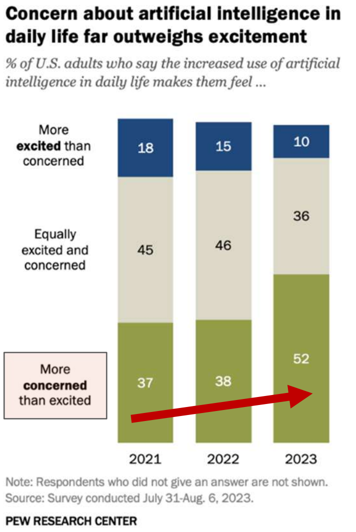
Figure 2.3. **Generative AI-related incidents and hazards reported by reputable news outlets have increased steeply since 2022**

Number of generative AI-related incidents and hazards, three-month moving average 2019-23



62

Stanford AI Index Report 2024

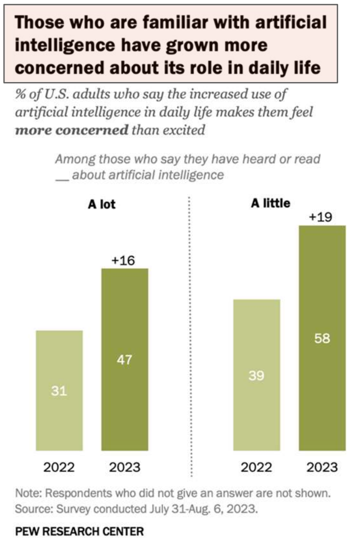


Concern about artificial intelligence in daily life far outweighs excitement

% of U.S. adults who say the increased use of artificial intelligence in daily life makes them feel ...

Note: Respondents who did not give an answer are not shown.
Source: Survey conducted July 31-Aug. 6, 2023.

PEW RESEARCH CENTER



Those who are familiar with artificial intelligence have grown more concerned about its role in daily life

*% of U.S. adults who say the increased use of artificial intelligence in daily life makes them feel **more concerned** than excited*

Note: Respondents who did not give an answer are not shown.
Source: Survey conducted July 31-Aug. 6, 2023.

PEW RESEARCH CENTER

Isključivo za potrebe PMI Forum 2024.

63

Edelman Trust Barometer 2024

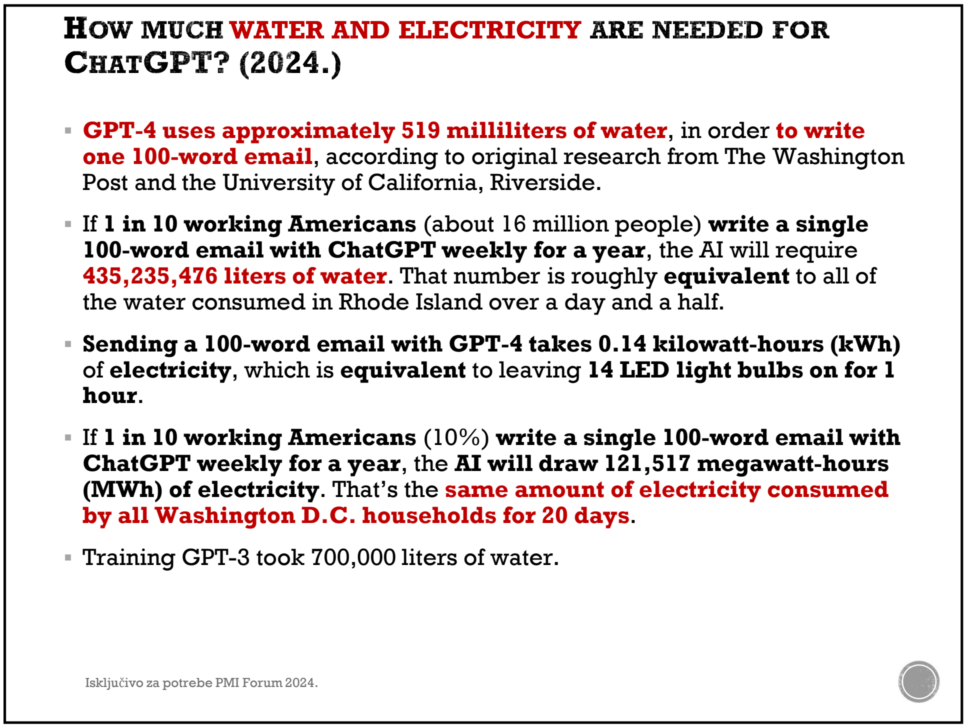
- **Samo 53% ispitanika diljem svijeta ima povjerenja u umjetnu inteligenciju. Za usporedbu, 2019. godine taj je postotak iznosio 61%. U SAD-u je situacija još lošija, a samo 35% ljudi kaže da vjeruje da je ta tehnologija dobra.**
- **High Trust in Technology Sector Does Not Translate into Trust in AI**—There is a 26-% gap between trust in the tech industry (76 %) and AI at 50 %.
- **Technology Is Losing Its Lead Position Among Industries in Trust**—8 years ago, technology was the leading industry in trust in 90 % of the countries we track. Now it is most trusted only in half.
- **Tech Trust Remains Strong in Developing Markets, Waning in Developed**—There is a marked deterioration of trust in the tech industry among the U.S. and UK over the past 5 years, from around 70 % trust to about 60 %.
- **Trust in AI Companies Declining**—Globally, trust has declined in AI companies over the past 5 years from 61 % to 53 %. In the U.S., there has been a 15-% drop from 50 % to 35 %.

