SKI
tag

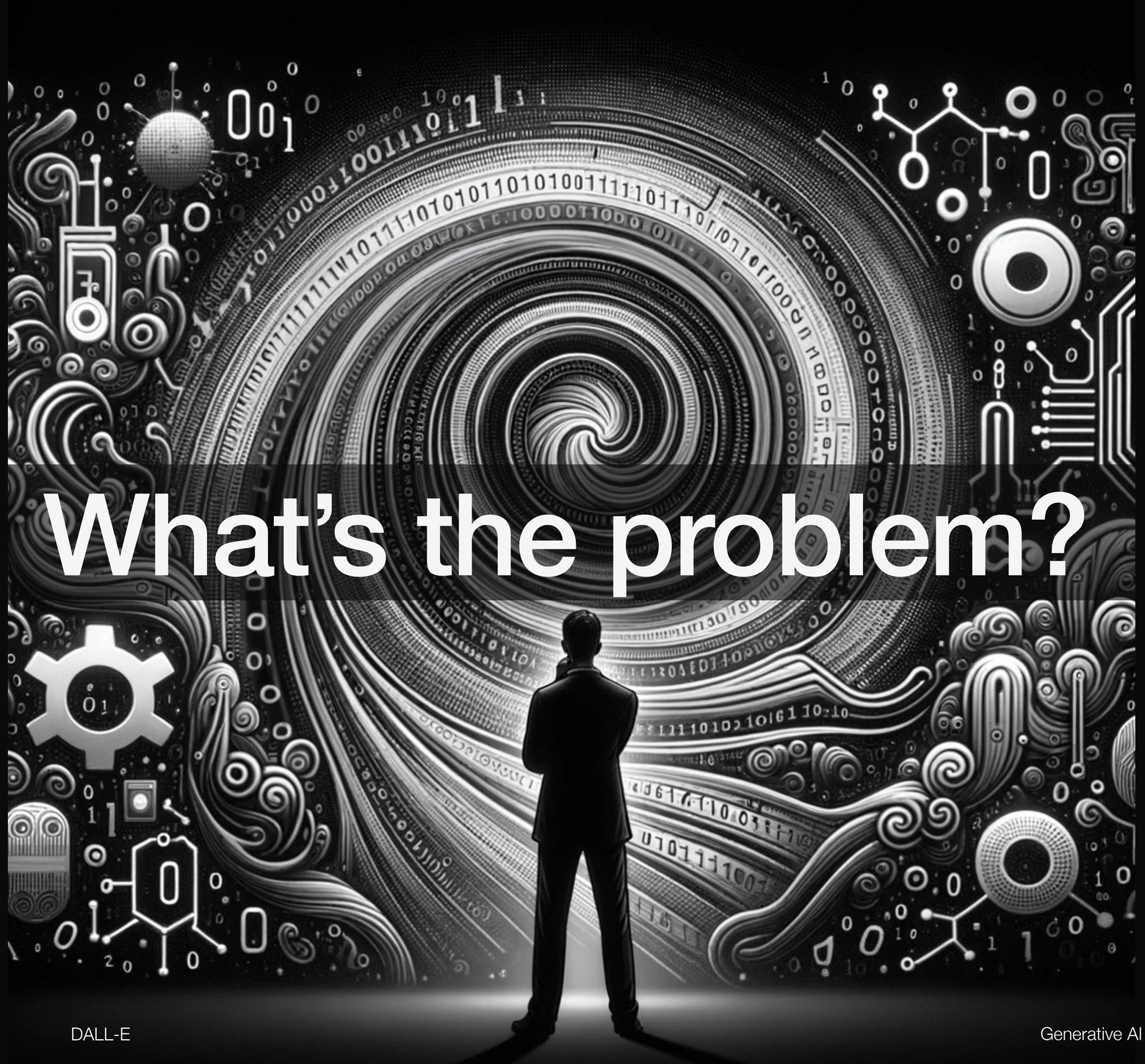
Generative AI

What’s the problem? Ethics and Regulation

January, 2024

SKI tag

What’s the problem?

