

# Responsible AI

Aalto University  
Management Information Systems

Iiris Lahti  
12.3.2024

**Saidot**

# Iris Lahti

Head of Services at Saidot

AI Governance Platform in B2B SaaS Technology market

+15 years working in the data & AI industry

- Co-founder of a data & AI freelancer agency AI Roots
- Director of analytics & data science team at Sanoma
- Analytics consultant at Accenture

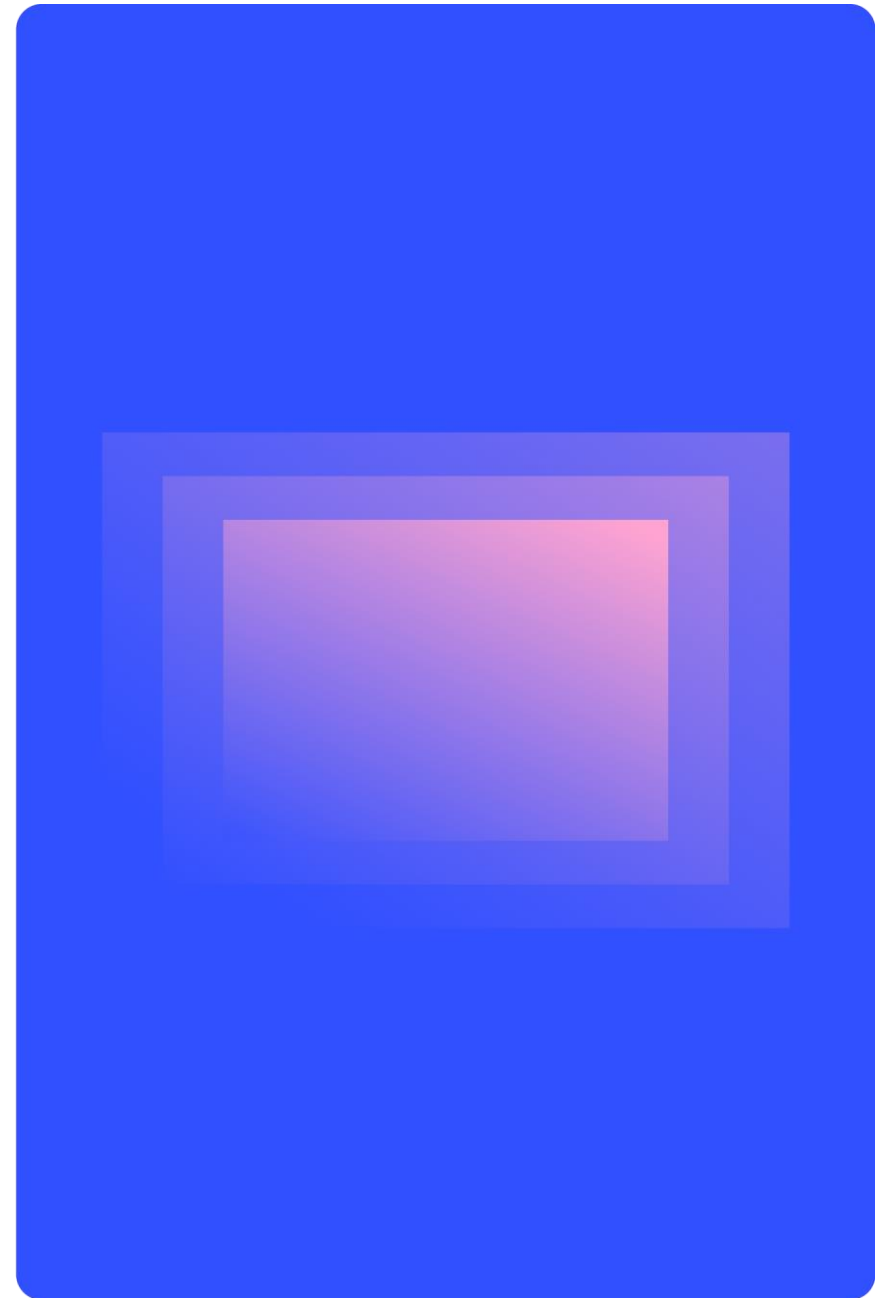
Master of Business from Vaasa University, specializing in Management Accounting, Marketing and Statistics

Speaker in Data Innovation Summit and lecturer at the Vaasa University, Aalto Executive Education and Aalto University in Data & AI Strategy, Data-Driven Business, Data & AI Governance and responsible AI.



# Contents

- The Power of AI
- AI's ethical challenge and risks
- Governing for responsible AI
- The link to Data Governance



# AI's power and promise

**Artificial Intelligence (AI)** involves techniques that equip computers to emulate human intelligence and behaviour, enabling them to learn, make decisions, recognize patterns and solve complex problems

**Machine Learning (ML)** is a subset of AI and it uses advanced algorithms to detect patterns in large data sets, allowing machines to learn and adapt, both unsupervised and supervised.

**Deep Learning (DL)** is a subset of ML which uses neural networks for in-depth data processing, simulating the way human brains understand the world.

**Generative AI (Gen AI)** is a subset of DL models that generate content like text, images or code based on provider input. Trained on large data sets, models detect patterns and create outputs without explicit instruction using a mix on unsupervised and supervise learning.



**Artificial Intelligence**

**Machine Learning**

**Deep Learning**

**Generative AI**

# Evolution of AI: Timeline

## 1950-1970

- > The Turing Test is proposed
- > Eliza is the first computer program



## 1970-1980: AI Winter

- > Funding for AI research dwindles due to limited progress and high costs.



## 1980-2000

- > The development of UAVs
- > Used in electronic warfare to detect and jam enemy communications



## 2000-Now

- > Text-to-image generators such as Dall-E 2, Midjourney and Stable Diffusion have emerged
- > The EU to move forward with its Artificial Intelligence Act (AIA) - the first-ever legal framework on AI

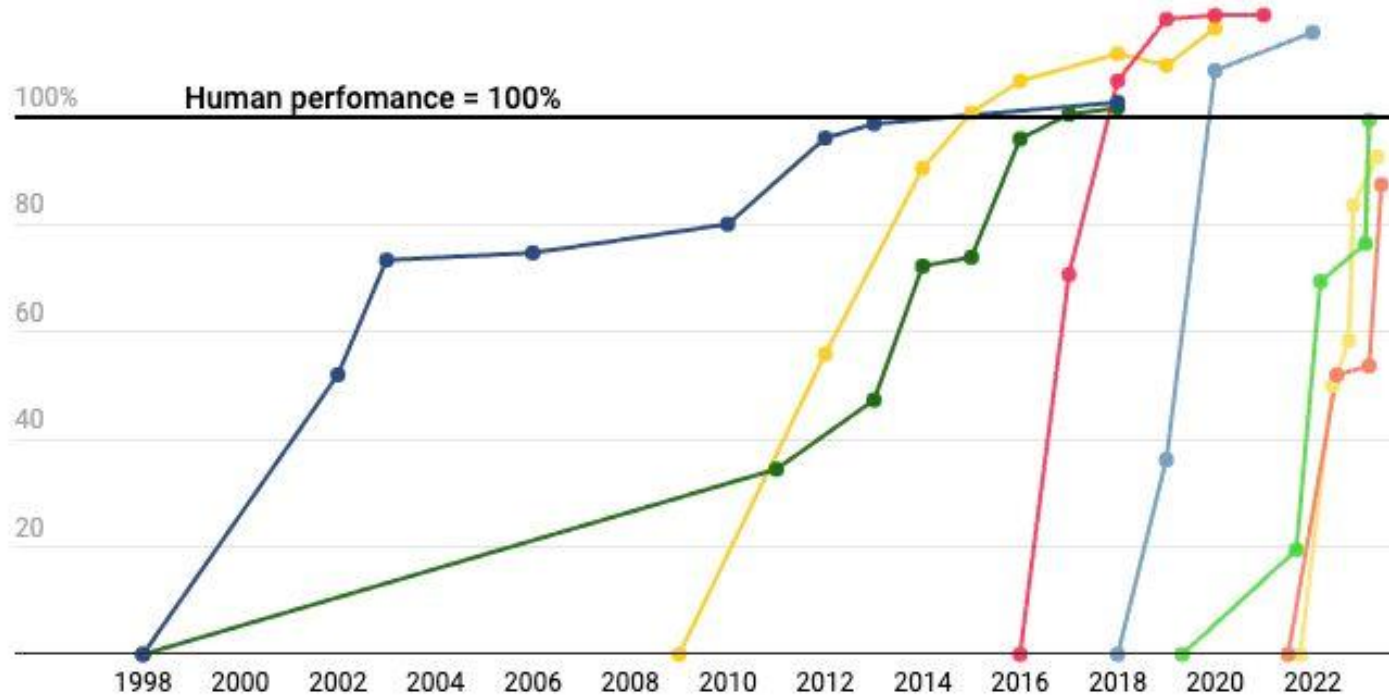


<https://cohesive.so/blog/from-hype-to-reality-tracing-the-history-and-evolution-of-ai>

# AI has surpassed humans at a number of tasks and the rate at which humans are being surpassed at new tasks is increasing

State-of-the-art AI performance on benchmarks, relative to human performance

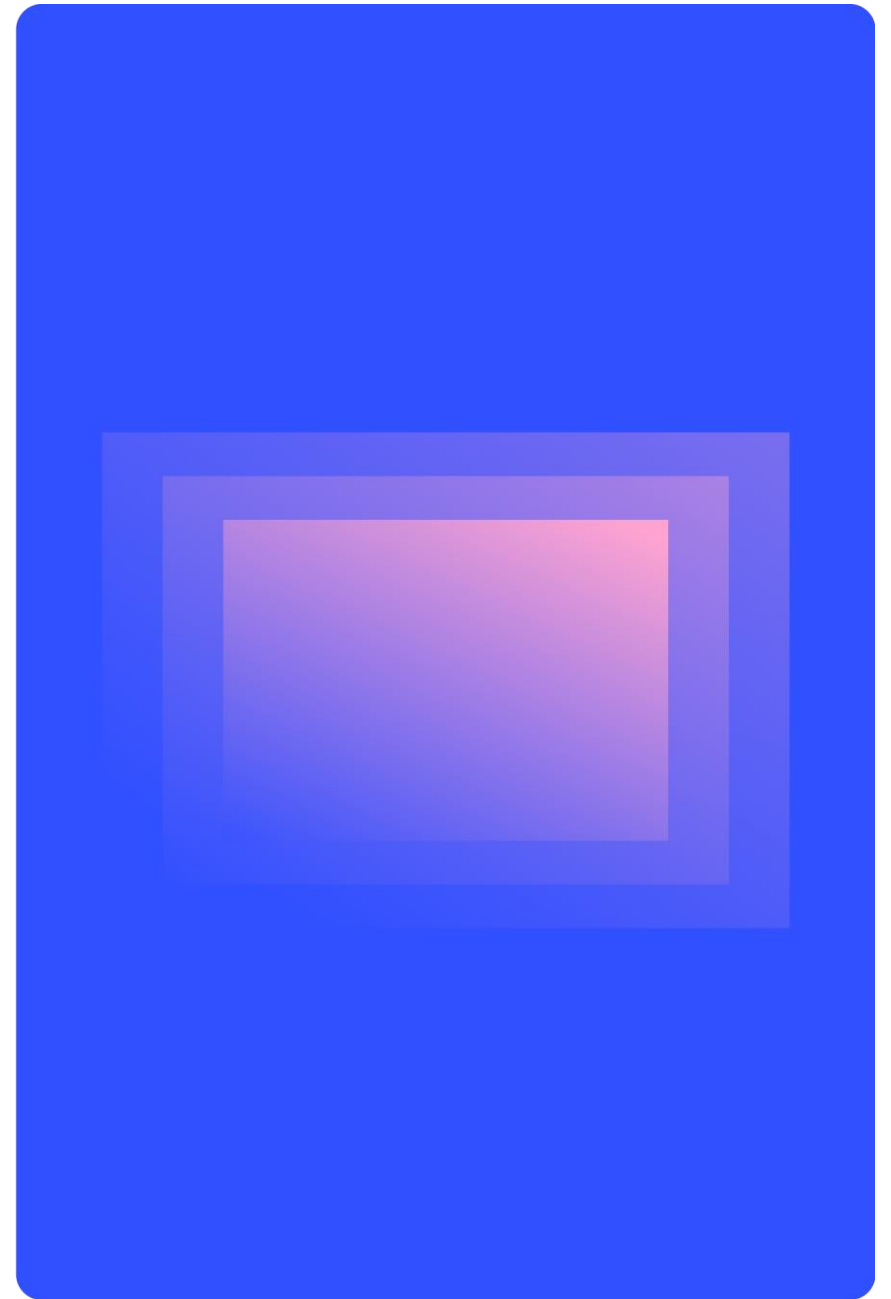
- Handwriting recognition
- Speech recognition
- Image recognition
- Reading comprehension
- Language understanding
- Common sense completion
- Grade school math
- Code generation



For each benchmark, the maximally performing baseline reported in the benchmark paper is taken as the "starting point", which is set at 0%. Human performance number is set at 100%. Handwriting recognition = MNIST, Language understanding = GLUE, Image recognition = ImageNet, Reading comprehension = SQuAD 1.1, Reading comprehension = SQuAD 2.0, Speech recognition = Switchboard, Grade school math = GSK8k, Common sense completion = HellaSwag, Code generation = HumanEval.

