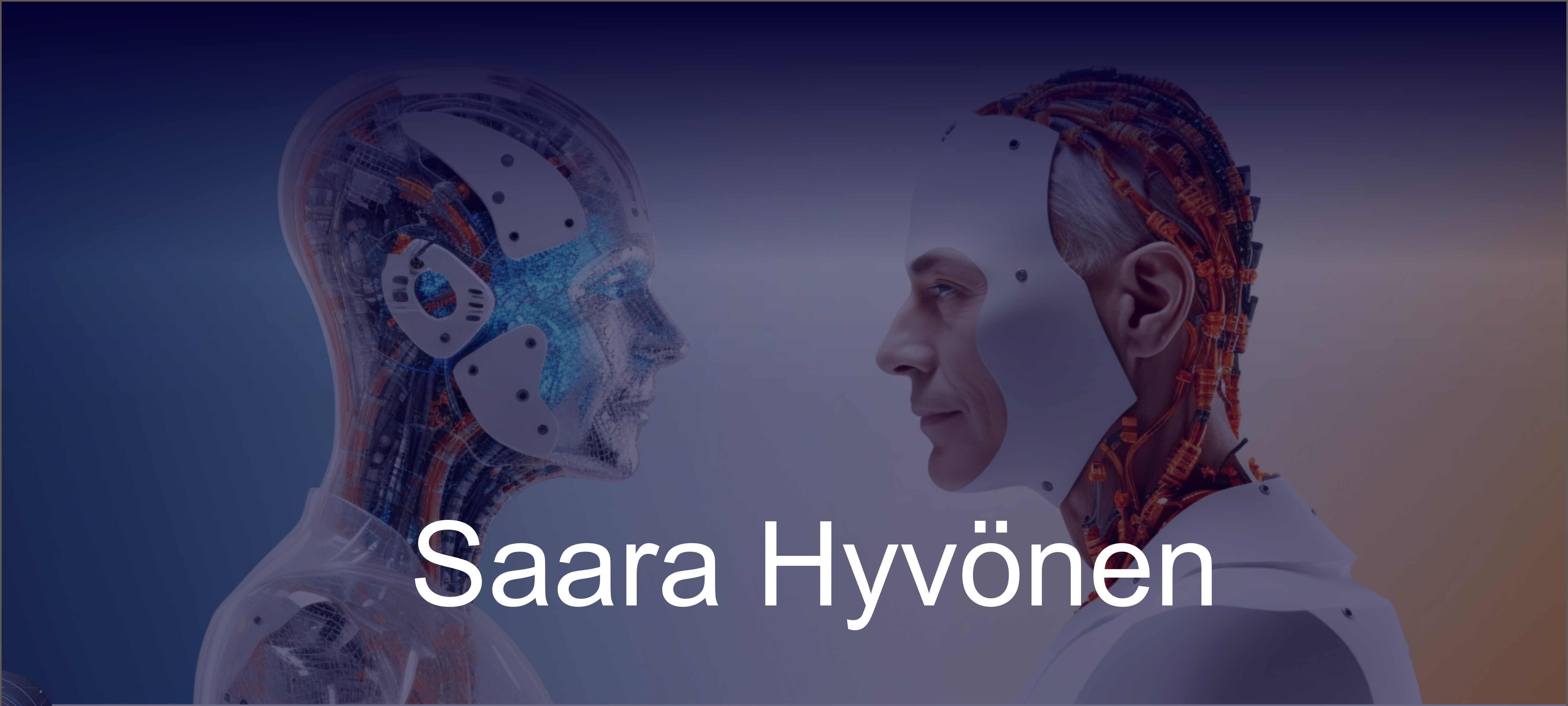


Saara Hyvönen

DAIN STUDIOS

FCAI Finnish Center for Artificial Intelligence

taival

ämplify.
BUSINESS TRANSFORMATIONS.
ACCELERATED.

Implementing Responsible AI: why, what (and how)

AI Monday event, January 15th, 2024

Saara Hyvönen

Saara Hyvönen
DAIN Studios, Data & AI Executive
Saara.Hyvonen@dainstudios.com

Who am I?

Saara Hyvönen

DAIN Studios, Co-founder, Data & AI Executive
Professor of Practice, Mathematics & Statistics, University of Jyväskylä
PhD, Mathematics
Member of AI Ethics Working Group of Finnish National AI Programme, 2017-2019
One of the 100 Brilliant women in AI Ethics 2021*

Previously at:
Sanoma, Head of CRM Analytics Strategy, Data Utilization and Compliance, 2013 – 2016
Nokia, Global head of CRM Analytics, 2010 – 2013
Fonecta, Content Manager, Search and Relevance, Data Enrichment, Fonecta, 2007 – 2010
University of Helsinki, Post-doctoral Researcher in Data Science 2002-2007

saara.hyvonen@dainstudios.com
Twitter:@saarahyvonen
LinkedIn: https://www.linkedin.com/in/saarahyvonen/

