TECHNOLOGIES
Bring the future to humanity

State of AI 2024,

20 January 2024

Introduction

This report is intended to summarize the advancements in AI and its media perception in the last 12 months (January- December 2023): the progress in research, specific AI news in the media and emerging trends (catalysts) which will impact the near future. Our understanding of the AI advancements and the trends allow us to make 5 predictions we believe will happens in the next 1-2 years (last 5 slides). We will review accordingly in one year time.

Author

Andrea Isoni

Chief AI Officer

Physicist and PhD from Imperial College, Former Founders Factory (FF), Chief AI Officer at AI Technologies. Experienced in leading edge AI algorithms and how to translate them into business applications (including pioneering Human Machine Interface systems). Advisor of Antler Venture Capital and writer in AI ('Machine Learning for the Web' published by Packt in English, Chinese and Korean and sold more than 10,000 copies). Committee member of ISO and IEEE for the AI standards and International AI speaker (London, New York, Dubai, Saudi Arabia). AI Newsletter 'Thoughts about AI by a Human' (2.4k+ subscribers).

TECHNOLOGIES
Bring the future to humanity

2023

Notes on the Report

*

Last year predictions' span was 1-2 years (by 2025). We were very strict: we considered wrong predictions that have 'time to happen'.

*

'Secular Trends' means trends that are developing in the last 1-2 decades or more. These trends are always impacting all predictions.

*

'Catalyst Trends' means trends and event developing in the last 5 years or less. These trends impact some predictions.

*

On prediction 1 and Trend Catalyst 1 we considered mostly EU AI Act. We believe similar situations will happen in other jurisdictions. There may be a delay.

*

The BONUS prediction is half a joke, half true. We believe many more will grasp some coding terms and that may naturally spill out into common conversations.

Technologies

2022

Catalyst Trends

Download 2022 report here

01

Governance AI: EU AI regulations are coming by 2024.

02

Fewer births + increased data volume.

03

Cloud adoption growth.

04

Adoption of AI transformers in music and images by 2024: Expected 1b revenue of chatGPT, Text2Music ‘good’ products coming in 2023, Language accents models are becoming mature, Sensational and controversial news of text2image.

05

Largest Models will surpass 1T parameters + raise synthetic data.

Checks

01

Happening. Regulations are coming but slower than expected.

02

Happening. This trends is certainly continuing if not worsening.

03

Happening. Fostered by AI and automation actually cloud adoption is raising even more.

04

Happening. Audio generation is now used by many users (MusicLM, MusicGen). chatGPT surpassed 1B revenue. We had Getty suing Stable Diffusion. We do see massive improvement on text2speech (and accents)

05

Partially happening. Models have surpassed 1T parameters, reasoning was limited this year and sometimes LLMs seem under-trained. Missed on synthetic data.

TECHNOLOGIES

Bring the future to humanity

2022

Predictions

Download 2022 report here

01

Adoption of simple models in industry.

02

Human machine hybridization, chatGPT just last tool.

03

MLOps will be combined with digital twins.

04

Emerging of awareness of privacy and **'personal'** IP.

05

Limit of current 'quantity' method in AI will be apparent.

Outcomes

01

Partially right. Mixed adoption of genAI tools (Gartner believes we are the peak of hype), unregulated industries (marketing) are using more (https://bit.ly/3OehebI). We cannot claim the AI regulations is the cause yet. It is early to know.

02

Partially right. chatGPT then perceived as 'super-human' will be reduced as 'just a tool'. We said 40% adoption by 2025 and there is 14% adoption by 2024 (US market) https://bit.ly/3U5FoJq. We missed on adoption but got perception right.

03

Wrong. We realised that it will take more time for digital twins to get traction especially in the consumer market. We may review it in 1 year or more but it will not happen in few years as we expected.

04

Right. **Personal** lawsuits over IP rights: on October 23 artists sued Stable Diffusion, MidJourney and DevianArt over infringement of copyrights.(http://bit.ly/42cYlfo). Also commercial cases (Shutterstock and NYT vs Open AI),

05

Right. We got the right understanding on why increasing the number of parameters was likely not improving performances (data scarcity, architecture). This has been confirmed by Sam Altman and other research published. (https://bit.ly/4b4Ct9Y)

TECHNOLOGIES
Bring the future to humanity

2023

Best AI advancements

01

‘QLoRA: Efficient Finetuning of Quantized LLMs’ --> Methodology to reduce a 70B model (LLama2) to fit in 40Gb GPU or less. Seminal to many quantization papers to run/train LLM in few GB of RAM space.

02

‘Direct Preference Optimization’ --> A method that shows Human Feedback (HF, used by Llama and gpt versions) can be overcome without losing performance.

03

‘Mistral 7B paper’ ---> It shows an open source with just 7B outperform Llama and comparable to gpt 3.5 (seminal to Zephyr and Mixtral and open source 'small' model with great performances).

04

‘ConvNets Match Vision Transformers at Scale’--> It shows that 'traditional' CNN are still competitive with transformers if trained with the same amount of data (Data more important than architecture).

05

‘Segment Anything’--> Advancement on image segmentation and lead to other projects SAM-based, like EdgeSam, MobileSAM, and EfficientSAM.

06

‘HyenaDNA’--> First LLM trained on human genome data to predict DNA profiles (with orders of magnitude fewer parameters and pretraining data).