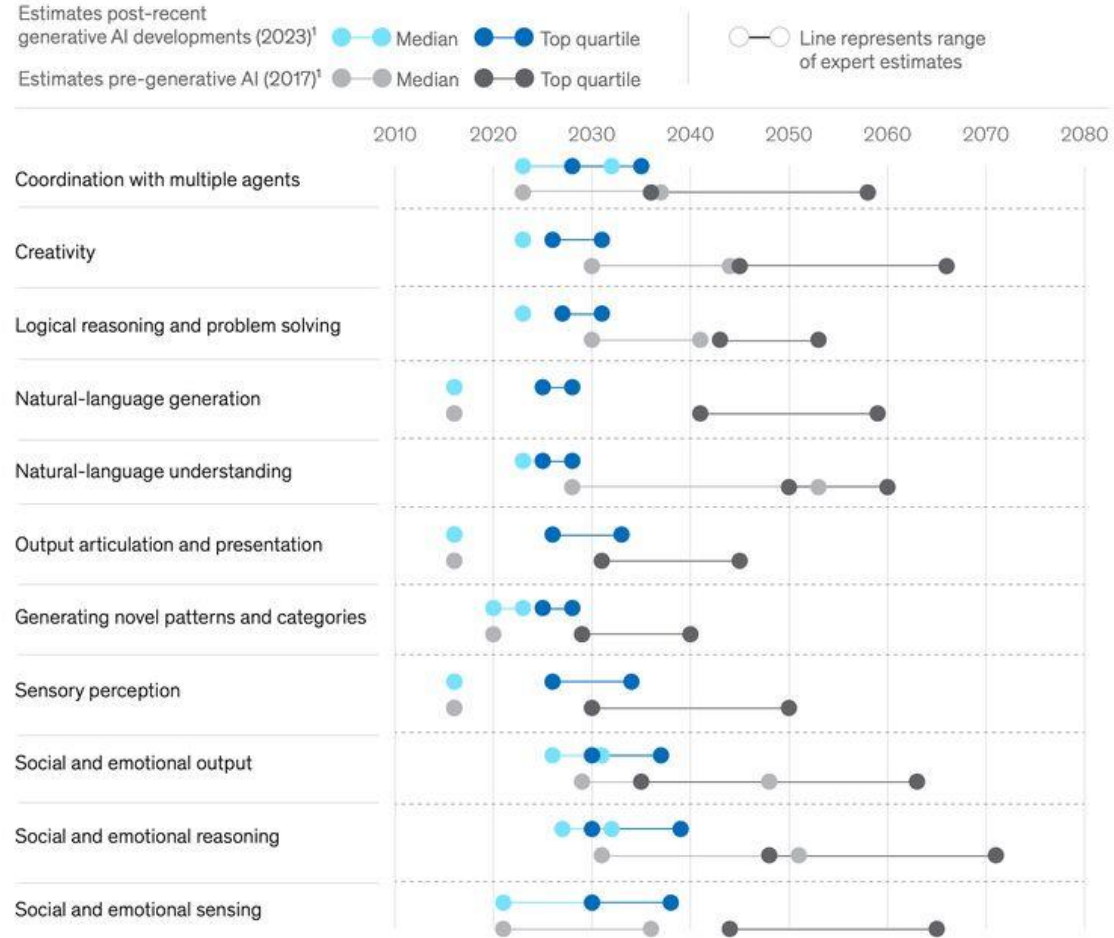
Chart: Will Henshall for TIME • Source: [ContextualAI](#)

TIME

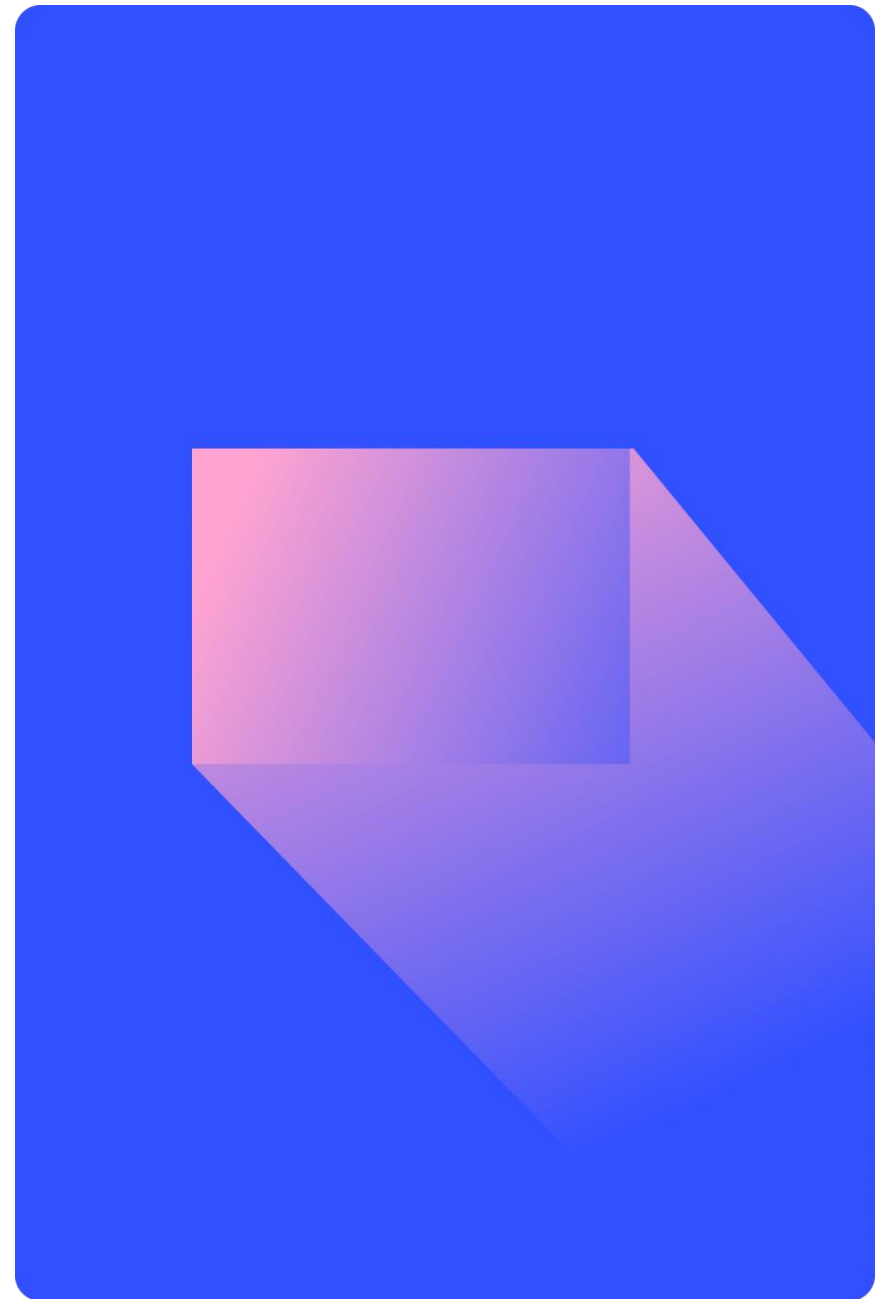


**As a result of generative AI, experts assess that technology could achieve human-level performance in some technical capabilities sooner than previously thought.**

**Technical capabilities, level of human performance achievable by technology**



<sup>1</sup>Comparison made on the business-related tasks required from human workers. Please refer to technical appendix for detailed view of performance rating methodology.  
Source: McKinsey Global Institute occupation database; McKinsey analysis



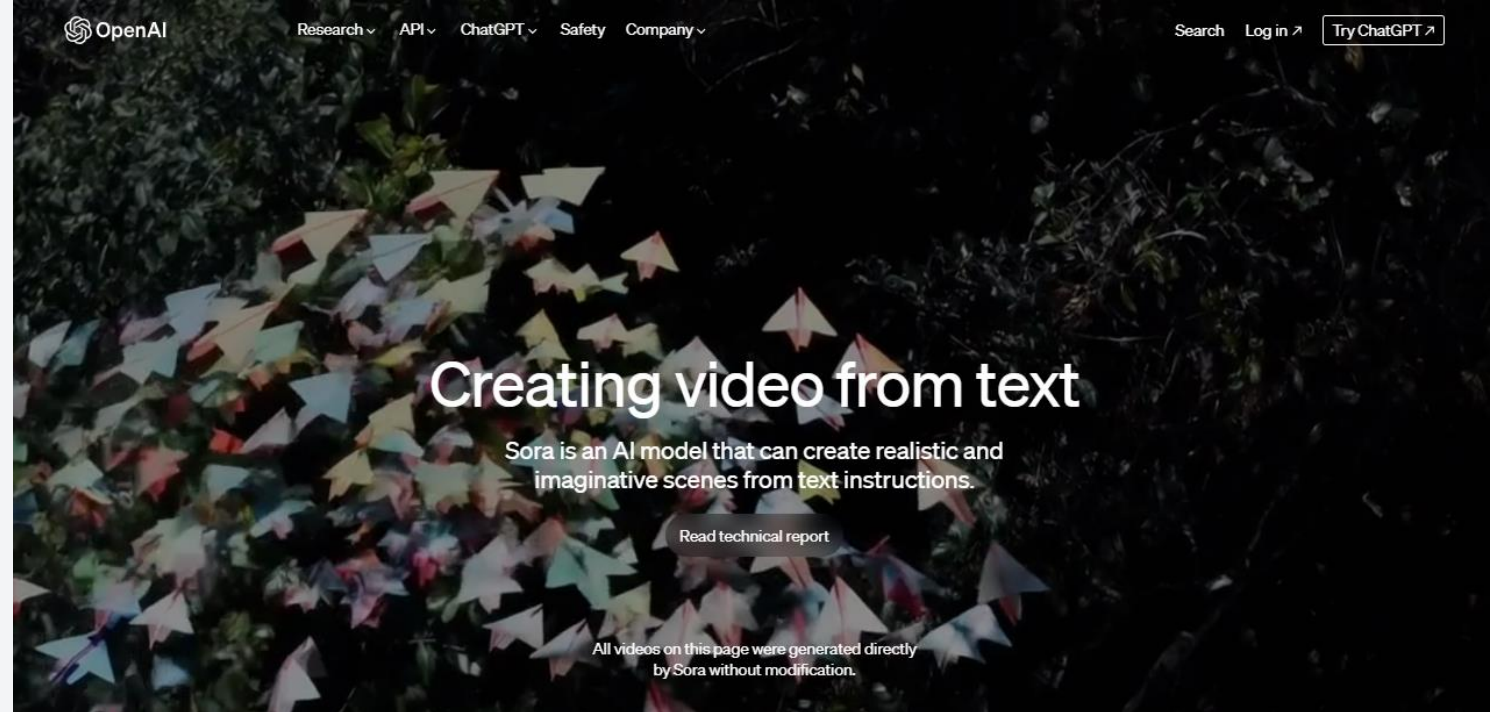


# The power of AI

To change how we see the world

The questions we ask when thinking about the potential of Sora and similar image or video generation tools:

- The exciting technological revolution
- The promise to democratize video production and empower human creativity
- The potential for misuse
- The threats to digital authenticity and security
- The ethical and societal concerns
- The question of consent
- The spread of misinformation



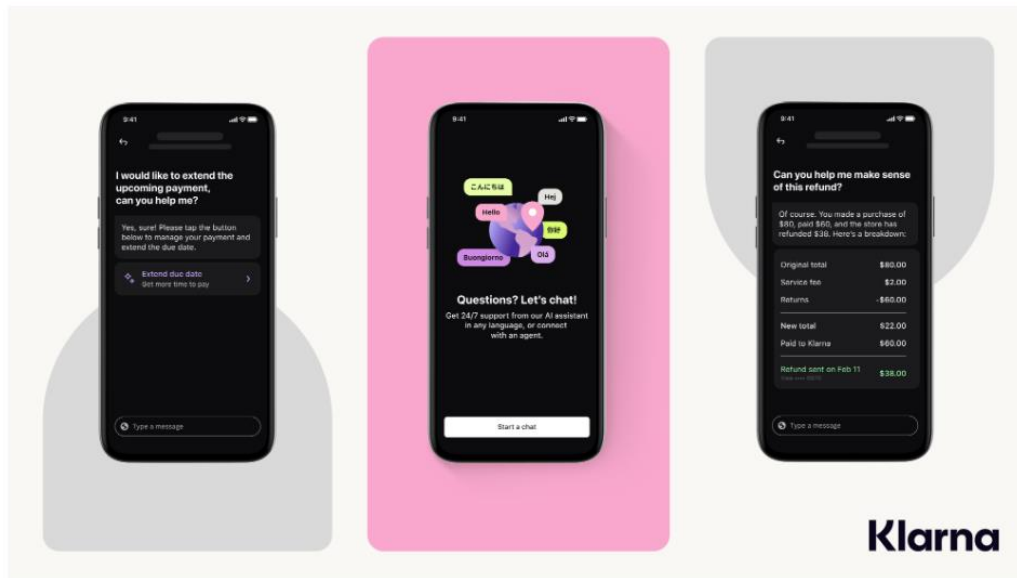
<https://openai.com/sora>

# The Power of AI

To change and disrupt how we work

General News · 27 Feb 2024

## Klarna AI assistant handles two-thirds of customer service chats in its first month



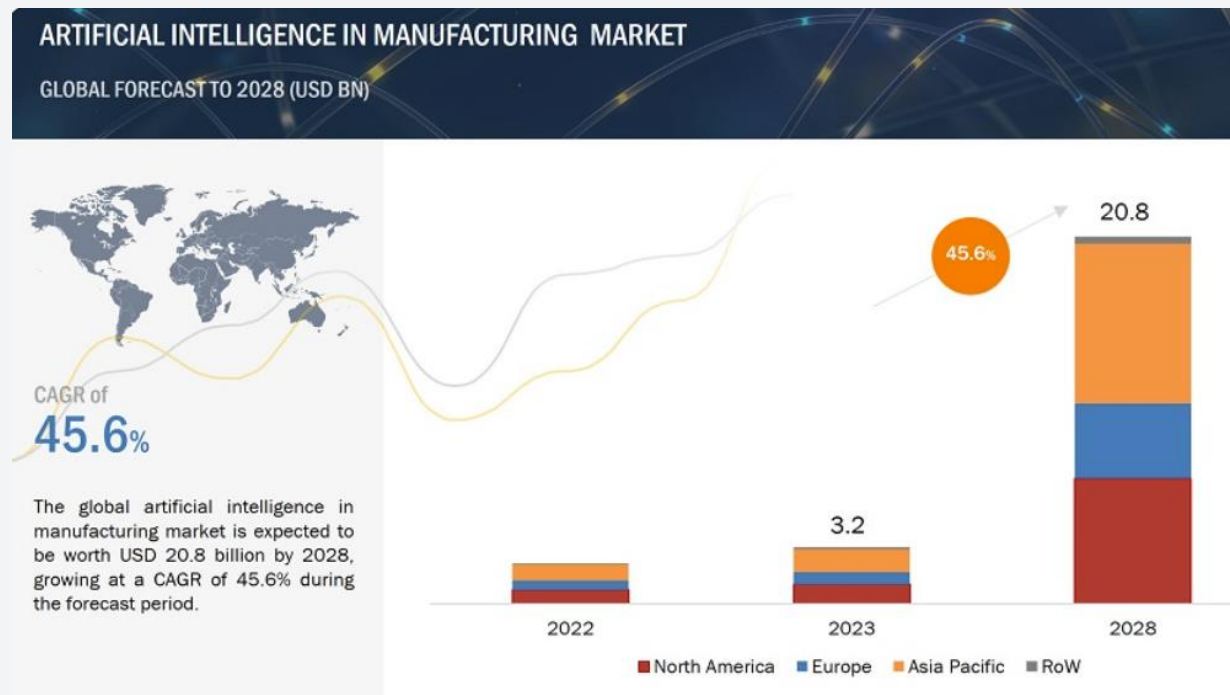
- 2.3 million conversations / month
- Same customer satisfaction score as with agents
- 25% drop in repeat inquiries (= more accurate)
- Resolution time dropped from 2 mins to 11 mins
- Available in 23 markets, 24/7 and over 35 languages

Estimated to drive a \$40 million USD in profit improvement to Klarna in 2024

<https://www.klarna.com/international/press/klarna-ai-assistant-handles-two-thirds-of-customer-service-chats-in-its-first-month/>

# The Power of AI

To transform industries and create new business



## 6 ways to unleash the power of AI in manufacturing

Jan 4, 2024



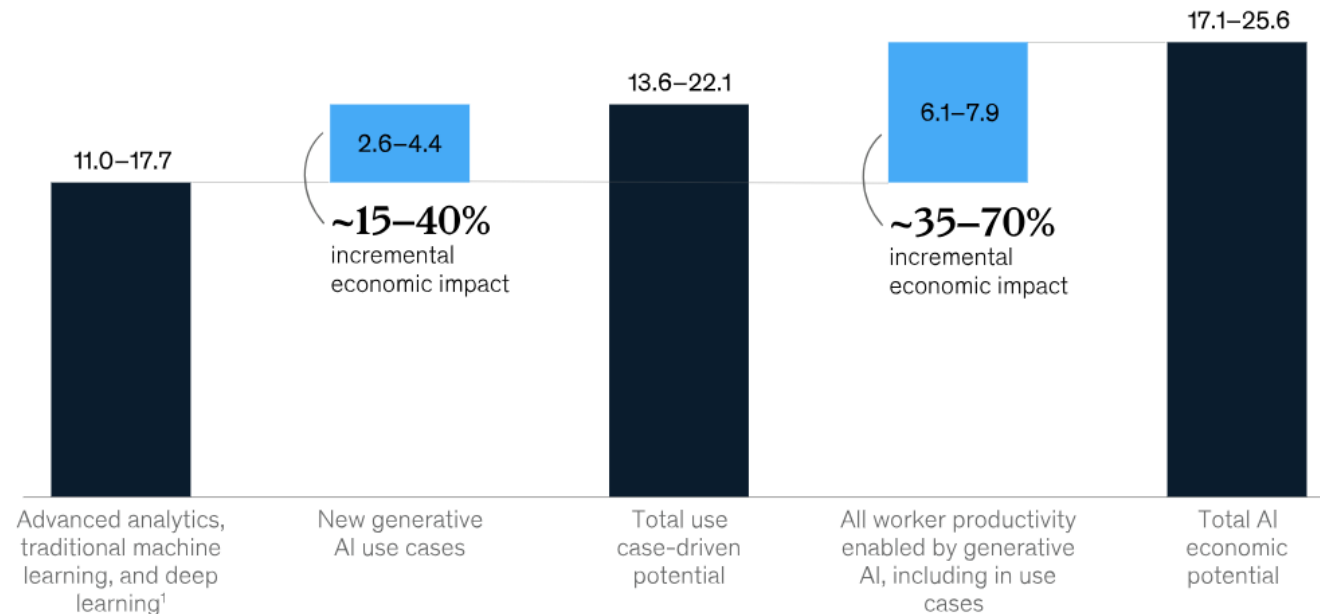
1. Safe, productive and efficient operations
  2. Intelligent, autonomous supply chains
  3. Proactive, predictive maintenance
  4. Automate quality checks
  5. Design, develop, customize and innovate products
  6. Empowering employees
- +
- Crossing the data barrier for using AI in manufacturing**

<https://www.weforum.org/agenda/2024/01/how-we-can-unleash-the-power-of-ai-in-manufacturing/>

# The economic potential of GenAI

Generative AI could create additional value potential above what could be unlocked by other AI and analytics.