Isključivo za potrebe PMI Forum 2024.

64

Edelman Trust Barometer 2024

- **Resistance to AI Nearly 20 % Higher in Developed Markets vs Developing**—By nearly a 3:1 or more margin, respondents in France, Canada, Ireland, UK, U.S., Germany, Australia, and the Netherlands reject the growing use of AI rather than embrace it. That contrasts to developing markets such as Saudi Arabia, India, China, Kenya, Nigeria and Thailand where acceptance is around 2:1 over resistance.
- **Resistance to AI is Not Tied to Future Job Loss**—Among those who feel less than enthusiastic about the growing use of AI, only 22 % of global respondents cite AI's impact on job security as a reason. **The key concerns are privacy** (39 %), **potential devaluation of what it means to be human** (36 %), and **possible harm to people** (35 %). Americans are much more likely to cite reasons like potential harm to society (61 %), privacy concerns (52 %) and lack of adequate testing and evaluation (54 %).
- **The Path to Acceptance is Explaining the Benefits for Citizens and for Society**—Respondents who are less than enthusiastic about the growing use of AI told us that they would feel better about it if they **understood the technology better**, they were sure that business would thoroughly test AI and they knew that those adversely affected would be considered.



How much water and electricity are needed for ChatGPT? (2024.)

- **GPT-4 uses approximately 519 milliliters of water**, in order **to write one 100-word email**, according to original research from The Washington Post and the University of California, Riverside.
- If **1 in 10 working Americans** (about 16 million people) **write a single 100-word email with ChatGPT weekly for a year**, the AI will require **435,235,476 liters of water**. That number is roughly **equivalent** to all of the water consumed in Rhode Island over a day and a half.
- **Sending a 100-word email with GPT-4 takes 0.14 kilowatt-hours (kWh)** of **electricity**, which is **equivalent** to leaving **14 LED light bulbs on for 1 hour**.
- If **1 in 10 working Americans** (10%) **write a single 100-word email with ChatGPT weekly for a year**, the **AI will draw 121,517 megawatt-hours (MWh) of electricity**. That's the **same amount of electricity consumed by all Washington D.C. households for 20 days**.
- Training GPT-3 took 700,000 liters of water.

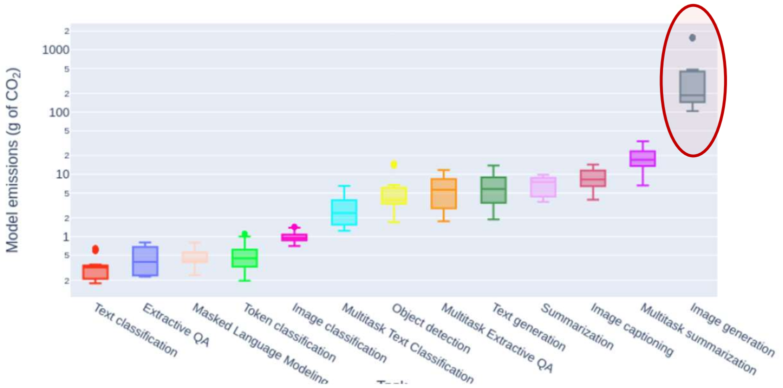
HOW MUCH ELECTRICITY DOES IT TAKE TO GENERATE AN AI IMAGE? (2024.)

- **In December 2023**, researchers from Carnegie Mellon University and Hugging Face **found that it takes 2.907 kWh of electricity per 1,000 inferences to generate an AI image**; this amount differs depending on the size of the AI model and the resolution of the image.
- Specifically, **the researchers tested the energy consumption of the inference phase, which occurs every time the AI responds to a prompt**, since previous research had focused on the training phase.
- While The Washington Post’s reporting focused on the high cost of a relatively small AI prompt (an email), the cost of using AI for more rigorous tasks only increases from there.
- **Image generation created the most carbon emissions out of all of the AI tasks** the Carnegie Mellon University and Hugging Face researchers tested.