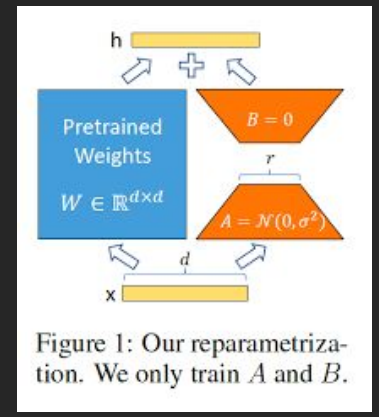
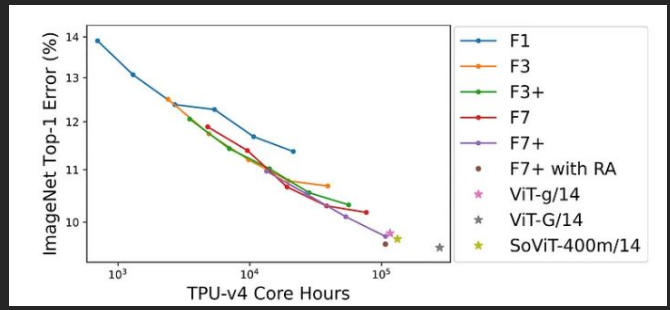
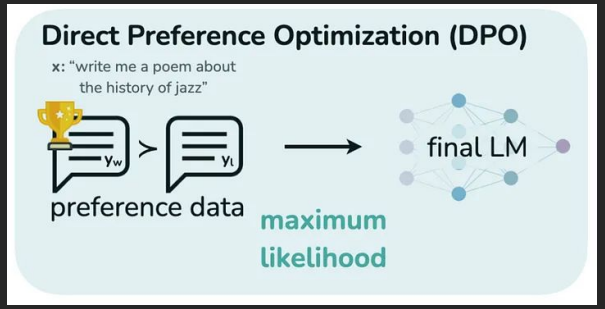
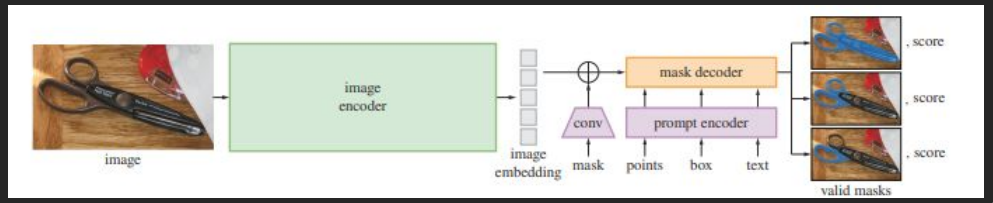


Figure 1: Our reparametrization. We only train A and B.







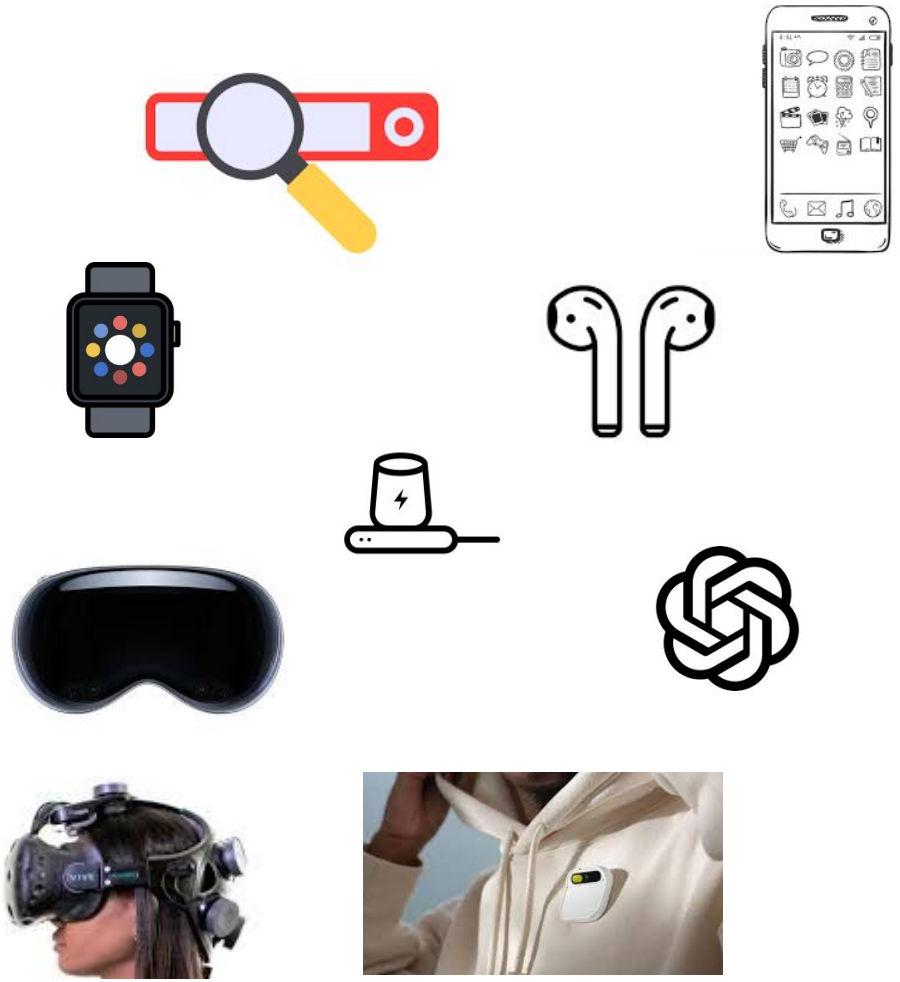
Technologies

Secular Trends 1

Human Machine Hybridization

Humans are progressively integrating with technology, hardware and software.