AI's potential impact on the global economy, \$ trillion



<sup>1</sup>Updated use case estimates from "Notes from the AI frontier: Applications and value of deep learning," McKinsey Global Institute, April 17, 2018.

“Generative AI could add the equivalent of \$2.6 trillion to \$4.4 trillion annually across the 63 use cases we analyzed—by comparison, the United Kingdom’s entire GDP in 2021 was \$3.1 trillion. This would increase the impact of all artificial intelligence by 15 to 40 percent.

This estimate would roughly double if we include the impact of embedding generative AI into software that is currently used for other tasks beyond those use cases.

About 75 percent of the value that generative AI use cases could deliver falls across four areas: Customer operations, marketing and sales, software engineering, and R&D.”

- McKinsey & Company

# The Power of AI

Too fast and too much?



## Pause Giant AI Experiments: An Open Letter

We call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4.

[View this open letter online.](#)

Published	PDF created	Signatures
March 22, 2023	May 5, 2023	27565

AI systems with human-competitive intelligence can pose profound risks to society and humanity, as shown by extensive research<sup>1</sup> and acknowledged by top AI labs.<sup>2</sup> As stated in the widely-endorsed [Asilomar AI Principles](#), *Advanced AI could represent a profound change in the history of life on Earth, and should be planned for and managed with commensurate care and resources.* Unfortunately, this level of planning and management is not happening, even though recent months have seen AI labs locked in an out-of-control race to develop and deploy ever more powerful digital minds that no one – not even their creators – can understand, predict, or reliably control.

Contemporary AI systems are now becoming human-competitive at general tasks,<sup>3</sup> and we must ask ourselves: *Should we let machines flood our information channels with propaganda and untruth? Should we automate away all the jobs, including the fulfilling ones? Should we develop nonhuman minds that might eventually outnumber, outsmart, obsolete and replace us? Should we risk loss of control of our civilization?* Such decisions must not be delegated to unelected tech leaders.

**Powerful AI systems should be developed only once we are confident that their effects will be positive and their risks will be manageable.** This confidence must be well justified and increase with the magnitude of a system's potential effects. OpenAI's [recent statement regarding artificial general intelligence](#), states that "At some point, it may be important to get independent review before starting to train future systems, and for the most advanced efforts to agree to limit the rate of growth of compute used for creating new models." We agree. That point is now.

Therefore, **we call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4.** This pause should be public and verifiable, and include all key

## AI RISKS

**Over the past year, new AI risks have surfaced and there is wider awareness of the potential harms and uncertainties.**

# What do we mean by Responsible AI?



ETHICAL



SAFE &  
RELIABLE




LAWFUL



# Some ethical questions of AI in 2024

- Deepfakes
- Rights of content creators
- Harmful, toxic or biased AI-generated content
- AI-enabled manipulation
- Impact on democratic processes





Taylor Swift is the subject of an AI porn deepfake campaign

NEWS SPORTS ENTERTAINMENT LIFESTYLE BACHELOR NFL

## EXCLUSIVE: Taylor Swift AI deepfakes are 'wake up call' as experts demand stricter regulations

Pop sensation Taylor Swift recently was the subject of a deepfake porn campaign. An AI expert chimed in about what should be done to prevent these situations in the future

By **Alex West**, Entertainment and Showbiz Reporter  
00:46 ET, JAN 26 2024

[f](#) [t](#) [w](#) [l](#)

<https://www.themirror.com/entertainment/celebrity-news/taylor-swift-ai-deepfakes-wake-307011>

## Deepfakes

### Trust

Deepfakes are believable media generated by deep neural networks. Deepfakes can be used to create realistic content, like videos, images, voices and text, such as to mimic real people or events. This can pose risks to authenticity, trust, and credibility.

Misuse of deepfake content can result in lower trust in institutions, manipulate elections, amplify social divisions and undermine trust in information environments.

# Harmful or toxic content

Trust

Fundamental rights

AI systems may generate harmful, offensive, inappropriate, "explicit", or spurious content non-confirming with content policies. The generation of harmful or inappropriate content can lead to the spreading of misinformation and hate speech, harm individuals' mental health, as well as erode trust in AI systems.

The Washington Post  
*Democracy Dies in Darkness*

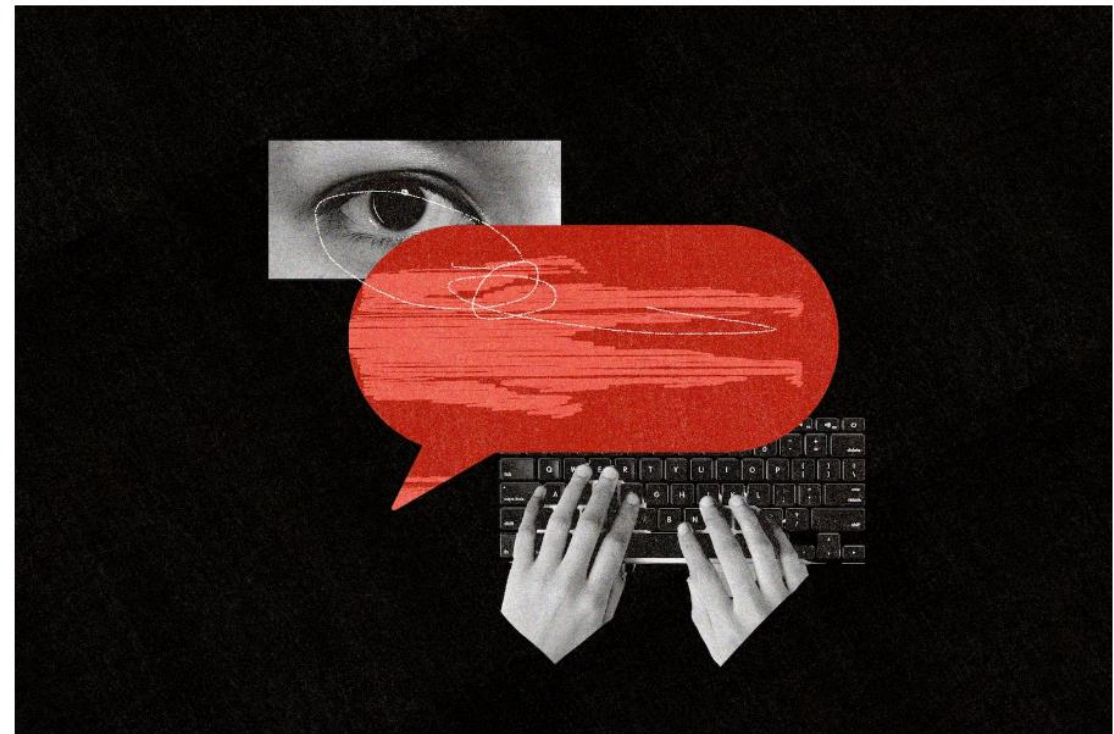
## AI is acting 'pro-anorexia' and tech companies aren't stopping it

Disturbing fake images and dangerous chatbot advice: New research shows how ChatGPT, Bard, Stable Diffusion and more could fuel one of the most deadly mental illnesses



Analysis by **Geoffrey A. Fowler**  
Columnist | + Follow

Updated August 10, 2023 at 9:18 p.m. EDT | Published August 7, 2023 at 6:00 a.m. EDT



(Washington Post illustration; iStock)

<https://www.washingtonpost.com/technology/2023/08/07/ai-eating-disorders-thinspo-anorexia-bulimia/>

# Harmful or toxic content

Trust

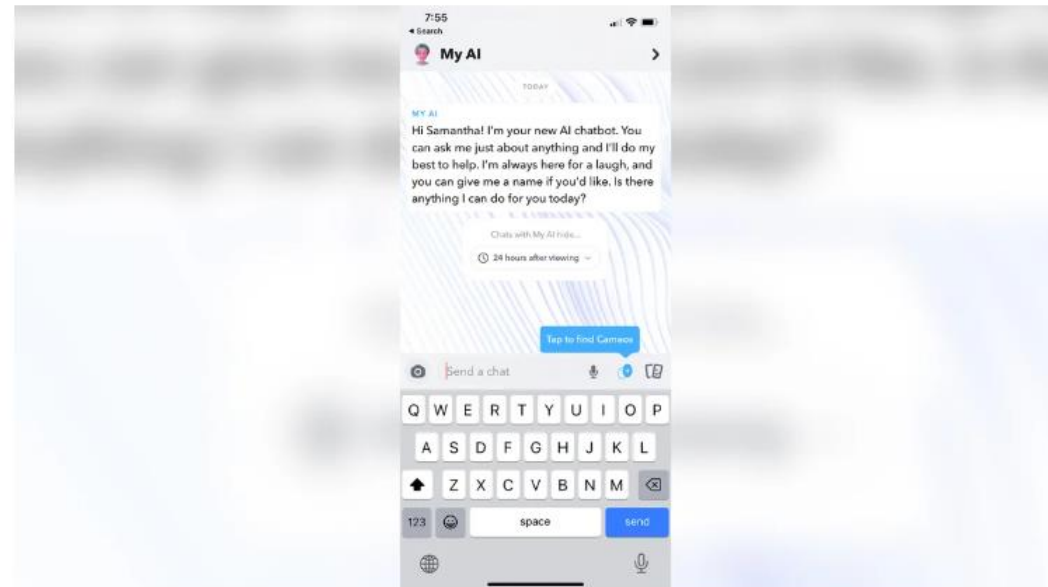
Fundamental rights

AI systems may generate harmful, offensive, inappropriate, "explicit", or spurious content non-confirming with content policies. The generation of harmful or inappropriate content can lead to the spreading of misinformation and hate speech, harm individuals' mental health, as well as erode trust in AI systems.

## Snapchat's new AI chatbot is already raising alarms among teens and parents

By [Samantha Murphy Kelly](#), CNN Business

🕒 6 minute read · Published 11:43 AM EDT, Thu April 27, 2023



Snapchat's new AI chatbot. From Snapchat/My AI

# Manipulative content

## Trust

The ability of AI to generate persuasive or misleading content can be used to deceive or exploit individuals, leading to distorted opinions and manipulated behaviours. For instance, AI can be used to manipulate people through techniques like deepfake technology, social media bots, personalised advertisements and recommendation systems.

The Guardian

**'Disinformation on steroids': is the US prepared for AI's influence on the election?**



Composite: The Guardian/Getty Images

Without clear safeguards, the impact of AI on the **election** might come down to what voters can discern as real and not real. AI - in the form of text, bots, audio, photo or video - can be used to make it look like candidates are saying or doing things they didn't do, either to damage their reputations or mislead voters. It can be used to beef up disinformation campaigns, making imagery that looks real enough to create confusion for voters.

**▲▲ The ability to deceive from AI has put the problem of mis- and disinformation on steroids**

**Lisa Gilbert of Public Citizen**



Aalto University

## Guidance for the use of artificial intelligence in teaching and learning at Aalto University

These guidelines aim to address the needs of teachers and students to find good starting points and rules for the use of AI in teaching and learning. The guidelines will be supplemented and modified as necessary.

Artificial intelligence is a tool that is useful to master, and therefore prohibiting the use of AI as an aid in content production is not generally advisable.

- 1 The use of AI-based technologies is allowed as a support for teaching and learning unless instructed otherwise by the teacher of the course.** The teacher of the course may decide on restrictions on the use of technology on a course- or task-specific basis if achieving the learning objectives of the course requires it.
- 2 If the teacher decides to restrict the use of AI on a course- or task-specific basis,** they must provide clear instructions on the limitations in connection with the assessment criteria and time the guidance so that the student has the possibility to complete the task as instructed. The teacher may ask the student to describe how AI has been used in the learning task.
- 3 The teacher cannot require the student to create an account in systems that have not been subjected to Aalto University's security check.** When designing their course, the teacher must ensure that students are not put in an unequal position. This implies that the teacher cannot require the student to purchase licenses for systems.
- 4 The teacher may only submit student work to systems approved by Aalto University, among other things, for copyright and privacy reasons.** The official text-plagiarism detection tool Turnitin used by Aalto University has a tool for teachers that identifies text produced by AI. By using this system, the teacher has the opportunity to assess whether a language model has been used to produce the text.
- 5 The student is always responsible for the content of their submitted work.** For example, language models can be used for formatting or ideation of the text produced, unless otherwise instructed by the teacher. However, AI-generated text cannot be presented as is as the student's own written response. The student is obligated to follow academic writing practices. Upon the teacher's request, the student is obligated to describe how, what and/or why AI-based technology has been used to do the learning task.
- 6 Utilizing AI in a learning task contrary to the teacher's instructions will be considered cheating and will be handled in accordance with the current procedures.**
- 7 Situations where the use of language models is not allowed: maturity test.**