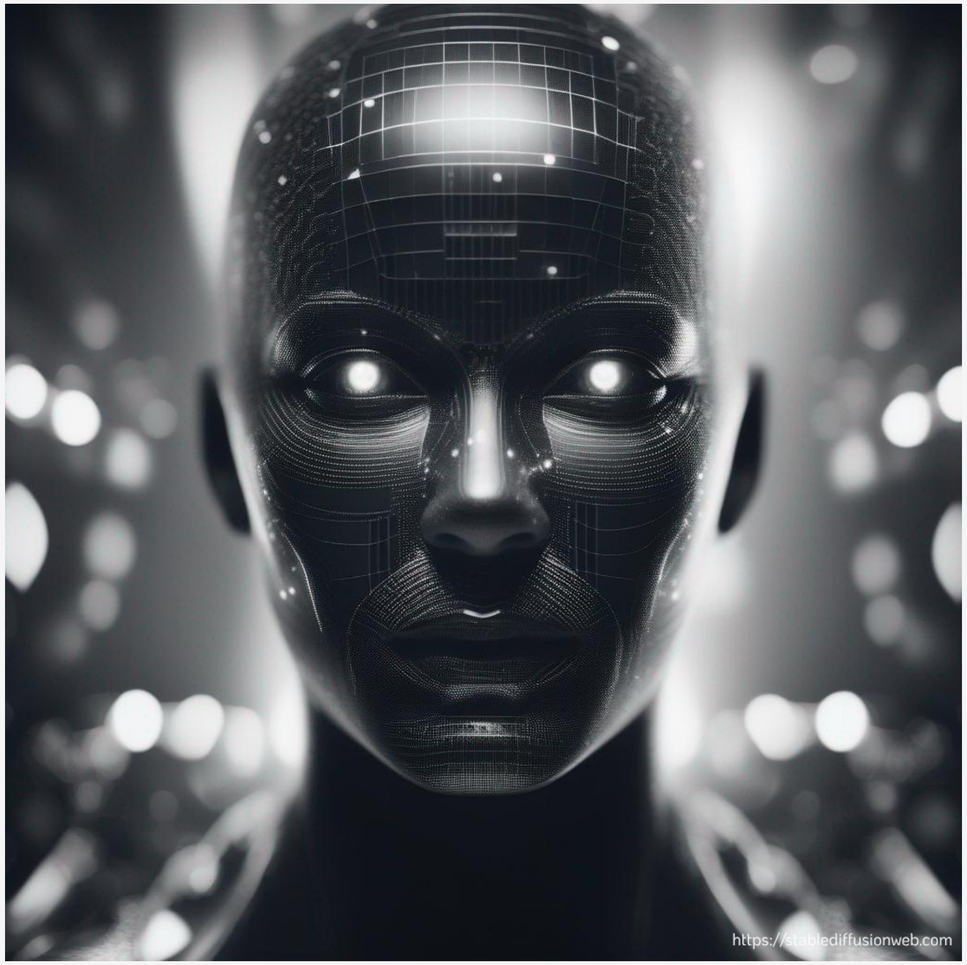


DALL-E

GenAI - Ethics and Regulations

Planning

- Debate: The Times against OpenAI.
- AI Systems.
- AI Lifecycle.
- Narrow AI.
- Generative AI.
- Read Teaming: GenAI Risk Assessment.
- General Purpose AI.
- Post Mortem: GPAI “Super-alignment”.



by Stable Diffusion XL

The Times Sues OpenAI and Microsoft

Debate

SKI tag

“The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work”, The New York Times, December, 27th 2023, https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html

