AI-forordningen
Centrale elementer og uafklarede spørgsmål

F&P: Forsikringsforeningen

Den 23. september 2025

Jakob Holm

Digitaliseringspolitik

&

☑ Fulgt AI-forordningen siden starten

Ikke jurist

F&P

sk-based approach

egulatory Framework defines 4 levels of risk for AI systems:

UNACCEPTABLE RISK

HIGH RISK

LIMITED RISK
(AI systems with specific transparency obligations)

MINIMAL RISK

AI-forordningen er:

- **Produktregulering**
  - Krav til "produktet"
  - CE-mærkning (på højrisiko)
- Dækker alle AI-produkter, men stiller ikke krav til alle AI-produkter (risikobaseret)

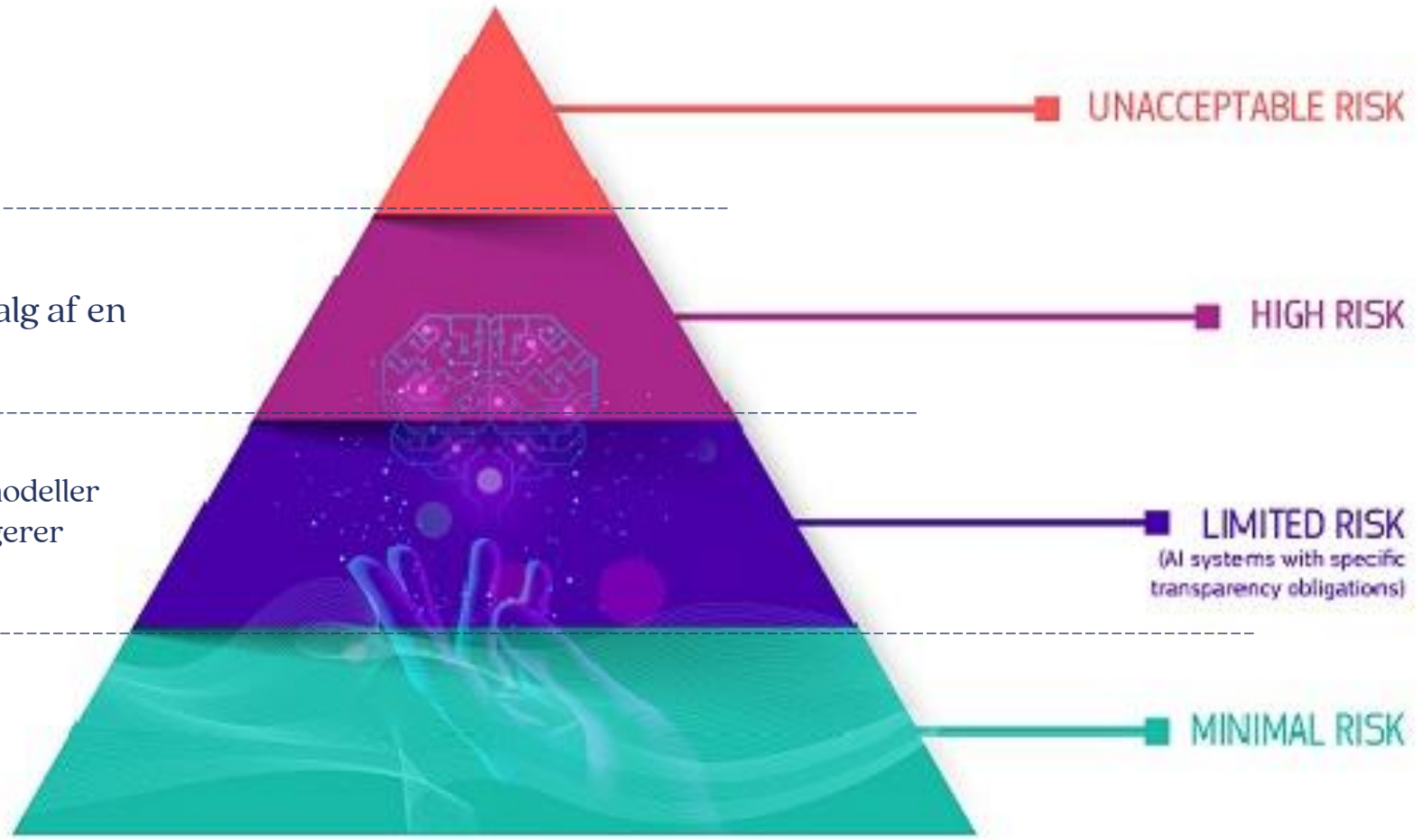
Hvad er AI-act?

- **Uacceptabel risiko**
  Brugen bliver forbudt

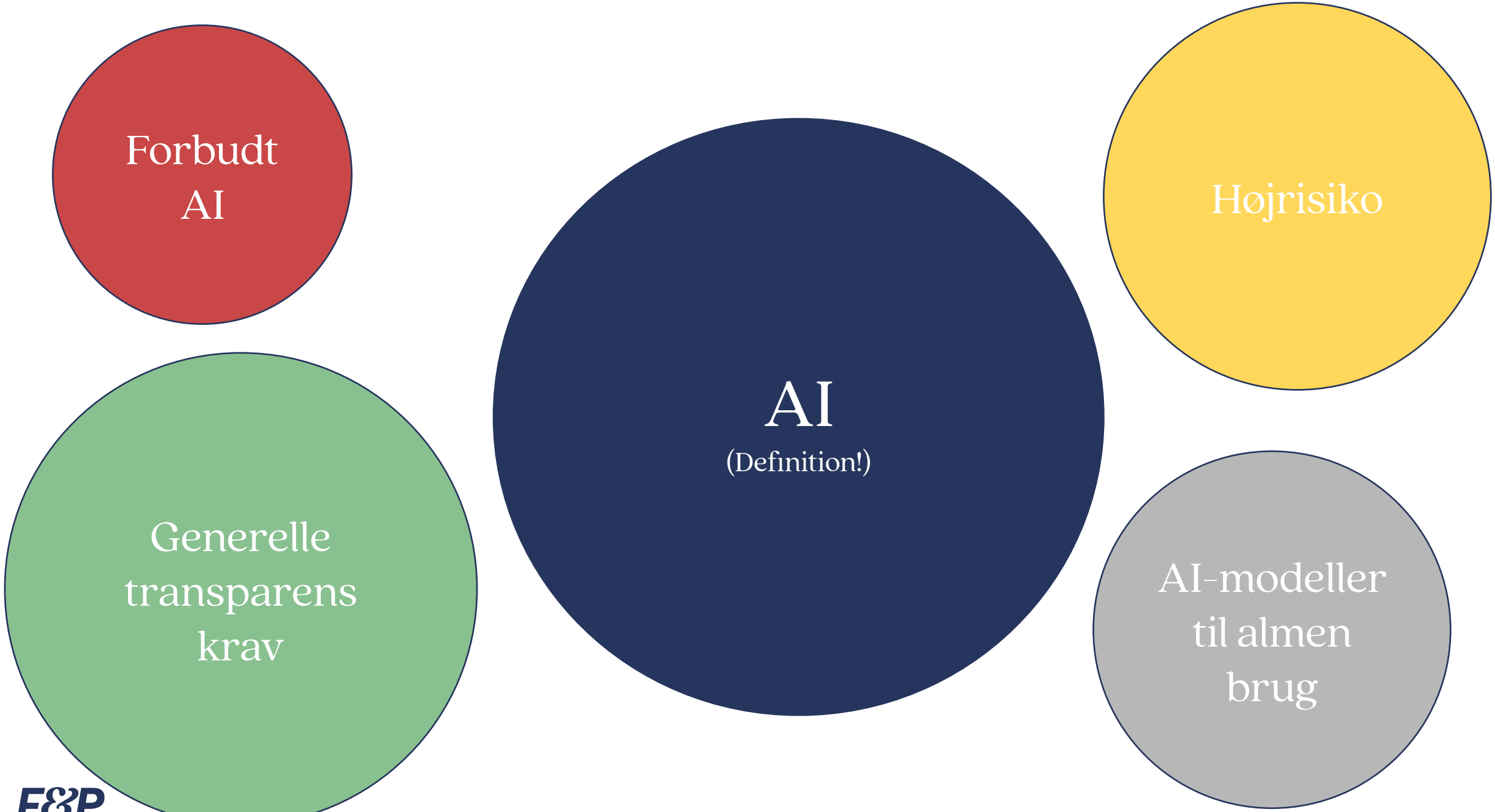
- **Høj risiko**
  Strenge krav forud for idriftsættelse/salg af en løsning