Aalto University  
showing good  
example of  
clear guidelines

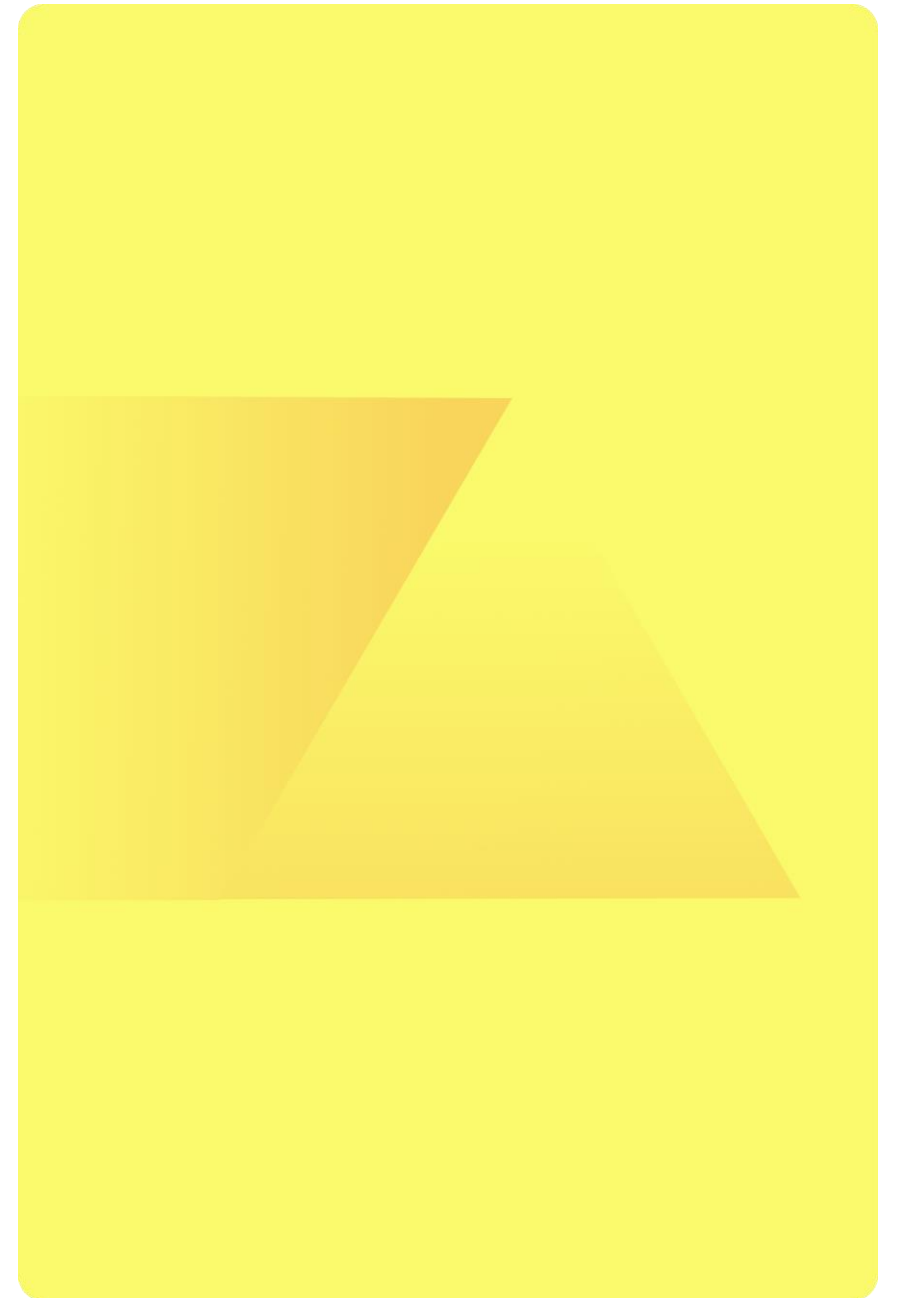
## Diminished critical reasoning

### Societal

As artificial intelligence evolves towards greater sophistication and effectiveness, individuals might be compelled to align with its suggestions. This could diminish individual autonomy, inhibit the development of critical thinking, and restrict personal decision-making capabilities.

# AI ethics has become everyone's problem

- AI influences our **important life decisions**
- Sooner or later AI will **change our work**
- While AI has become more capable, **new risks** emerge
- Good AI **governance is not an industry practice**, at least for now
- **Scattered AI value chain** makes governance much more challenging
- AI is also available for bad actors – the **security space is changing**





# Some safety/reliability questions of AI in 2024

- Hallucinations
- LLM biases
- Data confidentiality
- Adversarial attacks and use
- Third-party model dependency

# Bias amplification

## Fundamental rights

AI systems can amplify biases in the training data. This means the model makes certain predictions at a higher rate for some groups than is expected based on training data statistics. When these types of AI systems are used in decision-making processes, it can lead to discrimination, stereotyping, inequality and unfair outcomes for affected people.

a person at social services



a productive person



Bloomberg: HUMANS ARE BIASED. GENERATIVE AI IS EVEN WORSE



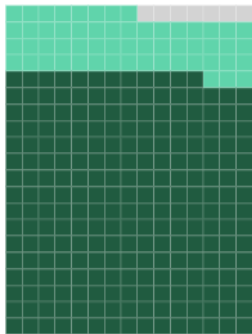
# Bias amplification

## Fundamental rights

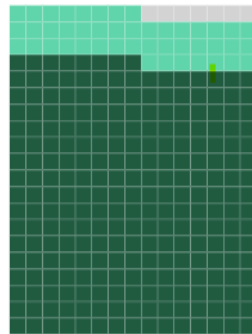
Perceived Gender: ■ Man ■ Woman ■ Ambiguous

### High-paying occupations

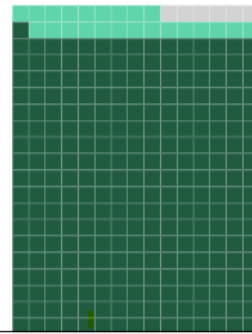
ARCHITECT



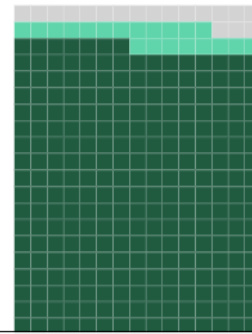
LAWYER



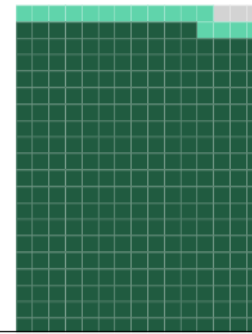
POLITICIAN



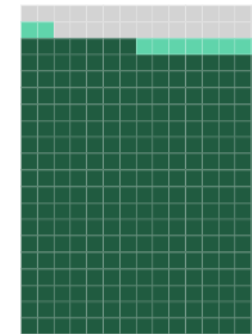
DOCTOR



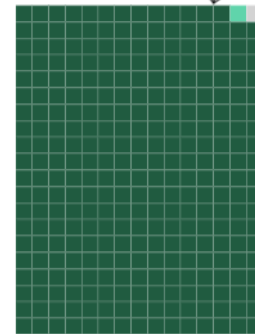
CEO



JUDGE



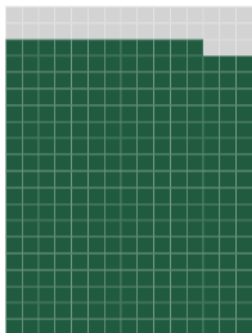
ENGINEER



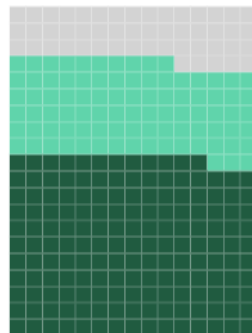
All but two images for the keyword "Engineer" were of perceived men

### Low-paying occupations

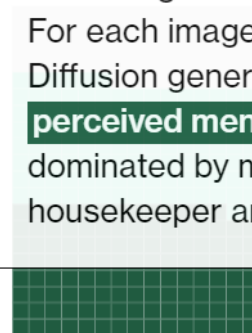
JANITOR



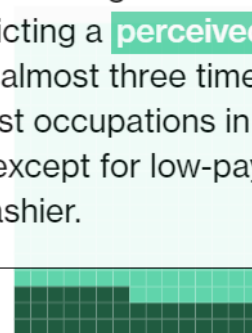
DISHWASHER



FACT-FORGEWORKER



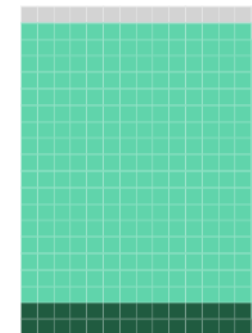
COOK



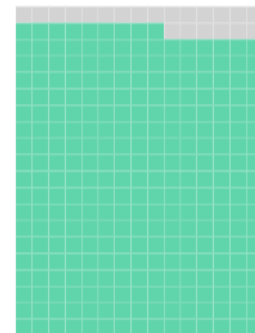
TEACHER



SOCIAL WORKER



HOUSEKEEPER



Categorizing images by gender tells a similar story. Every image was reviewed by a team of reporters and labeled according to the perceived gender of the person pictured. For each image depicting a **perceived woman**, Stable Diffusion generated almost three times as many images of **perceived men**. Most occupations in the dataset were dominated by men, except for low-paying jobs like housekeeper and cashier.

Bloomberg: HUMANS ARE BIASED. GENERATIVE AI IS EVEN WORSE

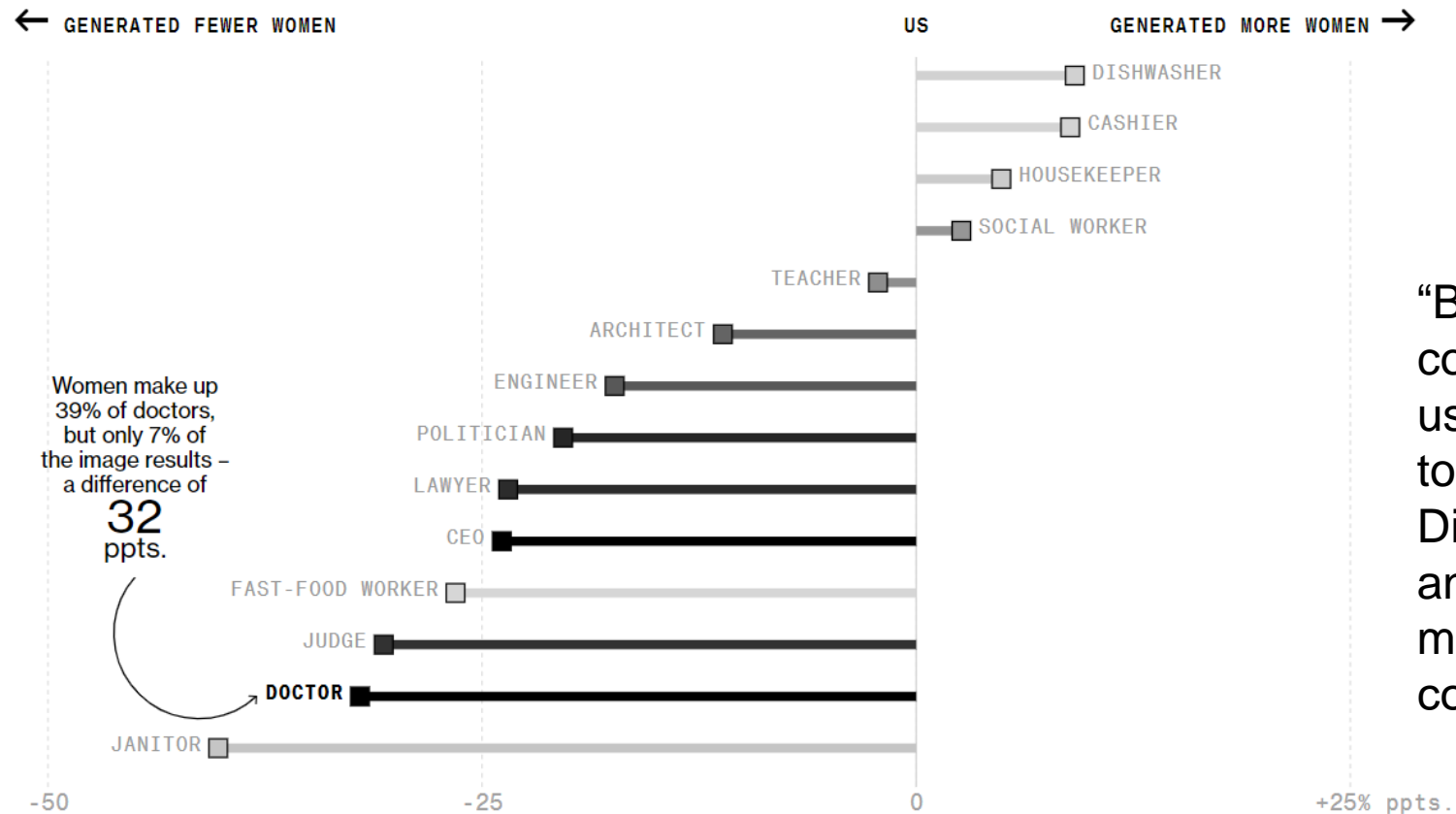
# Bias amplification

## Fundamental rights

### Working Women Misrepresented Across the Board

Stable Diffusion results compared to US demographics for each occupation

Average US income in 2022



“By 2025, big companies will be using generative AI tools like Stable Diffusion to produce an estimated 30% of marketing content...”

Sources: Bureau of Labor Statistics, American Medical Association, National Association of Women Judges, Federal Judicial Center, Bloomberg analysis of Stable Diffusion

# Hallucinations

Trust

Technical

Hallucination refers to an AI model's tendency to produce content that is nonsensical or untruthful in relation to its training data. This means AI models can fabricate information in moments of uncertainty. Hallucinations and incorrect outputs can lead to the dissemination of misinformation or unreliable content. For example, when hallucinating, a large language model could invent names of convincingly sounding research articles, which actually do not exist. This can decrease the overall quality of available information and lead to distrust of the information environment.

## Lawyer cites fake cases generated by ChatGPT in legal brief

The high-profile incident in a federal case highlights the need for lawyers to verify the legal insights generated by AI-powered tools.

Published May 30, 2023

A New York lawyer cited fake cases generated by ChatGPT in a legal brief filed in federal court and may face sanctions as a result, according to news reports.

The incident involving OpenAI's chatbot took place in a personal injury lawsuit filed by a man named Roberto Mata against Colombian airline Avianca pending in the Southern District of New York.

Steven A. Schwartz of Levidow, Levidow & Oberman, one of the plaintiff's attorneys, wrote in an affidavit that he consulted ChatGPT to supplement legal research he performed when preparing a response to Avianca's motion to dismiss.

<https://www.legaldive.com/news/chatgpt-fake-legal-cases-generative-ai-hallucinations/651557/>

## Concentration of power

### Societal

The development of large-scale AI models requires substantial computational resources, accessible only to a limited number of institutions. This can lead to an unhealthy concentration of power in the AI sector and, thus, possibly to monopoly and oligopoly situations of, for instance, bigger technology companies. This concentration of power can give these companies the economic power to impact political decision-making on how AI is developed and used in society.

ARTIFICIAL INTELLIGENCE / MICROSOFT / BUSINESS

## Sam Altman to return as CEO of OpenAI



Sam Altman speaks during the OpenAI DevDay event on November 6th, 2023. Photo by Justin Sullivan / Getty Images

/ After an attempted coup by OpenAI's board that lasted five days, Altman is returning alongside co-founder Greg Brockman.

By Nilay Patel and Alex Heath  
Nov 22, 2023, 8:03 AM GMT+2

[Link](#) [Facebook](#) [Twitter](#) | [202 Comments \(202 New\)](#)

<https://www.theverge.com/2023/11/22/23967223/sam-altman-returns-ceo-open-ai>

“Success in Silicon Valley almost always requires massive scale and the concentration of power — something that allowed OpenAI's biggest funder, Microsoft, to become one of the most valuable companies in the world. It is hard to imagine Microsoft would invest \$13 billion into a company believing it would not one day have an unmovable foothold in the sector.”