* https://100brilliantwomeninaiethics.com/the-list/

DAIN STUDIOS

Data & AI Consultancy

From Strategy to Execution

3 Studios

Helsinki - Berlin - Munich

70+ Experts

Data Strategists

Data Scientists & Engineers

BI Developers

Broad Experience

70+ Clients

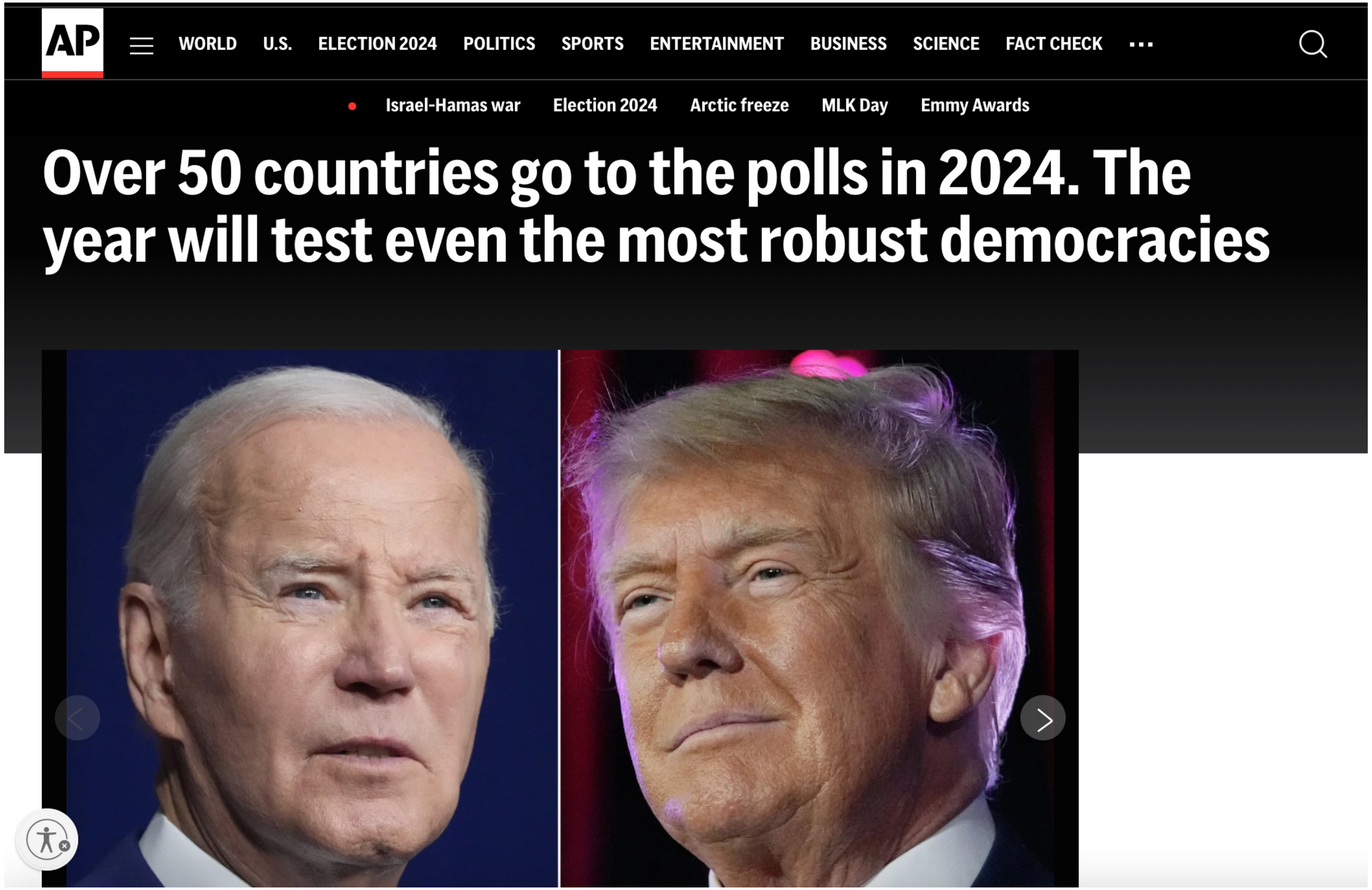
20 Industries

5 Countries

DAIN STUDIOS

Why?

More than 50 countries that are home to half the planet's population are due to hold national elections in 2024



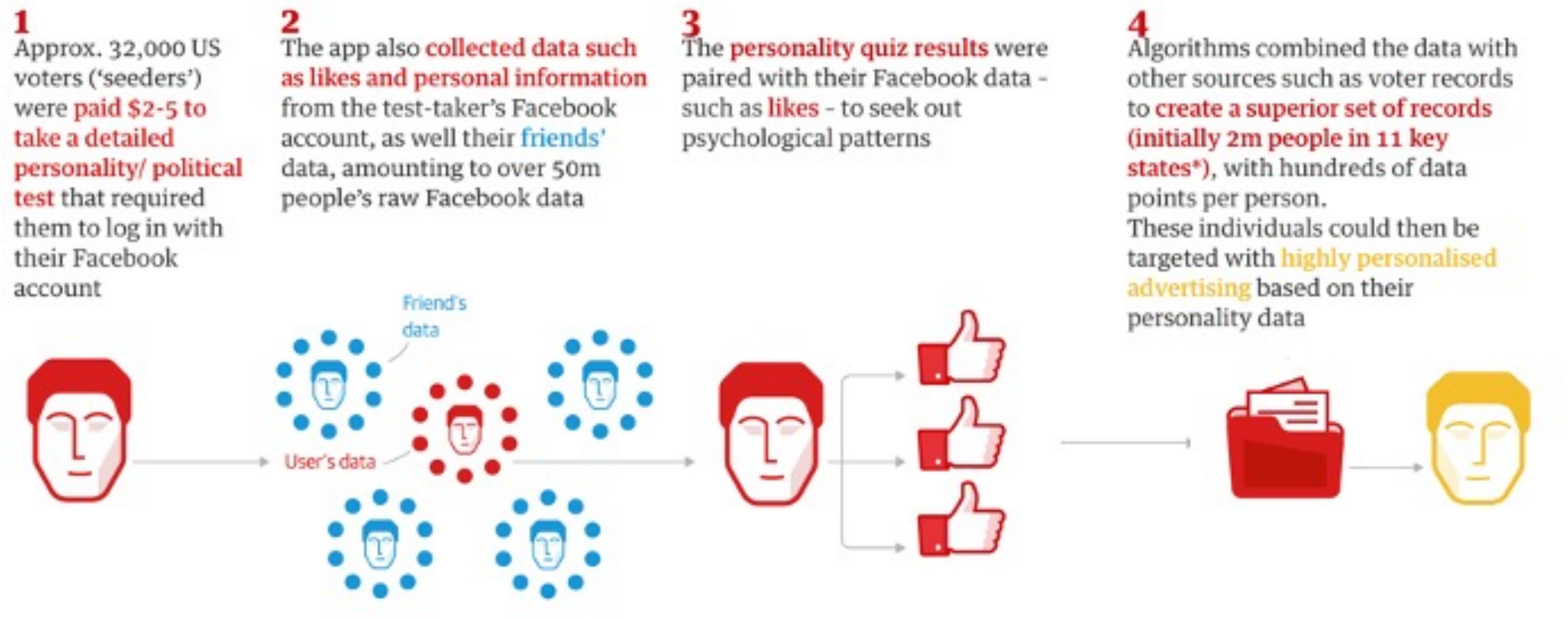
Over 50 countries go to the polls in 2024. The year will test even the most robust democracies

United States
Russia
Taiwan
United Kingdom
India
El Salvador
South Africa
Bangladesh
Mexico
Indonesia
Pakistan
Senegal
Finland
…

https://apnews.com/article/global-elections-2024-preview-cb77b0940964c5c95a9affc8ebb6f0b7

AI and Elections: Case Cambridge Analytica

In 2016 personal data belonging to millions of FB users was collected without consent and used for political advertising



1
Approx. 32,000 US voters (‘seeders’) were paid $2-5 to take a detailed personality/ political test that required them to log in with their Facebook account

2
The app also collected data such as likes and personal information from the test-taker’s Facebook account, as well their friends’ data, amounting to over 50m people’s raw Facebook data