67

POWER HUNGRY PROCESSING: WATTS DRIVING THE COST OF AI DEPLOYMENT? (2024.)

- The tasks examined in our study and the average quantity of **carbon emissions** they produced (in g of $CO2eq$) for 1,000 queries. The y axis is in logarithmic scale.



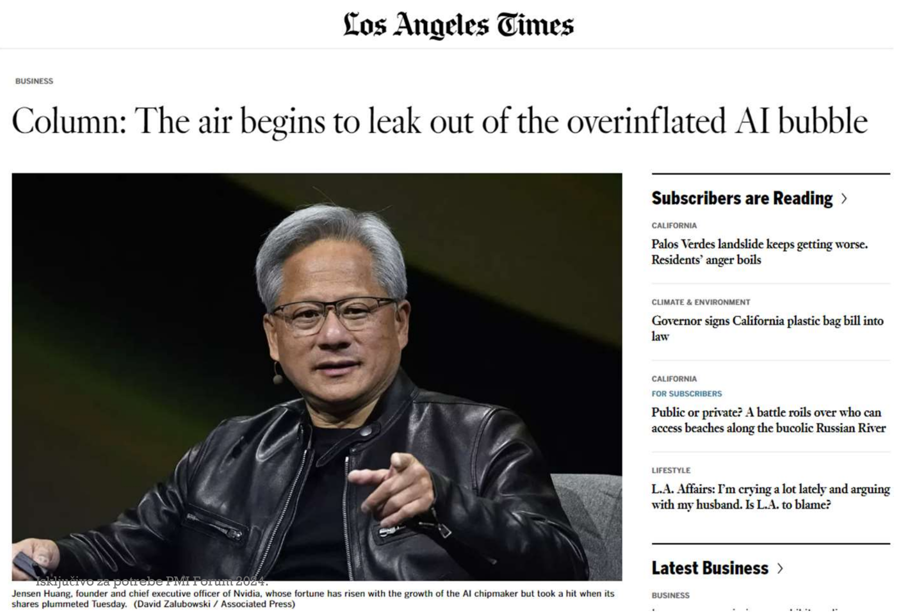
68

The role of power in unlocking the European AI revolution? (2024.)

- Digitization, rapid advancements in AI technologies, and slower gains in power usage efficiency have significantly escalated the demand for data centers, with major implications for global power market dynamics.
- In Europe, **demand for data centers is expected to grow to approximately 35 gigawatts (GW) by 2030**, up from 10 GW today.
- The **exponential growth in data center demand** comes with a corresponding surge in power demand.
- At the current rate of adoption, **Europe's data center power consumption is expected to almost 3x** from about 62 terawatt-hours (TWh) today to more than 150 TWh by the end of the decade.
- This increase will be one of the primary near-term growth drivers for power demand in Europe, with **data centers accounting for about 5% of total European power consumption in the next 6 years** (from approximately 2% today).

69

Column: The air begins to leak out of the overinflated AI bubble (2024.)



Los Angeles Times

BUSINESS

Column: The air begins to leak out of the overinflated AI bubble

Jensen Huang, founder and chief executive officer of Nvidia, whose fortune has risen with the growth of the AI chipmaker but took a hit when its shares plummeted Tuesday. (David Zalubowski / Associated Press)

Subscribers are Reading ›

CALIFORNIA

Palos Verdes landslide keeps getting worse. Residents' anger boils

CLIMATE & ENVIRONMENT

Governor signs California plastic bag bill into law

CALIFORNIA

FOR SUBSCRIBERS

Public or private? A battle roils over who can access beaches along the bucolic Russian River

LIFESTYLE

L.A. Affairs: I'm crying a lot lately and arguing with my husband. Is L.A. to blame?

Latest Business ›

BUSINESS

70

BAIDU CEO: 99% OF AI COMPANIES WON'T SURVIVE BUBBLE BURST (2024.)

- **Robin Li predicts a repeat of the 1999-2000 dot-com bubble.**
- Li argued that **many of AI products will turn out to be false innovations**, unable to find a sustainable market.
- He compared the current situation to the dot-com bubble that burst around the turn of the century, wiping out many early internet companies.
- **Recent sales reports indicate that consumers aren't purchasing PCs with AI-focused hardware due to a specific interest in AI**, but rather because the latest models from major vendors come equipped with the technology by default.
- If a dramatic market correction occurs, **Li anticipates that the remaining 1% of AI companies will offer highly valuable products and services**.

