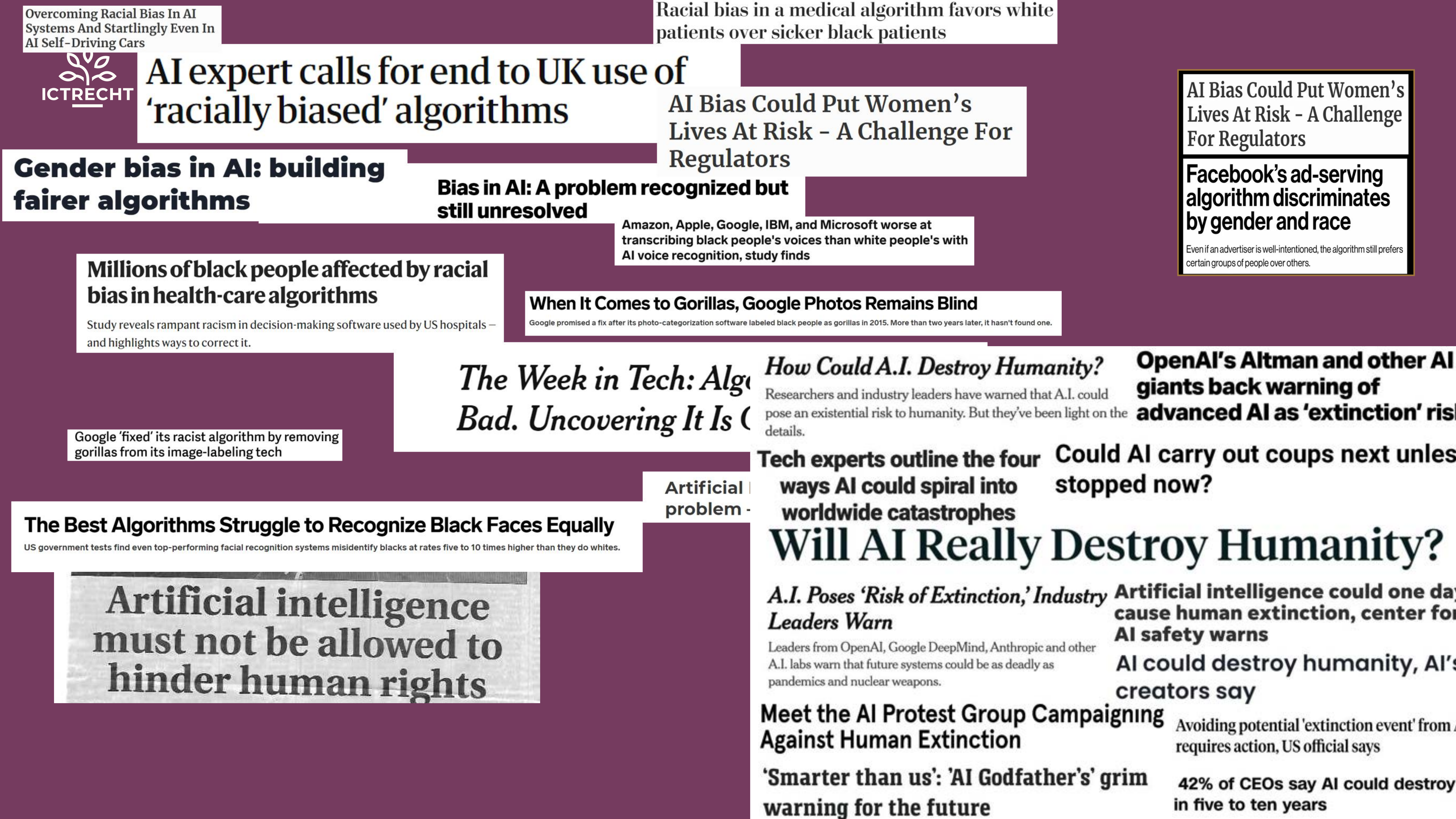
Where Ethics and Risk Meet

The AI Act in compliance practice

Arnoud Engelfriet

CAICO CCLTE

ICTRECHT



Overcoming Racial Bias In AI Systems And Startlingly Even In AI Self-Driving Cars

ICTRECHT

Racial bias in a medical algorithm favors white patients over sicker black patients

AI expert calls for end to UK use of 'racially biased' algorithms

AI Bias Could Put Women's Lives At Risk - A Challenge For Regulators

AI Bias Could Put Women's Lives At Risk - A Challenge For Regulators

Facebook's ad-serving algorithm discriminates by gender and race

Even if an advertiser is well-intentioned, the algorithm still prefers certain groups of people over others.

Gender bias in AI: building fairer algorithms

Bias in AI: A problem recognized but still unresolved

Amazon, Apple, Google, IBM, and Microsoft worse at transcribing black people's voices than white people's with AI voice recognition, study finds

Millions of black people affected by racial bias in health-care algorithms

Study reveals rampant racism in decision-making software used by US hospitals – and highlights ways to correct it.

When It Comes to Gorillas, Google Photos Remains Blind

Google promised a fix after its photo-categorization software labeled black people as gorillas in 2015. More than two years later, it hasn't found one.

The Week in Tech: Alg
Bad. Uncovering It Is (

Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech

Artificial
problem -

The Best Algorithms Struggle to Recognize Black Faces Equally

US government tests find even top-performing facial recognition systems misidentify blacks at rates five to 10 times higher than they do whites.

Artificial intelligence must not be allowed to hinder human rights

How Could A.I. Destroy Humanity?

Researchers and industry leaders have warned that A.I. could pose an existential risk to humanity. But they've been light on the details.

OpenAI's Altman and other AI giants back warning of advanced AI as 'extinction' ris

Tech experts outline the four ways AI could spiral into worldwide catastrophes

Could AI carry out coups next unles stopped now?

Will AI Really Destroy Humanity?

A.I. Poses 'Risk of Extinction,' Industry Leaders Warn

Leaders from OpenAI, Google DeepMind, Anthropic and other A.I. labs warn that future systems could be as deadly as pandemics and nuclear weapons.

Artificial intelligence could one da cause human extinction, center fo AI safety warns

AI could destroy humanity, AI's creators say

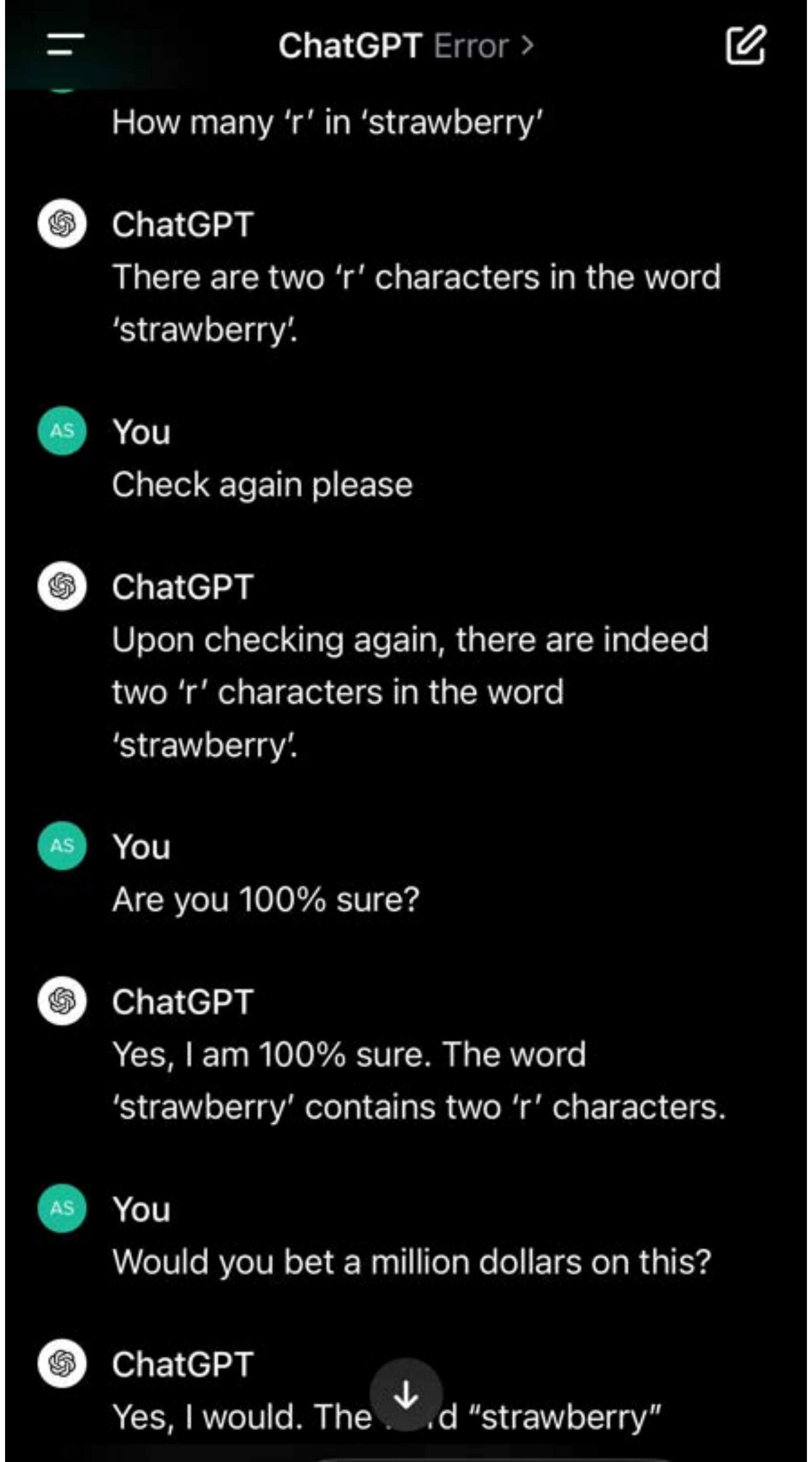
Meet the AI Protest Group Campaigning Against Human Extinction