- Mobile phones, search engines (2000s)
- Smart Watch (2010s), iPods
- Smart Home (2015s)
- **Now (2020s), (chatgpt) AI assistants, VR headset**
- Future.. wearable pins, Brain Computer Interface

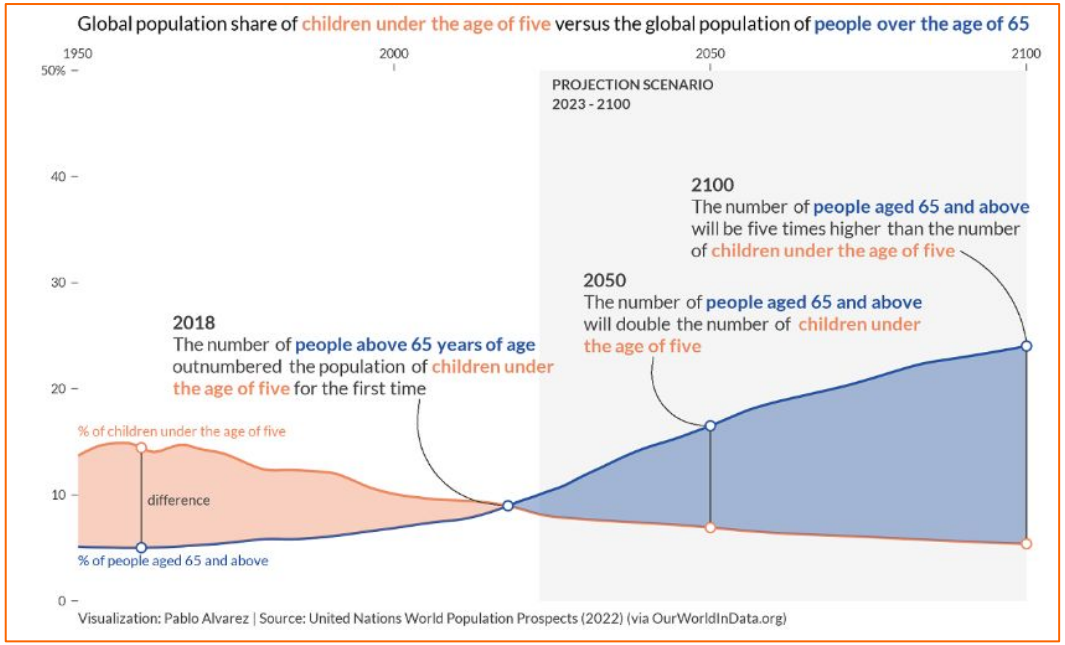


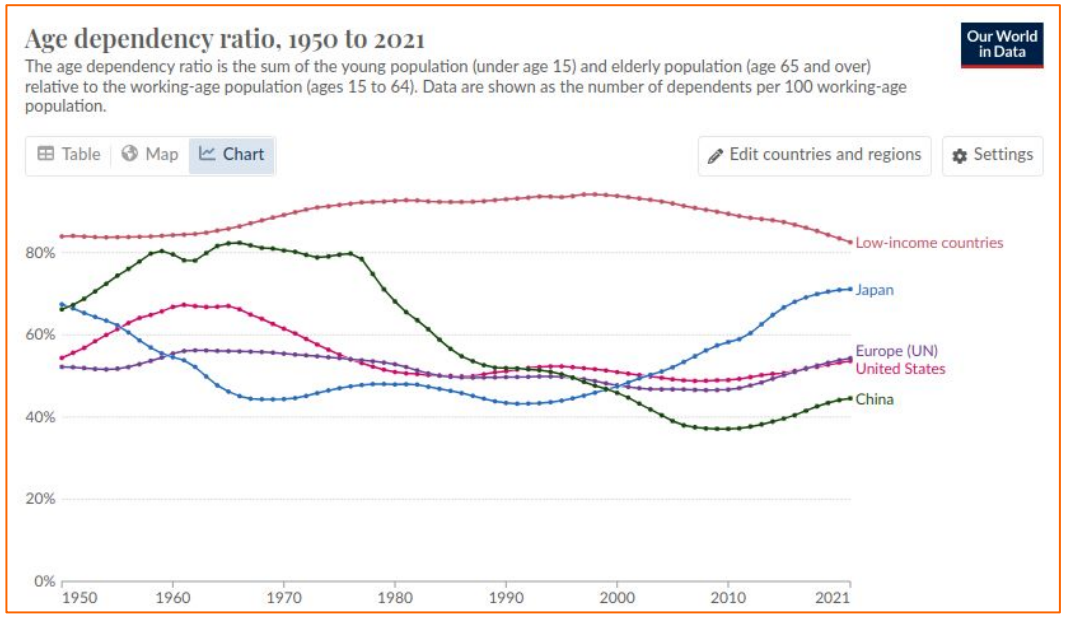
TECHNOLOGIES
Bring the future to humanity

Secular Trends 2

Aging

Most developed and advanced countries have an increasingly older population with births rates less than 2. Age dependency ratio (bottom right) is rising.