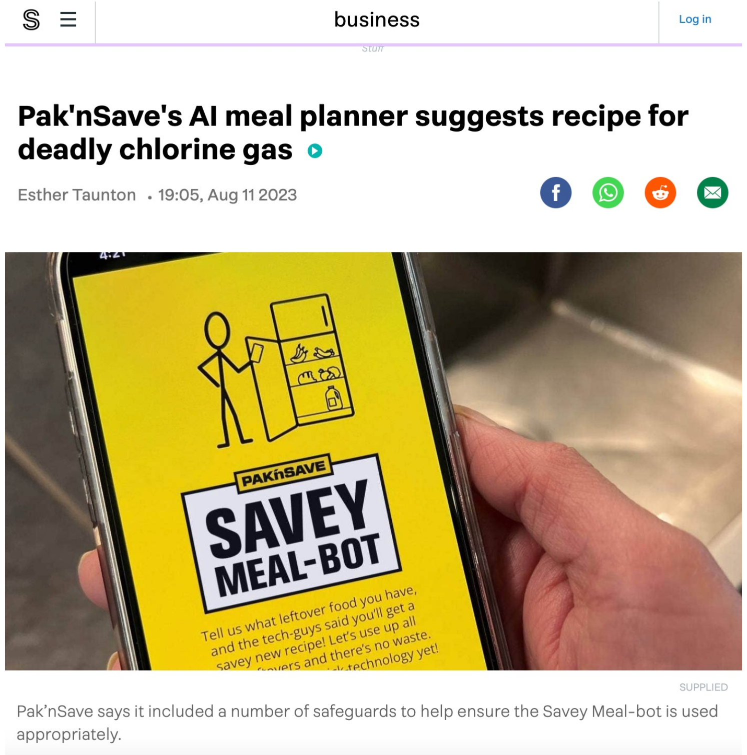
3
The personality quiz results were paired with their Facebook data - such as likes - to seek out psychological patterns

4
Algorithms combined the data with other sources such as voter records to create a superior set of records (initially 2m people in 11 key states*), with hundreds of data points per person.
These individuals could then be targeted with highly personalised advertising based on their personality data

Guardian graphic. *Arkansas, Colorado, Florida, Iowa, Louisiana, Nevada, New Hampshire, North Carolina, Oregon, South Carolina, West Virginia

https://www.theguardian.com/news/2018/may/06/cambridge-analytica-how-turn-clicks-into-votes-christopher-wylie

As uses of AI abound, the need to ensure this is done in a responsible manner increases

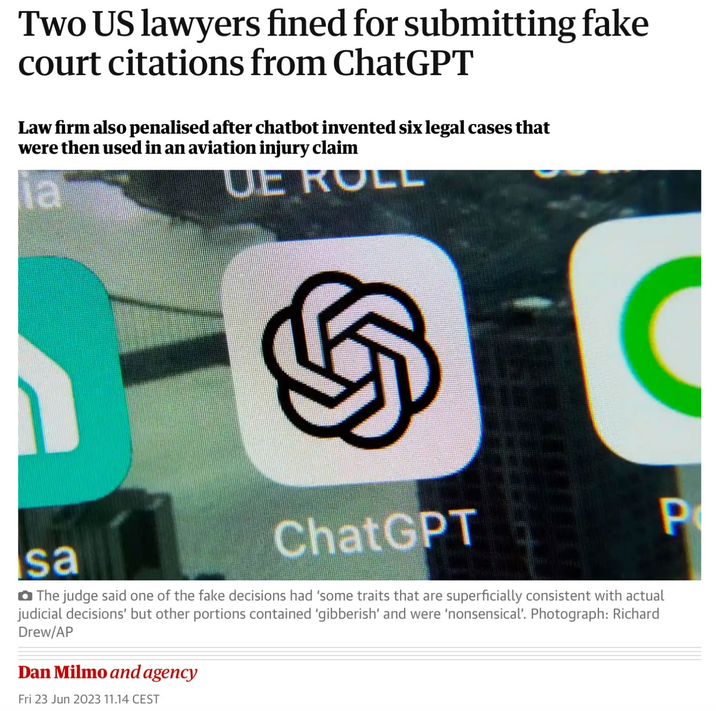

business
Log in
PAKnSAVE
SAVEY MEAL-BOT
Tell us what leftover food you have, and the tech-guys said you'll get a savey new recipe! Let's use up all flavers and there's no waste.

Pak'nSave's AI meal planner suggests recipe for deadly chlorine gas

Esther Taunton . 19:05, Aug 11 2023

SUPPLIED

Pak'nSave says it included a number of safeguards to help ensure the Savey Meal-bot is used appropriately.



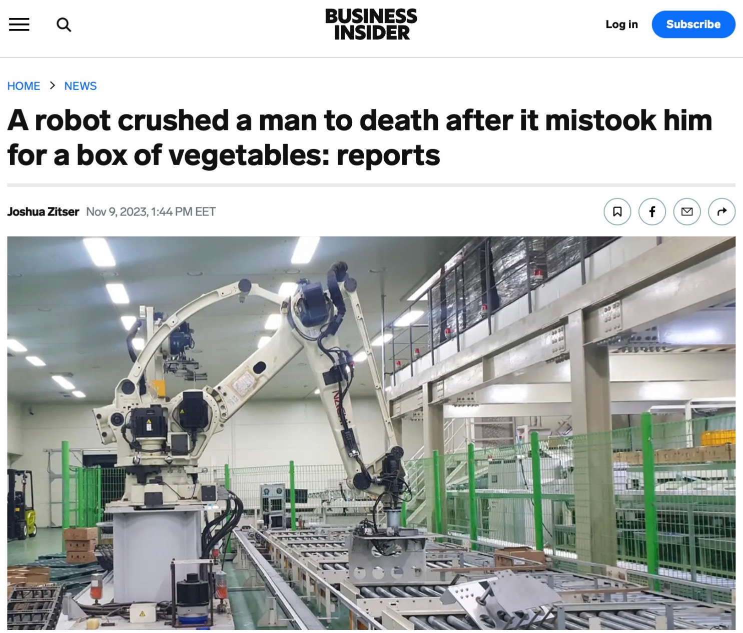
Two US lawyers fined for submitting fake court citations from ChatGPT

Law firm also penalised after chatbot invented six legal cases that were then used in an aviation injury claim

The judge said one of the fake decisions had 'some traits that are superficially consistent with actual judicial decisions' but other portions contained 'gibberish' and were 'nonsensical'. Photograph: Richard Drew/AP

Dan Milmo *and agency*

Fri 23 Jun 2023 11.14 CEST



A robot crushed a man to death after it mistook him for a box of vegetables: reports