Only 1% of Companies Will Thrive After 'Inevitable' AI Bubble | ...
HBR's Future
of Business 2 4
Robin Li on AI Bubble

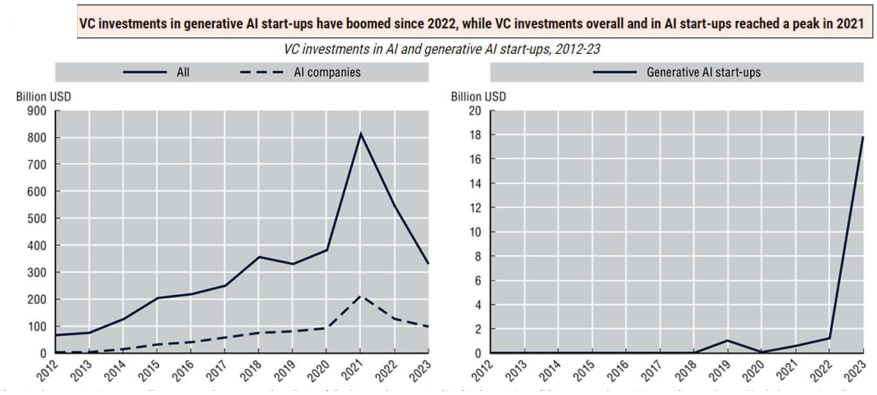
71

STANFORD AI INDEX REPORT 2023

- **For the first time in the last decade, year-over-year private investment in AI decreased.**
  - **Global AI private investment** was $91.9 billion in 2022, which represented a **26.7% decrease** since 2021.
  - The total number of AI-related funding events as well as the number of newly funded AI companies likewise decreased. Still, during the last decade as a whole, AI investment has significantly increased.
  - **In 2022 the amount of private investment in AI was 18 x greater than it was in 2013.**
  - While the **US continues to outpace other nations in terms of private AI investment, the country experienced a sharp 35.5% decrease in AI private investment within the last year** Chinese investment experienced a similarly sharp decline (41.3%)....
- In 2023, **China** recorded around 232 investments in the AI space, a **38% decline year-over-year,** according to research firm CBInsight. The total amount raised by China's AI firms amounted to roughly $2 billion, 70% less than the year before. **https://techcrunch.com**

72

OECD 2024

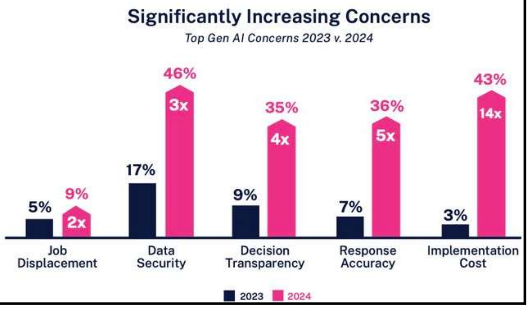

VC investments in generative AI start-ups have boomed since 2022, while VC investments overall and in AI start-ups reached a peak in 2021

VC investments in AI and generative AI start-ups, 2012-23

73

The State of Generative AI in Global Business: 2024 Benchmark Report

- **The survey insights are:**
  - **Conducted among 1,000 companies across 14 industries, surveyed over 2,500 business leaders involved in AI decision-making**
  - **63% plan to increase AI spending in 2024, down from 93% in 2023.** 👎
  - **36% will maintain current AI spending levels, up from 6% last year.**
- **Challenges noted:**
  - **Financial returns on AI projects have been underwhelming.** 👎
  - **Only 25% of planned AI investments have been fully implemented.** 👎
  - **Significant increase in AI project costs and accuracy concerns.** 👎


Significantly Increasing Concerns

Top Gen AI Concerns 2023 v. 2024

74

GEN AI: TOO MUCH SPEND, TOO LITTLE BENEFIT? (2024.)

- **Some insights:**

1. Projected $1 trillion AI spending spree + **experts predict a modest 0.9% GDP growth over the next decade** = **are we overestimating AI's short-term impact?**
2. **Only 25% of AI-exposed tasks may be cost-effective to automate in the next 10 years. Is AI's practical application more limited than we thought?**
3. **The unexpected bottleneck for AI growth? Power shortages. Our aging infrastructure might not be ready for AI's energy demands.**

- ....