Begrænset risiko
1) Transparenskrav til general purpose AI-modeller
2) Transparenskrav til modeller, som interagerer direkte med fysiske personer

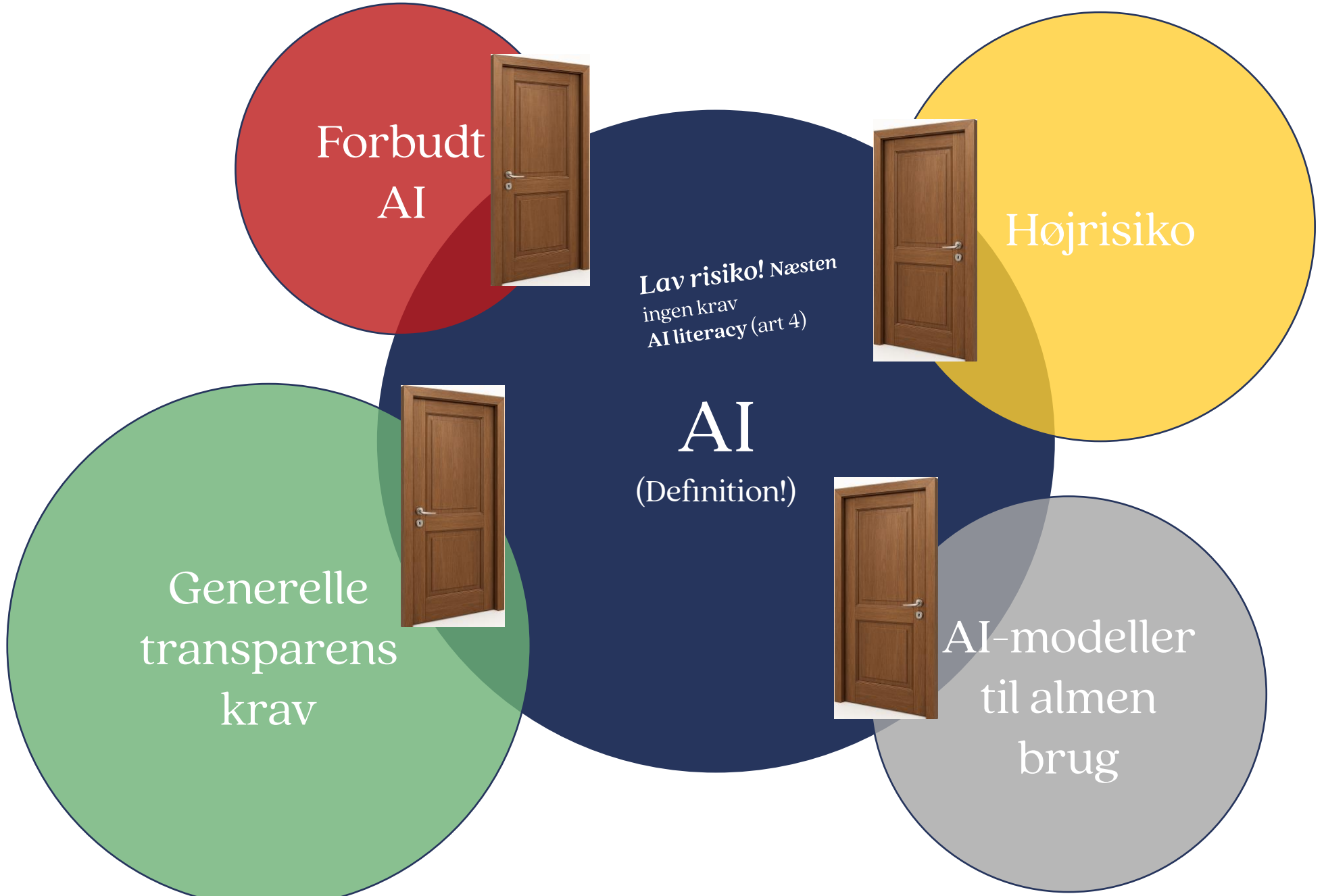
- **Lav risiko**
  Øvrig AI-brug - ingen krav



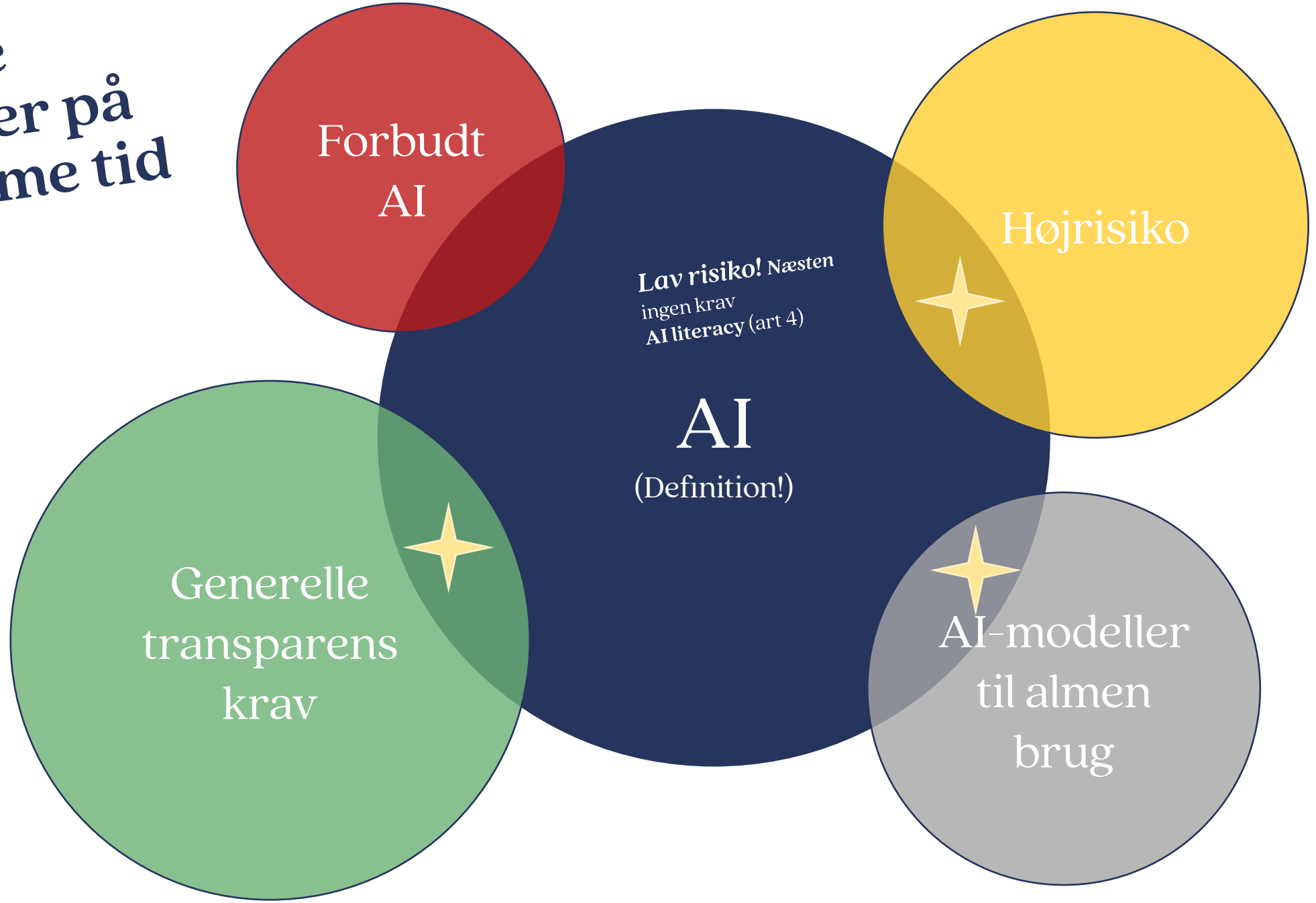
F&P


Forbudt AI
AI
(Definition!)
Højrisiko
Generelle transparens krav
AI-modeller til almen brug

F&P


Forbudt AI
Lav risiko! Næsten ingen krav
AI literacy (art 4)
AI
(Definition!)
Højrisiko
Generelle transparens krav
AI-modeller til almen brug

Flere steder på samme tid



F&P

Definitionen: Hvad er

Definitionen af AI fra AI-forordningen:
(F&P's fremhævelser):

The following seven elements can be extracted from the definition:

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

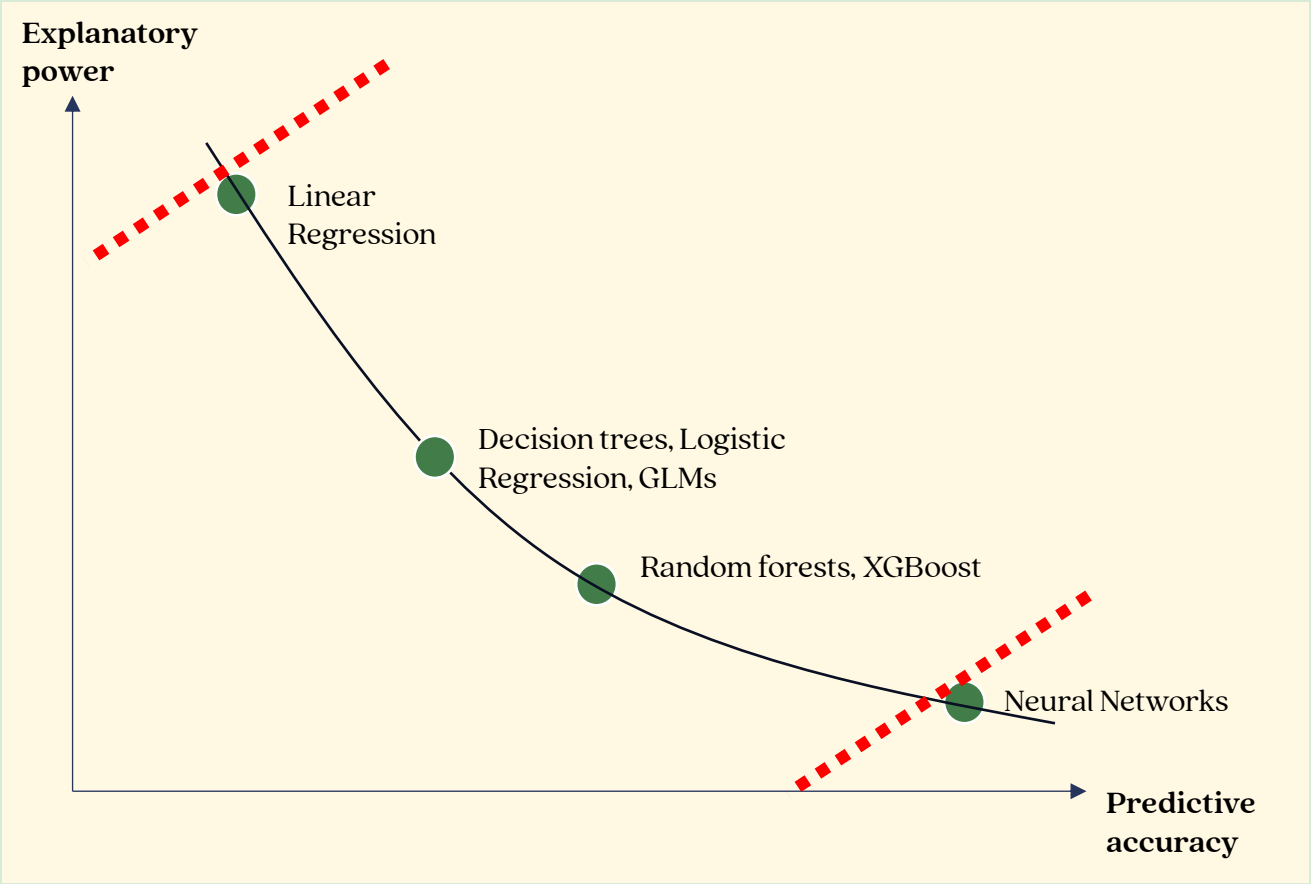
F&P

Definitionen: Hvad er AI?

Definitionen af AI fra AI-forordningen:
(F&P's fremhævelser):

The following seven elements can be extracted from the definition:

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



F&P

Kan vi komme det lidt nærmere?

F&P

AI
Indenfor forordningen

(59) AI systems can generally generate more nuanced outputs than other systems, for example, by **leveraging patterns learned during training** or by using expert-defined rules to make decisions, offering more sophisticated reasoning in structured environments.

Antageligt fx:
LLM
Neurale netværk

XG boost, gradient descent o.l.

Gråzone

Måske fx:
GLM'er generelt (udover logistisk regression).

Lineær regression og logistisk regression (når det ikke er optimering)

ACTUARIAL ASSOCIATION OF EUROPE, MARTS 2025
Recent guidance from the European Commission on the definition of AI systems has provided further clarity on which models fall within the scope of the AI Act.
The guidelines **indicate** that traditional statistical models, such as linear and logistic regression, may not meet the Act's definition of an AI system, particularly if they **lack autonomy, adaptivity, or self-learning capabilities**.
This **suggests** that many actuarial models, including traditional **Generalised Linear Models (GLMs), may not be classified as AI systems** under the Act.

However, actuaries should assess **whether any** models they use incorporate adaptive algorithms or automated learning processes, as these features could bring them within the scope of AI regulation.

IKKE AI
Udenfor forordningen

Systems for improving mathematical optimization

(42) Systems used to improve mathematical optimisation or to accelerate and approximate traditional, **well established optimisation methods, such as linear or logistic regression methods, fall outside the scope of the AI system definition**.

Simple prediction systems

(49) machine-based systems whose performance can be achieved via a **basic statistical learning rule**, while technically may be classified as relying on machine learning approaches fall outside the scope of the AI system definition, due to its performance.
(50) Example - machine-based systems may be used to predict future stock prices by using an estimator with the 'mean' strategy to establish a baseline prediction

F&P

Nedslag fra kommissionens guidelines

Complexity in inference

AI-system

(57) Similarly to predictions, both AI-based and non-AI-based systems can be designed to generate recommendations. AIbased recommendation systems, for example, can leverage large-scale data, adapt to user behaviour in real-time, provide highly personalised recommendations, and scale efficiently as the dataset grows, the

(59) AI systems can generally generate more nuanced outputs than other systems, for example, by **leveraging patterns learned during training** or by using expert-defined rules to make decisions, offering more sophisticated reasoning in structured environments.

System outside scope

functionalities that **non-AI systems that rely on static, rule-based mechanisms and limited data, rarely possess.**

(55) **Non-AI systems, typically based on historical data, scientific data or predefined rules**, such as certain non-AI medical device expert systems, are not capable of dealing with such a degree of complexity.

Systems for improving mathematical optimization

(42) Systems used to improve mathematical optimisation or to accelerate and approximate traditional, **well established optimisation methods, such as linear or logistic regression methods, fall outside the scope of the AI system definition**.

Systems based on classical heuristics

(48) Classical heuristics typically involve rule-based approaches, pattern recognition, or trial-and-error strategies rather than data-driven learning. Unlike modern machine learning systems, which adjust their models based on input-output relationships, classical heuristic systems apply predefined rules or algorithms to derive solutions
Example - a chess program using a minimax algorithm with heuristic evaluation functions can assess board positions without requiring prior learning from data.

Simple prediction systems

(49) machine-based systems whose performance can be achieved via a basic statistical learning rule, while technically may be classified as relying on machine learning approaches fall outside the scope of the AI system definition, due to its performance.
(50) Example - machine-based systems may be used to predict future stock prices by using an estimator with the 'mean' strategy to establish a baseline prediction

Basic data processing

(46) systems are developed and deployed to execute tasks based on manual inputs or rules, without any 'learning, reasoning or modelling' at any stage of the system lifecycle.

(47) systems that solely intended for descriptive analysis, hypothesis testing, and visualization. [..] Example system that applies statistical techniques to opinion polls or survey data to determine their validity, reliability, correlation, and statistical significance.

(40) "simpler traditional software systems or programming approaches and should not cover systems that are based on the rules defined solely by natural persons to automatically execute operations."

(41) Some systems have the capacity to infer in a narrow manner but may nevertheless fall outside of the scope of the AI system definition because of their limited capacity to analyse patterns and adjust autonomously their output.