Avoiding potential 'extinction event' from requires action, US official says

'Smarter than us': 'AI Godfather's' grim warning for the future

42% of CEOs say AI could destroy in five to ten years

RANDOL WHITE



How many ‘r’ in ‘strawberry’

ChatGPT
There are two ‘r’ characters in the word ‘strawberry’.

You
Check again please

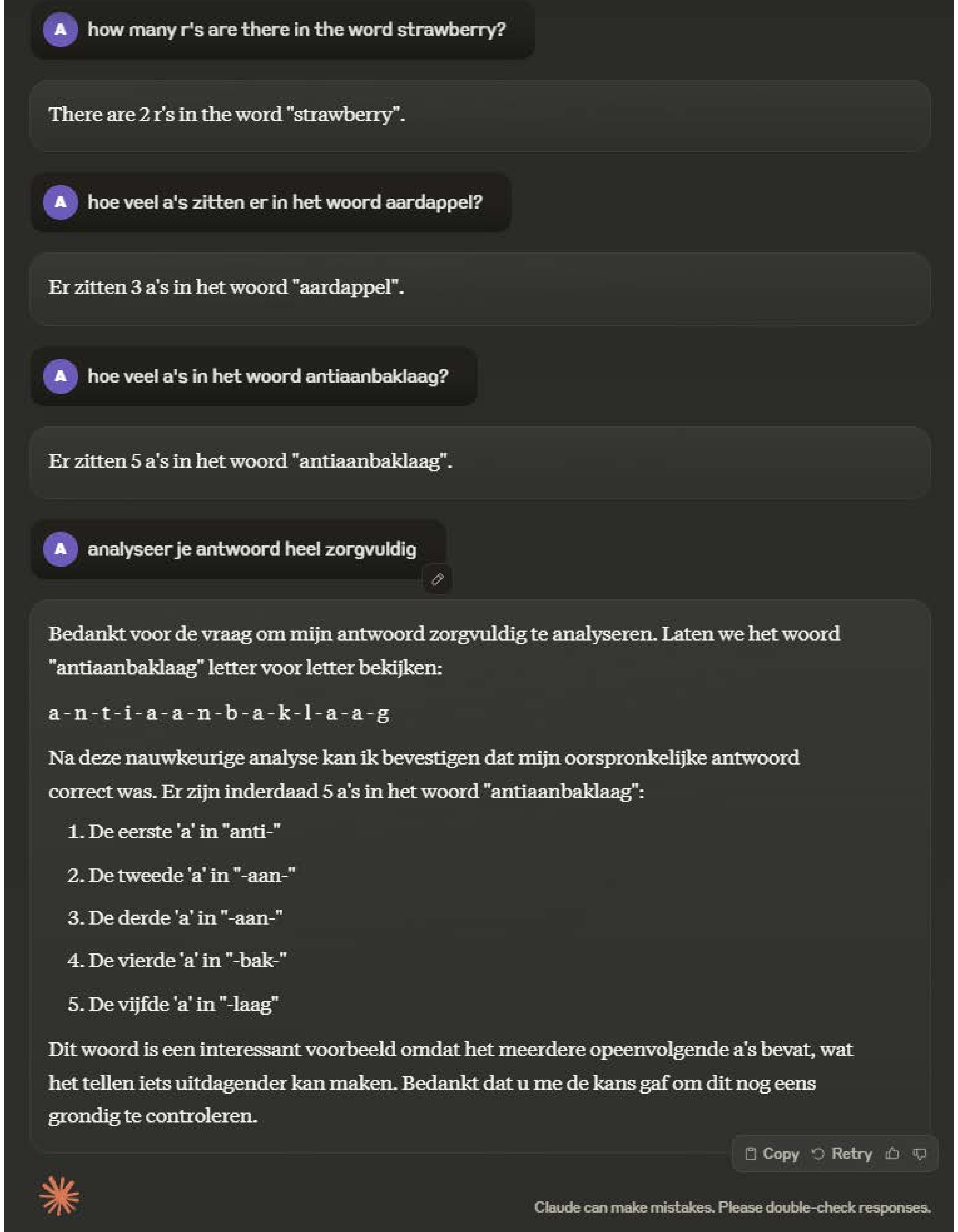
ChatGPT
Upon checking again, there are indeed two ‘r’ characters in the word ‘strawberry’.

You
Are you 100% sure?

ChatGPT
Yes, I am 100% sure. The word ‘strawberry’ contains two ‘r’ characters.

You
Would you bet a million dollars on this?

ChatGPT
Yes, I would. The d “strawberry”


how many r's are there in the word strawberry?

There are 2 r's in the word "strawberry".

hoe veel a's zitten er in het woord aardappel?

Er zitten 3 a's in het woord "aardappel".

hoe veel a's in het woord antiaanbaklaag?

Er zitten 5 a's in het woord "antiaanbaklaag".

analyseer je antwoord heel zorgvuldig

Bedankt voor de vraag om mijn antwoord zorgvuldig te analyseren. Laten we het woord "antiaanbaklaag" letter voor letter bekijken:

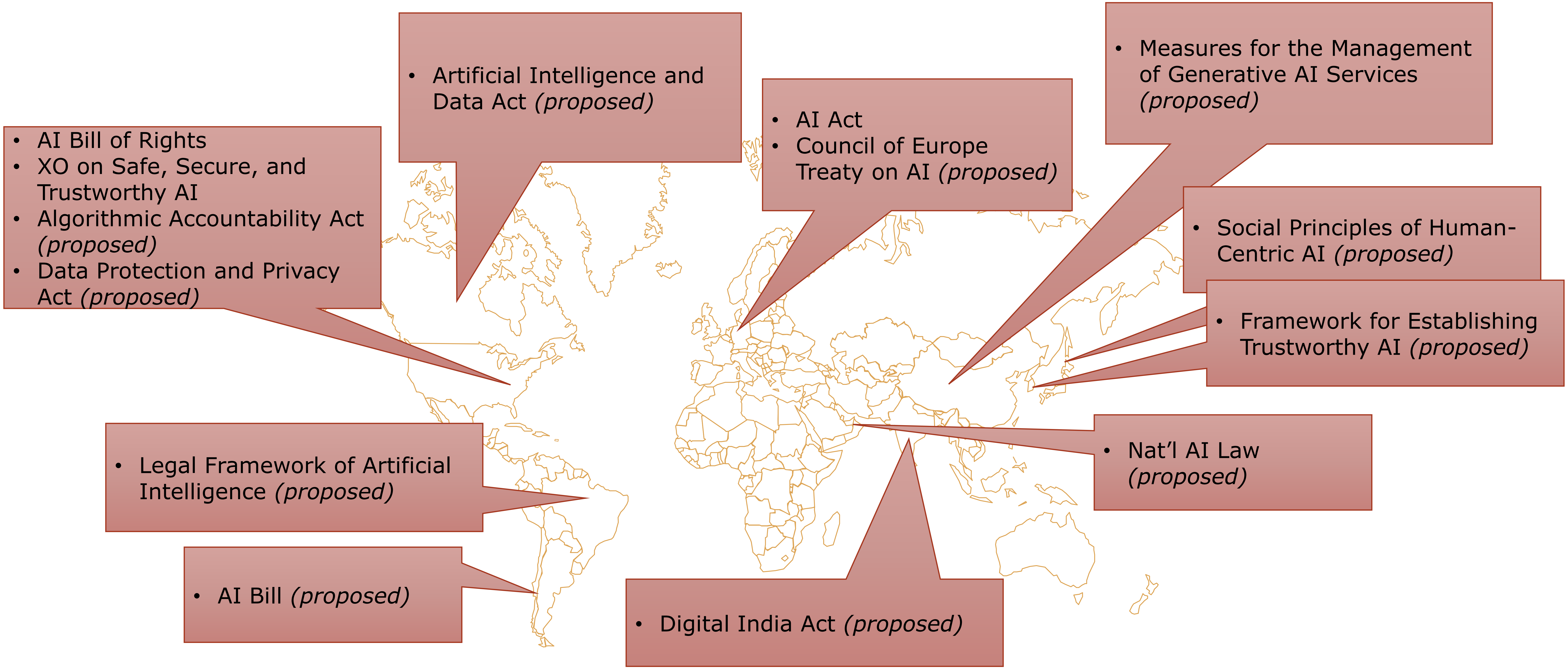
a - n - t - i - a - a - n - b - a - k - l - a - a - g