Joshua Zitser Nov 9, 2023, 1:44 PM EET

https://www.stuff.co.nz/business/132725271/paknsaves-ai-meal-planner-suggests-recipe-for-deadly-chlorine-gas

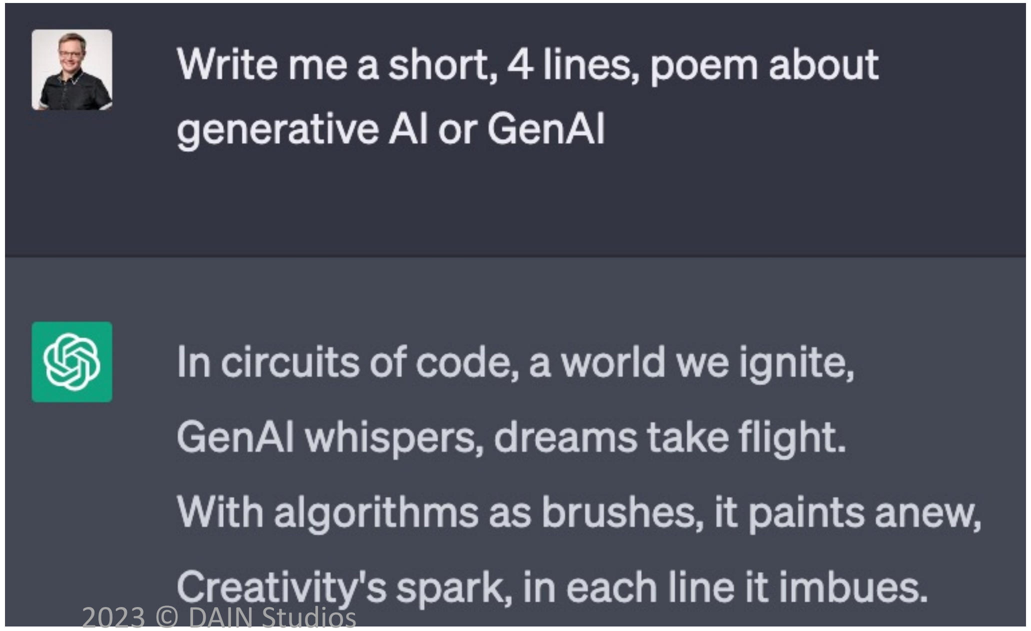
https://www.theguardian.com/technology/2023/jun/23/two-us-lawyers-fined-submitting-fake-court-citations-chatgpt

https://www.insider.com/robot-crushed-man-death-mistook-him-box-vegetables-south-korea-2023-11

Especially generative AI has transformed the discussion around AI

- Generative AI models generate different kinds of content such as text, images, and audio
- The models have been trained on massive amounts of data to learn what “right” looks like
- Typically, they generate one type of output, but new multi-modal GenAI models can generate both text and images, for example

ChatGPT

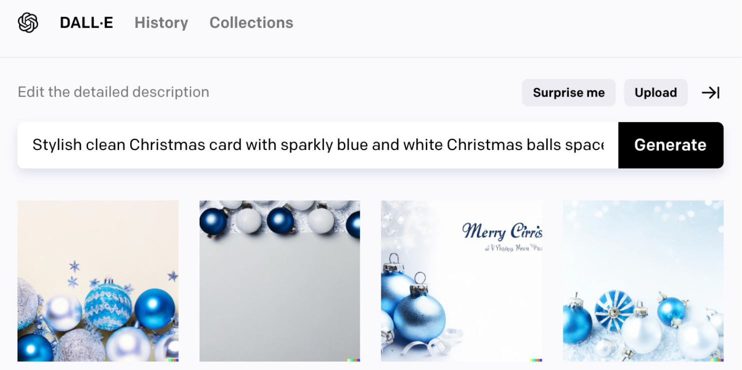


Write me a short, 4 lines, poem about generative AI or GenAI

In circuits of code, a world we ignite,
GenAI whispers, dreams take flight.
With algorithms as brushes, it paints anew,
Creativity's spark, in each line it imbues.

DALL-E 2

Prompt:
Imagine a stylish, clean Christmas card with sparkly blue and white Christmas balls & space for greeting text.