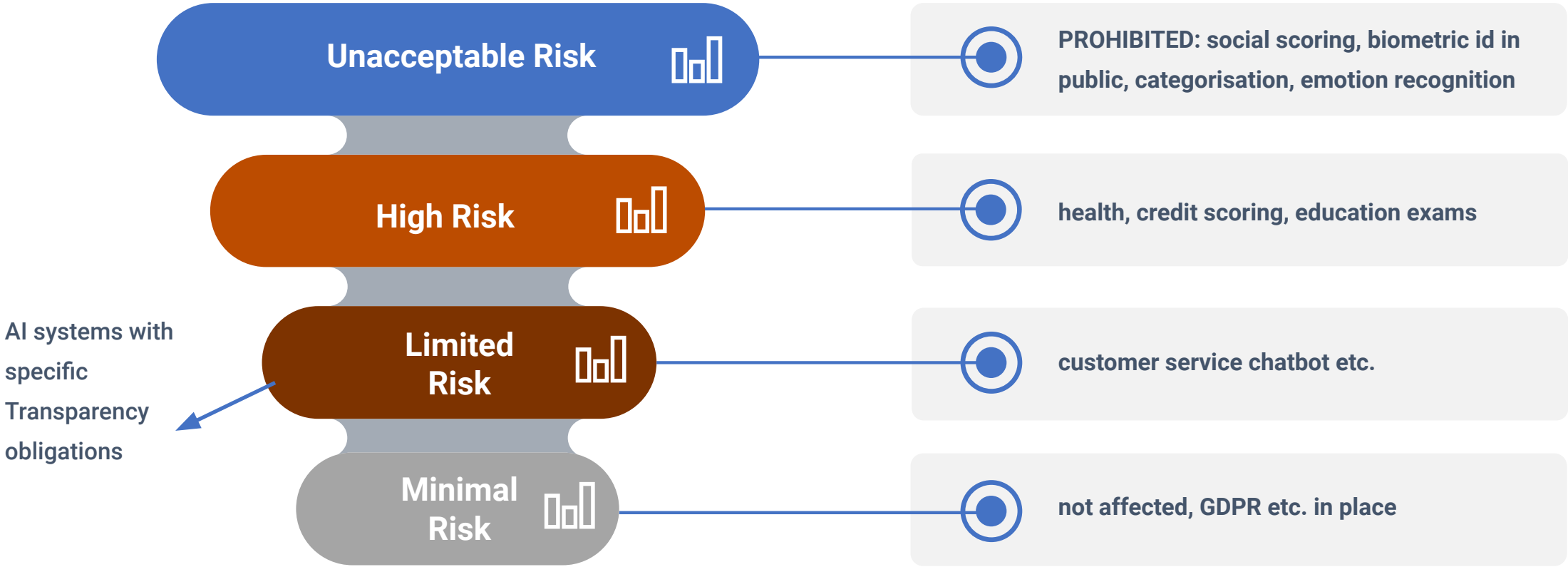




Trends Catalyst 1

governance AI: EU (and others) AI regulations coming in 2024

AI Risk Categories



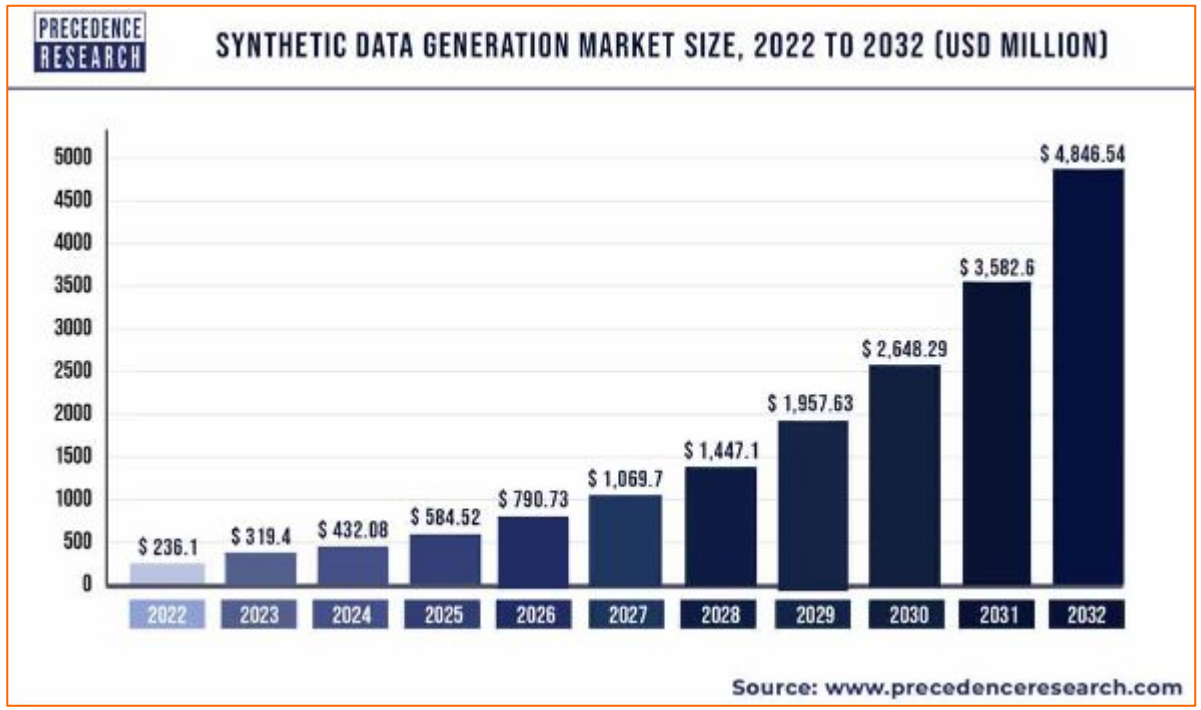
NOTE: it will take 6 (prohibited) to 24 (general purposes) months to enter into force

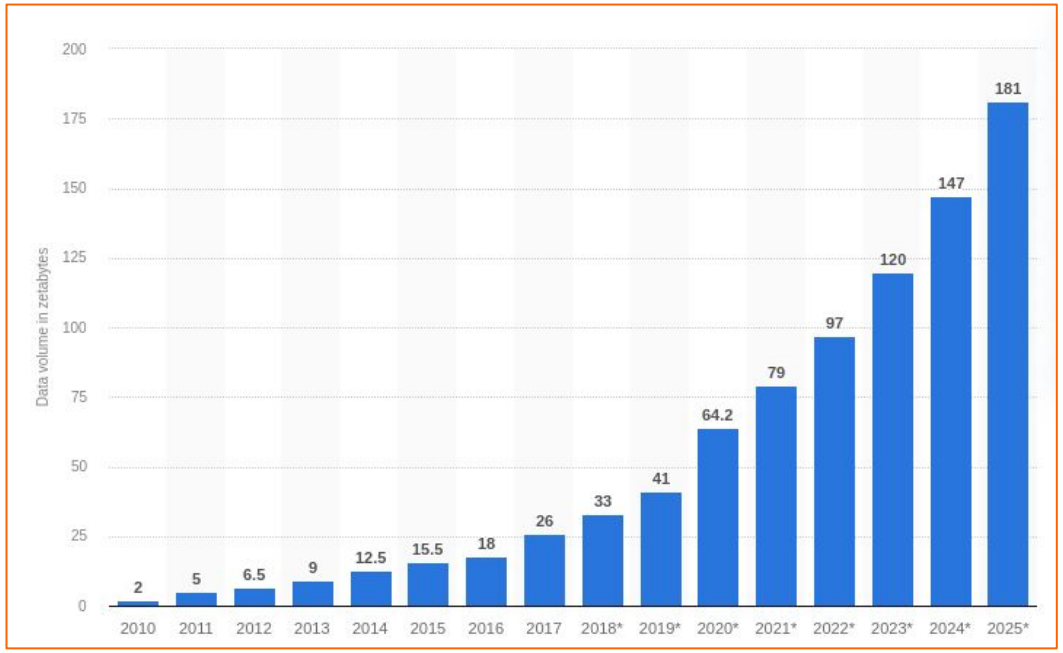
TECHNOLOGIES
Bring the future to humanity

Trends Catalyst 2

synthetic VS increase data generation

Human Data generation is increasing as well as synthetic data. Both may be enough to train LLMs (since models' datasets are not open source we cannot state if the data was synthetic/not and in which ratio).