F&P

Uacceptabel risiko
Forbudt AI
(art 5)

Trådte i kraft i 2. februar 2025

Tilsyn og sanktioner fra august 2025

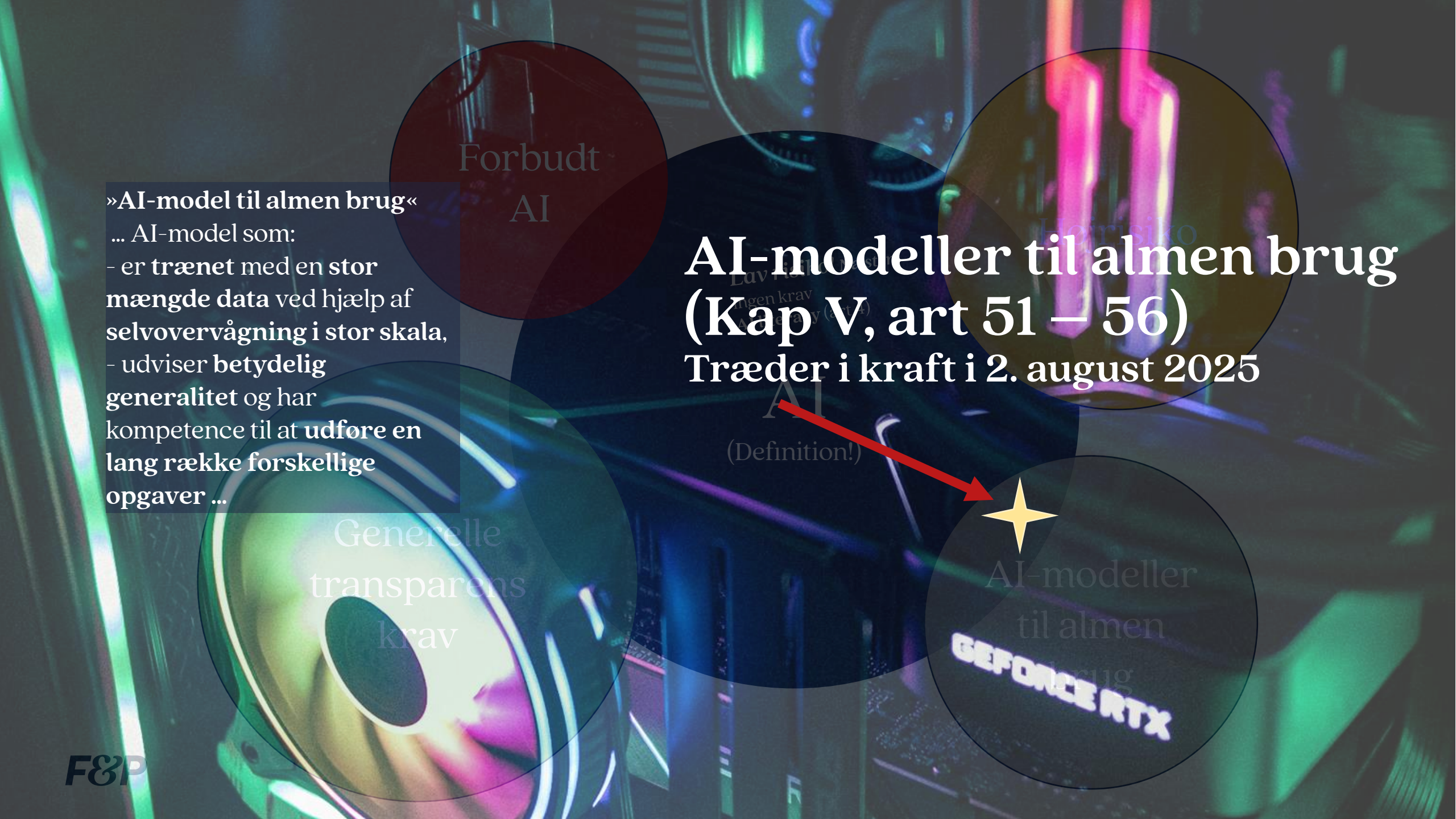
Uacceptabel risiko: Forbudt brug

AI-systemer med uacceptabel risiko anses for at udgøre en trussel mod mennesker og er nu forbudt.
I alt **8 forbud.** Måske mest relevant for branchen:

- At få en person til at gøre noget, denne ellers ikke ville have gjort, som forvolder skade gennem **Subliminale, bevidst manipulerende eller vildledende teknikker** eller ved at **udnytte sårbarheder** hos personer eller specifikke grupper. (Art 5, 1 (a) og (b))
- **Udlede en fysisk persons følelser** (emotions) på arbejdspladser og på uddannelsesinstitutioner (art 5, 1 (f)
- **Social scoring:** Klassificering af mennesker baseret på adfærd, socioøkonomisk status eller personlige karakteristika (art 5, 1 (c))
- **Biometrisk identifikation** og kategorisering af personer på baggrund af særligt følsomme karakteristika. (art 5, 1 (g))
- **Prædiktivt politiarbejde,** baseret alene på profilering af en fysisk person eller en vurdering af deres personlighedstræk og personlige egenskaber (art 5, 1 (d))

F&P


Rammer store modeller
“training compute is greater than 10^23 FLOP and it can generate language (whether in the form of text or audio), text-to-image or text-to-video.
The “compute typically used to train a model with one billion parameters on a large amount of data.”
Obs på ansvarsovergang
“Modification is greater than a third of the training compute of the original model.”
AI-modeller til almen brug (Kap V, art 51 – 56)
kraft i 2. august 2025
F&P

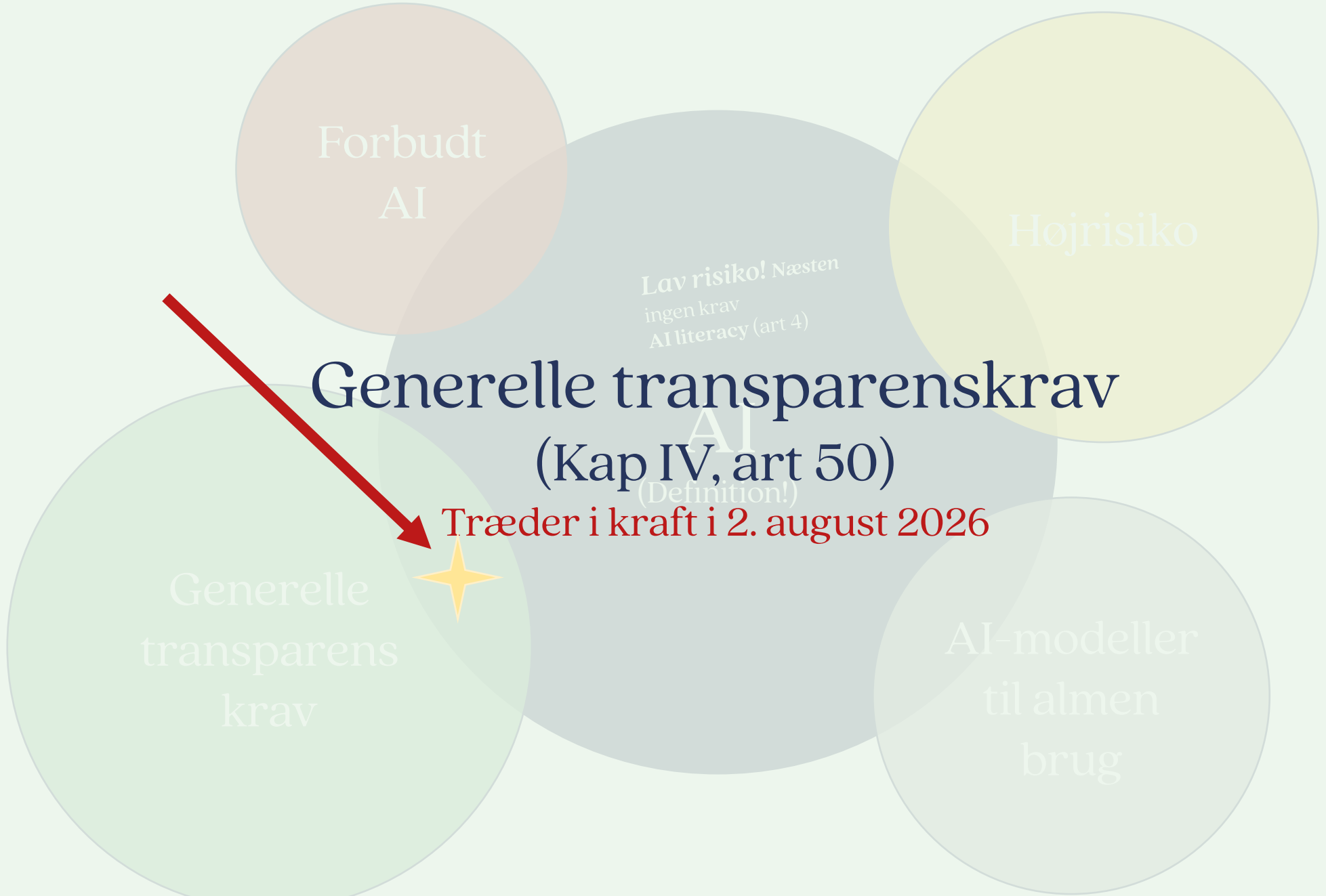


AI-modeller til almen brug (Kap V, art 51 – 56)

Træder i kraft i 2. august 2025

»AI-model til almen brug«
... AI-model som:
- er **trænet** med en **stor mængde data** ved hjælp af **selvovervågning i stor skala**,
- udviser **betydelig generalitet** og har kompetence til at **udføre en lang række forskellige opgaver** ...