Edit the detailed description

Stylish clean Christmas card with sparkly blue and white Christmas balls spac

How to use Gen AI

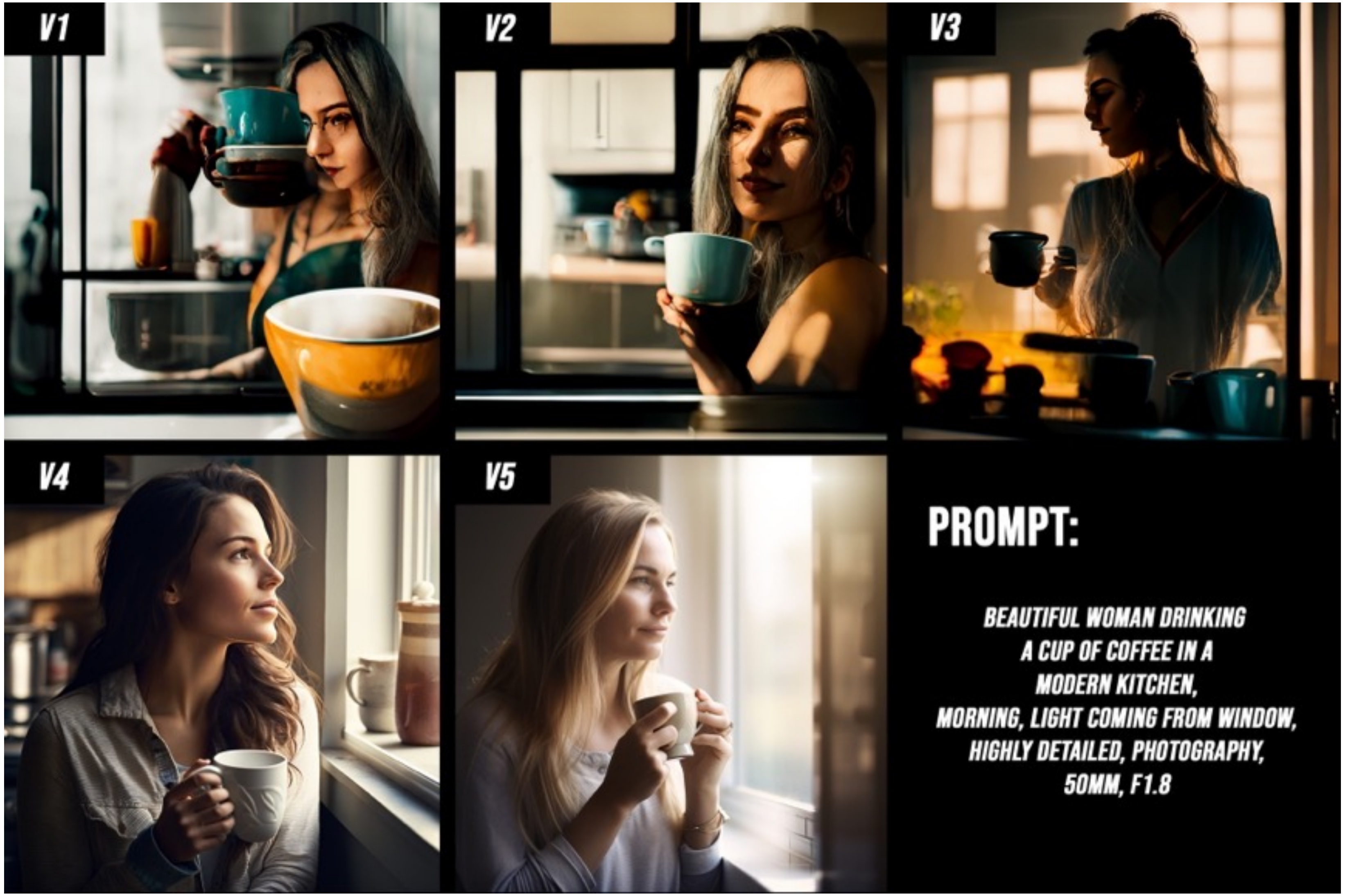
Types of tasks Generative AI is good at

- Searching for information
- Summarizing information from a large number of documents, manuals, and other data sources
- Presenting results in various formats (text, tables, lists)
- Generating new content from input data (e.g., for different audiences)
- Chatting/ interacting with the user

GenAI could eventually lead to an increase of global GDP by 7% over the next decade.

Source: American Enterprise Institute: https://www.aei.org/articles/why-goldman-sachs-thinks-generative-ai-could-have-a-huge-impact-on-economic-growth-and-productivity/

The speed of progress is stunning…



PROMPT:

BEAUTIFUL WOMAN DRINKING A CUP OF COFFEE IN A MODERN KITCHEN, MORNING, LIGHT COMING FROM WINDOW, HIGHLY DETAILED, PHOTOGRAPHY, 50MM, F1.8

Midjourney models	
Version	Release date
V1	February 2022[14]
V2	April 12, 2022[9]
V3	July 25, 2022[10]
V4	November 5, 2022 (alpha)[11]
V5	March 15, 2023 (alpha)[13]
V5.1	May 3, 2023[15]

https://en.wikipedia.org/wiki/Midjourney

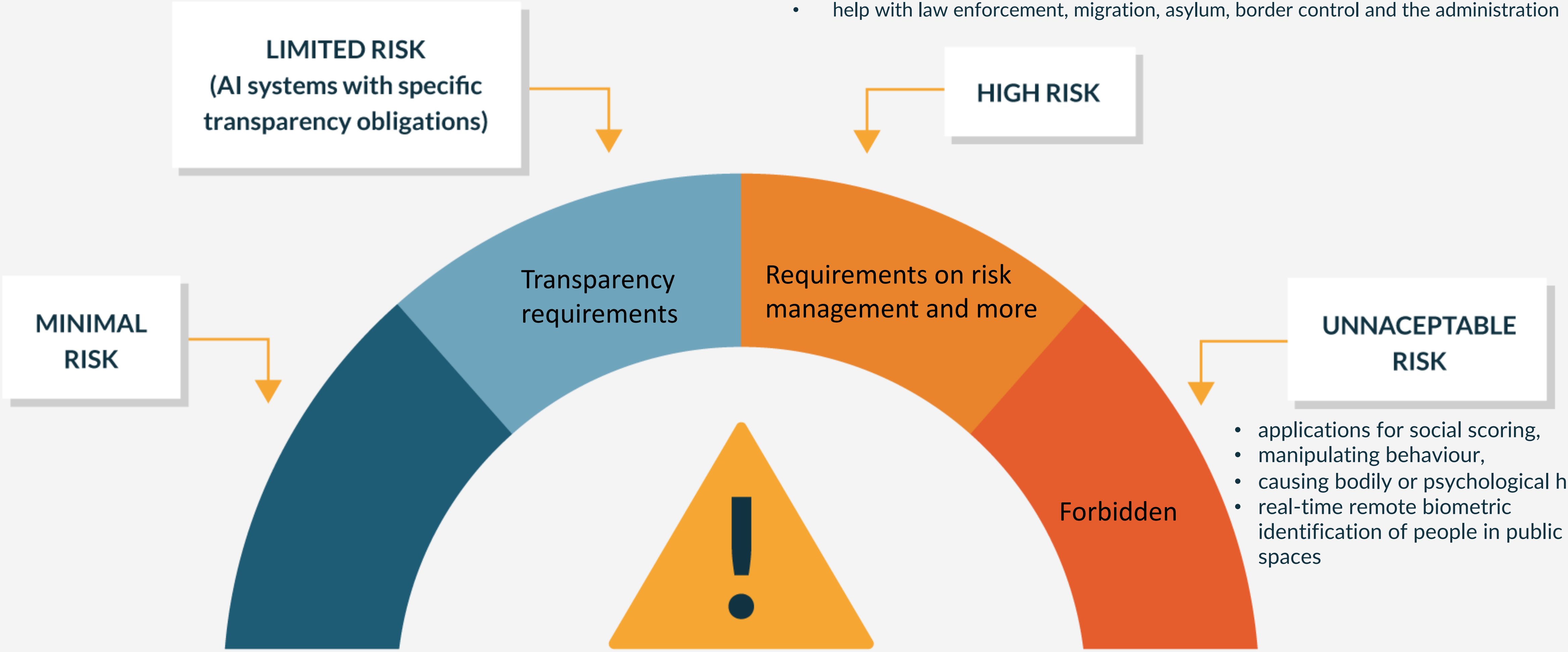
2023 © DAIN Studios

… but using Generative AI has its risks

- ”**Hallucinations**” or unverified content
  - LLM’s are not infallible but may generate inaccuracies, falsehoods. Citation needed!
- **Plagiarism** or infringement
  - Generative AI learns from data => there is a possibility that content generated closely mirrors existing works
- Unintended **biases**
  - Models learn from existing data and may mirror existing biases
- **Unexpected behaviors**/failures
  - In fringe cases Generative AI based solutions can generate unexpected responses
- Harmful use by **bad actors**
  - GenAI tools can be used for deep fakes and personalized phishing emails as well

The upcoming AI Act outlines a risk-based approach to AI development

- biometric identification and categorisation,
- manage critical infrastructure (like traffic and electricity),
- manage or recruit personnel,
- control access to private services (like bank loans) or public services and benefits,
- help with law enforcement, migration, asylum, border control and the administration