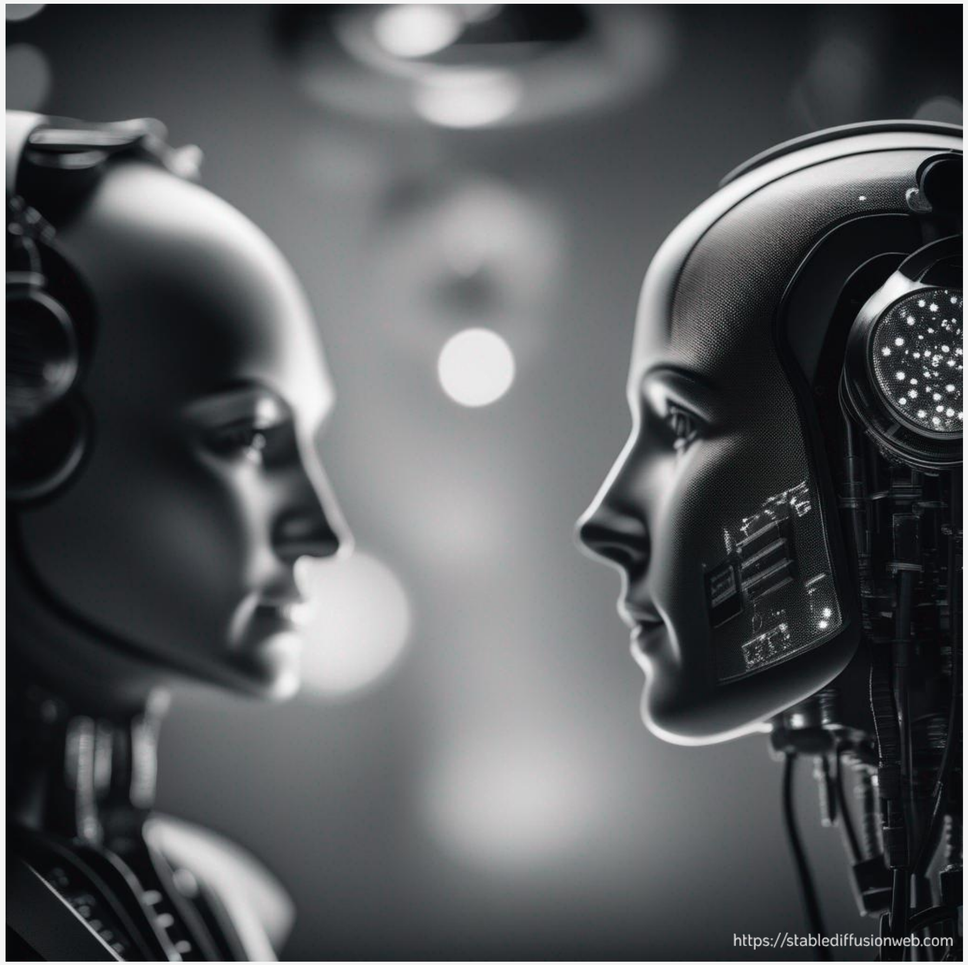
by DALL-E

AI Systems

Definition

SKI tag

"An AI system is a machine-based system 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. Different AI systems vary in their levels of autonomy and adaptiveness after deployment." **OECD**