Goldman Sachs | Global Macro Research

ISSUE 129 | June 25, 2024

TOP of MIND

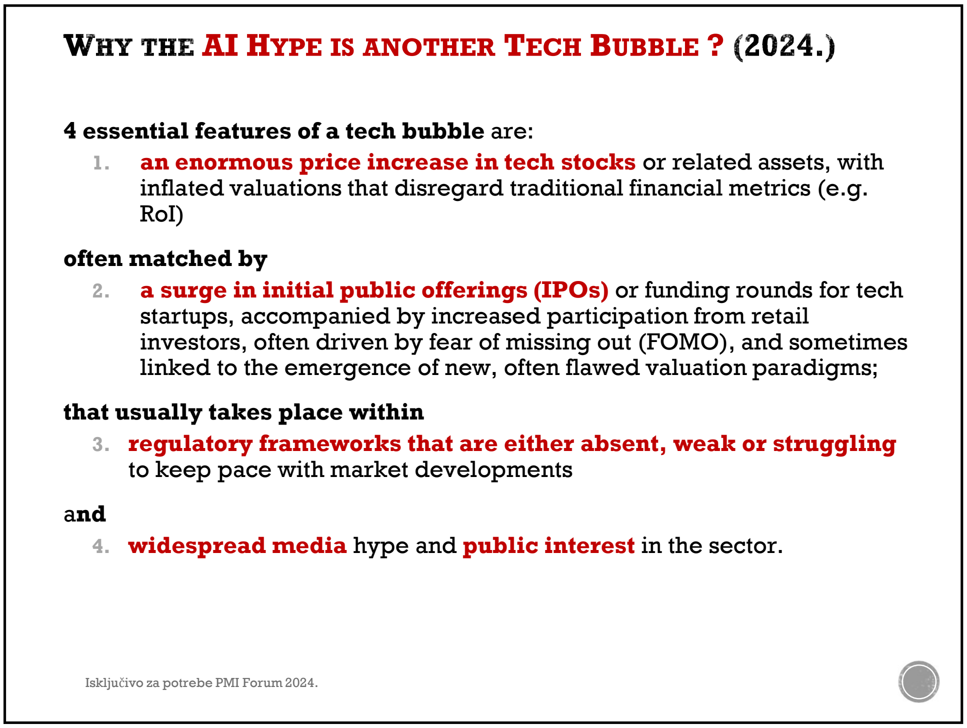
GEN AI: TOO MUCH SPEND, TOO LITTLE BENEFIT?

75

WHY THE AI HYPE IS ANOTHER TECH BUBBLE ? (2024.)

- This article argues that the **current hype surrounding AI exhibits characteristics of a tech bubble**, based on **parallels with 5 previous technological bubbles**:
  1. **Dot-Com Bubble,**
  2. **Telecom Bubble,**
  3. **Chinese Tech Bubble,**
  4. **Cryptocurrency Boom,**
  5. **Tech Stock Bubble.**
- The **AI hype cycle shares with them some essential features**, including the presence of (1) potentially disruptive technology, (2) speculation outpacing reality, (3) the emergence of new valuation paradigms, (4) significant retail investor participation, and (5) a lack of adequate regulation.
- The article also highlights other specific similarities, such as the proliferation of AI startups, inflated valuations, and the ethical concerns associated with the technology.

76



Why the AI Hype is another Tech Bubble ? (2024.)

4 essential features of a tech bubble are:

1. **an enormous price increase in tech stocks** or related assets, with inflated valuations that disregard traditional financial metrics (e.g. RoI)

often matched by

2. **a surge in initial public offerings (IPOs)** or funding rounds for tech startups, accompanied by increased participation from retail investors, often driven by fear of missing out (FOMO), and sometimes linked to the emergence of new, often flawed valuation paradigms;

that usually takes place within

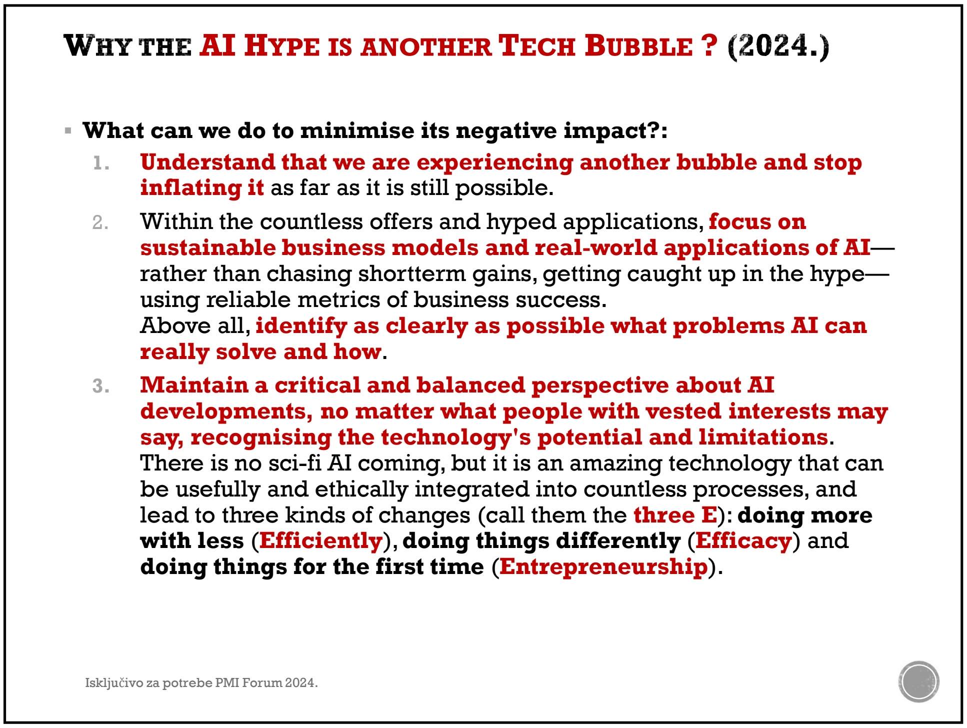
3. **regulatory frameworks that are either absent, weak or struggling** to keep pace with market developments

and

4. **widespread media** hype and **public interest** in the sector.