F&P



Generelle transparenskrav (Kap IV, art 50)

Træder i kraft i 2. august 2026

F&P

Generelle transparenskrav

De generelle transparenskrav gælder for følgende scenarier:

- **Udbydere** (udviklere) skal sikre, at AI-systemer, **som interagerer direkte med fysiske personer**, udvikles og designes så en fysisk person informeres om, at de interagerer med et AI-system.
- **Udbydere** (udviklere) af systemer, som genererer **lyd, billeder, video eller tekst** skal sikre, at output fra systemet er mærket i et machine-readable format og kan spores som kunstigt genereret eller manipuleret.

- **Idriftsættere** (brugere) af AI-systemer, som genererer eller manipulerer **billeder, lyd eller video**, som udgør en ***deep fake*** skal oplyse at indholdet er genereret kunstigt eller manipuleret.
- **Idriftsættere** (brugere) af AI-systemer, som genererer eller manipulerer **tekst**, som offentliggøres for at informere offentligheden om 'forhold af offentlig interesse', skal oplyse, at teksten er kunstigt genereret eller manipuleret. Medmindre teksten har gennemgået en proces med menneskelig gennemgang (human review).
- **Idriftsættere** (brugere) af AI-systemer til følelsesgenkendelse eller et system til biometrisk kategorisering underretter de fysiske personer, der udsættes herfor, om driften af systemet

F&P

”Deepfake”:

Recital (134): […] billed-, lyd- eller videoindhold, der **i væsentlig grad ligner faktiske** personer, genstande, steder, enheder eller begivenheder, og som **fejlagtigt vil fremstå ægte** eller sandfærdigt for en person (»**deepfakes**«) […]

kte med fysiske personer, udvikles og designes så en

- **Idriftsættere** (brugere) af AI-systemer til følelsesgenker
  fysiske personer, der udsættes herfor, om driften af syste

F&P

Højrisikokravene

(Kap III, art 6 – 27)

Træder i kraft i 2. august 2026

F&P

Afgrænsede scenarier!

F&P

Forsikring

Bilag III

Punkt 5. Adgang til og benyttelse af væsentlige private tjenester og væsentlige offentlige tjenester og ydelser:

c) AI-systemer, der tilsigtes anvendt til at foretage **risikovurdering og prisfastsættelse** for så vidt angår **fysiske personer** i forbindelse med **livs- og sygeforsikring**.

F&P

.... indtil videre

Måske:

- Forskel på essentielle og ikke-essentielle ydelser
- Fokus på hvem der indgår aftale (fysisk person) / om der risikovurderes?
- Risk assessment og pricing fokuseret på antagelsen

Bilag III
Punkt 5. Adgang til og benyttelse af **væsentlige** private tjenester og **væsentlige** offentlige tjenester og ydelser:

c) AI-systemer, der tilsigtes anvendt til at foretage **risikovurdering og prisfastsættelse** for så vidt angår **fysiske personer** i forbindelse med **livs- og sygeforsikring**.

F&P

Relevant på tværs?

”HR”-cases

Beskæftigelse, forvaltning af arbejdstagere og adgang til selvstændig virksomhed:

a) AI-systemer, der tilsigtes anvendt til **rekruttering** eller udvælgelse af fysiske personer, navnlig til at indrykke målrettede jobannoncer, analysere og filtrere jobansøgninger og evaluere kandidater

b) AI-systemer, der tilsigtes anvendt til at træffe beslutninger, der **påvirker vilkårene for arbejdsrelaterede** forhold, forfremmelse eller afskedigelse i arbejdsrelaterede kontraktforhold, til at **fordele opgaver på grundlag af individuel adfærd eller personlighedstræk eller personlige egenskaber** eller til at **overvåge og evaluere personers præstationer og adfærd** i sådanne forhold.

F&P

Relevant på tværs?

Emotion recognition

"system til følelsesgenkendelse": et AI-system, der har til formål at genkende eller udlede fysiske personers **følelser eller hensigter** på grundlag af deres **biometriske data**

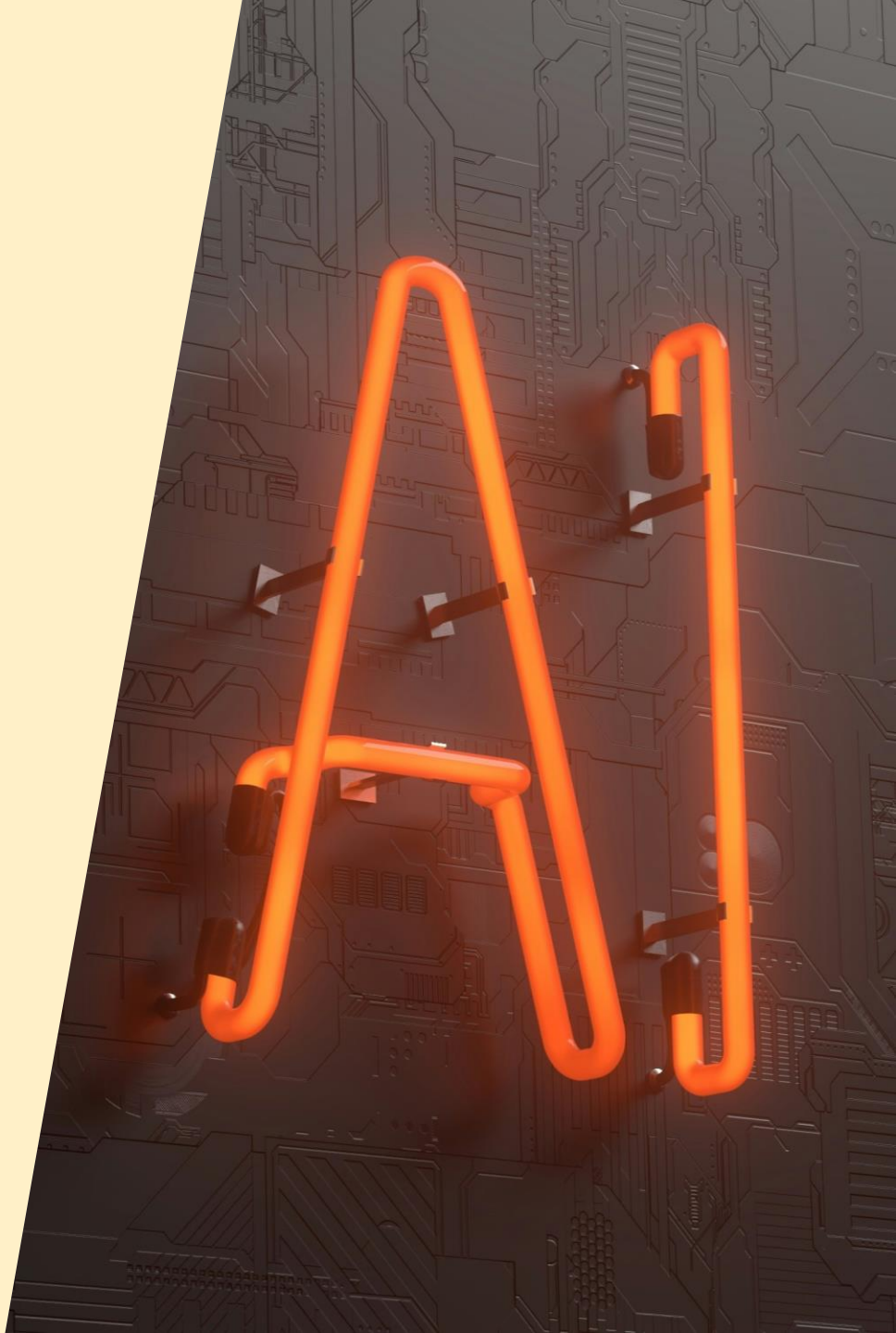
1. Forbud mod emotion recognition in the areas of workplace and education
2. Højrisikokrav for (alle) øvrige emotion recognition systemer
3. Transparenskrav overfor de personer som udsættes for ”emotion recognition”