by Stable Diffusion XL

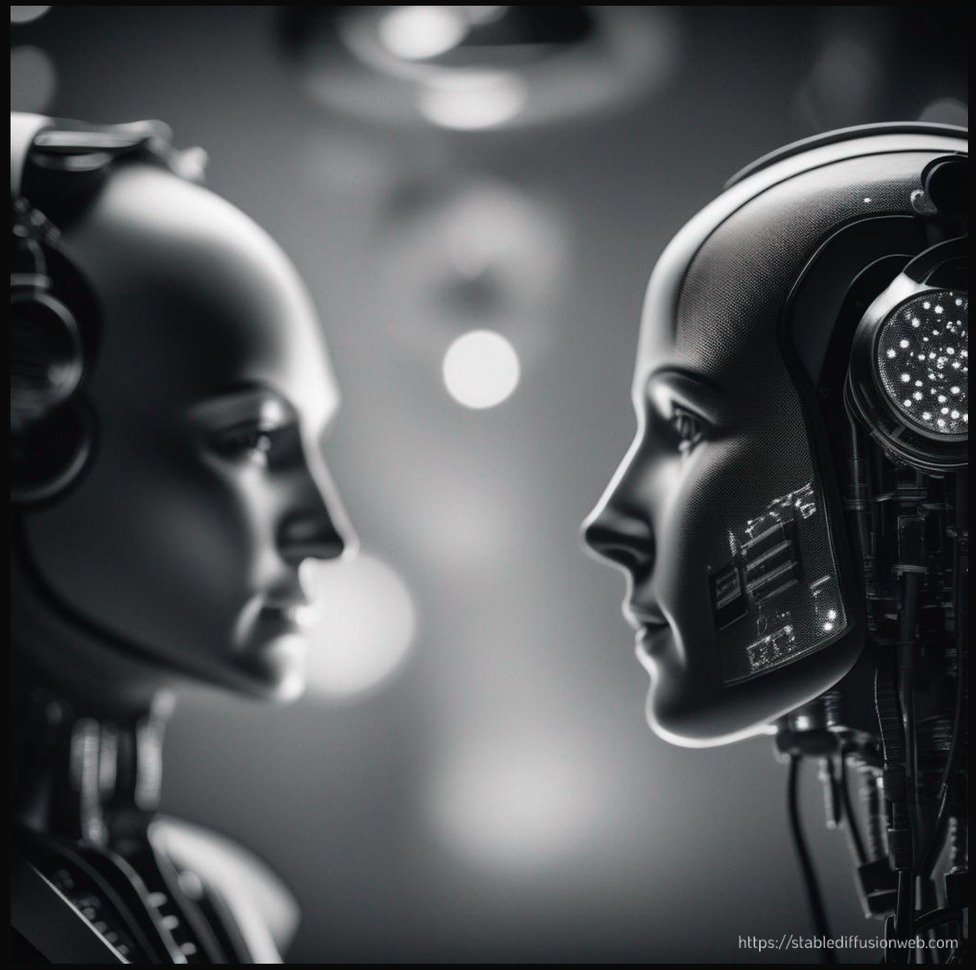
AI Systems

Human-centric. Autonomy.

Promote a human-centric and trustworthy interaction with a widely adopted autonomous AI.

AI has positive effects, like improve welfare, fosters open innovation, increase productivity, and solve global challenges.

But can also have negative effects, like augment socio economic inequalities (e.g. labor market), undesired outcomes, and challenge digital privacy and security.