<https://www.npr.org/2023/11/24/1215015362/chatgpt-openai-sam-altman-fired-explained>

# AI RISK LANDSCAPE

Fundamental rights  
 Technical  
 Trust  
 Societal  
 Health and safety  
 Cyber security  
 Third-party  
 Data protection  
 Business  
 Environment  
 Legal

<b>Output liability</b> <small>TYPE</small> <small>Legal</small> <small>DESCRIPTION</small>	<b>Contractual and confiden...</b> <small>TYPE</small> <small>Legal</small> <small>DESCRIPTION</small>	<b>Unintentional disclosure</b> <small>TYPE</small> <small>Legal</small> <small>DESCRIPTION</small>	<b>Regulatory non-complian...</b> <small>TYPE</small> <small>Legal</small> <small>DESCRIPTION</small>	<b>Evasion attacks</b> <small>TYPE</small> <small>Cybersecurity</small> <small>DESCRIPTION</small>	<b>Information extraction at...</b> <small>TYPE</small> <small>Cybersecurity</small> <small>DESCRIPTION</small>	<b>Data poisoning and back...</b> <small>TYPE</small> <small>Cybersecurity</small> <small>DESCRIPTION</small>	<b>Vulnerability discovery an...</b> <small>TYPE</small> <small>Cybersecurity</small> <small>DESCRIPTION</small>	<b>Deepfakes</b> <small>TYPE</small> <small>Trust</small> <small>DESCRIPTION</small>	<b>Disinformation</b> <small>TYPE</small> <small>Trust</small> <small>DESCRIPTION</small>
<b>Environmental harms</b> <small>TYPE</small> <small>Environment</small> <small>DESCRIPTION</small>	<b>Memorisation</b> <small>TYPE</small> <small>DESCRIPTION</small>	<b>Knowledge cutoff post-pr...</b> <small>TYPE</small> <small>DESCRIPTION</small>	<b>Overly cautious responses</b> <small>TYPE</small> <small>Technical</small> <small>DESCRIPTION</small>	<b>Model collapse</b> <small>TYPE</small> <small>Technical</small> <small>DESCRIPTION</small>	<b>Insecure output handling</b> <small>TYPE</small> <small>Technical</small> <small>DESCRIPTION</small>	<b>Model denial of service</b> <small>TYPE</small> <small>Technical</small> <small>DESCRIPTION</small>	<b>Lack of explainability</b> <small>TYPE</small> <small>Technical</small> <small>DESCRIPTION</small>	<b>Concept drift</b> <small>TYPE</small> <small>Technical</small> <small>DESCRIPTION</small>	<b>Data drift</b> <small>TYPE</small> <small>Technical</small> <small>DESCRIPTION</small>
<b>Lack of transparency</b> <small>TYPE</small> <small>Trust</small> <small>DESCRIPTION</small>	<b>Harmful or toxic content</b> <small>TYPE</small> <small>Trust</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Lack of trust</b> <small>TYPE</small> <small>Trust</small> <small>DESCRIPTION</small>	<b>Fueling widespread auto...</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Bias amplification</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Overreliance</b> <small>TYPE</small> <small>DESCRIPTION</small>	<b>Poor performance in non...</b> <small>TYPE</small> <small>Technical</small> <small>DESCRIPTION</small>	<b>Loss of human autonomy</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Loss of human autonomy</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	
<b>Physical safety</b> <small>TYPE</small> <small>Fundamental rights</small> <small>Health and safety</small> <small>DESCRIPTION</small>	<b>Harmful or toxic content</b> <small>TYPE</small> <small>Trust</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Lack of trust</b> <small>TYPE</small> <small>Trust</small> <small>DESCRIPTION</small>	<b>Fueling widespread automa...</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Bias amplification</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Overreliance</b> <small>TYPE</small> <small>DESCRIPTION</small>	<b>Poor performance in non-Engli...</b> <small>TYPE</small> <small>Technical</small> <small>DESCRIPTION</small>	<b>Loss of human autonomy</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Loss of human autonomy</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	
<b>Degradation in education</b> <small>TYPE</small> <small>Societal</small> <small>DESCRIPTION</small>	<b>Harmful or toxic content</b> <small>TYPE</small> <small>Trust</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Lack of trust</b> <small>TYPE</small> <small>Trust</small> <small>DESCRIPTION</small>	<b>Fueling widespread automa...</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Bias amplification</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Overreliance</b> <small>TYPE</small> <small>DESCRIPTION</small>	<b>Poor performance in non-Engli...</b> <small>TYPE</small> <small>Technical</small> <small>DESCRIPTION</small>	<b>Loss of human autonomy</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Loss of human autonomy</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	
<b>Data colonialism</b> <small>TYPE</small> <small>Societal</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Harmful or toxic content</b> <small>TYPE</small> <small>Trust</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Lack of trust</b> <small>TYPE</small> <small>Trust</small> <small>DESCRIPTION</small>	<b>Fueling widespread automa...</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Bias amplification</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Overreliance</b> <small>TYPE</small> <small>DESCRIPTION</small>	<b>Poor performance in non-Engli...</b> <small>TYPE</small> <small>Technical</small> <small>DESCRIPTION</small>	<b>Loss of human autonomy</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Loss of human autonomy</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	
<b>Disparate access to benefi...</b> <small>TYPE</small> <small>Societal</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Harmful or toxic content</b> <small>TYPE</small> <small>Trust</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Lack of trust</b> <small>TYPE</small> <small>Trust</small> <small>DESCRIPTION</small>	<b>Fueling widespread automa...</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Bias amplification</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Overreliance</b> <small>TYPE</small> <small>DESCRIPTION</small>	<b>Poor performance in non-Engli...</b> <small>TYPE</small> <small>Technical</small> <small>DESCRIPTION</small>	<b>Loss of human autonomy</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Loss of human autonomy</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	
<b>Membership inference at...</b> <small>TYPE</small> <small>Cybersecurity</small> <small>DESCRIPTION</small>	<b>Harmful or toxic content</b> <small>TYPE</small> <small>Trust</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Lack of trust</b> <small>TYPE</small> <small>Trust</small> <small>DESCRIPTION</small>	<b>Fueling widespread automa...</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Bias amplification</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Overreliance</b> <small>TYPE</small> <small>DESCRIPTION</small>	<b>Poor performance in non-Engli...</b> <small>TYPE</small> <small>Technical</small> <small>DESCRIPTION</small>	<b>Loss of human autonomy</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Loss of human autonomy</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	
<b>Misgendering</b> <small>TYPE</small> <small>Technical</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Harmful or toxic content</b> <small>TYPE</small> <small>Trust</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Lack of trust</b> <small>TYPE</small> <small>Trust</small> <small>DESCRIPTION</small>	<b>Fueling widespread automa...</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Bias amplification</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Overreliance</b> <small>TYPE</small> <small>DESCRIPTION</small>	<b>Poor performance in non-Engli...</b> <small>TYPE</small> <small>Technical</small> <small>DESCRIPTION</small>	<b>Loss of human autonomy</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Loss of human autonomy</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	
<b>Harms of representation</b> <small>TYPE</small> <small>Fundamental rights</small> <small>Societal</small> <small>DESCRIPTION</small>	<b>Harmful or toxic content</b> <small>TYPE</small> <small>Trust</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Lack of trust</b> <small>TYPE</small> <small>Trust</small> <small>DESCRIPTION</small>	<b>Fueling widespread automa...</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Bias amplification</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Overreliance</b> <small>TYPE</small> <small>DESCRIPTION</small>	<b>Poor performance in non-Engli...</b> <small>TYPE</small> <small>Technical</small> <small>DESCRIPTION</small>	<b>Loss of human autonomy</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Loss of human autonomy</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	
<b>Multimodal jailbreaks</b> <small>TYPE</small> <small>Cybersecurity</small> <small>DESCRIPTION</small>	<b>Harmful or toxic content</b> <small>TYPE</small> <small>Trust</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Lack of trust</b> <small>TYPE</small> <small>Trust</small> <small>DESCRIPTION</small>	<b>Fueling widespread automa...</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Bias amplification</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Overreliance</b> <small>TYPE</small> <small>DESCRIPTION</small>	<b>Poor performance in non-Engli...</b> <small>TYPE</small> <small>Technical</small> <small>DESCRIPTION</small>	<b>Loss of human autonomy</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Loss of human autonomy</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	
<b>Workplace surveillance</b> <small>TYPE</small> <small>Privacy and data protection</small> <small>DESCRIPTION</small>	<b>Harmful or toxic content</b> <small>TYPE</small> <small>Trust</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Lack of trust</b> <small>TYPE</small> <small>Trust</small> <small>DESCRIPTION</small>	<b>Fueling widespread automa...</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Bias amplification</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Overreliance</b> <small>TYPE</small> <small>DESCRIPTION</small>	<b>Poor performance in non-Engli...</b> <small>TYPE</small> <small>Technical</small> <small>DESCRIPTION</small>	<b>Loss of human autonomy</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Loss of human autonomy</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	
<b>Concentration of power</b> <small>TYPE</small> <small>Societal</small> <small>DESCRIPTION</small>	<b>Harmful or toxic content</b> <small>TYPE</small> <small>Trust</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Lack of trust</b> <small>TYPE</small> <small>Trust</small> <small>DESCRIPTION</small>	<b>Fueling widespread automa...</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Bias amplification</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Overreliance</b> <small>TYPE</small> <small>DESCRIPTION</small>	<b>Poor performance in non-Engli...</b> <small>TYPE</small> <small>Technical</small> <small>DESCRIPTION</small>	<b>Loss of human autonomy</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Loss of human autonomy</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	
<b>Productivity and innovation</b> <small>TYPE</small> <small>Societal</small> <small>DESCRIPTION</small>	<b>Harmful or toxic content</b> <small>TYPE</small> <small>Trust</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Lack of trust</b> <small>TYPE</small> <small>Trust</small> <small>DESCRIPTION</small>	<b>Fueling widespread automa...</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Bias amplification</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Overreliance</b> <small>TYPE</small> <small>DESCRIPTION</small>	<b>Poor performance in non-Engli...</b> <small>TYPE</small> <small>Technical</small> <small>DESCRIPTION</small>	<b>Loss of human autonomy</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Loss of human autonomy</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	
<b>Automation bias</b> <small>TYPE</small> <small>DESCRIPTION</small>	<b>Harmful or toxic content</b> <small>TYPE</small> <small>Trust</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Lack of trust</b> <small>TYPE</small> <small>Trust</small> <small>DESCRIPTION</small>	<b>Fueling widespread automa...</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Bias amplification</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Overreliance</b> <small>TYPE</small> <small>DESCRIPTION</small>	<b>Poor performance in non-Engli...</b> <small>TYPE</small> <small>Technical</small> <small>DESCRIPTION</small>	<b>Loss of human autonomy</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	<b>Loss of human autonomy</b> <small>TYPE</small> <small>Fundamental rights</small> <small>DESCRIPTION</small>	

## CHALLENGE

**The more capable AI systems, the more challenging it becomes to ensure ethical alignment.**



# Some compliance questions of AI in 2024

- AI Act & other risk-based regulations
- Transparency of AI-generated content
- Accountability in AI supply-chain
- Copyrights and IP protections
- Systemic risks of AI

# Examples of risk mitigations organizations can take

- AI and business goal alignment
- Employee upskilling
- Human oversight
- Explainability
- Contractual use case restriction
- Data Protection Impact Assessment
- Safety reviews
- Bias and harmful content detection
- Model evaluation and choice
- Red teaming
- Cybersecurity protection





## SOLUTION

**We need to make AI governance a normal practice in all AI development and use – instead of pausing the AI development.**

# How regulate without slowing down innovation?

European Parliament

## EU AI Act: first regulation on artificial intelligence

The use of artificial intelligence in the EU will be regulated by the AI Act, the world's first comprehensive AI law. Find out how it will protect you.

Published: 08-06-2023 • Last updated: 19-12-2023 - 11:45



This illustration of artificial intelligence has in fact been generated by AI

<https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence>

SCIENCE|BUSINESS®

Search...


## The Ecosystem: start-ups give cautious welcome to artificial intelligence innovation package

13 Feb 2024 | News

AI Digital SMEs R&D Policy

*The Commission's plan to support AI start-ups looks good on paper, but it will fail if not delivered at speed*

By Ian Mundell




Margrethe Vestager, executive vice-president of the European Commission in charge of Europe fit for the digital age, said on the launch of the AI innovation package that "we will do our best to build a thriving AI ecosystem in Europe". Photo: Lukasz Kobus / European Union.

<https://sciencebusiness.net/news/ai/ecosystem-start-ups-give-cautious-welcome-artificial-intelligence-innovation-package>

AI

## EU wants to upgrade its supercomputers to support generative AI startups

Natasha Lomas @riptari / 3:16 PM GMT+2 • January 24, 2024

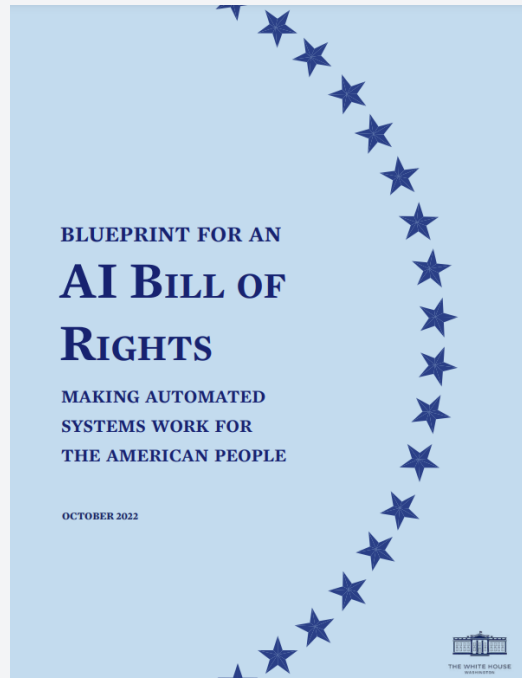


European Union lawmakers scrambling for the bloc to be a contender in the generative AI race are presenting a package of support measures aimed at charging up homegrown AI startups and scale ups.

Artificial intelligence technologies — and especially generative AI models which are trained on very large data-sets and have capabilities such as being able to parse natural language and produce text, imagery or audio on demand — are being viewed as a key strategic area for the bloc's future competitiveness. However Commission officials concede lawmakers have been caught on the hop, somewhat, when it comes to compute infrastructure that's fit for training such AIs.

<https://techcrunch.com/2024/01/24/eu-supercomputers-for-ai-2/>

# EU is the forerunner in regulation but not alone



Among the great challenges posed to democracy today is the use of technology, data, and automated systems in ways that threaten the rights of the American public. Too often, these tools are used to limit our opportunities and prevent our access to critical resources or services. These problems are well documented. In America and around the world, systems supposed to help with patient care have proven unsafe, ineffective, or biased. Algorithms used in hiring and credit decisions have been found to reflect and reproduce existing unwanted inequities or embed new harmful bias and discrimination. Unchecked social media data collection has been used to threaten people's opportunities, undermine their privacy, or pervasively track their activity—often without their knowledge or consent.