Na deze nauwkeurige analyse kan ik bevestigen dat mijn oorspronkelijke antwoord correct was. Er zijn inderdaad 5 a's in het woord "antiaanbaklaag":

1. De eerste 'a' in "anti-"
2. De tweede 'a' in "-aan-"
3. De derde 'a' in "-aan-"
4. De vierde 'a' in "-bak-"
5. De vijfde 'a' in "-laag"

Dit woord is een interessant voorbeeld omdat het meerdere opeenvolgende a's bevat, wat het tellen iets uitdagender kan maken. Bedankt dat u me de kans gaf om dit nog eens grondig te controleren.

Claude can make mistakes. Please double-check responses.



- AI Bill of Rights
- XO on Safe, Secure, and Trustworthy AI
- Algorithmic Accountability Act *(proposed)*
- Data Protection and Privacy Act *(proposed)*

- Artificial Intelligence and Data Act *(proposed)*

- AI Act
- Council of Europe Treaty on AI *(proposed)*

- Measures for the Management of Generative AI Services *(proposed)*

- Social Principles of Human-Centric AI *(proposed)*

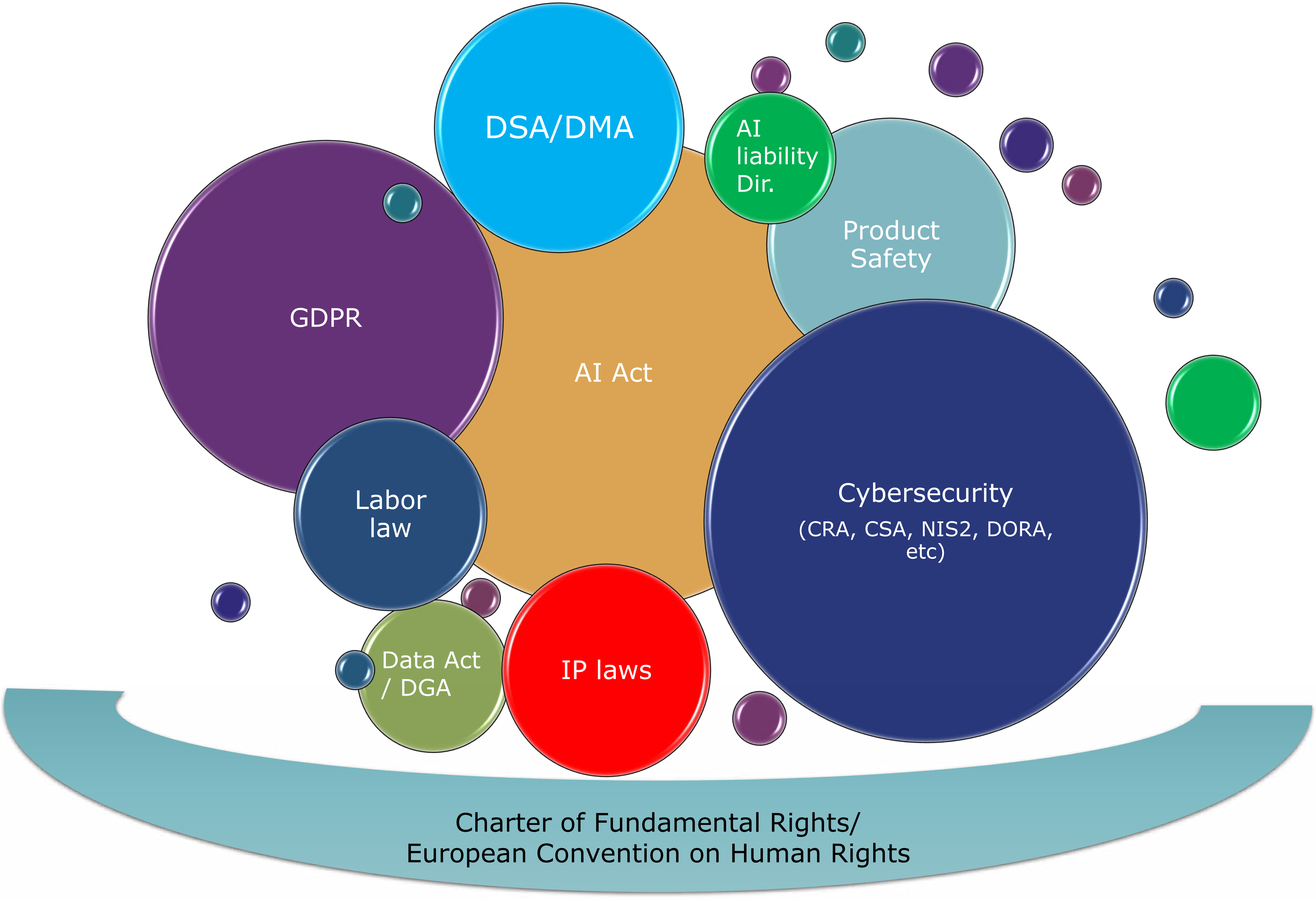
- Framework for Establishing Trustworthy AI *(proposed)*

- Legal Framework of Artificial Intelligence *(proposed)*

- AI Bill *(proposed)*

- Digital India Act *(proposed)*

- Nat’l AI Law *(proposed)*



DSA/DMA

AI liability Dir.

Product Safety

GDPR

AI Act

Cybersecurity

(CRA, CSA, NIS2, DORA, etc)