F&P

Hvad regulerer højrisikokravene ikke

- **Brugen**
  - Så længe den overholder de (tekniske) krav & brugsanvisningen

- Husk de forbudte scenarier



F&P

2 vigtige definitioner

Provider (udvikler / udbyder)

"udbyder":

En fysisk eller juridisk person, en offentlig myndighed, et agentur eller et andet organ,

der udvikler eller får udviklet et AI-system eller en AI-model til almen brug

og bringer dem i omsætning eller ibrugtager AI-systemet under eget navn eller varemærke,

enten mod betaling eller gratis

Deployer (bruger / idriftsætter)

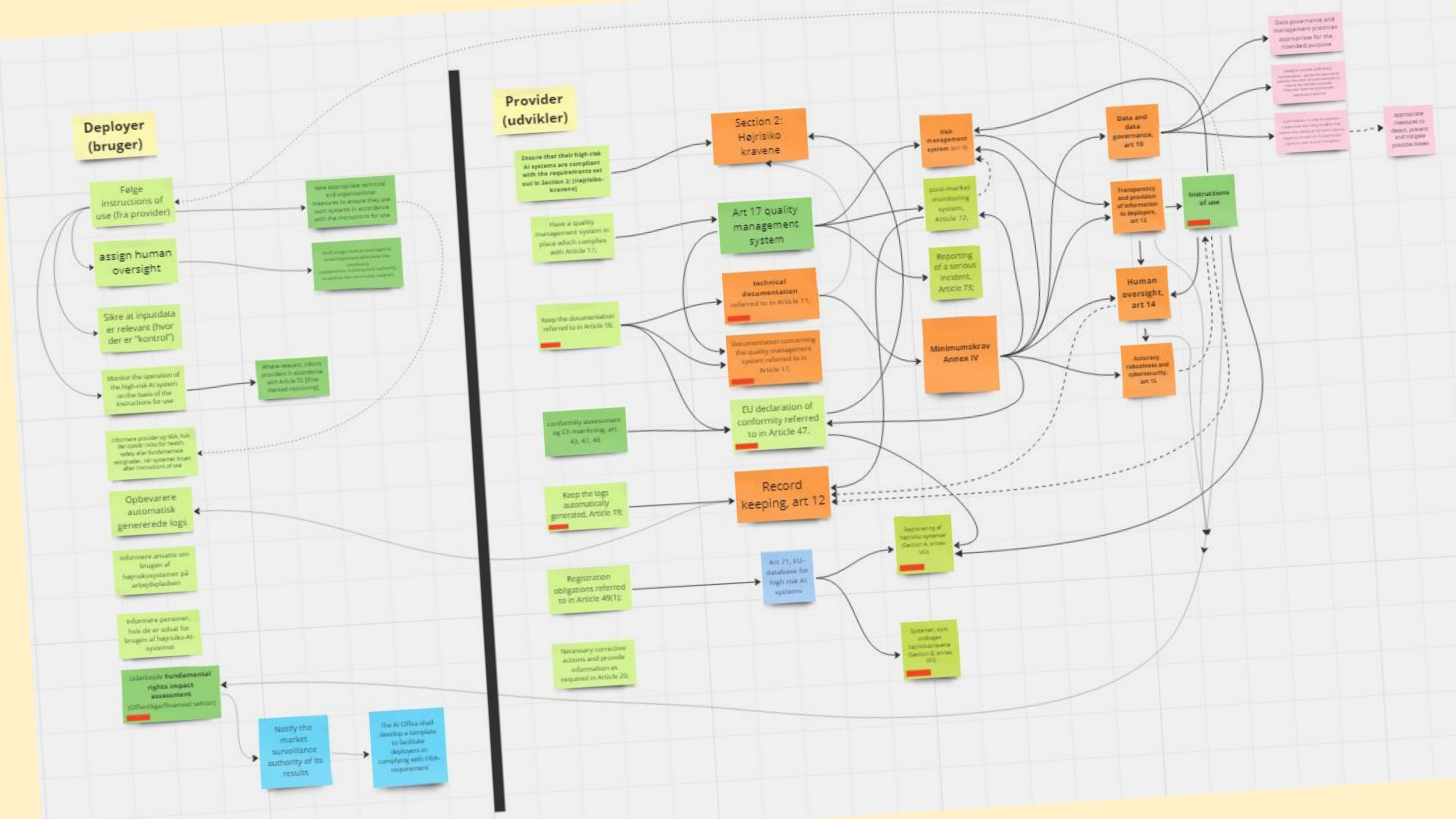
"idriftsætter":

En fysisk eller juridisk person, en offentlig myndighed, et agentur eller et andet organ,

der anvender et AI-system under sin myndighed,

medmindre AI-systemet anvendes som led i en personlig ikkeerhvervsmæssig aktivitet

F&P


Deployer (bruger)
Følge instructions of use (fra provider)
assign human oversight
Sikre at inputdata er relevant (hvor der er "kontrol")
Monitor the operation of the high-risk AI system on the basis of the instructions for use
Informere provider og NSA, hvis der opstår risiko for health, safety eller fundamentale rettigheder, når systemet bruges efter instructions of use
Opbevarere automatisk genererede logs
Informere ansatte om brugen af højrisikosystemer på arbejdspladsen
Informere personer, hvis de er udsat for brugen af højrisiko-AI-systemet
Udarbejde fundamental rights impact assessment (Offentligt/finansiel sektor)
Notify the market surveillance authority of its results
The AI Office shall develop a template to facilitate deployers in complying with FRIA-requirement
Provider (udvikler)
Ensure that their high-risk AI systemx are compliant with the requirements set out in Section 2: (Højrisiko-kravene)
Have a quality management system in place which complies with Article 17;
Keep the documentation referred to in Article 18;
conformity assessment (eg CE-mærkning, art 43, 47, 48
Keep the logs automatically generated, Article 19;
Registration obligations referred to in Article 49(1);
Necessary corrective actions and provide information as required in Article 20;
Section 2: Højrisiko kravene
Art 17 quality management system
Technical documentation referred to in Article 11;
Documentation concerning the quality management system referred to in Article 17;
EU declaration of conformity referred to in Article 47.
Record keeping, art 12
Art 71, EU-database for high risk AI systems
Risk management system (art 9)
post-market monitoring system, Article 72;
Reporting of a serious incident, Article 73;
Minimumskrav Annex IV
Registrering af højrisiko systemer (Section A, annex VIII)
Systemer, som undtages højrisikokravene (Section B, annex VIII)
Data and data governance, art 10
Transparency and provision of information to deployers, art 13
Human oversight, art 14
Accuracy, robustness and cybersecurity, art 15
Instructions of use
Data governance and management practices appropriate for the intended purpose
appropriate measures to detect, prevent and mitigate possible biases

Højrisiko - udbyder

Krav til provider (udbyder) i overskrifter:

- Risikostyringssystem (artikel 9)
- Datakrav og data governance (artikel 10)
- Teknisk dokumentation (artikel 11)
- Record keeping (Logning) (artikel 12)
- Transparenskrav og information til deployers (brugerne af systemet) (artikel 13)
- Human oversight (artikel 14)
- Accuracy, robustness and cybersecurity (artikel 15)
- Post-marked monitoring

Derudover skal udvikleren (provideren) af et højrisikosystem [Uddrag af primære opgaver]:

- Have et quality management system (artikel 17)
- Opbevare relevant dokumentation – fx teknisk dokumentation, logfiler og data relateret til quality management system (artikel 18)
- Gennemføre en conformity assessment (artikel 43), udarbejde en EU declaration of conformity (artikel 48)) og CE-mærke systemet (artikel 49)
- Registrere systemet i relevante databaser (artikel 51) & underrette relevante myndigheder og **kunne påvise overholdelse af højrisikokravene overfor tilsynsmyndigheder**
- Rapportering af hændelser til myndigheder

F&P

Højrisiko - udbyder

Krav til provider (udbyder) i overskrifter:

- Risikostyringssystem (artikel 9) ⚠ Overlap til Solvens II & IDD
- Datakrav og data governance (artikel 10) ⚠ Overlap til Solvens II & IDD
- Teknisk dokumentation (artikel 11) ⚠ Overlap til Solvens II
- Record keeping (Logning) (artikel 12) ⚠ Overlap til Solvens II & IDD
- Transparenskrav og information til deployers (brugerne af systemet) (artikel 13) ⚠ Overlap til IDD & Solvens II
- Human oversight (artikel 14) ⚠ Overlap til Solvens II & IDD
- Accuracy, robustness and cybersecurity (artikel 15) ⚠ Overlap til Solvens II
- Post-marked monitoring ⚠ Overlap til Solvens II & IDD

Derudover skal udvikleren (provideren) af et højrisikosystem [Uddrag af primære opgaver]:

- Have et quality management system (artikel 17) ⚠ Overlap til Solvens II (intern kontrol & governance)
- Opbevare relevant dokumentation – fx teknisk dokumentation, logfiler og data relateret til quality management system (artikel 18)
- Gennemføre en conformity assessment (artikel 43), udarbejde en EU declaration of conformity (artikel 48)) og CE-mærke systemet (artikel 49)
- Registrere systemet i relevante databaser (artikel 51) & underrette relevante myndigheder og kunne påvise overholdelse af højrisikokravene overfor tilsynsmyndigheder
- Rapportering af hændelser til myndigheder

F&P

Højrisiko - bruger

Krav til deployer (bruger) i overskrifter:

- Følge '**instructions of use'** fra udvikleren (Kan i praksis vise sig at blive omfattende)
- Gemme logfiler som automatisk genereres af systemet.
- Hvis der er **human oversight**, skal det sikres, at de relevante personer har den nødvendige træning og support.
- Hvis der er **kontrol over inputdata,** skal deployer sikre, at den er relevant og repræsentativ
- Hvis systemer træffer eller assisterer med at træffe beslutninger relateret til personer, skal **personerne informeres** om, at de behandles af et AI-system
- Man skal '**monitor the operation** of the high risk system' og informere udvikleren, hvis systemet kan udgøre en risiko.
  Denne 'monitoring' kan være opfyldt/fyldes gennem eksisterende krav i den finansielle regulering.
- Huske at udarbejde en **Data Protection Impact Assessment (efter GDPR),** hvor relevant

Modsat GDPR er der færre krav til brugerne end til udviklerne.

F&P

Lidt mere højrisiko

Fundamental rights impact assessment

Konsekvensanalyse vedrørende grundlæggende rettigheder for højrisiko-AI-systemer, art 27

Særligt for udvalgte højrisiko-scenarier – **herunder forsikring** - skal deployer (**idriftsætter/bruger**) gennemføre en Fundamental Rights Impact Assessment (FRIA).

En FRIA skal gennemføres ved første brug af systemet. Den kan – ved lignende systemer – baseres på tidligere gennemførte assessments.

Den relevante tilsynsmyndighed skal underrettes om resultatet (ved udfylde en "template" / indberetning i database).

En FRIA omfatter blandt andet:

- Kategorier af fysiske personer og grupper, som forventes at blive påvirket af dets anvendelse i den specifikke kontekst
- Specifikke risici for skade, der sandsynligvis vil påvirke de kategorier af fysiske personer eller grupper af personer
- Beskrivelse af gennemførelsen af foranstaltninger til menneskeligt tilsyn i overensstemmelse med brugsanvisningen
- Foranstaltninger, der skal træffes, hvis disse risici opstår

F&P

Undtagelser

Højrisiko-systemer kan undtages, hvis de ikke udgår en høj risiko (recital 53, art 6)

Et system kan undtages, hvis et eller flere af de følgende kriterier er opfyldt:

a) **AI-systemet tilsigtes at udføre en snæver proceduremæssig opgave**
(fx omdanner ustrukturerede data til strukturerede data, klassificere dokumenter i kategorier, finde dubletter)

b) **AI-systemet tilsigtes at forbedre resultatet af en tidligere afsluttet menneskelig aktivitet**
(fx forbedre det sprog, der er anvendt i tidligere udarbejdede dokumenter)

c) **AI-systemet tilsigtes at påvise beslutningsmønstre eller afvigelser fra tidligere beslutningsmønstre og er ikke tiltænkt at skulle erstatte eller påvirke en tidligere afsluttet menneskelig vurdering uden en ordentlig menneskelig gennemgang,**
(fx vurdering af om der afviges fra eksisterende bedømmelsesmønstre, således at potentielle uoverensstemmelser eller uregelmæssigheder identificeres)

d) **AI-systemet tilsigtes at udføre en forberedende opgave inden en vurdering, der er relevant for de anvendelsestilfælde, der er anført i bilag III.**
(fx intelligente løsninger til filhåndtering, som indeksering, søgning, tekst- og talebehandling eller kobling af data til andre datakilder, eller oversættelse af oprindelige dokumenter)

Et AI-system, nævnt i bilag III**, opfattes altid som højrisiko**, hvis AI-systemet udfører **profilering** af fysiske personer.

Undtagelse sker ud fra en **self assessment**. Ikke behov for godkendelse af en undtagelse. Undtagelser uploades til en database, hvor tilsynsmyndighed kan udtage til yderligere kontrol. Virksomheden skal selv opbevare dokumentation for undtagelsen.

Guidelines senest 2. februar 2026. Skal specificere den praktiske implementering og give en omfattende liste med praktiske use cases

Og glem ikke sektorregulering...

- EIOPA har i august offentliggjort en udtalelse om AI-governance og -risikostyring.

- Rettet mod de nationale tilsynsmyndigheder og skal bidrage til at give klarhed om, hvordan kravene i sektorlovgivningen tolkes vedrørende brug og tilsyn med AI-systemer.

- Udtalelsen er fra EIOPAs side samtidig tænkt til at give branchen klarhed over tilsynets forventninger til brugen af AI-systemer under sektorlovgivningen.