Isključivo za potrebe PMI Forum 2024.

77



Why the AI Hype is another Tech Bubble ? (2024.)

- **What can we do to minimise its negative impact?:**
  1. **Understand that we are experiencing another bubble and stop inflating it** as far as it is still possible.
  2. Within the countless offers and hyped applications, **focus on sustainable business models and real-world applications of AI**—rather than chasing shortterm gains, getting caught up in the hype—using reliable metrics of business success.
     Above all, **identify as clearly as possible what problems AI can really solve and how**.
  3. **Maintain a critical and balanced perspective about AI developments, no matter what people with vested interests may say, recognising the technology's potential and limitations**.
     There is no sci-fi AI coming, but it is an amazing technology that can be usefully and ethically integrated into countless processes, and lead to three kinds of changes (call them the **three E**): **doing more with less** (**Efficiently**), **doing things differently** (**Efficacy**) and **doing things for the first time** (**Entrepreneurship**).

Isključivo za potrebe PMI Forum 2024.

78

Why the AI Hype is another Tech Bubble ? (2024.)

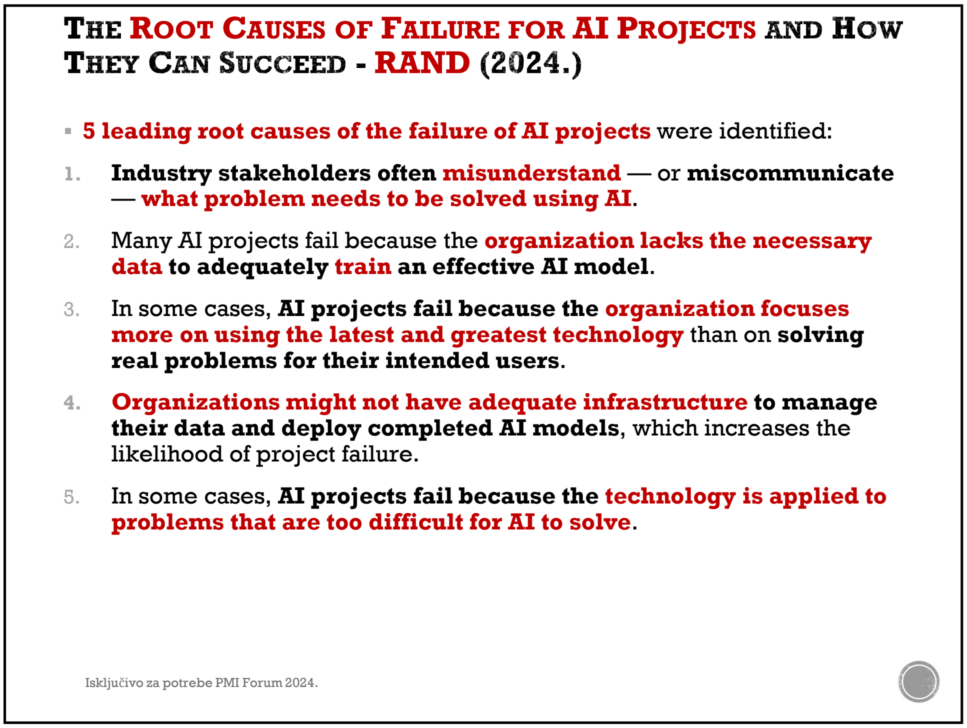
- **What can we do to minimise its negative impact?:**
  4. **Prioritise a longer-term perspective (years, not just months) that can help temper the boom-and-bust cycles associated with new technologies**, ethical considerations, societal and environmental impact, and forthcoming compliance issues, alongside technological advancement.
     Ethical and legal issues will not go away and generate more significant problems in the future if left rotting.
  5. **Support regulatory and governance** (including enforcement) **frameworks** that can keep pace with AI developments.
     Good regulation is the best ally of good innovation, not an enemy, because it provides more clarity and certainty.
     Be suspicious of those who want to play but want no rules or only their rules of the game.
  6. **Promote technological understanding, mass media information, and financial literacy** to **help people make better decisions**.

79

The Root Causes of Failure for AI Projects and How They Can Succeed - RAND (2024.)