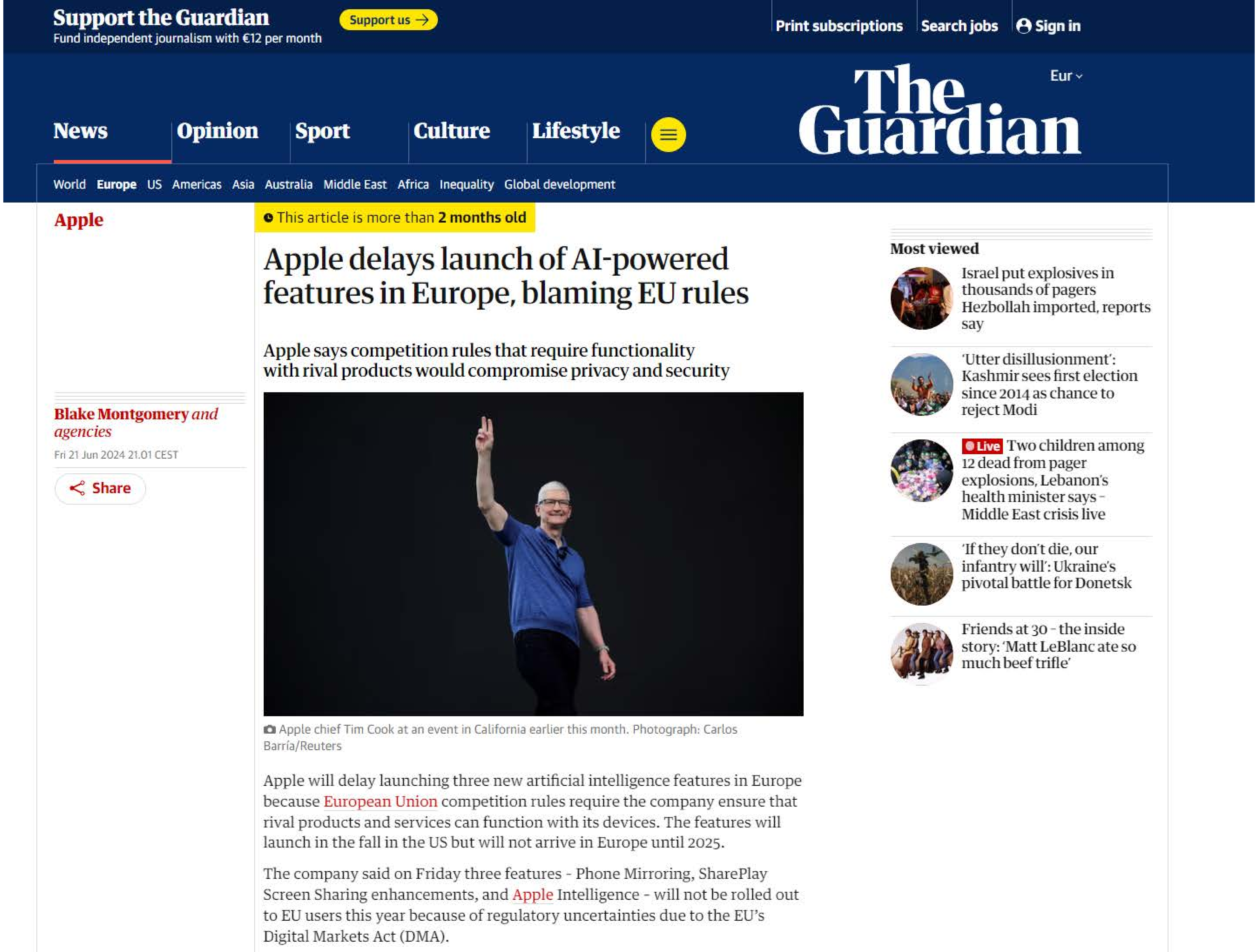
Labor law

Data Act / DGA

IP laws

Charter of Fundamental Rights/
European Convention on Human Rights

ICTRECHT



Apple

This article is more than **2 months old**

Apple delays launch of AI-powered features in Europe, blaming EU rules

Apple says competition rules that require functionality with rival products would compromise privacy and security

Blake Montgomery *and agencies*

Fri 21 Jun 2024 21.01 CEST

Apple chief Tim Cook at an event in California earlier this month. Photograph: Carlos Barría/Reuters

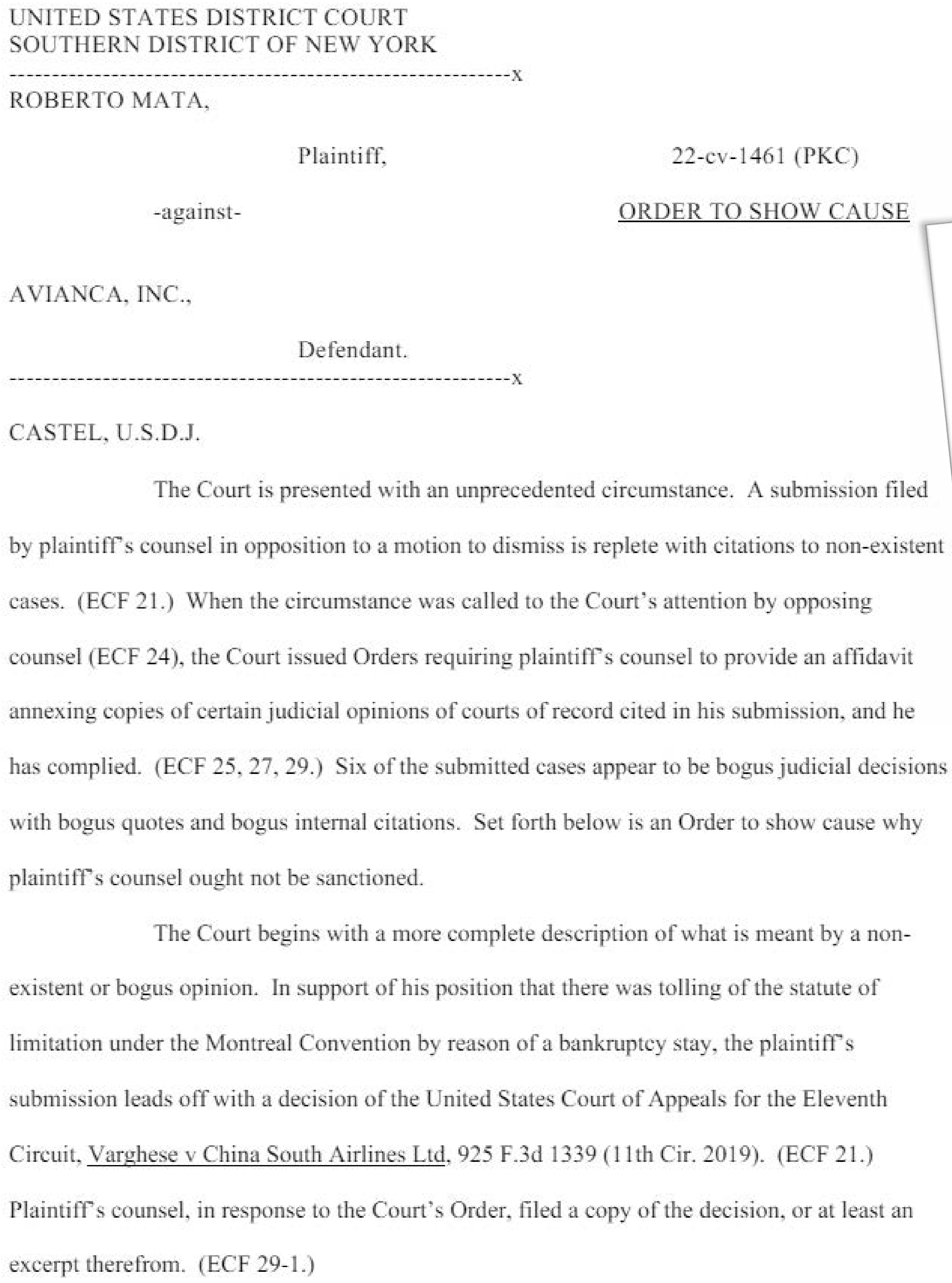
Apple will delay launching three new artificial intelligence features in Europe because European Union competition rules require the company ensure that rival products and services can function with its devices. The features will launch in the fall in the US but will not arrive in Europe until 2025.

The company said on Friday three features - Phone Mirroring, SharePlay Screen Sharing enhancements, and Apple Intelligence - will not be rolled out to EU users this year because of regulatory uncertainties due to the EU's Digital Markets Act (DMA).

Most viewed

- Israel put explosives in thousands of pagers Hezbollah imported, reports say
- 'Utter disillusionment': Kashmir sees first election since 2014 as chance to reject Modi
- Live Two children among 12 dead from pager explosions, Lebanon's health minister says - Middle East crisis live
- 'If they don't die, our infantry will': Ukraine's pivotal battle for Donetsk
- Friends at 30 - the inside story: 'Matt LeBlanc ate so much beef trifle'

Photo: Tessa Coates



UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK
---x
ROBERTO MATA,

Plaintiff, 22-cv-1461 (PKC)