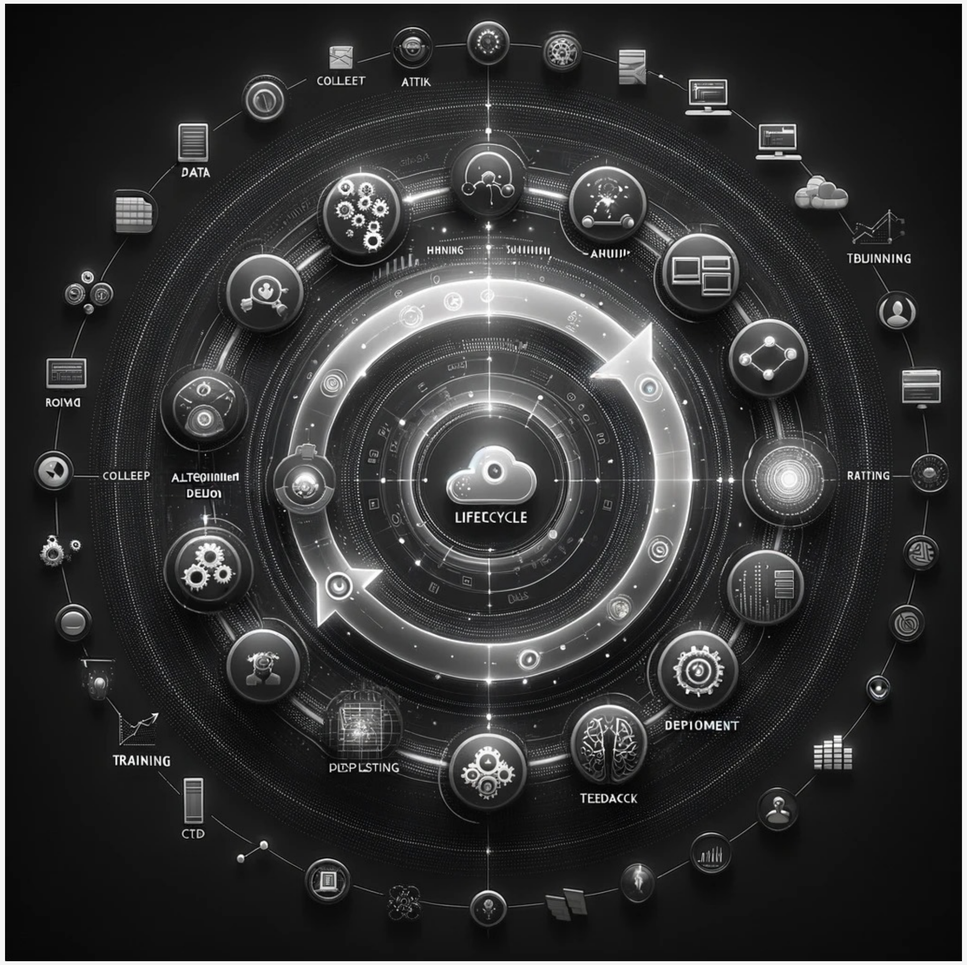


by Stable Diffusion XL

AI Lifecycle

Iterative process

- Data collection, processing, and labelling (if necessary).
- Machine Learning model training, validation, and test (NIST TEVV, Test, Evaluation, Verification and Validation). Alignment.
- Technology Deployment, and Adoption. Human-Machine interaction and decision making. Accountability.
- Monitoring. Real-time feedback. Reinforcement Learning. Human-in-the-loop. Traceability. Update and upgrade. Model and data drift assessment. Resilience (e.g. data poisoning resistant).



by DALL-E

AI Lifecycle

Security, Privacy, and Risk-based Decision making

Adoption of Ethics Guidelines and Regulations. Implementation of Regulatory Sandboxes (or Testbeds). Read Teaming.

OECD. World Economic Forum. European Commission. United Kingdom. United States. National Institute of Standards and Technology.

- Human centred values and fairness
- Transparency and Explainability
- Robustness, security and safety
- Accountability
- Environmental and Social wellbeing
- “Sandboxing” and Red Teaming

by DALL-E

Narrow AI

Security and Privacy

Data Privacy and Security oriented (e.g. Personally Identifiable Information).

Privacy-enhancing technologies PET (e.g. differential privacy, de-identification, and aggregation).

Bias (NIST: systemic, computational and Statistical, and human-cognitive). Synthetic Input.

by DALL-E

Narrow AI

Risk-based Decision making

Human intervention depending on the level of risk exposure to an automated AI-based decision.

- Low (e.g. Recommender Systems).
- Medium (e.g. Payment Fraud).
- High (e.g. Impersonation).
- Unacceptable (e.g. Surveillance).

Risk Management Framework (e.g. NIST RMF) for High risk AI systems. Judicial, human-in-the-loop, and forensic analysis for unacceptable.

Green Light. Yellow Light. Red Light.

by DALL-E

Generative AI

Definition

SKI tag

"*Generative AI refers to the type of artificial intelligence that is designed to create new content, ideas, or data that are similar but not identical to the original inputs. It learns from a wide range of examples and can generate text, images, music, and other types of content, often being used in machine learning and deep learning to produce complex and creative outputs*." **GPT4**

by DALL-E

Generative AI

Concerns

- Hallucinations
- Output: Discriminatory. Harmful. Unforeseen, or undesirable behaviour. Misuse.
- Content Provenance
- Copyrighted work.
- Model Ownership. OpenAI or Gemini Developer.
- Decentralisation.

by DALL-E

Generative AI

Read Teaming

Simulate adversarial attacks to tests the GenAI performance (i.e. outputs), and assess their impact on the ecosystem.

Red Team: Find vulnerabilities. Crack the system.

- What would you test?
- What kind of request would you ask the GenAI?
- What are the expected (or not) outputs?

Blue Team: Risk Management and Assessment. Resilience.

- How to responde to adversarial attacks.
- What kind of outputs will the GenAI give.
- How would you rate the risks?