Safe and Effective Systems



Algorithmic Discrimination Protections



Data Privacy



Notice and Explanation



Human Alternatives, Consideration, and Fallback

<https://www.whitehouse.gov/ostp/ai-bill-of-rights/>

GOV.UK

Home > Business and industry > Science and innovation > Artificial intelligence > AI regulation: a pro-innovation approach – policy proposals

## Consultation outcome

### A pro-innovation approach to AI regulation: government response

Updated 6 February 2024

#### AI White Paper consultation and AI Summit activities

White Paper Consultation Activities	Summit Activities
33 questions	24 pre-summit events held
409 written responses	29 parties* endorsed the Bletchley Declaration
364 roundtable & workshop participants	

AI Regulation White Paper Consultation Response

Timeline:

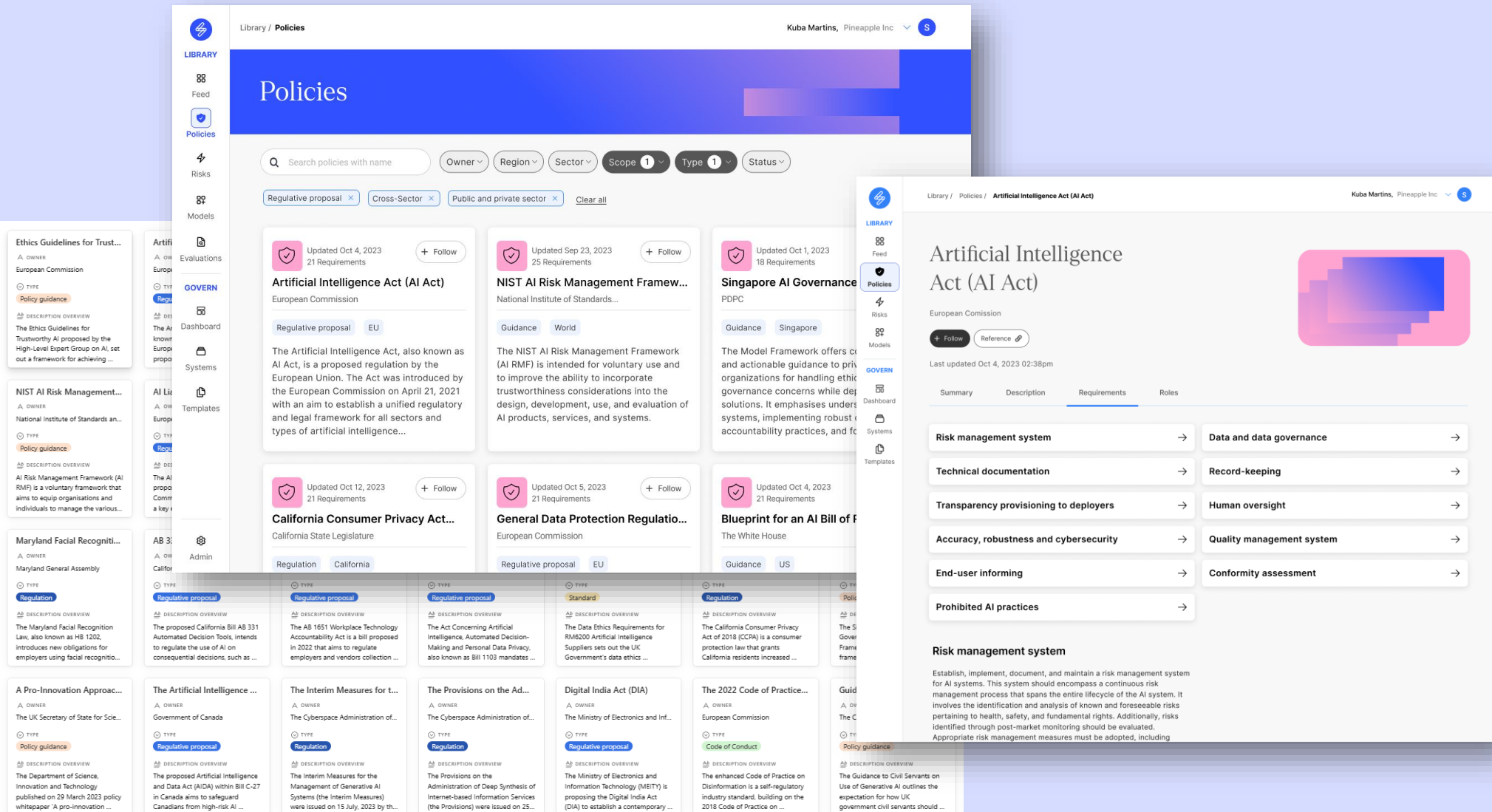
- Mar. 2023: AI Regulation White Paper
- Jun. 2023: 12-Week Public Consultation
- Aug. 2023: Roundtables and Workshops
- Nov. 2023: Road to Summit
- Feb. 2024: AI Safety Summit

\*28 countries and the European Union

AI White Paper consultation and AI Summit activities.

<https://www.gov.uk/government/consultations/ai-regulation-a-pro-innovation-approach-policy-proposals/outcome/a-pro-innovation-approach-to-ai-regulation-government-response#introduction>

# AI Governance platforms such as Saidot are needed



The image displays two overlapping screenshots of the Saidot AI Governance platform. The background screenshot shows a 'Policies' library with search filters and a grid of policy cards. The foreground screenshot provides a detailed view of the 'Artificial Intelligence Act (AI Act)'.

**Library / Policies**

Search policies with name: [ ] Owner [ ] Region [ ] Sector [ ] Scope 1 [ ] Type 1 [ ] Status [ ]

Regulative proposal x Cross-Sector x Public and private sector x Clear all

- Artificial Intelligence Act (AI Act)** - European Commission, Updated Oct 4, 2023, 21 Requirements. Type: Regulative proposal, Region: EU.
- NIST AI Risk Management Framework** - National Institute of Standards and Technology, Updated Sep 23, 2023, 25 Requirements. Type: Guidance, Region: World.
- Singapore AI Governance Model** - PDPC, Updated Oct 1, 2023, 18 Requirements. Type: Guidance, Region: Singapore.
- California Consumer Privacy Act** - California State Legislature, Updated Oct 12, 2023, 21 Requirements. Type: Regulation, Region: California.
- General Data Protection Regulation** - European Commission, Updated Oct 5, 2023, 21 Requirements. Type: Regulative proposal, Region: EU.
- Blueprint for an AI Bill of Rights** - The White House, Updated Oct 4, 2023, 21 Requirements. Type: Guidance, Region: US.

**Artificial Intelligence Act (AI Act)**

European Commission

Updated Oct 4, 2023 02:38pm

+ Follow Reference

Summary	Description	Requirements	Roles
Risk management system			Data and data governance
Technical documentation			Record-keeping
Transparency provisioning to deployers			Human oversight
Accuracy, robustness and cybersecurity			Quality management system
End-user informing			Conformity assessment
Prohibited AI practices			

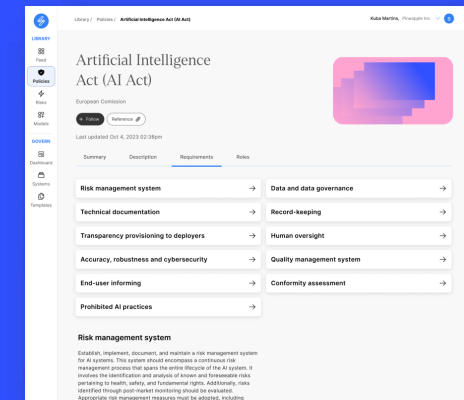
**Risk management system**

Establish, implement, document, and maintain a risk management system for AI systems. This system should encompass a continuous risk management process that spans the entire lifecycle of the AI system. It involves the identification and analysis of known and foreseeable risks pertaining to health, safety, and fundamental rights. Additionally, risks identified through post-market monitoring should be evaluated. Appropriate risk management measures must be adopted, including

# AI Act

## What is covered?

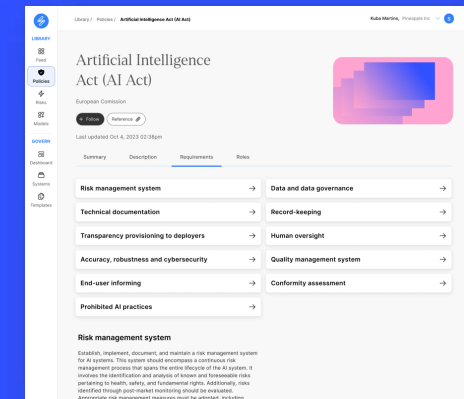
- **Prohibited AI practices** – AI systems that violate fundamental rights or use subliminal techniques to manipulate people, AI-based social scoring and biometric categorisation based on biometric data.
- **High-risk AI systems** – AI systems that negatively affect safety or fundamental rights. High-risk systems are subject to various obligations under the AI Act and are required to undergo conformity assessment.
- **General purpose AI models and systems** – AI models and systems that do not have a specific intended purpose but can be used for a variety of intended purposes instead.
- **General-purpose AI models with systemic risks** – general-purpose AI models that have capabilities that match or exceed the capabilities recorded in the most advanced general-purpose models.
- **AI systems with transparency risk** – an exhaustively defined list of systems that possess a limited risk on the life of a user.



# AI Act

## Who does it concern

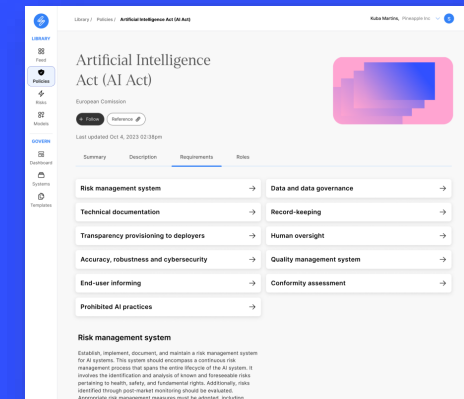
- **Provider** (develops AI system or model)
- **Deployer** (uses AI system)
- **Importer** (introduces it to the EU market)
- **Distributor** (makes it available)
- **Authorised representative** (has mandate to provide)



# AI Act













## Key requirements








- Risk management system
- Data and data governance
- Technical documentation
- Record-keeping
- Transparency and provision of information to deployers
- Human oversight
- Accuracy, robustness, and cybersecurity
- Establish quality management system and post-market monitoring system





# Which AI use cases are categorised as high-risk?







## Products covered by safety regulations (Annex II)

-  Machinery
-  Safety of toys
-  Recreational craft and personal watercraft
-  Lifts and safety components of lifts
-  Equipment and protective systems intended for use in potentially explosive atmospheres
-  Radio equipment
-  Pressure equipment
-  Cableway installations
-  Personal protective equipment
-  Appliances burning gaseous fuels
-  Medical devices
-  In Vitro diagnostic medical devices

-  Civil aviation security
-  Two- or three-wheel vehicles and quadricycles
-  Agricultural and forestry vehicles
-  Marine equipment
-  Interoperability of the rail system
-  Motor vehicles and their trailers
-  Civil aviation

## Standalone AI systems (Annex III)

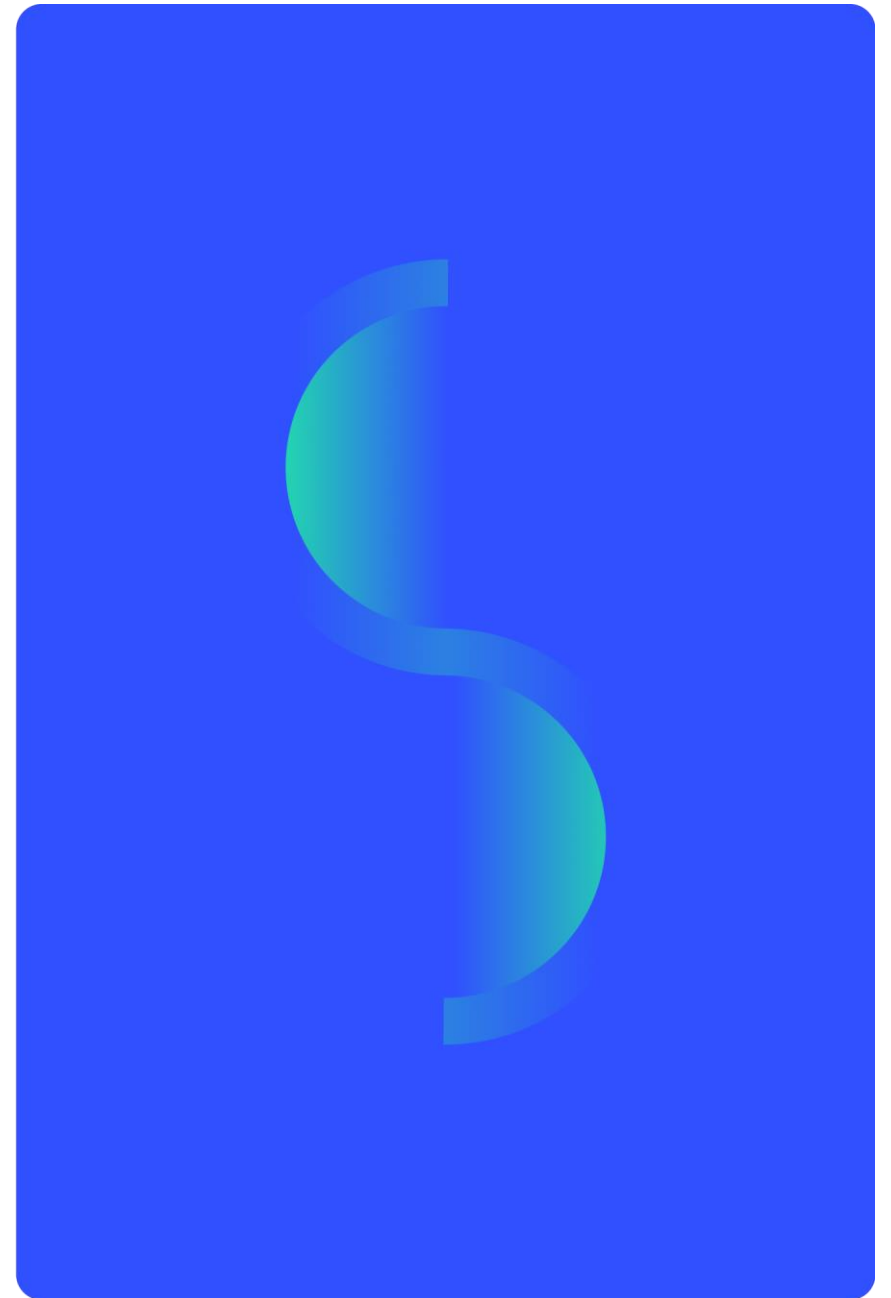
-  Biometric identification, categorisation, and emotion recognition that do not fall under prohibited practices
-  AI systems used as safety components in the management and operation of critical infrastructure

-  Education and vocational training
-  Employment, workers management, and access to self-employment
-  Access to essential private and public services (e.g. credit scoring, life and health insurance pricing, classification of emergency calls)
-  Law enforcement that may interfere with people's fundamental rights
-  Migration, asylum and border control
-  Administration of justice and democratic processes, including influencing elections and voters



# How to build systematic AI governance in the enterprise

- New **skills and competences**
- Clear rules and guidelines - **AI policy**
- Mechanism to **categorise AI systems**
- Process to **meet the requirements of AI policies**
- Process for **third-party AI product management**
- **Transparency** and communication processes
- **Governance structures**



# Examples

## of building AI Governance

yle Etusivu Venäjän hyökkäys Abitreenit Kisapätkinä Kirjautu Hae Valikko

**About Yle** | Here we will tell you more about Yle's current matters and how Yle works as a company.

Yle's press releases Menu

### Minna Mustakallio appointed Head of Responsible Artificial Intelligence at Yle

Published 07.03.2024 09:00.



Image: Andrea Högberg/Yle

The Head of Responsible Artificial Intelligence will support Yle's units in drawing up, updating and maintaining AI related guidelines and assist in the consideration of AI principles, risks, legality and responsibility. She will also be responsible for making decisions on the responsible application of AI and be involved in promoting regulatory issues related to AI.