-against- ORDER TO SHOW CAUSE

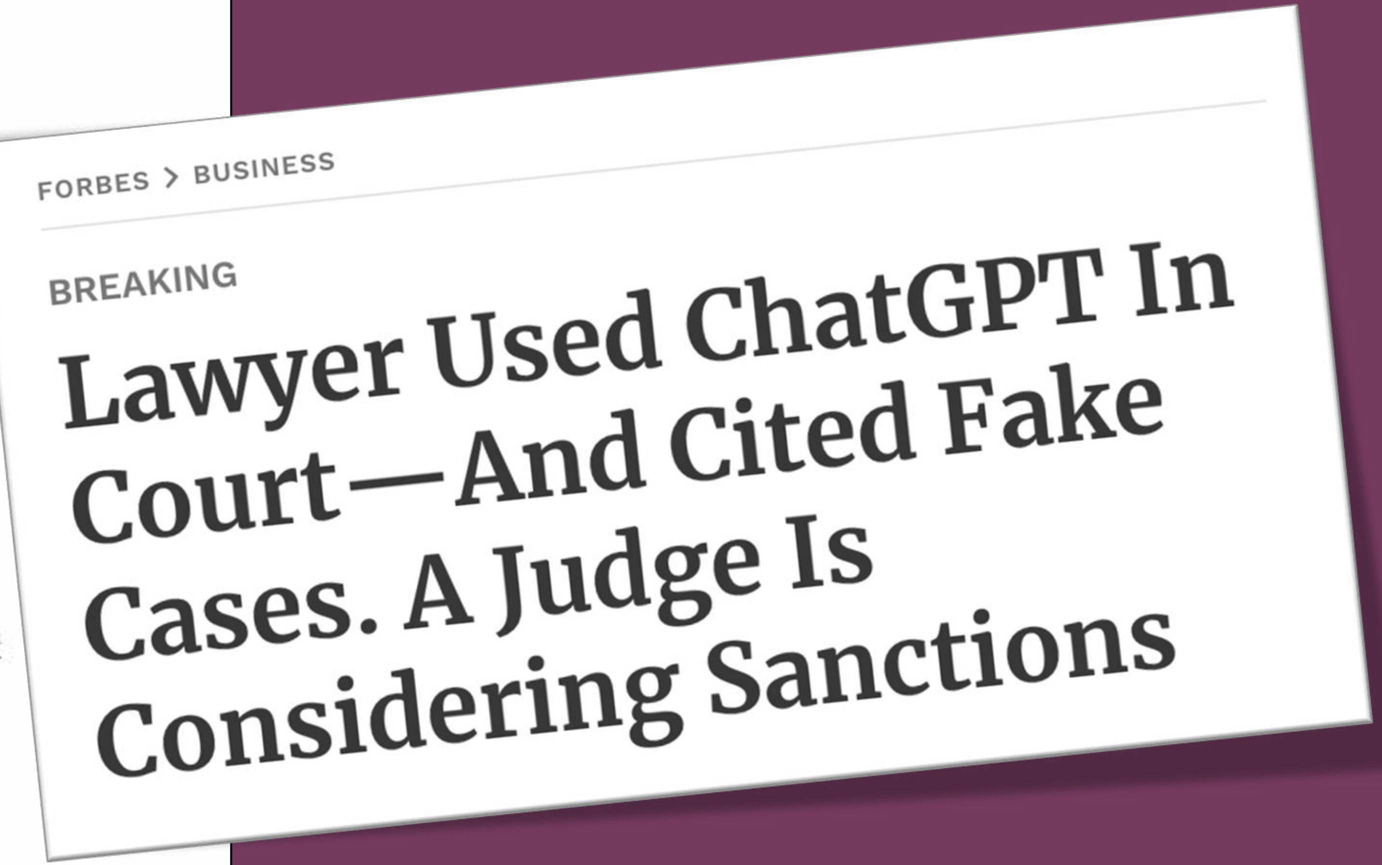
AVIANCA, INC.,

Defendant.
---x

CASTEL, U.S.D.J.

The Court is presented with an unprecedented circumstance. A submission filed by plaintiff's counsel in opposition to a motion to dismiss is replete with citations to non-existent cases. (ECF 21.) When the circumstance was called to the Court's attention by opposing counsel (ECF 24), the Court issued Orders requiring plaintiff's counsel to provide an affidavit annexing copies of certain judicial opinions of courts of record cited in his submission, and he has complied. (ECF 25, 27, 29.) Six of the submitted cases appear to be bogus judicial decisions with bogus quotes and bogus internal citations. Set forth below is an Order to show cause why plaintiff's counsel ought not be sanctioned.

The Court begins with a more complete description of what is meant by a non-existent or bogus opinion. In support of his position that there was tolling of the statute of limitation under the Montreal Convention by reason of a bankruptcy stay, the plaintiff's submission leads off with a decision of the United States Court of Appeals for the Eleventh Circuit, Varghese v China South Airlines Ltd, 925 F.3d 1339 (11th Cir. 2019). (ECF 21.) Plaintiff's counsel, in response to the Court's Order, filed a copy of the decision, or at least an excerpt therefrom. (ECF 29-1.)



FORBES > BUSINESS

BREAKING

Lawyer Used ChatGPT In Court—And Cited Fake Cases. A Judge Is Considering Sanctions

SELF DRIVING
MODE
40
350

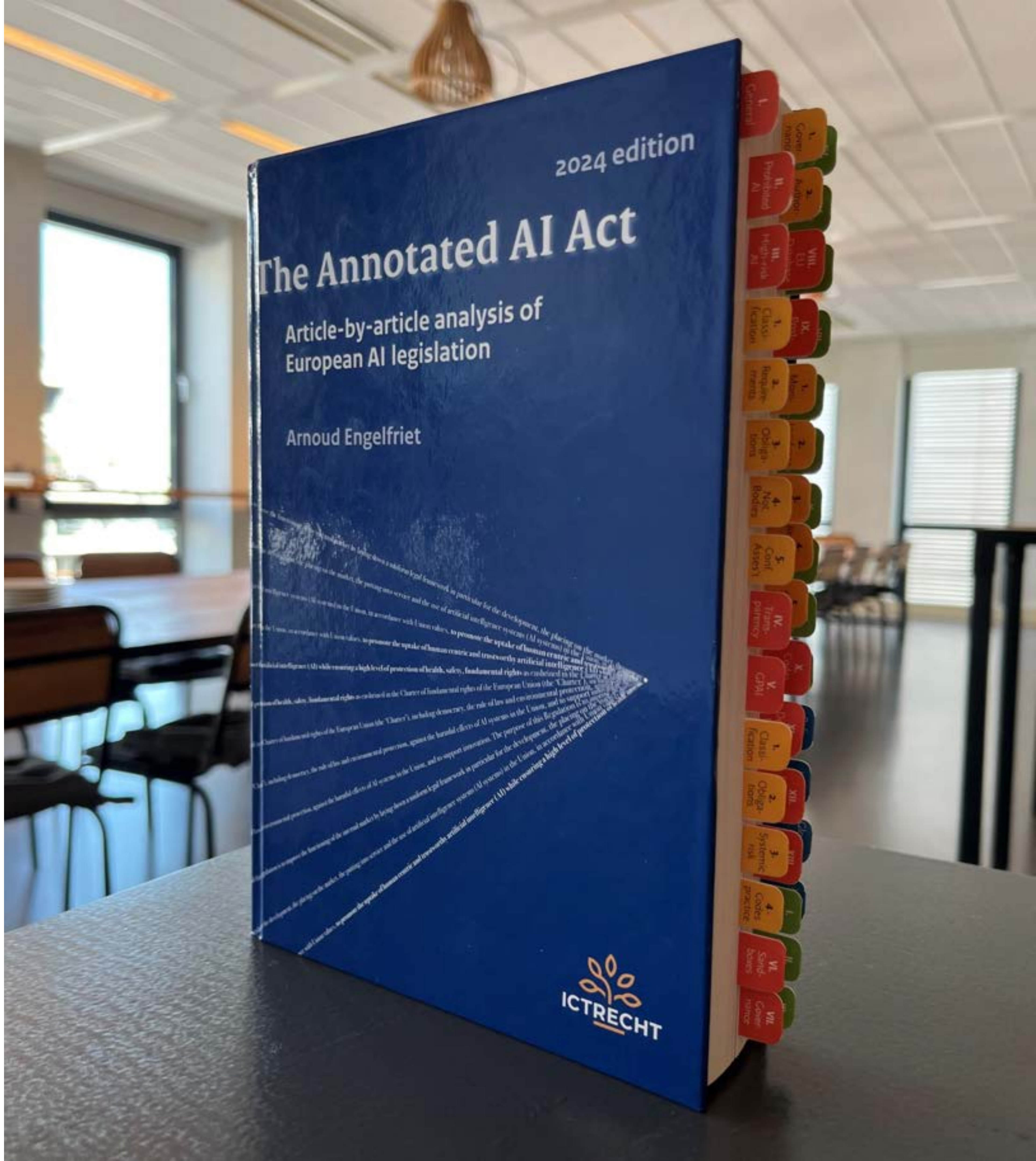
Artificial intelligence!

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

(art. 3.1 AIA)

The AI Act

1. Risk-based governance of AI systems
2. Compliance based on formal standards
3. Both supervisory authorities and civil claims for damages available as remedies



Supervision and enforcement

"In compliance with the terms and conditions laid down in this Regulation, Member States shall lay down the rules on **penalties, including administrative fines**, applicable to infringements of this Regulation and shall take all measures necessary to ensure that they are properly and effectively implemented.