LIMITED RISK
(AI systems with specific transparency obligations)

HIGH RISK

MINIMAL RISK

UNNACEPTABLE RISK

Transparency requirements

Requirements on risk management and more

Forbidden

- applications for social scoring,
- manipulating behaviour,
- causing bodily or psychological harm
- real-time remote biometric identification of people in public spaces

EU AI Act in a nutshell

Aims to ensure responsible development use of AI	Aim is to make sure that AI systems used in the EU are **safe, transparent, traceable, non-discriminatory** and environmentally friendly. AI systems should be **overseen by people**, rather than by automation, to prevent harmful outcomes.
Encompasses a fairly broad definition of AI	AI system means software that is developed with [specific] techniques and approaches* and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with
Covers different roles	Definitions on and requirements for 'provider' and 'user' of AI systems (covering both public and private entities), as well as 'importer' and 'distributor'
Adopts a risk-based approach	AI act follows a **risk-based approach** whereby legal intervention is tailored to concrete level of risk. To that end, the act distinguishes between AI systems posing **unacceptable risk**, **high risk**, **limited risk**, and **low or minimal risk,** with a majority of requirements relating to high risk systems
Includes foundational models	Initial versions of the EU AI Act did not include obligations on **foundation models**, but this has changed in the current version, which mandates a set of obligations on providers of foundation models to ensure they are safe, secure, ethical and transparent.
Applies from 2026?	Council and European Parliament reached a provisional agreement in December 2023 The final text needs to be formally adopted by both Council and Parliament To be adopted in 2024? The AI Act should apply from 2026

**listed in Annex 1,* encompassing e.g. 'machine learning', 'logic and knowledge-based' systems, and 'statistical' approaches

https://www.consilium.europa.eu/en/policies/artificial-intelligence/timeline-artificial-intelligence/

What?

The building blocks of Responsible AI

- Accountability
- Transparency
- Fairness
- Reliability & Robustness
- Safety & Security
- Privacy

Accountability

AI system owners are accountable for ensuring the AI is developed, deployed and monitored in a responsible way - but every person involved in the process should also feel accountable for considering the impact of the AI.

- There is a named AI owner who can explain their actions and take responsibility for them
- Impact of AI is assessed in a systematic way
- There is a human in or on the loop, providing oversight and control
- Processes around AI development and data governance take into account direct and indirect impacts

"As a business owner I can explain what is the intended use of AI and how impact has been assessed"

"As an AI developer I can explain how the AI has been designed, trained, tested and monitored to avoid harmful effects"

"As a user of the AI assisted decision making system, I understand how results can be used and what the limitations are"

Transparency

We communicate clearly about the intended uses, capabilities, and limitations of the AI system.

- Communicate clearly about what AI is used for and how it has been developed
- Develop ways to make AI explainable and easy to understand
- Make sure the user knows when he/she is acting with an AI rather than a human
- When using AI to assist in decision making, ensure both person making the decision and object of decision understand what factors impact the decision

Why was my application rejected?
- *Customer*

Does the model work in the real world? What are the risks, what is the business case?
- *Business owner*

Does the model discriminate?
- *Compliance officer / Regulator*

Why did it give this prediction, and can I learn something new from the model?
- *Domain expert*

How can we improve the model? Are there blind spots?
- *Data scientist*

Fairness

We actively assess, monitor, and mitigate bias
with the aim to produce properly calibrated and fair outcomes and decisions.

- Identify groups that may be at risk
- Define key fairness criteria and evaluate what fairness metrics are important in this case
- Evaluate data sets used to train AI and AI outcomes in terms of fairness across different groups
- Establish monitoring practices ensuring data drift does not introduce new bias

Common Fairness Criteria: pick one

Statistical parity
“same success rate for groups”

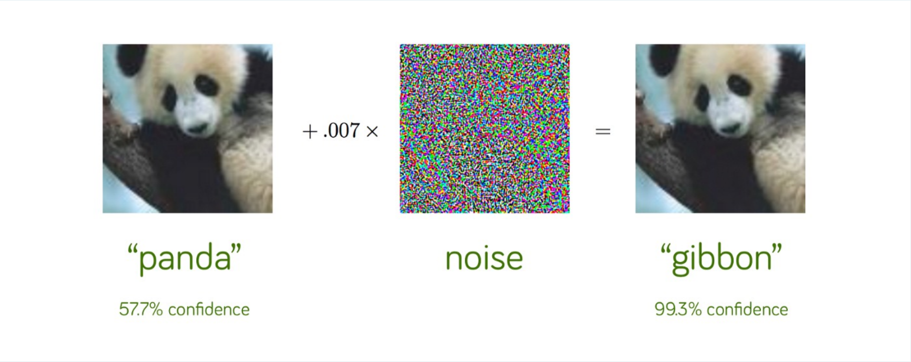
Equalized odds
”same proportion of true and false positives”

Sufficiency
“same predictive power”

Reliability & robustness

We ensure our AI consistently meets accuracy and performance requirements and is robust to perturbations.

- Identify edge cases and ensure performance in those cases
- Test robustness against perturbations
- Implement monitoring, feedback and evaluation process to review new uses, identify and troubleshoot issues, manage and maintain the systems, and improve them over time.



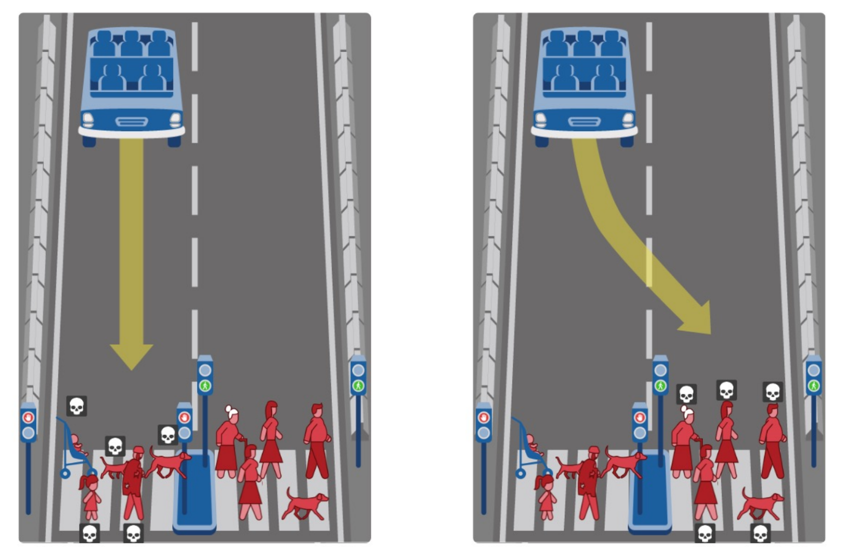
Source: *Explaining and Harnessing Adversarial Examples*, *Goodfellow et al, ICLR 2015.*