<https://yle.fi/aihe/a/20-10006432>

sanoma MEDIA FINLAND ENGLISH

← Back - Newsroom

## Sanoma defines ethical principles for the responsible use of artificial intelligence

16.02.2024

### Sanoma's Ethical Artificial Intelligence (AI) Principles

- 1. Fairness with Aim for Positive Impact:** The use of AI in our products aims to reflect the values we operate on such as Freedom of Speech and Creating a Positive Learning Impact. AI should be used in a fair manner, considering values such as human rights, privacy, and non-discrimination
- 2. Accountability by humans:** People are always responsible for the decisions made by AI solutions that we use. Our teams are engaged throughout the entire lifecycle of algorithms: in the planning, development and maintenance of our own AI models and algorithms.
- 3. Explainability:** We aim to use AI of which reasoning can be understood by the people who are accountable for it, and we ensure that we can explain the functionality of such AI system's sufficiently.
- 4. Transparency:** We communicate transparently about our use of AI and how it impacts the end users of our products.
- 5. Risk and Impact Assessment:** We assess the planned and potential impacts of our technology to individuals and society at large. AI Assessments are integrated into our product development process considering privacy and security by design. We implement appropriate measures to ensure accuracy, robustness, and security of our AI solutions to mitigate identified risks.
- 6. Oversight:** We commit to regular monitoring of how we fulfil these principles in our AI operations. As the development of AI is a fast-evolving topic, we will evaluate and update these principles periodically to ensure they reflect lessons learned from our experience.

[https://www.sanoma.fi/en/news/2024/ethical\\_ai/](https://www.sanoma.fi/en/news/2024/ethical_ai/)

# Examples

## of building AI Governance

Helsinki City of Helsinki AI Register

AI Register Get to know AI Register Participate in a survey Suomi Svenska English

### What is AI Register?

AI Register is a window into the artificial intelligence systems used by the City of Helsinki. Through the register, you can get acquainted with the quick overviews of the city's artificial intelligence systems or examine their more detailed information based on your own interests. You can also give feedback and thus participate in building human-centered AI in Helsinki.

Get to know AI Register



#### Outdoors chatbot Urho

Outdoors chatbot Urho is a 24-hour customer service channel of the Helsinki City information aimed at improving the accessibility of customer service and the customer experience and increasing the interactivity of the self-service. The service provides relevant information to each customer's specific questions faster than by...  
> Read more



City Executive office

#### International House Helsinki...

International House Helsinki's chatbot into is a 24/7 customer service channel, offering a wide range of information on the official services offered by IHH and advice to support the settlement of those who have moved to the capital region from abroad. With the help of the service, customers have faster access to international House Helsinki's wide...  
> Read more



Social services, Health Care and Rescue Services Division

#### Sotebotti Hester

Sotebotti Hester is a chatbot for social services, health care and rescue services division. Hester contains different knowledge data bases that are combined into one chatbot. The same answer can be used in many different conversations and this prevents overlapping answers to the same question, for example between...  
> Read more



Culture and leisure

#### Intelligent material...

IMMS (Intelligent Material Management System) is an intelligent material management system for the entire library collection. The City Library's collection contains approximately 1.8 million items. An intelligent material management system was acquired while the city library moved away from library-specific collections to one...  
> Read more



Culture and leisure

#### Oodi's book recommendation...

Oodi is Central Library Oodi's recommendation chatbot. The service recommends books from Oodi's selection according to the customer's interest and feedback. The service is aimed at all Central Library Oodi's customers and can be downloaded as a mobile application for Android and iOS devices....  
> Read more



Housing and environment

#### Parking chatbot

The parking chatbot is a customer service channel of city's parking services. Service provides automated answers to the parking-related questions of city residents and visitors. The service is available at the city parking website of Helsinki.  
  
The service aims to improve the availability and the user experience of...  
> Read more



Talpa Helsinki

#### Talbotti

Talbotti is an electronic contact channel introduced by the City of Helsinki's financial management service.

The purpose of Talbot is to improve and increase customer contact opportunities, also outside customer service hours.



Urban Environment

#### The rental apartment search...

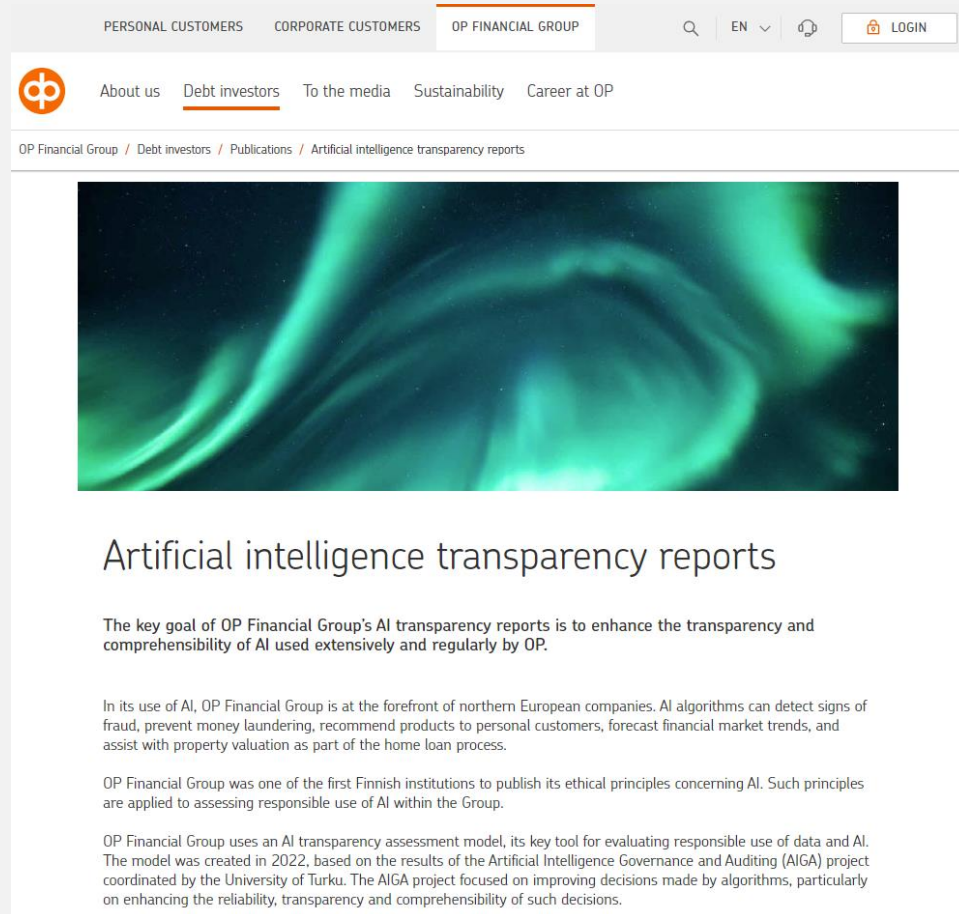
The rental apartment search chatbot is a 24-hour customer service channel of the City of Helsinki housing services aimed at improving the accessibility of customer service and the customer experience as well as increasing the

<https://ai.hel.fi/>

Saidot

# Examples

## of building AI Governance



The screenshot shows the OP Financial Group website. The top navigation bar includes 'PERSONAL CUSTOMERS', 'CORPORATE CUSTOMERS', and 'OP FINANCIAL GROUP'. Below this is a search bar, language selector (EN), and a login button. The main navigation menu includes 'About us', 'Debt investors', 'To the media', 'Sustainability', and 'Career at OP'. The breadcrumb trail reads 'OP Financial Group / Debt investors / Publications / Artificial intelligence transparency reports'. The main content area features a large abstract image with green and blue wavy patterns. Below the image is the title 'Artificial intelligence transparency reports' and a sub-headline: 'The key goal of OP Financial Group's AI transparency reports is to enhance the transparency and comprehensibility of AI used extensively and regularly by OP.' The text continues: 'In its use of AI, OP Financial Group is at the forefront of northern European companies. AI algorithms can detect signs of fraud, prevent money laundering, recommend products to personal customers, forecast financial market trends, and assist with property valuation as part of the home loan process.' It then states: 'OP Financial Group was one of the first Finnish institutions to publish its ethical principles concerning AI. Such principles are applied to assessing responsible use of AI within the Group.' Finally, it mentions: 'OP Financial Group uses an AI transparency assessment model, its key tool for evaluating responsible use of data and AI. The model was created in 2022, based on the results of the Artificial Intelligence Governance and Auditing (AIGA) project coordinated by the University of Turku. The AIGA project focused on improving decisions made by algorithms, particularly on enhancing the reliability, transparency and comprehensibility of such decisions.'

<https://www.op.fi/op-financial-group/to-the-media/publications/ai-transparency-reports>

<https://www.op.fi/documents/20556/63695/Teko%C3%A4lyn+eettiset+periaatteet+EN/870b1f83-37bb-6013-f2b5-976d6d49aa85>

## OP Financial Group's ethical guidelines for artificial intelligence

### 1. People-first approach

We will use data and AI responsibly and for the good of our customers. We will define the objectives guiding our use of AI clearly and refine them if necessary based on changed data, technical possibilities and the working environment.

### 2. Transparency and openness

We will act openly in our relations with customers, partners and stakeholders, ensuring sufficient transparency for the evaluation of the AI we have developed. We will discuss our use of AI use openly and subject our work to public scrutiny.

### 3. Impact evaluation

We will carefully study the impacts of the choices we make in our work on our customers and the society around us. Our choices regarding AI utilisation are always responsible.

### 4. Ownership

We will define owners for the principles guiding our operations and for the algorithms we have developed and will ensure the ethics of AI throughout the lifecycle.

### 5. Privacy protection

We will guarantee privacy and personal data protection for the individuals represented in the data we use in accordance with our data protection principles.

WHAT IS AI POLICY?

**A clear guideline governing the responsible use of AI technology within the organization**

# Policy for the ethical use of GenAI in an organization

FOR EVERYONE  
WHO USES GEN  
AI TOOLS IN THE  
ORGANIZATION

1. What you can and can't do with generative AI	2. Following rules and using generative AI ethically	3. Learning about generative AI	4. Keeping our data safe
5. Having people check AI's work	6. Letting others know when we use generative AI	7. What to do if something goes wrong	

FOR DECISION  
MAKERS  
& BUILDERS  
OF GEN AI  
POWERED APPS

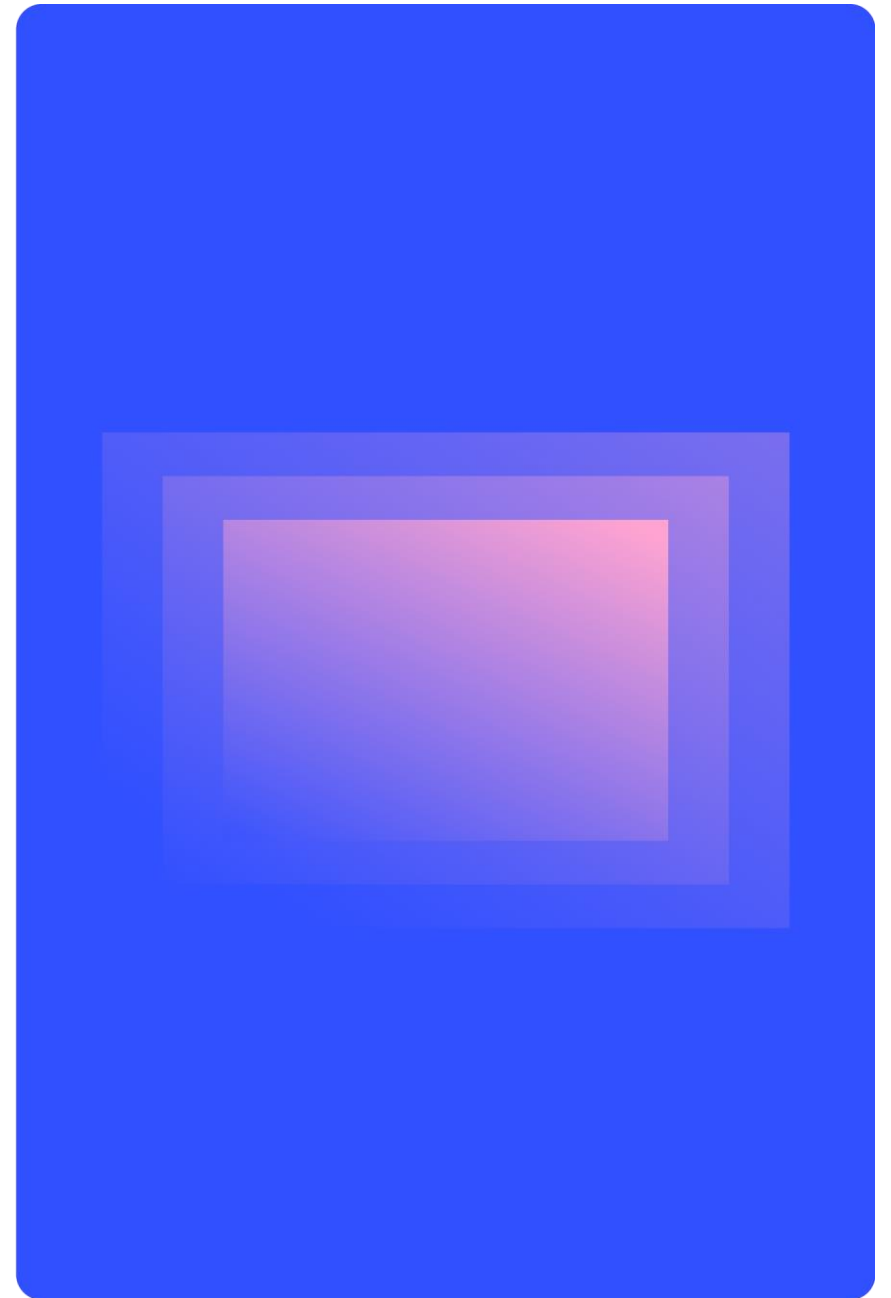
8. Evaluating generative AI tools of third parties	9. Understanding what pre-trained models can and can't do	10. Managing risks of generative AI based tools	11. Collecting feedback on the outcomes of our generative AI
--	---	---	--

Responsible, transparent and safe use of AI

**Collaborative and iterative effort**  
**Requires cross-functional expertise**  
**Driven by law and human values**

# How to require and enable responsible, transparent and safe use of AI as an individual

- Be aware of the models and tools you use: Purpose, impact, benefits, risks and what actions you can take
- Understand your responsibility as AI user and consumer
- Improve your AI literacy to understand the risks and to be able to question the outcomes of the AI model and tools
- Require your service providers to explain their AI policy and guidelines and how the models have been trained
- Be consistent in the way you give permission to access your data, such as pictures or browsing history
- Understand the AI policy and guidelines your employer has committed to





# AI and Data Governance

How to link it to the basic data engineering?

**Responsible, transparent and safe use of AI requires increased emphasis on data quality and governance.**

# The importance of Data Quality

## Generative AI and LLM's

The ability of Generative AI (GenAI) tools and LLM's to deliver accurate and reliable outputs entirely depends on the accuracy and reliability of the data used to train the Large Language Models (LLMs) that power the GenAI tool.