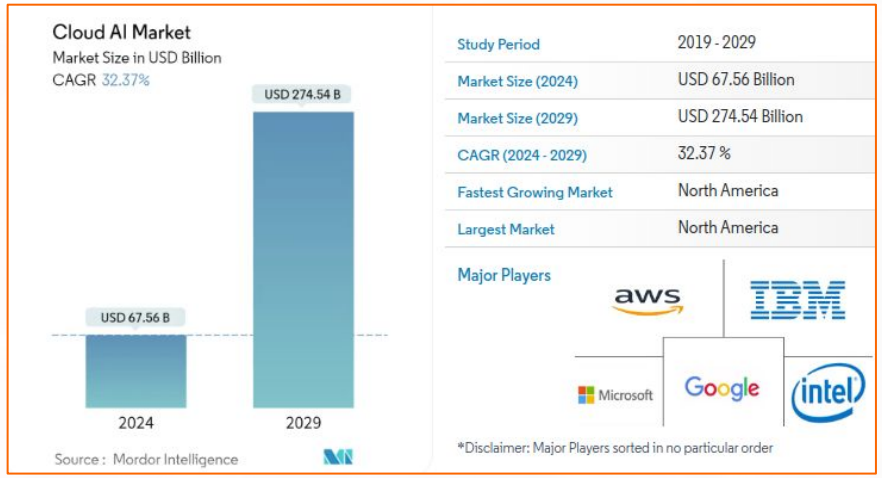
Volume of data/information created
(statista.com)

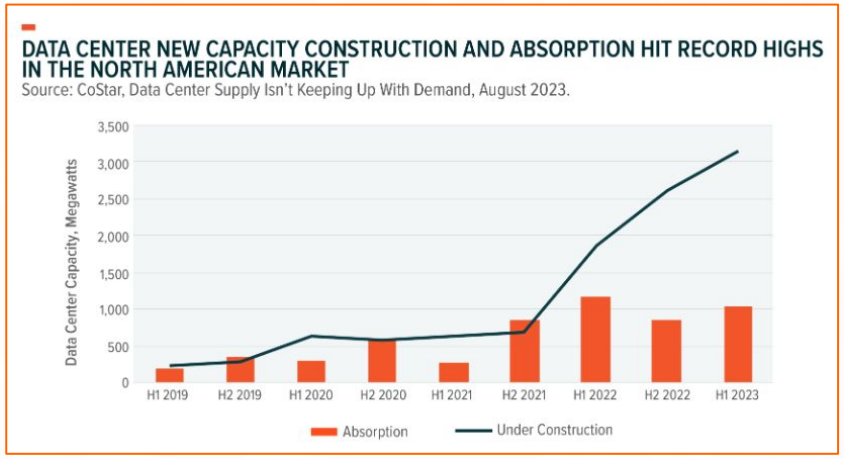
TECHNOLOGIES
Bring the future to humanity

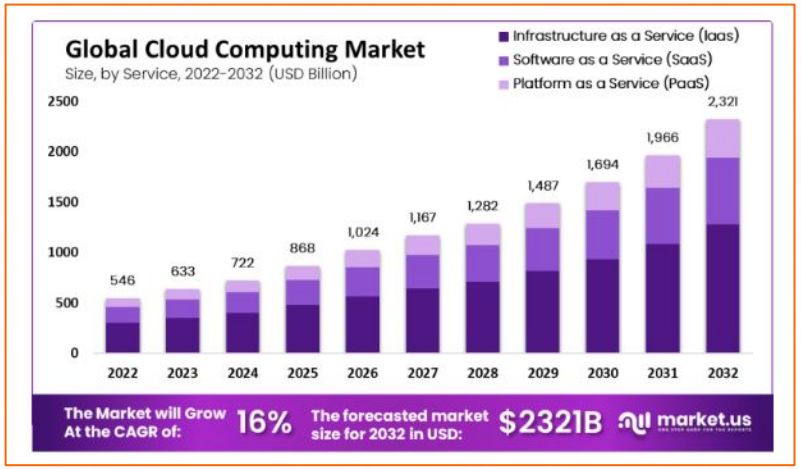
Trends Catalyst 3

cloud adoption growth

Cloud infrastructure is growing and becoming more ‘sophisticated’ (AI, automation, data platforms etc.) with more services on top of the infrastructure. Demand is stronger than ever.