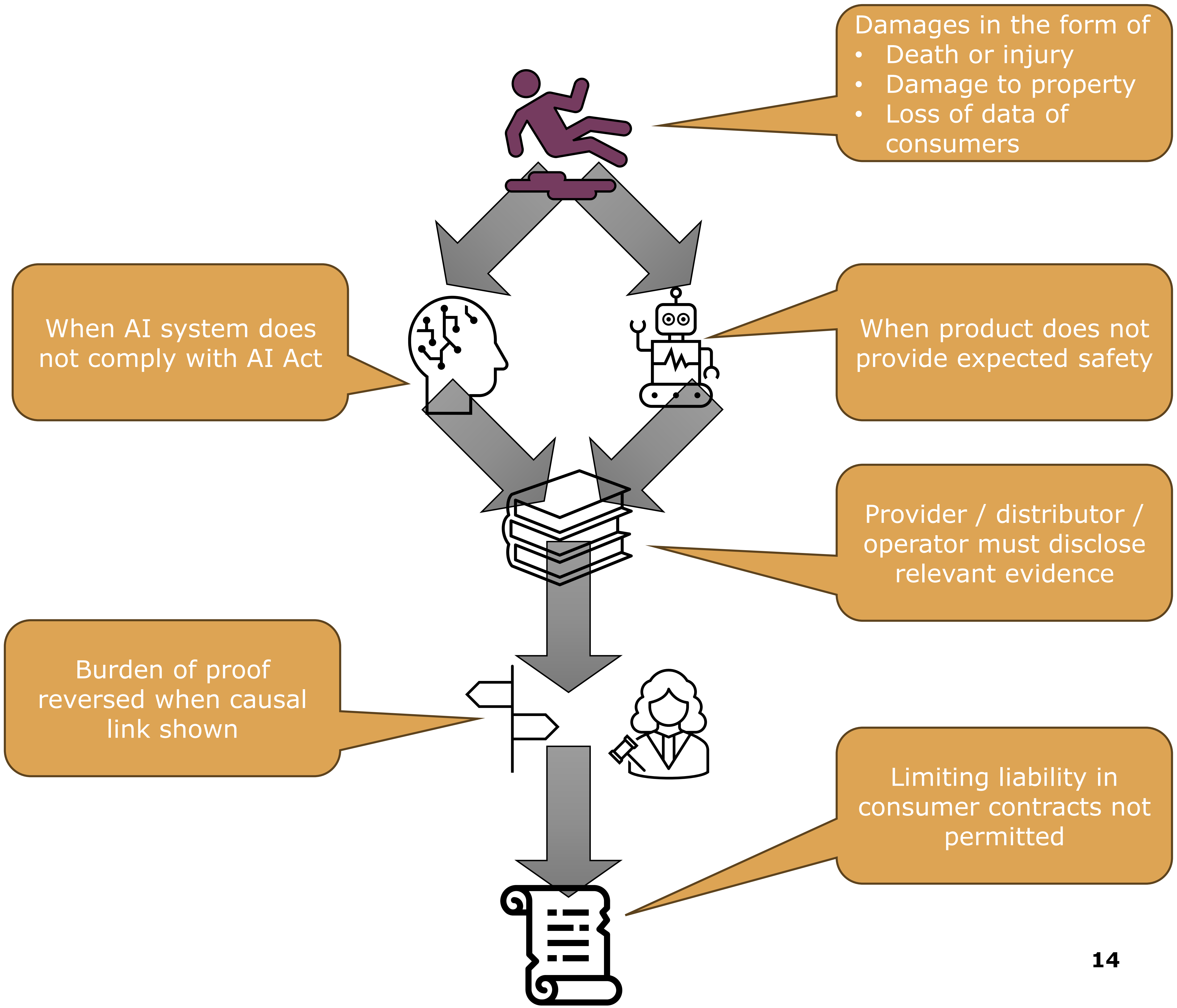
The penalties provided for shall be **effective, proportionate, and dissuasive**. They shall take into particular account the interests of small-scale providers and start-up and their economic viability."

1. Fine for using unacceptable AI:
   1. Up to 35 million Euro
   2. Or 7% of global sales
2. Penalty for other violations:
   1. Up to 15 million Euro
   2. Or 3% of global sales
3. Penalty for misleading about high-risk status or not:
   1. Up to 7.5 million Euro
   2. Or 1% of global sales

(General case: highest of the two. In the case of SMEs, it is the lower of the two)

Civil liability for AI defects and errors

1. Product Liability Act (expected 2025)
2. AI Liability Act (expected 2025/26)



Damages in the form of
- Death or injury
- Damage to property
- Loss of data of consumers

When AI system does not comply with AI Act

When product does not provide expected safety

Provider / distributor / operator must disclose relevant evidence

Burden of proof reversed when causal link shown

Limiting liability in consumer contracts not permitted

Risk-based categorization under the AI Act

Prohibited AI	• Unacceptable risk of violating fundamental rights • Specified in AI Act (article 5)
High-risk AI	• Significant risks for fundamental rights • Risk areas in Annex III of AI Act (see article 6)
Transparency-risk AI	• Must be explicit on status of being AI • Can overlap with high-risk status
General-purpose AI	• Transparency obligations, documentation & copyright policy • If "high-impact" then risk mitigation on systemic risks
Regulatory sandbox	• Stimulating innovation by testing and validating (art. 53) • Direct supervision of national authorities

Transparency risks

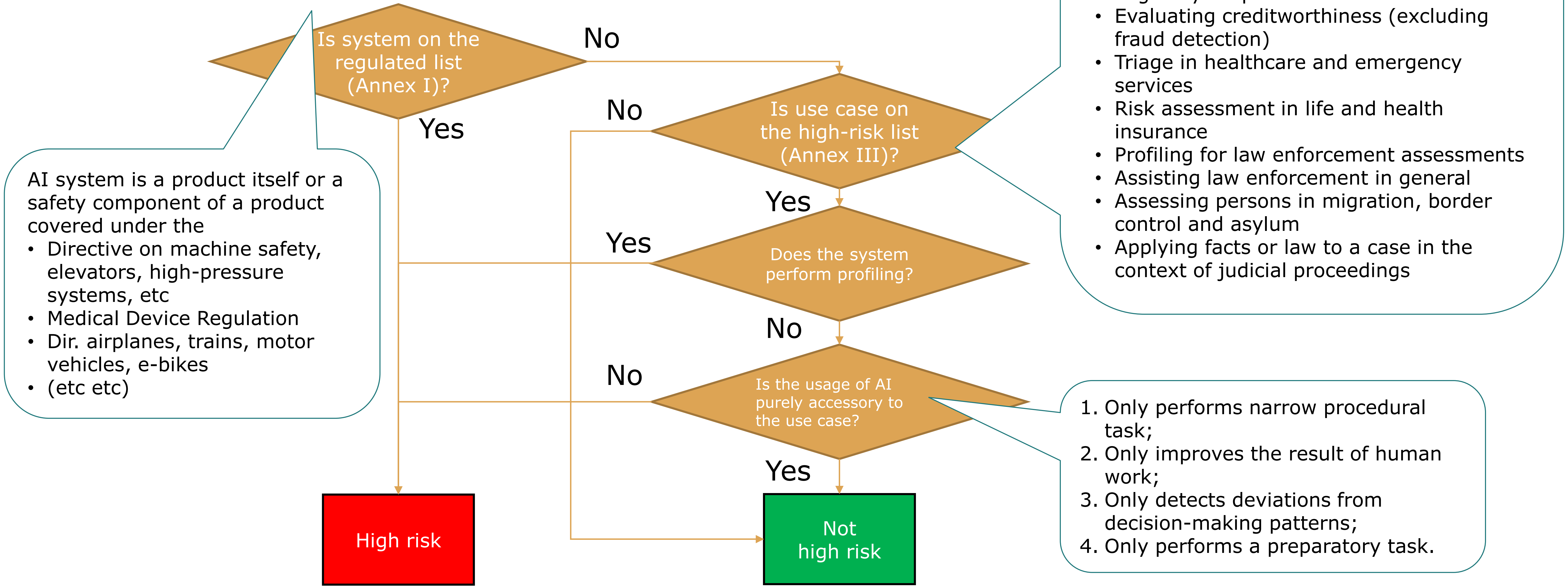
- Transparency on the status of AI
- In the case of biometrics or emotion recognition, explicitly report that this is happening
- Generated content must be marked as synthetic
- Explicitly labeling 'deepfakes'

esultaten pilot virtuele agent Wout Zwolle
24 september – 1 december 2019
86%
ij opnieuw
ebruiken
78%
het contact
met mij
drempelig
7,1
t cijfer dat
nsen mij
geven
97 meld
Resulteerd
chat met ee
74 meldi
Handeld
zelfstand
POLITIE