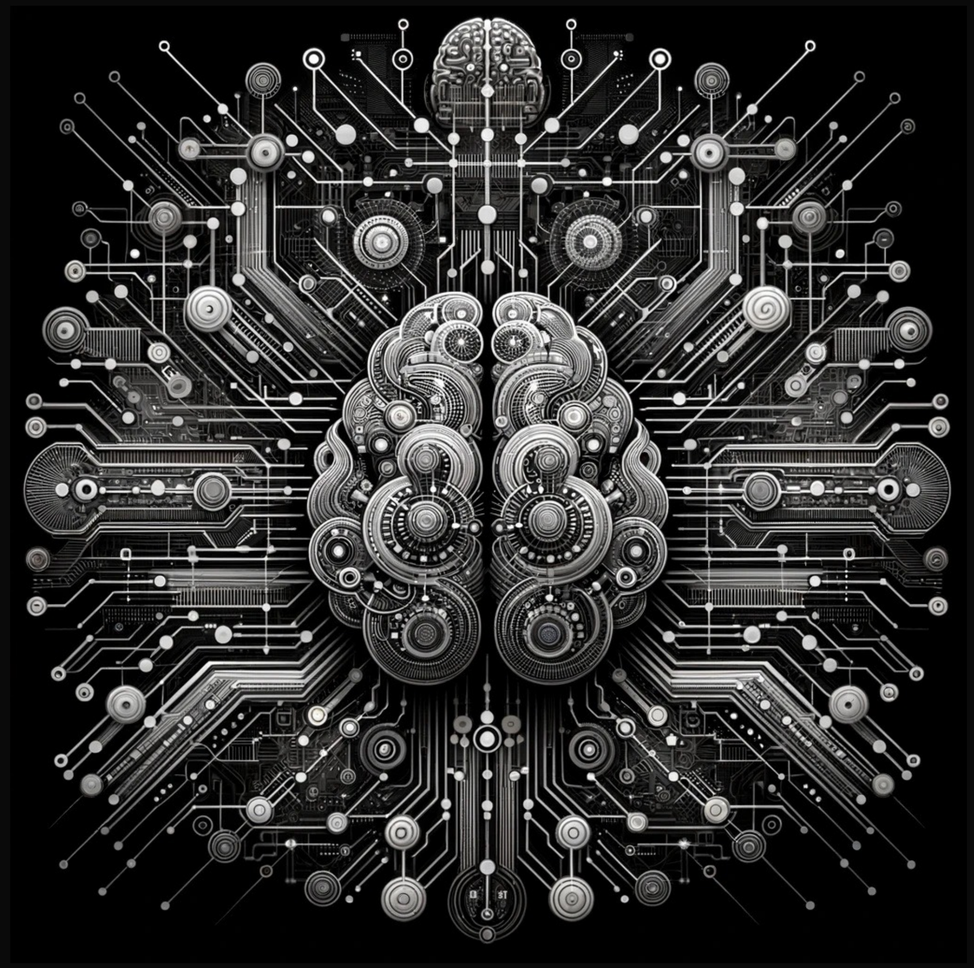
by DALL-E

SKI tag

General Purpose AI

Definition

"*General Purpose AI refers to a type of artificial intelligence that is capable of performing any intellectual task that a human being can. Unlike narrow or weak AI, which is designed to perform specific tasks or solve particular types of problems, General Purpose AI, also known as strong AI or Artificial General Intelligence (AGI), can understand, learn, and apply knowledge in a wide variety of contexts. It is adaptable, flexible, and has the potential to handle a vast range of tasks, including those it wasn't specifically programmed for. This level of AI is still theoretical and represents a significant goal in the field of AI research.*" **GPT4**