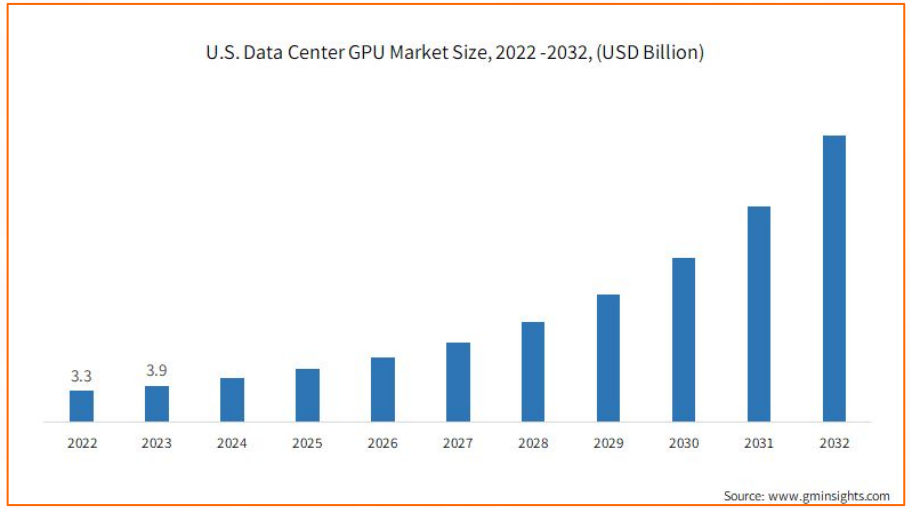


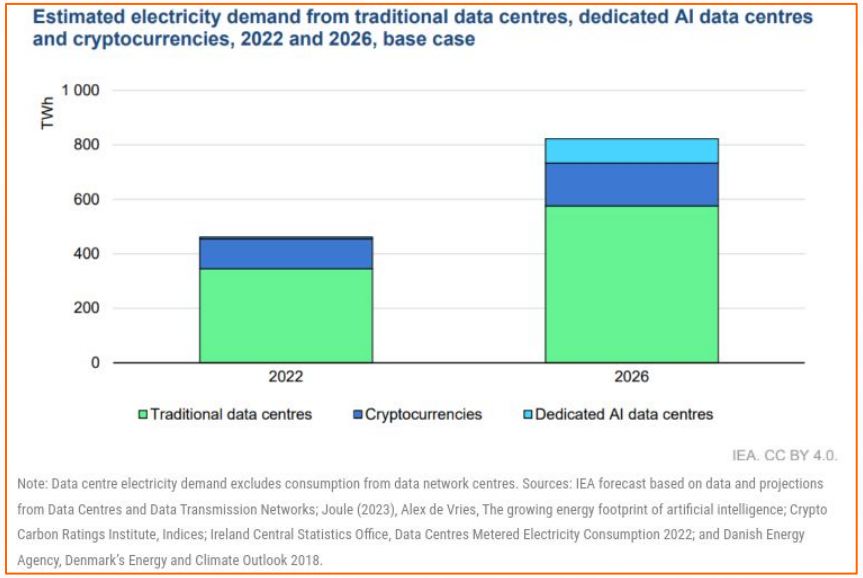


TECHNOLOGIES
Bring the future to humanity

Trends Catalyst 4

AI/crypto hardware demand and energy consumption





As crypto, AI research and applications are raising hardware demand and energy accordingly:

- By end 2024 just Meta will own 350k/out of 2.5m H100s GPUs, 14% of the market.https://www.ft.com/content/c7e9cfa9-3f68-47d3-92fc-7cf85bcb73b3
- Chatgpt has 2-3% search market and inference cost 10-1000x more today than search query. As users use more LLMs (inference) than search (query), energy consumption increases
- LLMs additional functionalities like multimodality and the longer token limits ('context') will contribute to increase energy consumption
- Bitcoin ETF approval, Bitcoin halving (April 24) , potential other ETFs will increase price and make mining profitable again

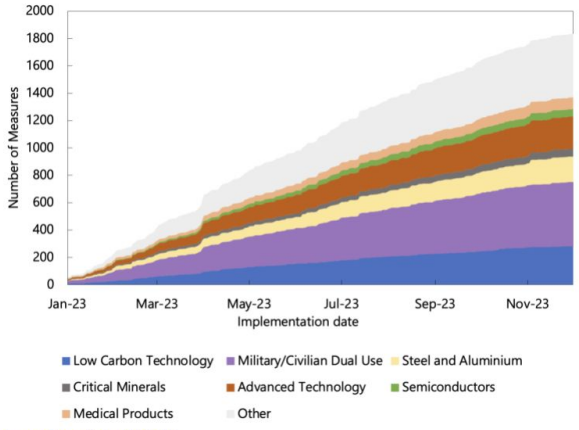
TECHNOLOGIES
Bring the future to humanity

Trends Catalyst 5

Supply chain localisation and workforce

Countries trying to be less vulnerable are localising supply chains and implemented ‘industrial policies’.

Figure 5 Trade distortive industrial policies by sector



Source: Evenett et al. (2024).

Rank	Jobs	People	Examples
1.	Office & Admin	13	Receptionists, Clerks, Customer service, Secretaries.
2.	Health	11	Doctors, Nurses, Paramedics, Vets, Orderlies, Personal care aides.
3.	Transport	9	Warehouse workers, Packagers, Pilots, Ambulance, Bus, Truck, Taxi drivers, Ship captains.
4.	Sales	9	Sales representatives, Counter clerks.
5.	Food	8	Food preparers & servers, Bartenders, Dishwashers, Hosts.
6.	Management	6	Legislators, Chief executives, Directors, General & Operations managers.
7.	Business & Finance	6	Accountants, Auditors, Financial analysts, Logisticians.

After the Pandemic and given the political worldwide instability , supply chain is getting localised or shifted to ‘nearshoring’ countries:

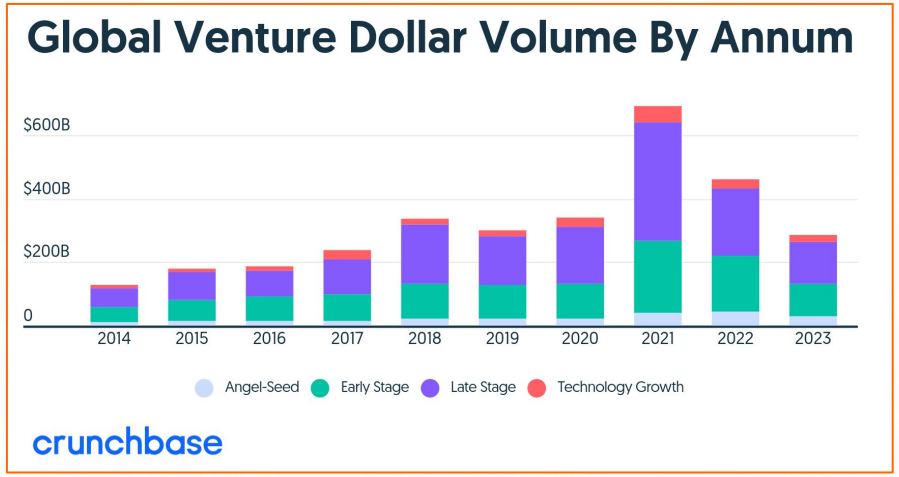
- GPU manufacturer TSM (Taiwanese company) opening factories in USA
- AI and data regulation implies data storage and processing within a country/economic area
- US and other countries bans on export of chips and related technology
- the largest workforce in US is employed in Admin (13%) while manufacturing is 6% https://bit.ly/3SeT81J