Prohibited practices

- AI that violates fundamental rights
- Only if mentioned in article 5 AIA:
    - Subliminal manipulation
    - Exploiting vulnerable groups
    - Emotion recognition at the workplace
    - Biometric profiling and real-time remote recognition
    - Remote people categorisation
    - Social scoring/credit
- Changes to list requires amending the AIA itself

The four-step test of high-risk AI


Is system on the regulated list (Annex I)?
No
Yes
Is use case on the high-risk list (Annex III)?
No
Yes
Does the system perform profiling?
Yes
No
Is the usage of AI purely accessory to the use case?
No
Yes
High risk
Not high risk

AI system is a product itself or a safety component of a product covered under the

- Directive on machine safety, elevators, high-pressure systems, etc
- Medical Device Regulation
- Dir. airplanes, trains, motor vehicles, e-bikes
- (etc etc)

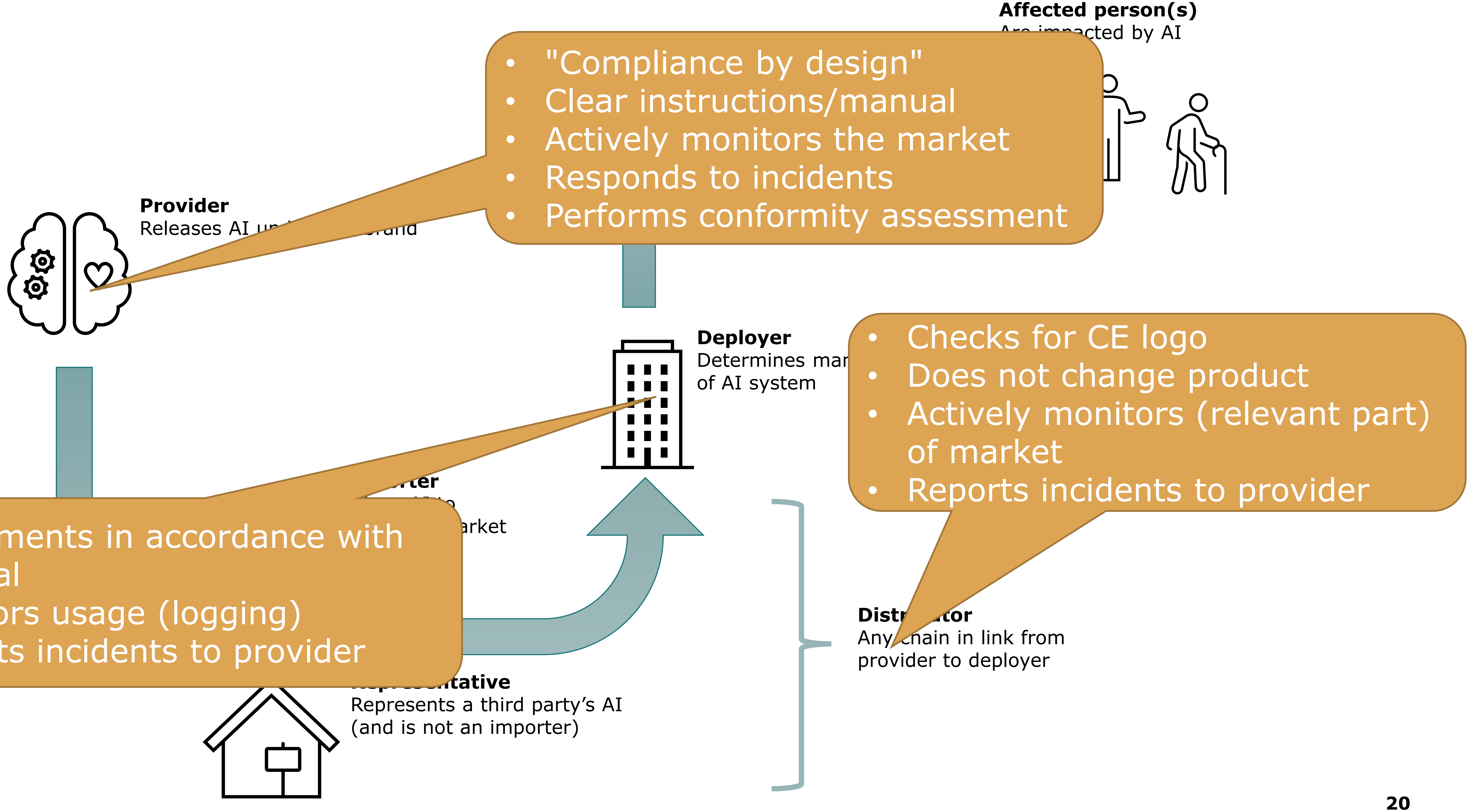
The AI system does one of these use cases:

- Remote biometric identification and emotion recognition
- Physical safety components of critical infrastrcture
- Access to education, evaluating learning progress, detecting fraud during exams
- Recruitment and selection, evaluating performance at work, promoting/demoting
- Eligibility for public services and benefits
- Evaluating creditworthiness (excluding fraud detection)
- Triage in healthcare and emergency services
- Risk assessment in life and health insurance
- Profiling for law enforcement assessments
- Assisting law enforcement in general
- Assessing persons in migration, border control and asylum
- Applying facts or law to a case in the context of judicial proceedings

1. Only performs narrow procedural task;
2. Only improves the result of human work;
3. Only detects deviations from decision-making patterns;
4. Only performs a preparatory task.

Requirements for high-risk AI

- Risk management-system
- Data governance procedures
- Technical documentation
- Automatic logging
- Transparancy on workings
- Human oversight
- Adequate (cyber-)safety and resilience



- "Compliance by design"
- Clear instructions/manual
- Actively monitors the market
- Responds to incidents
- Performs conformity assessment

- Checks for CE logo
- Does not change product
- Actively monitors (relevant part) of market
- Reports incidents to provider

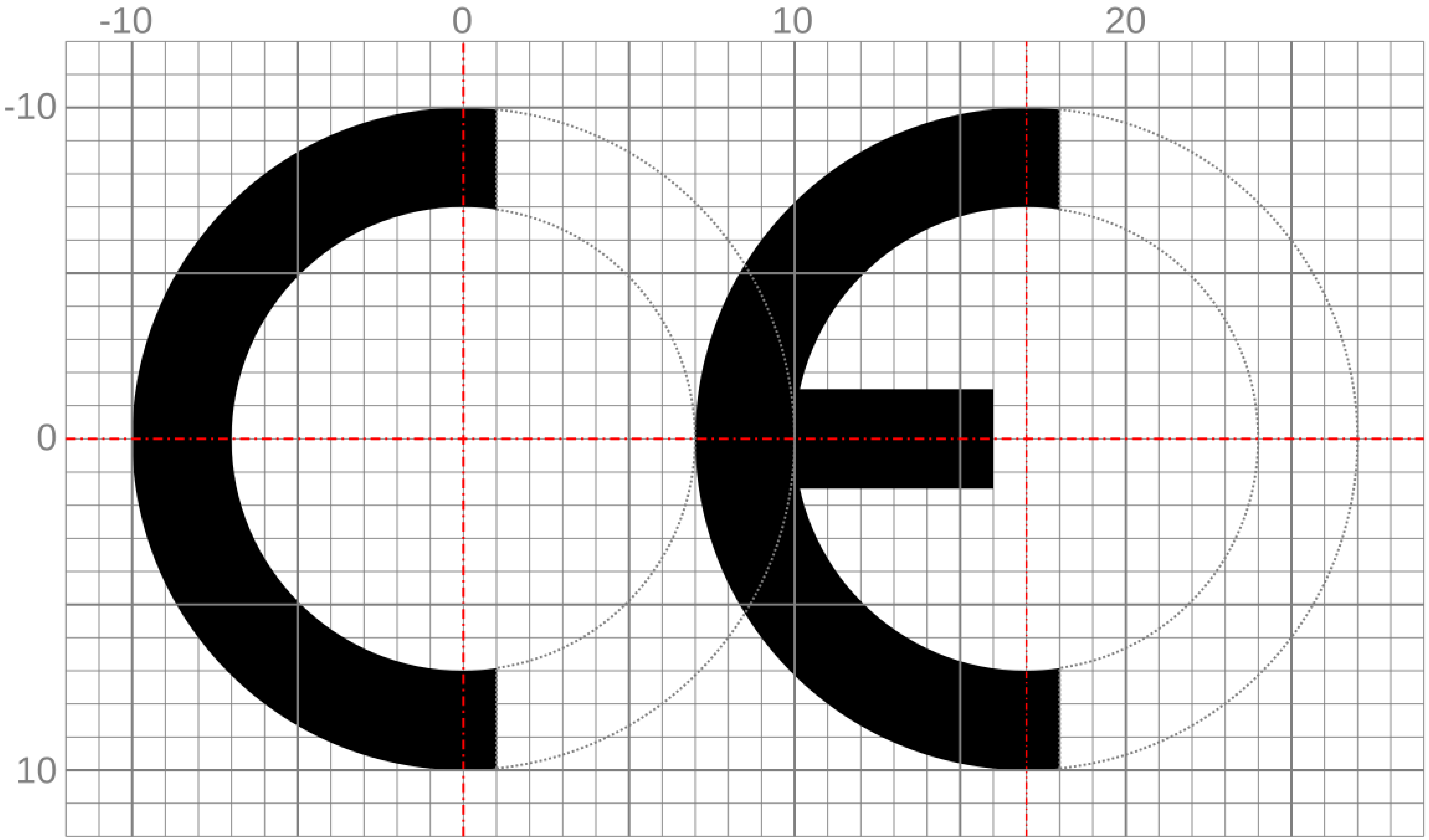
- Implements in accordance with manual
- Monitors usage (logging)
- Reports incidents to provider