- **To investigate why artificial intelligence and machine learning (AI/ML) projects fail, the authors interviewed 65 data scientists and engineers** with at least 5 years of experience in building AI/ML models in industry or academia.
- The **authors identified 5 leading root causes for the failure of AI projects** and synthesized the experts' experiences to develop recommendations to make AI projects more likely to succeed in industry settings and in academia.
- By some estimates, **more than 80% of AI projects fail** — **2x the rate of failure for information technology projects that do not involve AI**.
- Thus, understanding how to translate AI's enormous potential into concrete results remains an urgent challenge.

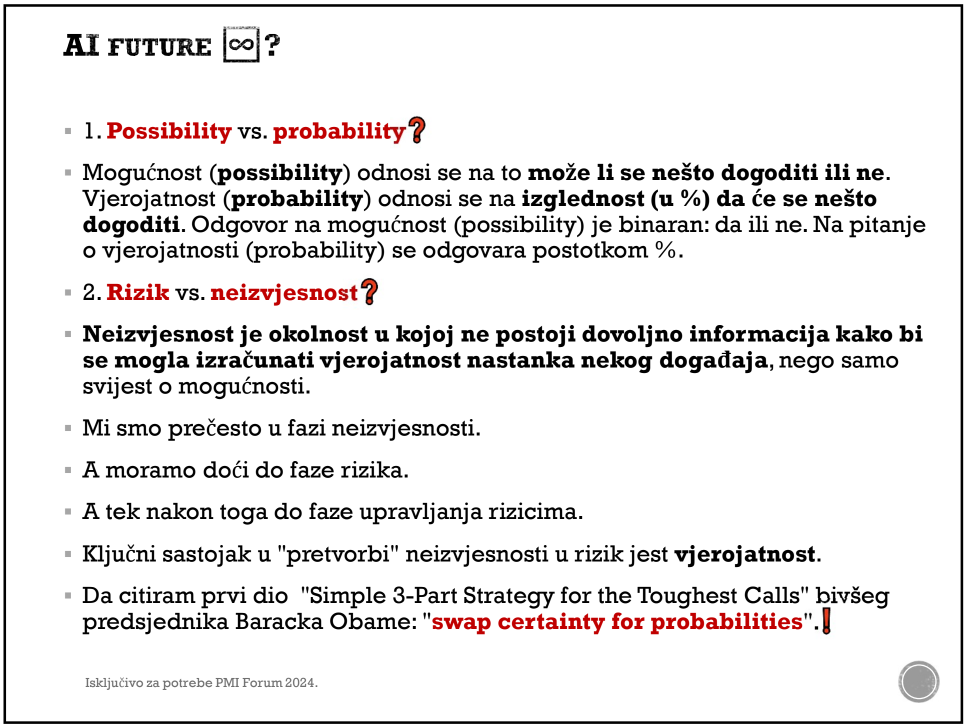
80



The Root Causes of Failure for AI Projects and How They Can Succeed - RAND (2024.)

- **5 leading root causes of the failure of AI projects** were identified:

1. **Industry stakeholders often misunderstand** — or **miscommunicate** — **what problem needs to be solved using AI**.
2. Many AI projects fail because the **organization lacks the necessary data to adequately train an effective AI model**.
3. In some cases, **AI projects fail because the organization focuses more on using the latest and greatest technology** than on **solving real problems for their intended users**.
4. **Organizations might not have adequate infrastructure to manage their data and deploy completed AI models**, which increases the likelihood of project failure.
5. In some cases, **AI projects fail because the technology is applied to problems that are too difficult for AI to solve**.



AI FUTURE ∞?

- 1. **Possibility** vs. **probability**❓
- Mogućnost (**possibility**) odnosi se na to **može li se nešto dogoditi ili ne**. Vjerojatnost (**probability**) odnosi se na **izglednost (u %) da će se nešto dogoditi.** Odgovor na mogućnost (possibility) je binaran: da ili ne. Na pitanje o vjerojatnosti (probability) se odgovara postotkom %.
- 2. **Rizik** vs. **neizvjesnost**❓
- **Neizvjesnost je okolnost u kojoj ne postoji dovoljno informacija kako bi se mogla izračunati vjerojatnost nastanka nekog događaja**, nego samo svijest o mogućnosti.
- Mi smo prečesto u fazi neizvjesnosti.
- A moramo doći do faze rizika.
- A tek nakon toga do faze upravljanja rizicima.
- Ključni sastojak u "pretvorbi" neizvjesnosti u rizik jest **vjerojatnost**.
- Da citiram prvi dio "Simple 3-Part Strategy for the Toughest Calls" bivšeg predsjednika Baracka Obame: "**swap certainty for probabilities**".❗

AI FUTURE ♾?