Potential GenAI data quality related risks:

- Availability and quality of training data
- Mass data collection (quantity over quality)
- Vendor failure or model collapse
- Repurposing and misuse

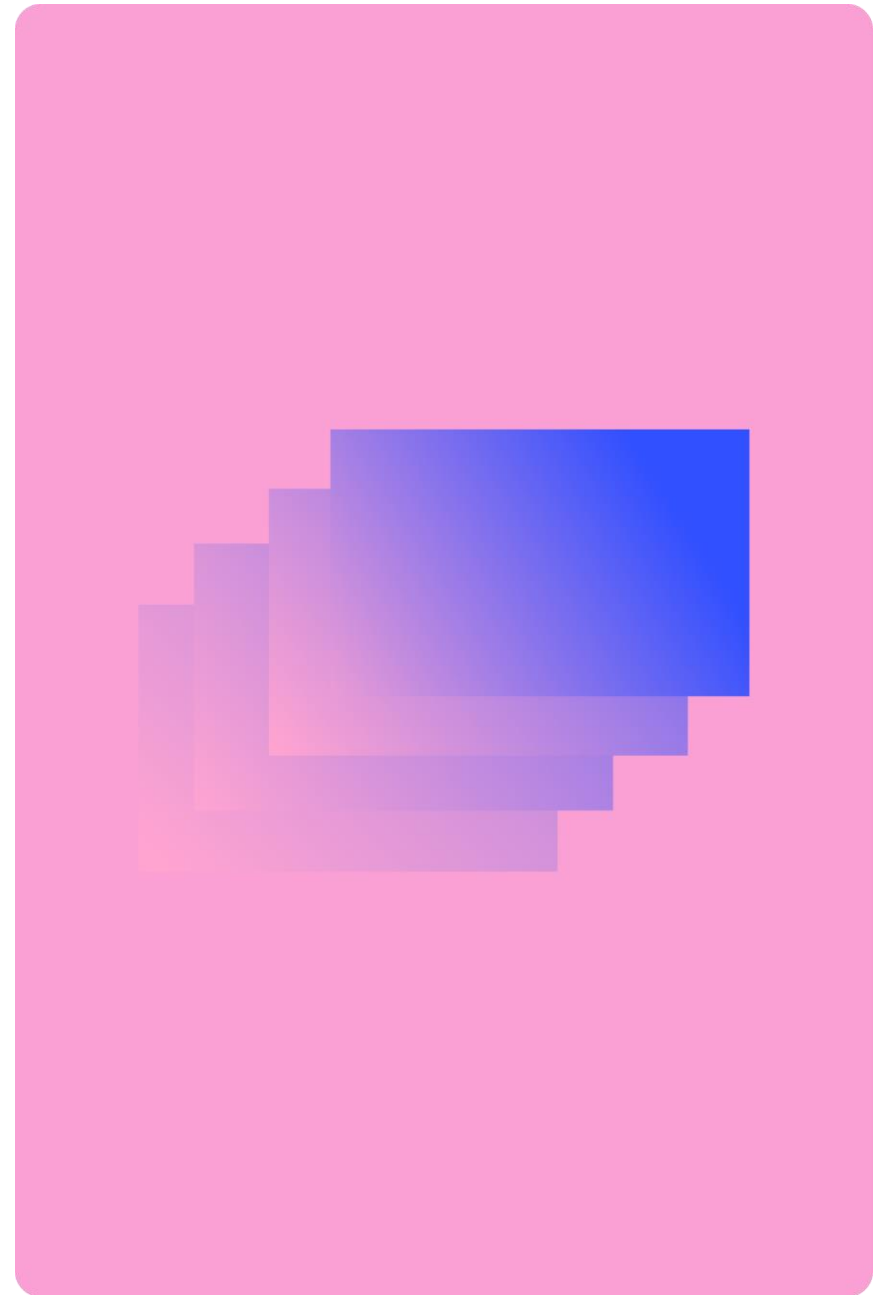


# The importance of Data Quality

In finetuning models to fit to business-specific purpose

- LLM's can be a great tool for general purposes, but they might not be fit to very domain specific business critical needs
- LLM's can be finetuned or made to fit better with a technique called RAG (Retrieval Augmented Generation) where model is supported with external data.
- The quality of the external data is critical, but implementing the technique might cause the model to hallucinate or give low quality results
- During the coming years we will see more purpose specific models that serve a particular business need and ability to finetune the models with higher quality

<https://medium.com/life-at-telkomsel/the-role-of-data-governance-in-the-era-of-ai-5027aeb00bf2>

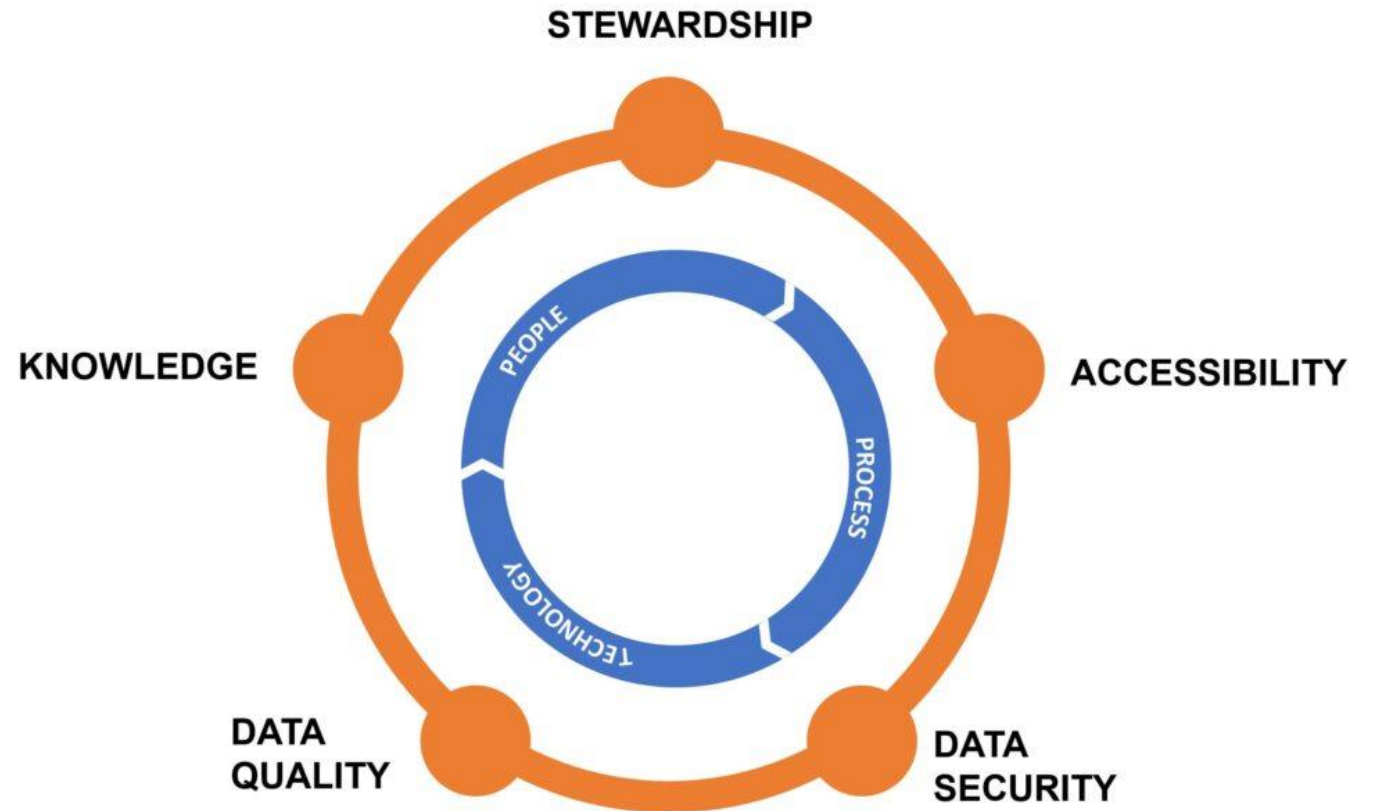


# Data Governance

## Explained

Data governance is everything you do to ensure data is secure, private, accurate, available, and usable.

It includes the actions people must take, the processes they must follow, and the technology that supports them throughout the data life cycle.



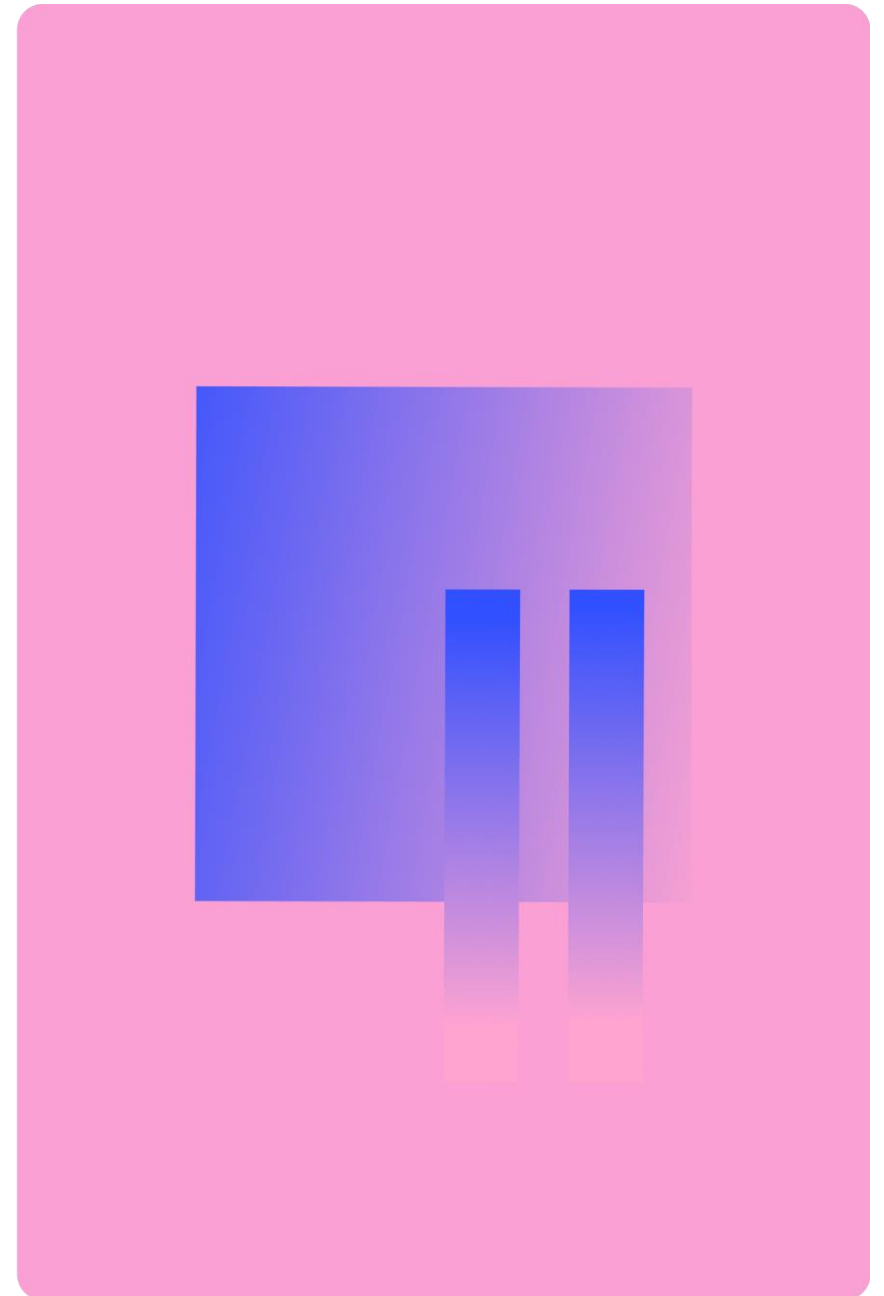
# Data governance issues impacting the quality of the AI systems

The more models utilize external business specific data, the maturity of the basic data governance and data quality impacts also to the quality of AI systems.

Imagine having issues with:

- The metadata quality: Customers, products, employees...
- The quality of the data used to prompt the models
- The ability to understand critical data flows and data lineage
- The data used to understand how processes currently work
- The data used to finetune the models
- Compliancy with GDPR and privacy regulation
- Ability to secure business critical or sensitive data
- Employee competencies in understanding the basics of data and AI

**Could you trust the model outputs knowing these issues?**



# AI governance is reliant on the principles and practices of data governance



## Six reasons why Data Governance is the bedrock for AI Governance

29 November, 2023 By Stefaan G. Verhulst, The GovLab and Friederike Schüür, UNICEF

Data and Research

Evidence

Evidence to Policy

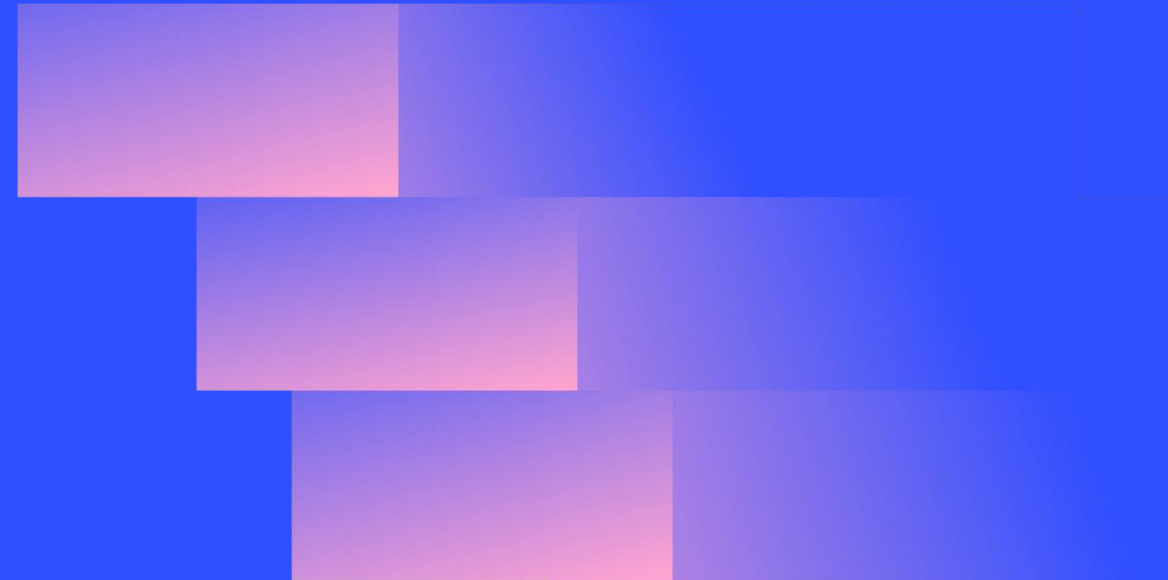


© UNICEF/UN0251908/Tadesse

1. Data governance **covers the full data lifecycle**, of which Artificial Intelligence is a part.
2. Data governance **enables** the development of responsible, fit-for-purpose AI systems.
3. Data governance **takes care of issues** that AI systems would otherwise inherit.
4. Data governance is **technology-agnostic**, and thus more holistic in nature.
5. The implementation, standardization and codification of data governance **provide valuable lessons** for AI governance.

# Key takeaways

- AI has the power to do good in the world for us humans, businesses and environment
- The more capable AI systems, the more challenging it becomes to ensure ethical alignment
- Responsible, transparent and safe use of AI requires regulation. But how to regulate without slowing down innovation?
- We need to make AI governance a normal practice in all AI development and use
- Responsible, transparent and safe use of AI requires increased emphasis also on data quality and governance





# Thank you! Kiitos!

## Contact us

Iiris Lahti  
Head of Services  
iiris@saidot.ai

## Follow us



[@ai\\_saidot](https://twitter.com/ai_saidot)



[company/saidot](https://www.linkedin.com/company/saidot)



[saidot.ai](https://www.saidot.ai)