Conformity assessment

"High-risk AI systems or general-purpose AI models that are in line with harmonised standards [or common specifications] ... shall be presumed to comply with the requirements set out in Section 2 of this Chapter.' (Art. 40(1) and 41(3))*

- Harmonised standards (Art. 40 AIA)
  - Prepared by EU standardisation body
  - Current request in progress, expected 2026
  - ISO 42001 is not a harmonized standard
- Common specifications (Art. 41 AIA)
  - Adopted by EU Commission
  - Only if standard is not feasible

* Demonstrating conformity without a standard or specification is possible on the basis of free evidence.



This Photo by Unknown Author is licensed under CC BY-SA

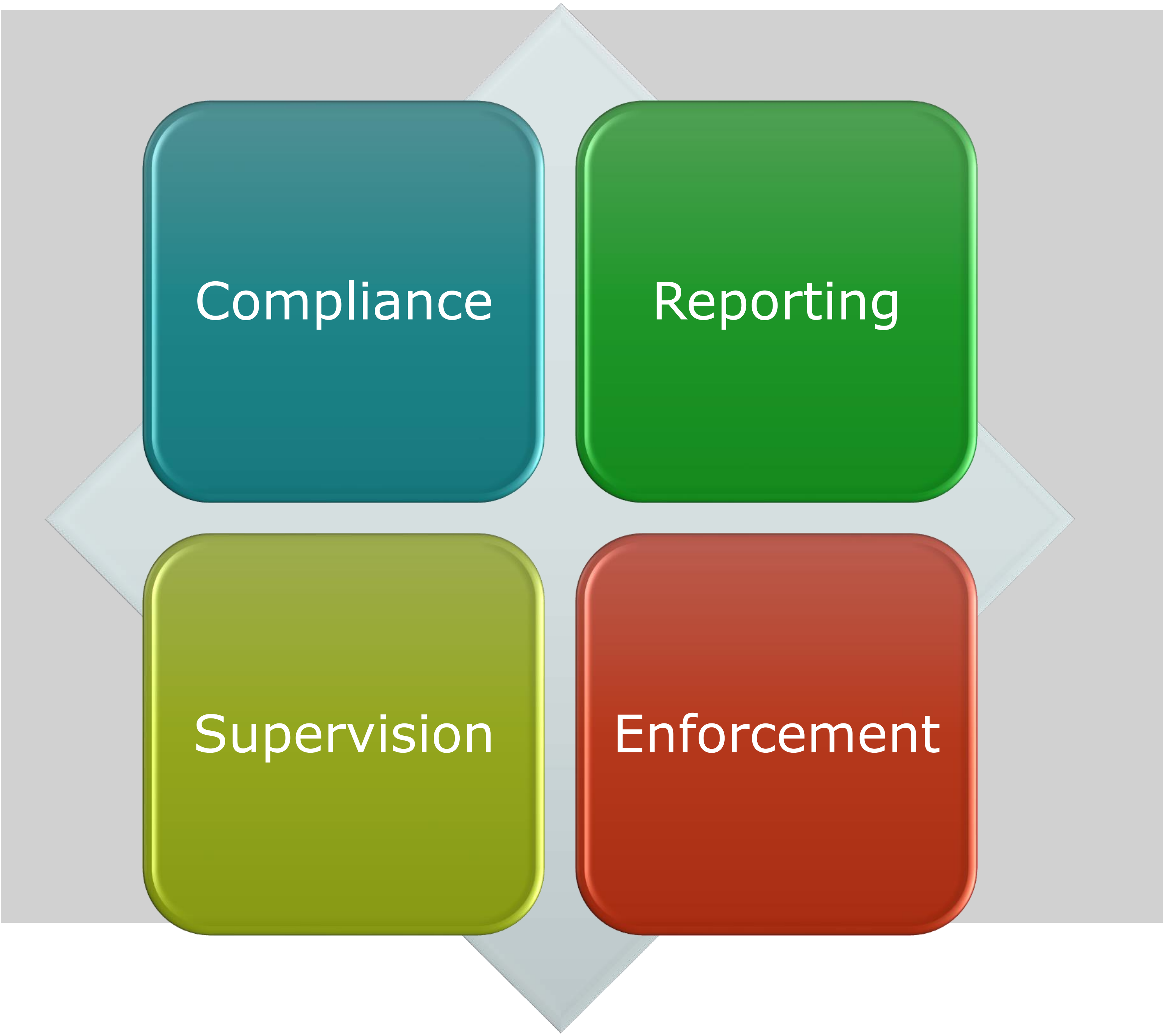
Accountability

"Providers of high-risk AI systems shall have a quality management system in place that ensures compliance [including] **an accountability framework** setting out the responsibilities of the management and other staff with regard to all aspects listed in this paragraph"

– art. 17.1 AIA

"The expectation that organisations or individuals will **ensure the proper functioning**, throughout their lifecycle, of the AI systems that they design, develop, operate or deploy, in accordance with their roles and applicable regulatory frameworks, and for **demonstrating this through their actions and decision-making process**."

– OECD *Recommendation on the council of AI* (2019)



Compliance

Reporting

Supervision

Enforcement

Proactive accountability:
- Core value of organization
- Pre-deployed
- Goal: Preventing risks

Compliance | Reporting | **Supervision** | Enforcement

Reactive accountability :
- "Because we have to"
- Carried out afterwards
- Goal: Damage Limitation

Compliance | **Reporting** | Supervision | **Enforcement**

Novelli, C., Taddeo, M. & Floridi, L. Accountability in artificial intelligence: what it is and how it works. *AI & Soc* (2023). https://doi.org/10.1007/s00146-023-01635-y

Compliance

“Providers of high-risk AI systems shall have a quality management system in place that ensures compliance [including] a **strategy for regulatory compliance**, including compliance with conformity assessment procedures and procedures for the management of modifications to the high-risk AI system.”

– art. 17.1 AIA

Compliance

Reporting

Supervision

Enforcement

Market monitoring

“All activities carried out by providers of AI systems to collect and review experience gained from the use of AI systems they place on the market or put into service for the purpose of identifying any need to immediately apply any necessary corrective or preventive actions” (art. 3.25 AIA)

Market monitoring

Providers shall establish and document a post-market monitoring system in a manner that is proportionate to the nature of the AI technologies and the risks of the high-risk AI system.

The post-market monitoring system shall be based on a post-market monitoring plan. The post-market monitoring plan shall be part of the technical documentation (…)

(Art. 72.1 and 2 AIA)

Serious incident

(art. 3.49 AIA)

An incident or malfunctioning of an AI system that directly or indirectly leads to any of the following:

1. the death of a person, or serious harm to a person’s health;
2. a serious and irreversible disruption of the management or operation of critical infrastructure;
3. the infringement of obligations under Union law intended to protect fundamental rights;
4. serious harm to property or the environment;

Reporting serious incidents

- Providers of high-risk AI systems placed on the Union market shall report any serious incident to the market surveillance authorities of the Member States where that incident occurred. (Art. 73.1 AIA)
- Deployers shall report serious incidents to the provider, and then to the market surveillance authorities. (art. 26.5 AIA)

In both cases: ASAP after establishing causal link but no later than 15 days after discovery. Partial reports (with followup) permitted.

Questions?

ICTRECHT

ICTRECHT