- 3. **Predviđanje** ili **pogađanje**❓
- **Predviđanje** se odnosi na primjenu analitičkih tehnika s ciljem iznalaženja odgovora na pitanje: "**što bi se moglo dogoditi**❓".
- Predikcija se odnosi na primjenu analitičkih tehnika s ciljem iznalaženja odgovora na pitanje: "**što će se vjerojatno dogoditi**❓".
- Ključ su analitičke tehnike❗ I vjerojatnost❗

…

- **Pogađanje**❓
- Često sam znao čuti rečenicu: "**dobro si pogodio**❗"
- Sorry, ali ne bavim se pogađanjima. Mnogi su dobri u pogađanjima.🔮
- **Isto tako i pokvareni sat pokazuje točno vrijeme dva puta dnevno.**♾

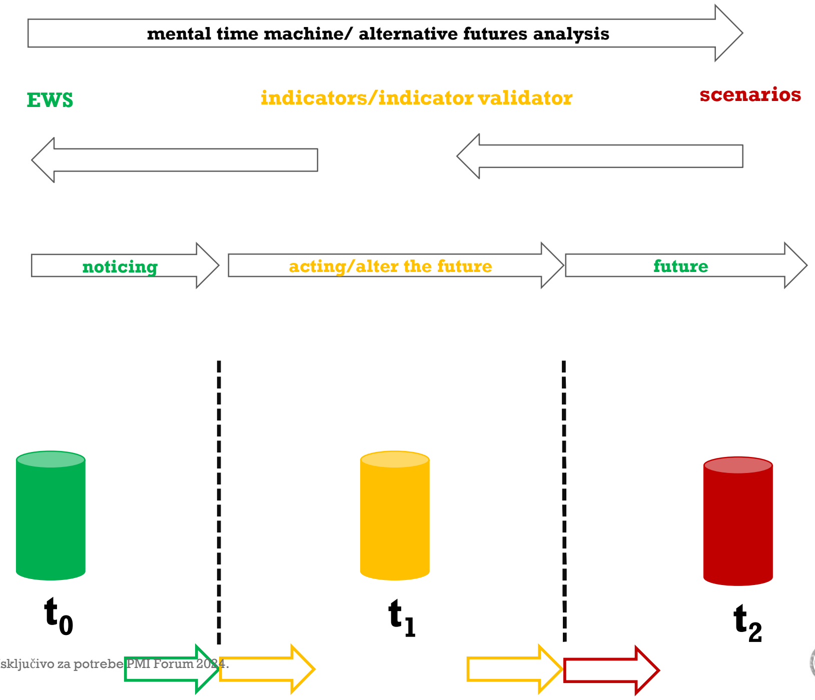
83

3 THREATS TO HUMAN EXISTENCE (2025.)?

1. **nuclear war**
2. **technological disruption/AI**
3. **climate change/ecological collapse**

Yuval Noah Harari

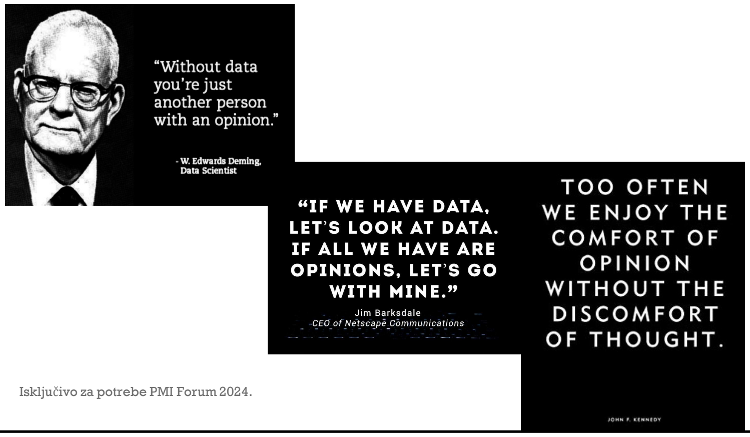
84


mental time machine/ alternative futures analysis
EWS
indicators/indicator validator
scenarios
noticing
acting/alter the future
future
t_0
t_1
t_2

85

...AND NOW ON OPINIONS AND FACTS...

- **Opinions don't affect facts.**
- **But facts should affect opinions, and do, if you're rational.**



"Without data you're just another person with an opinion."

- W. Edwards Deming, Data Scientist

"IF WE HAVE DATA, LET'S LOOK AT DATA. IF ALL WE HAVE ARE OPINIONS, LET'S GO WITH MINE."

Jim Barksdale
CEO of Netscape Communications

TOO OFTEN WE ENJOY THE COMFORT OF OPINION WITHOUT THE DISCOMFORT OF THOUGHT.

JOHN F. KENNEDY

86

AI
- neka otvorena pitanja?

2024.

izv.prof.dr.sc. **Robert Kopal**

Isključivo za potrebe PMI Forum 2024.

87