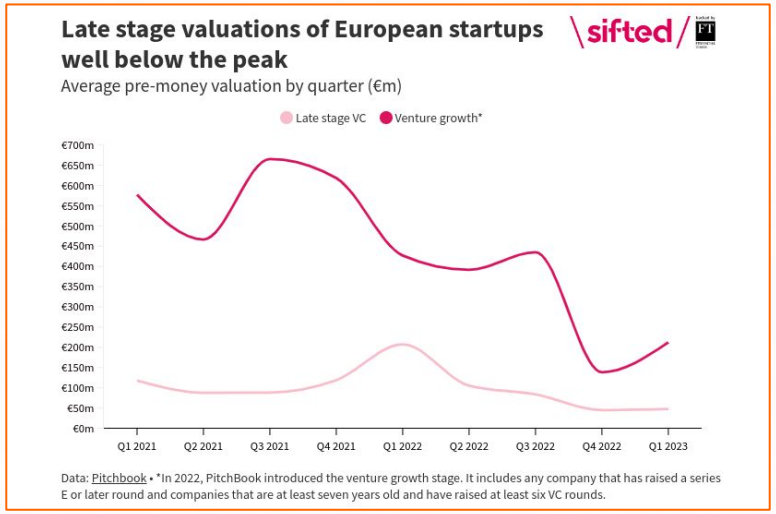
TECHNOLOGIES
Bring the future to humanity

Trends Catalyst 6

AI investment

Due to rising interest rates and other factors, tech investments have been cut.





Venture capital investments in tech startups continue to slow down:

- Global startup investment in 2023 reached $285 billion: a 38% decline year over year, down from the $462 billion invested in 2022
- Per Crunchbase, AI funding was up 9% in 2023 to 50b BUT 36% (18b) went to 3 foundation model companies (OpenAI, Anthropic, InflectionAI). If you take these 3 companies out, the investment in AI will be less (32b in 2023 compared to 45.8b in 2022)
- Evaluation of startups are decreasing. Per Pitchbook data and Sifted, there is decline in EU valuations and the bump in Q4 23 is due to 1 deal with solar startup Enpal. Same per Crunchbase worldwide

TECHNOLOGIES
Bring the future to humanity

Trends Catalyst 7

adoption of genAI in music and images, videos

Models and services to generate, music, images and videos keep improving:

- Expected continuing ‘consolidation’ on market share of images generation by DALL-E, Midjourney and few others (no major new player expected)
- Text2Music keeps progressing : in 2023 MusicLM (Google), MusicGen (Meta) were state of the art
- Text2Video based on transformer or diffusion models improved with faster training and better resolution (VideoLDM, MAGVIT)
- 3D from 2D images: while NeRF improved, a new method achieved better quality and faster training (‘3D Gaussian Splatting for Real-Time Radiance Field Rendering’)
- Video editing advancements continue like InstructPix2Pix, Gen-1,Drag Your GAN. Commercially there is a slight advantage by Runway, Pika but it is expected other players will emerge

TECHNOLOGIES
Bring the future to humanity

Adoption of AI models in industry→STILL ON FROM 2023

Trend Prediction 1

- AI regulations (EU) will impact in 2025-26
- Most of the models are not explainable neither transparent (see picture below)
- AI drone (swift) already won races https://bit.ly/3UixbBE
- Mild/none public interest in Global AI regulations
- Mixed adoption of genAI tools: Gartner believes we are the peak of hype, unregulated industries (marketing) are using more https://bit.ly/3Oehebl

↓

- More regulated the industry AND less explainable models→ more governance costs
- BigTech 'mild' genAI indemnification on liabilities

Foundation Model Transparency Index Scores by Major Dimensions of Transparency, 2023

Source: 2023 Foundation Model Transparency Index

Major Dimensions of Transparency	Meta Llama 2	BigScience BLOOMZ	OpenAI GPT-4	stability.ai Stable Diffusion 2	Google PaLM 2	ANTHROPIC Claude 2	cohere Command	AI21labs Jurassic-2	Inflection Inflection-1	amazon Titan Text	Average
Data	40%	60%	20%	40%	20%	0%	20%	0%	0%	0%	20%
Labor	29%	86%	14%	14%	0%	29%	0%	0%	0%	0%	17%
Compute	57%	14%	14%	57%	14%	0%	14%	0%	0%	0%	17%
Methods	75%	100%	50%	100%	75%	75%	0%	0%	0%	0%	48%
Model Basics	100%	100%	50%	83%	67%	67%	50%	33%	50%	33%	63%
Model Access	100%	100%	67%	100%	33%	33%	67%	33%	0%	33%	57%
Capabilities	60%	80%	100%	40%	80%	80%	60%	60%	40%	20%	62%
Risks	57%	0%	57%	14%	29%	29%	29%	29%	0%	0%	24%
Mitigations	60%	0%	60%	0%	40%	40%	20%	0%	20%	20%	26%
Distribution	71%	71%	57%	71%	71%	57%	57%	43%	43%	43%	59%
Usage Policy	40%	20%	80%	40%	60%	60%	40%	20%	60%	20%	44%
Feedback	33%	33%	33%	33%	33%	33%	33%	33%	33%	0%	30%
Impact	14%	14%	14%	14%	14%	0%	14%	14%	14%	0%	11%
Average	57%	52%	47%	47%	41%	39%	31%	20%	20%	13%	

Scores for 10 major foundation model developers across 13 major dimensions of transparency.