by DALL-E

SKI tag

General Purpose AI

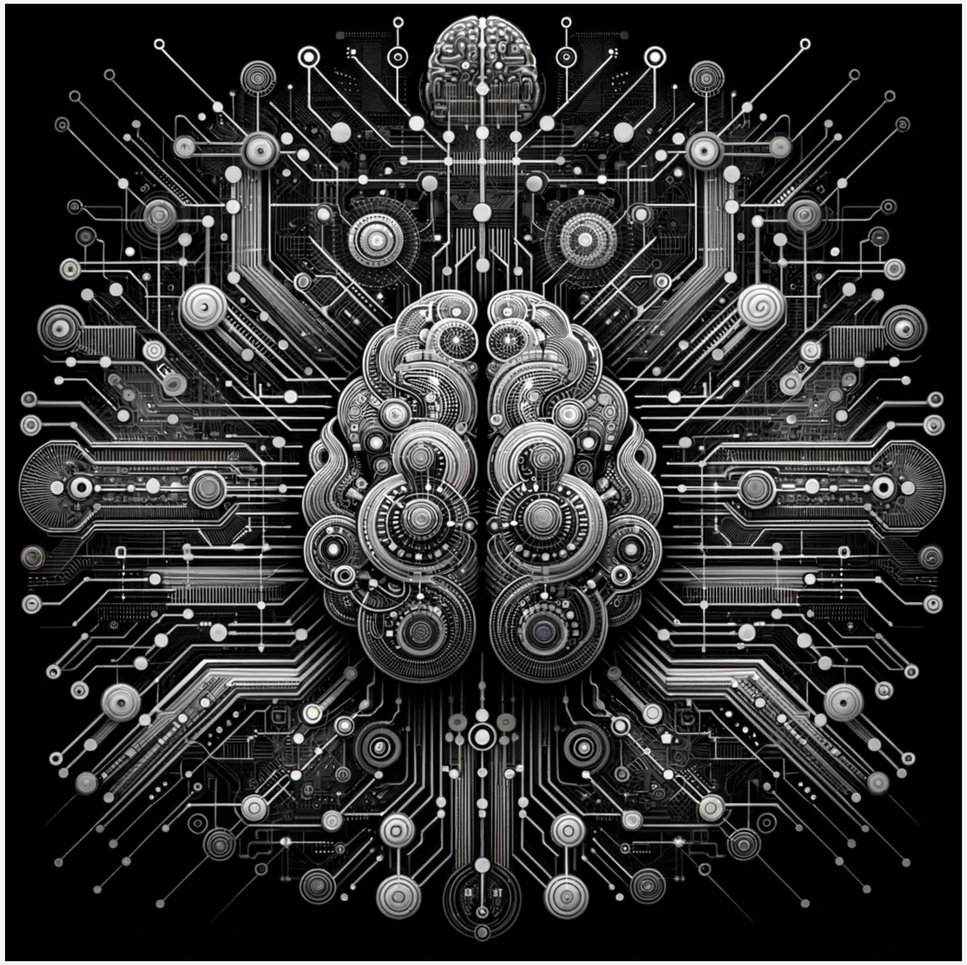
Paradigm Shift

AlphaZero (DeepMind).

Human-Machine AI-based interaction. Super-Humans.

"*Foundation models are an emerging type of general purpose AI that are trained on vast quantities of data and can be adapted to a wide range of tasks*" UK. "[…] *think abstractly and adapt to new situations*" EP. Dual-use foundation models (US).

Transparency, and strict assessment to mitigate [yet unknown] risk.



by DALL-E

General Purpose AI

Post Mortem Analysis

- "*Five Days of Chaos: How Sam Altman Returned to OpenAI*", The New York Times, Nov. 22, 2023, https://www.nytimes.com/2023/11/22/technology/how-sam-altman-returned-openai.html
- "*Now we know what OpenAI's superalignment team has been up to.*", MIT Technology Review, December 14, 2023, https://www.technologyreview.com/2023/12/14/1085344/openai-super-alignment-rogue-agi-gpt-4/

by DALL-E

SKI tag

References

OECD, 2023, Recommendations of the Council on Artificial Intelligence, https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449

WEF, 2023, The Presidio Recommendations on Responsible AI, https://www3.weforum.org/docs/WEF_Presidio_Recommendations_on_Responsible_Generative_AI_2023.pdf

EU, 2021, Artificial Intelligence Act, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52021PC0206

EU, 2023, Artificial Intelligence Act Deal, https://www.consilium.europa.eu/en/press/press-releases/2023/12/09/artificial-intelligence-act-council-and-parliament-strike-a-deal-on-the-first-worldwide-rules-for-ai/ and https://www.europarl.europa.eu/news/en/press-room/20231206IPR15699/artificial-intelligence-act-deal-on-comprehensive-rules-for-trustworthy-ai

ES, 2021, Carta Derechos Digitales, https://www.lamoncloa.gob.es/presidente/actividades/Documents/2021/140721-Carta_Derechos_Digitales_RedEs.pdf

UK, 2023, A pro-innovation approach to AI regulation, https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach

US, 2022, Blueprint for an AI Bill of Rights, https://www.whitehouse.gov/ostp/ai-bill-of-rights/

US, 2023, Executive Order on the Safe, Secure and Trustworthy Development and Use of Artificial Intelligence, https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

NIST, 2023, Artificial Intelligence Risk Management Framework, https://www.nist.gov/itl/ai-risk-management-framework

Reid Blackman, 2022, Ethical Machines, Harvard Business Review Press, https://www.reidblackman.com/ethical-machines/

PWC, 2019, A practical guide to Responsible Artificial Intelligence (AI), https://www.pwc.com/gx/en/issues/data-and-analytics/artificial-intelligence/what-is-responsible-ai/responsible-ai-practical-guide.pdf

WIRED, 2023, *The Myth of a Superhuman AI*, https://www.wired.com/2017/04/the-myth-of-a-superhuman-ai/