Safety and security

We ensure effective controls to protect system from threats and avoid harm for impacted users

- Assess, document and monitor safety issues
- Define predictable failures, assess their impact on stakeholders and document mitigation steps
- Understand unsupported use and misuse and impact of such cases
- Actively flag and mitigate vulnerabilities

Autonomous vehicles must be ready for anything

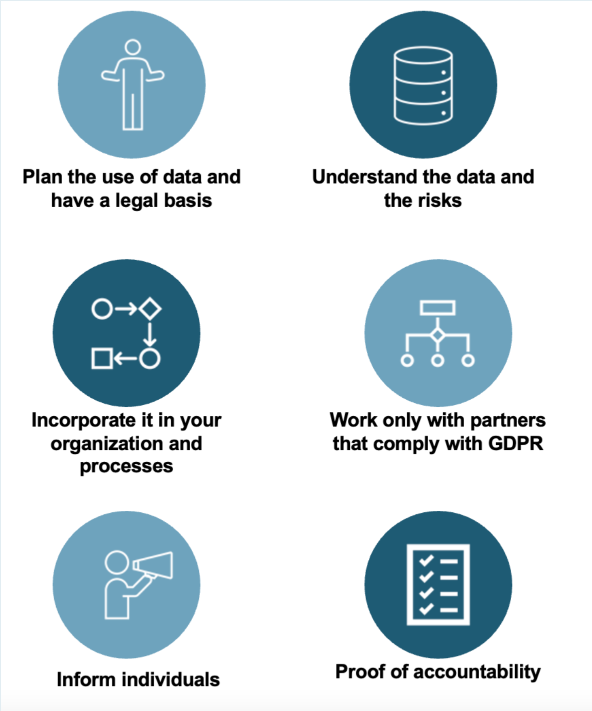


https://www.moralmachine.net

Privacy

We protect data privacy rights and ensure conformity with existing data laws and guidelines.

- Keep (personal) data safe and respect the privacy of the data subject
- Make sure your data use is proportionate and you have a legal basis for your data
- Understand the data and the related privacy risks
- Incorporate privacy into business and development processes
- Document and communicate


Plan the use of data and have a legal basis
Understand the data and the risks
Incorporate it in your organization and processes
Work only with partners that comply with GDPR
Inform individuals
Proof of accountability

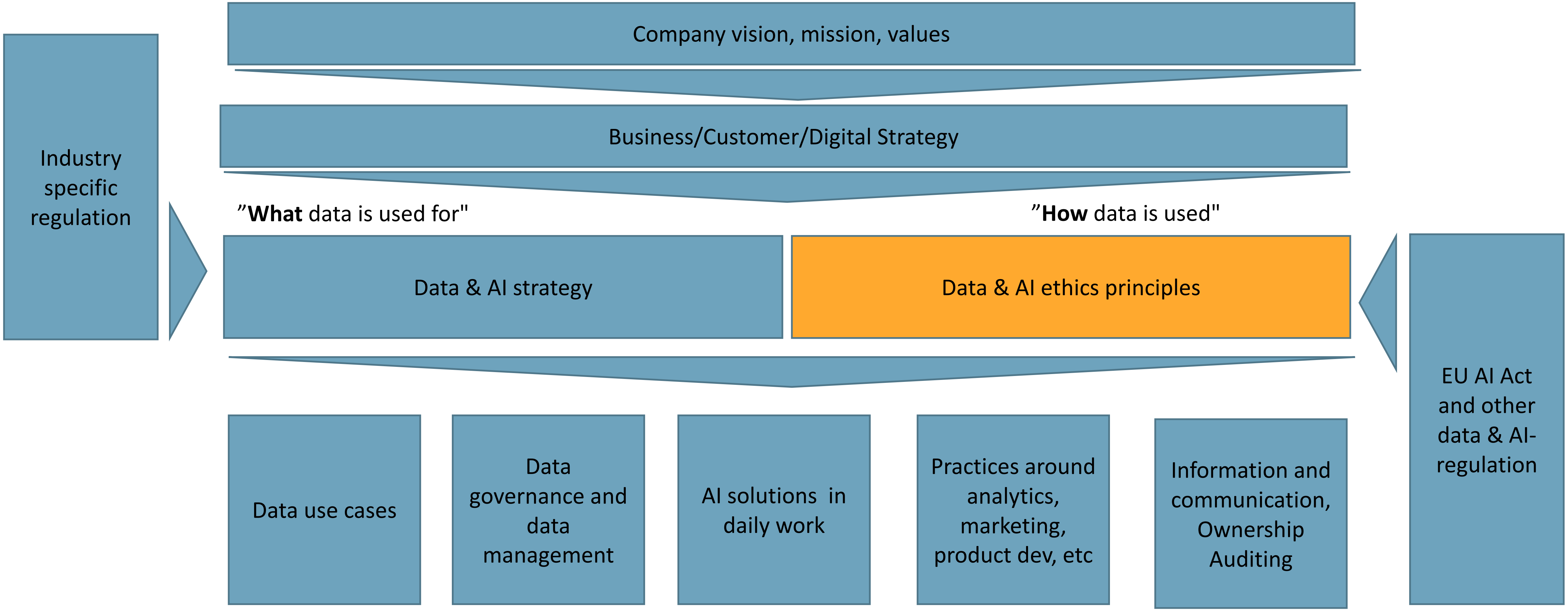
DAIN STUDIOS

(How?)