→

By 2025 :

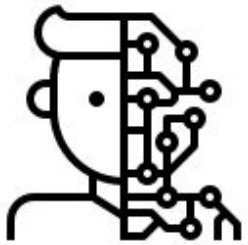
1. Commercial AI will be widely limited to 'simple/small' AI models in any applications classified as high and limited risks (sensitive data and human safety tiers)
2. We would see sophisticated AI models only on industries with low exposure to high costs of governance (and/or fines)
3. AI investment will be focused only on industry less affected by regulations and market leading foundation models
4. First AI drone controlled device used in battleground

→ **Triggered from Catalysts:** 1,3,5,6

Technologies

AI startup folded or bought at low price

Trend Prediction 2



AI startups on specific tasks are in jeopardy after the release of chatgpt.

- Chatgpt showed excellent performances in many tasks (commoditization)
- Investment in AI reduced and the ‘consolidation’ of investment in few ‘winners’
- Microsoft research showed that small specialised models not as good as large ones (“overwhelmed” if trained on uncurated datasets)
  https://arxiv.org/pdf/2305.07759.pdf

→

By 2025 : one ‘narrow’ AI related startup which

- raised $100m or more
- has 1B valuation or more

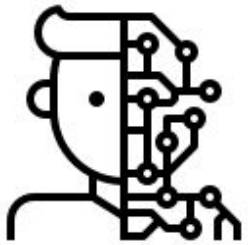
will be bought at less than 50% of last round or shut down (examples: a grammar checking one?, write assistant etc.)

→ **Triggered from Catalysts:** 4,6,7

Technologies

AI misuse drives cybersecurity and AI related spending

Trend Prediction 3



AI is be used for malicious activities and it will trigger counter measures that also leverage AI

- Deepfakes (text, images, videos, audio) will not be distinguishable from human generated data
- Cyber attacks are increasing https://bit.ly/3SqDNOr
- AI can easily boost cyber attacks (phishing etc.) https://bit.ly/3SsMwhH
- AI in cybersecurity is expected to more than 38B by 2026 https://www.statista.com/statistics/1291380/ai-in-cyber-security-market-size/

→

By 2025 :

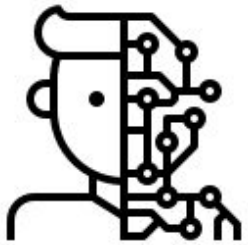
- AI fake detection tools start failing often and fade out of radar
- at least 15% of new art/images will have AI watermark to protect originality of work and IP (like Google SynthID)
- A genAI startup cybersecurity vendor will be acquired by a large cyber player for more than $100m

→ **Triggered from Catalysts:** 1,2,6,7

Technologies

adoption of LLMs/GenAI tools by consumers

Trend Prediction 4



Human and machines are going to integrate themselves and there will be AI assistants to advance in this direction and increase productivity

- 14% of US population used at least once genAI https://bit.ly/3U5FoJq
- Population is aging and increasing portion is 65+
- After US, India and Philippines have the largest traffic of chatGPT (Colombia is there too for now) https://bit.ly/42czqbl
- Productivity increases using genAI tools especially for less experienced professionals https://bit.ly/498yVSn
- Emails sent and received (so workload) will increase by 13% in 2026 https://bit.ly/498DnAp

→

By 2025 :

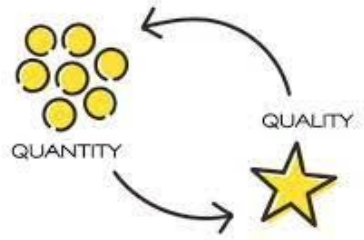
- chatGPT and competitors combined used at least once by 30% of the US workforce
- Top 3 non first world (and excl. China) countries with large, outsources workforces (like India etc.) will have a combined 20% of traffic of chatGPT and competitors
- US admin workforce will not decrease substantially due to AI
- GDP due to increased AI productivity will be calculated by at least 1% for the year

→ **Triggered from Catalysts:** 2,3,5,7

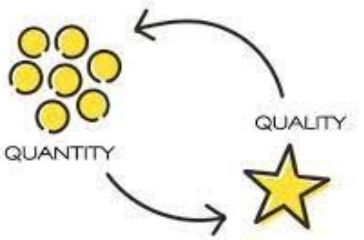
Technologies

Big Tech lead foundation, labs efficiency and fine tuning

Trend Prediction 5



Advancement in AI performances will be from the usual Big Tech while labs and small teams will make efficient, 'usable' models



- AI GPUs and money is in the hands of few large companies (Meta, Microsoft, OpenAI etc.)
- Some Big Tech may open source their code and data (Meta open source Llama but not the data, OpenAI none)
- New general LLMs will be multimodal, longer tokens
- Labs (Mistral etc.) have brains but less GPUs resources available

→

By 2025 :

- the leading performance and breakthroughs will be from Big Tech models and:
  - RLHF will be matched by some other methods
  - a model which is **not** transformer based reaches gpt-3.5 level
- a Labs will reach GPT-3.5 benchmarks but with :
  - 100GB of GPU memory or less
  - less than 20B parameters

→ **Triggered from Catalysts:** 2,4,6,7

Technologies

BONUS: We will talk Pynglish

Trend Prediction *Bonus* (semi-serious one)

Non coders will start picking up some code language

- Developers are adopting copilot and alike AI software development tools quickly https://survey.stackoverflow.co/2023/#ai
- Interns and others software related professionals (project managers etc.) learn becomes proficient in coding faster using AI tools https://bit.ly/498yVSn
- Demand for developers overall will not shrink (apart from front end)

→

By 2027:

- 10% of non-coders professionals will at least try genAI tools to make their one software scripts
- non-coders will pick up developers jargon in common life
- An authoritative English dictionary (like Oxford English Dictionary) will add a software word as a new recognised word

Technologies

CONTACT US

andrea@aitechnologies.co

Andrea Isoni, Chief AI Officer, PhD

AI Technologies

TECHNOLOGIES

Bring the future to humanity