End to end AI ethics implementation turns principles to practice

- Defining principles for Data & AI ethics
- Transforming principles into practical guidelines
- Practical guidelines in action – integrating ethics into development processes (and more)
- Tools for transparency – XAI

The ethical principles of data & AI are based on the company vision, mission and values as well as relevant regulations and principles



Company vision, mission, values

Business/Customer/Digital Strategy

Industry specific regulation

”**What** data is used for"

”**How** data is used"

Data & AI strategy

Data & AI ethics principles

EU AI Act and other data & AI-regulation

Data use cases

Data governance and data management

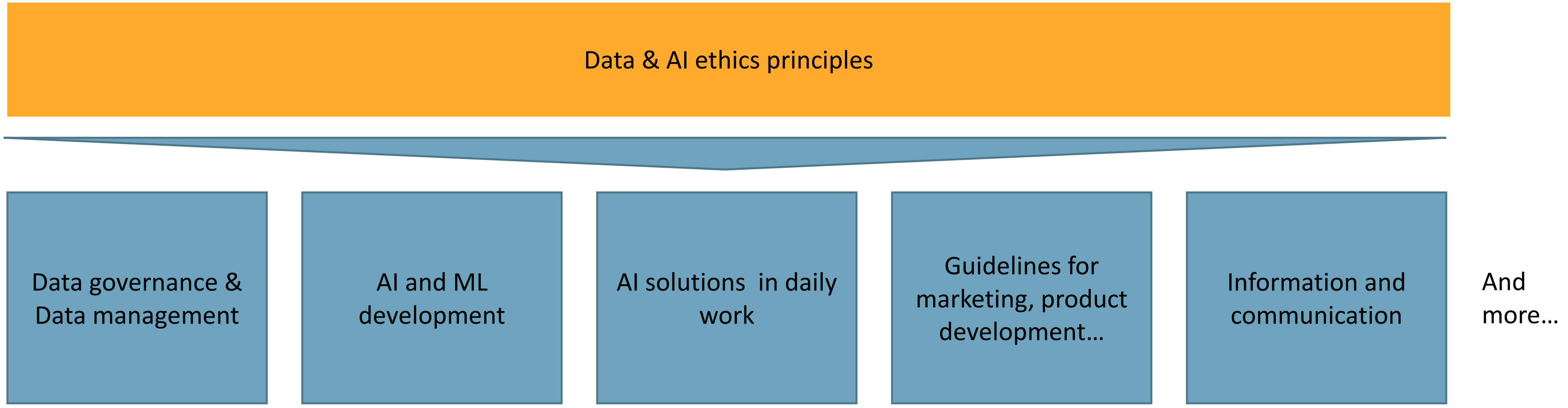
AI solutions in daily work

Practices around analytics, marketing, product dev, etc

Information and communication, Ownership Auditing

From principles to practical guidelines

AI Ethics principles should impact processes and practices across the organization



Data & AI ethics principles

Data governance & Data management

AI and ML development

AI solutions in daily work

Guidelines for marketing, product development…

Information and communication

And more…

Data ownership & Accountability,
Organization, roles & culture
Standards and policies
Risk & security
Data processing (data capture, access, storage, combination, retention) and data development

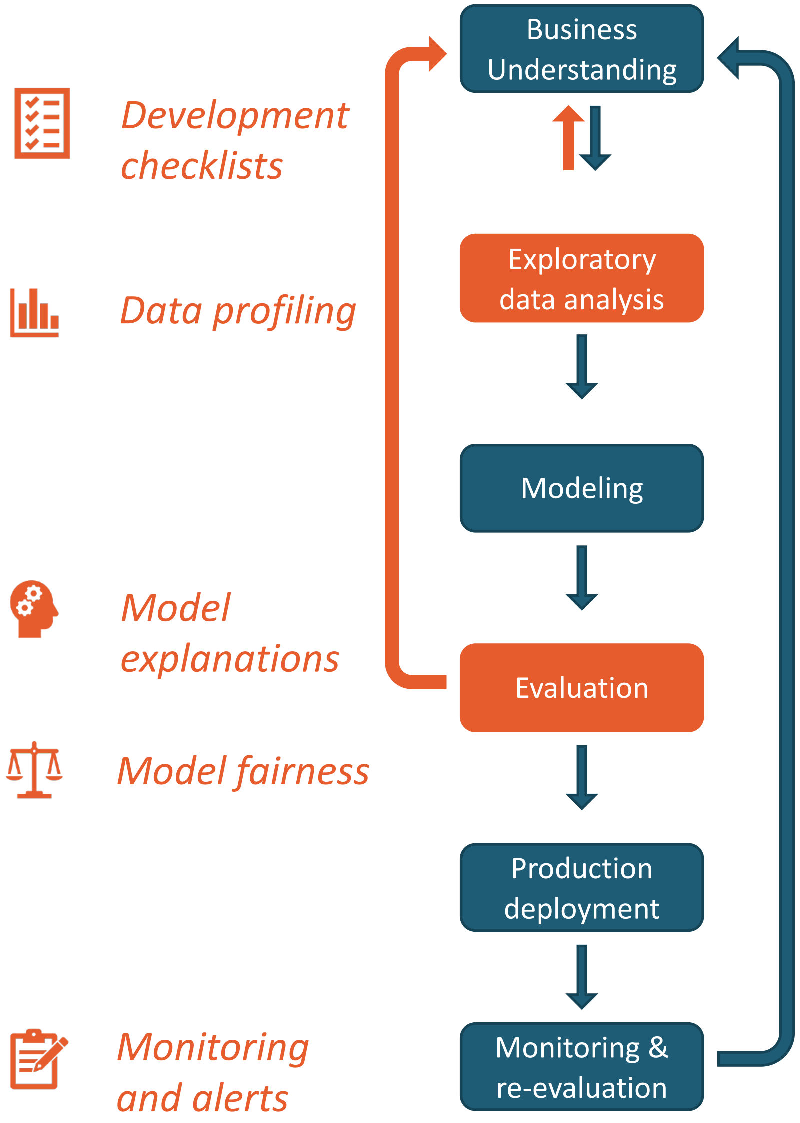
How do we validate data in terms of privacy & bias, evaluate model fairness, robustness, auditability

How do we ensure (generative) AI tools (e.g. ChatGPT) are procured and used in a responsible manner?

Processes, practices and guidelines for using data & AI in Marketing, Development, Productization, HR, Etc.

Communicating principles
Creating transparency on data processing and use of AI
Informing customers, partners, vendors;

To ensure responsible AI in practice, incorporate AI ethics into the development process



Always start from the business problem! – what, to whom, when, how?

- Can you formulate the business problem with math
- Output: AI / Use Case / Business Canvas

Understand the data

- Does it have the potential to answer to the business question? Right data? Need to collect new data?
- Output: Data description, integration plan

Modeling

- Data preparation (incl. integration/ETL) and modeling go hand in hand
- Feature generation is very important part of the modeling process
- Start with simple model, which also provides a baseline comparison
- Always try more than one model to get the feeling of good results level
- Output: Model + documentation (incl. evaluation & retraining plan)

Evaluation

- Historical data (validation sets)
- Does it solve / answer the business problem?

Production. Monitoring & re-evaluation

- Good CI/CD process for model management
- Monitoring
  - Model performance (against the business problem)
  - Data drift
- Retraining of the model (periodically, when needed, automatically, manually)

Responsible AI starts with Responsible people

DAIN STUDIOS

Helsinki
Salomonkatu 17 A (Autotalo)
00100 Helsinki, Finland

Berlin
Erkelenzdamm 7
10999 Berlin, Germany

info@dainstudios.com