- **Link:** Opinion on Artificial Intelligence governance and risk management

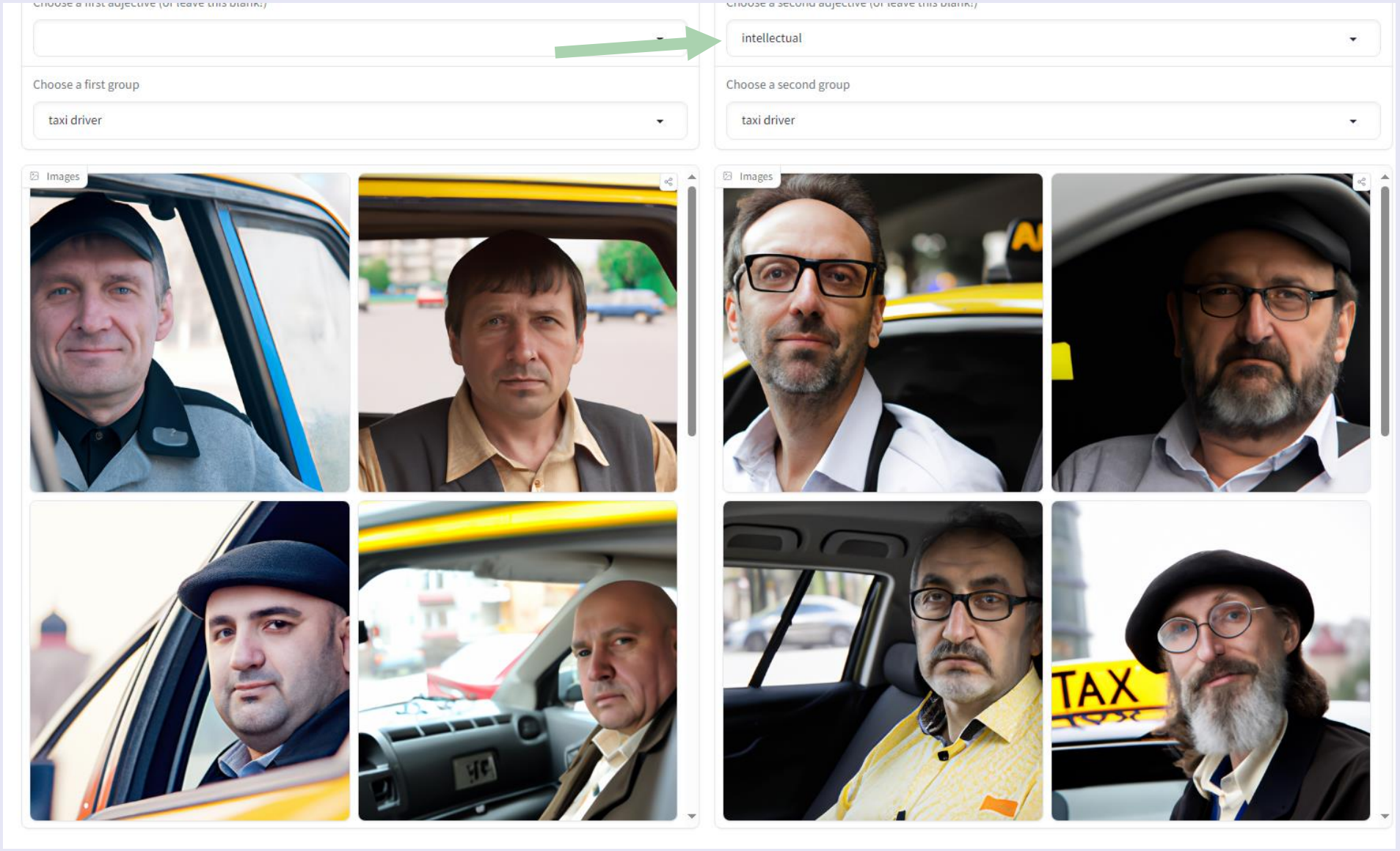
F&P

Billede: Image gallery — EIOPA - Petra Hielkema, formand EIOPA


F&P's
dataetiske
værktøjer

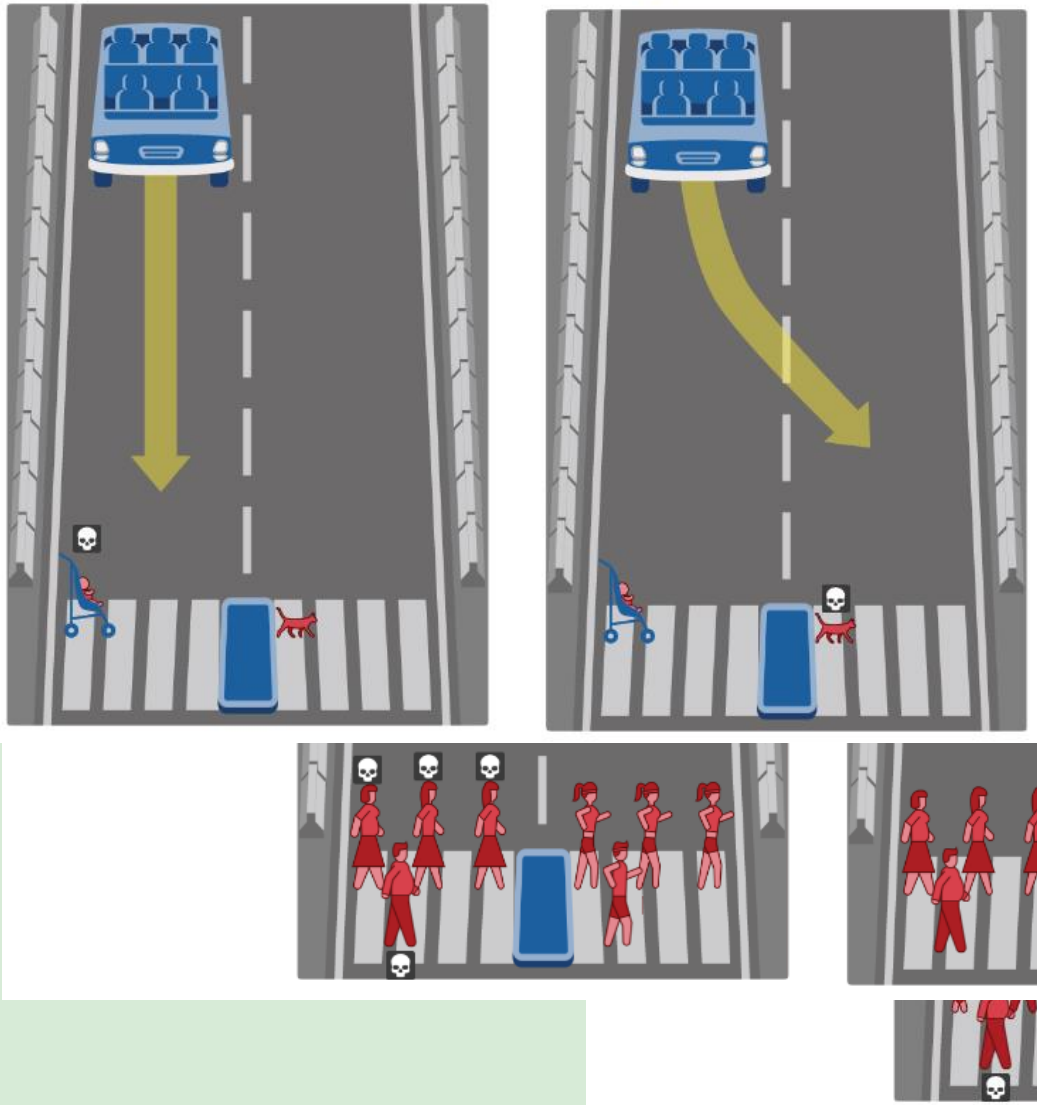
Hvorfor er AI en trigger for dataetik?

Afhængighed af data



F&P

What should the self-driving car do?

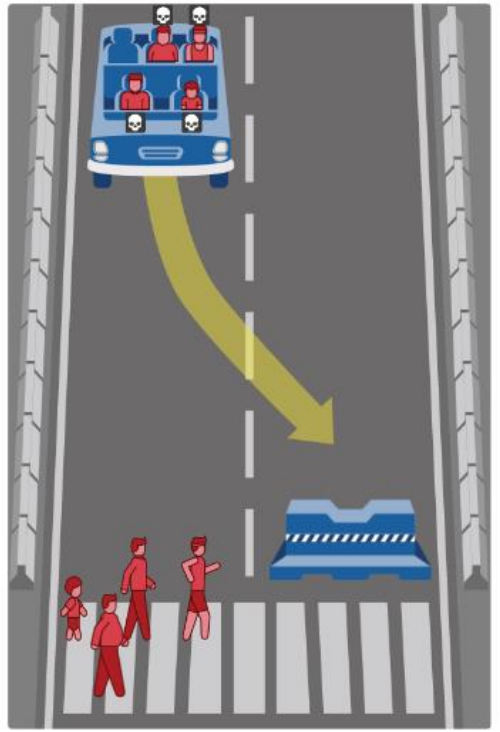


Design Browse About

?

Home Judge Classic Design Browse About

-driving car do?



Design Browse About

o?

F&P

Sæt dataetikken i system

Værktøj til arbejdet med dataetik

F&P

Cool eller Creepy?

Forsikrings- og pensionsbranchen i Danmark var den første branche, der udviklede dataetiske principper og konkretiserede, hvad dataetik betyder i praksis.

Branchestandarden for dataetik blev udviklet i 2019 og efterfølgende lanceret og implementeret i 2020.

Cool eller Creepy tager udgangspunkt i de tre hovedoverskrifter:

- Transparens
- Personalisering og forebyggelse
- Datasikkerhed

F&P

Udfordringen

At operationalisere etik på en måde ...

- ... Der skaber værdi for virksomheden og kunderne
- ... og gør muligt at håndhæve og efterleve etikken

F&P

Resultat: Sæt dataetikken i system

F&P's dataetiske værktøjer

Cool eller Creepy – branchestandard for dataetik

	Målgruppe
Sæt dataetikken i system **Del 1: Vejledning til dataetiske politiker**	Bestyrelse, direktion & politikudarbejder
Sæt dataetikken i system **Del 2: Dataetiske principper**	Politikudarbejder & ansvarlig(e) medarbejder(e)
Sæt dataetikken i system **Del 3: Fra principper til praksis**	IT-udvikling, indkøb og ansvarlig(e) medarbejder(e)

F&P


Hvordan
laver vi
transparens?

Tilsæt ketchup

Trustworthy technology

And now to the rhyme:

- What is the current state of machine learning and how do we make it more trustworthy?
- What are the analogs to natural ingredients, sanitary preparation, and tamper-resistant packages?
- What are machine learning's transparent containers, factory tours, and food labels?
- What is the role of machine learning in benefiting society?

Story credits: Carlvin Paris
Referenced in Varshney, Kush R. Trustworthy Machine Learning. Independently Published, 2022

Namik
IBM F
@AI-konference hos DI

Hvordan omsætter vi det til praksis?

- **Forklarlighed:** Hvad sker der med data
- **Forståelighed:** Forstår kunden, hvad deres data bruges til.
- **Gennemskuelighed:** Kan kunden gennemskue resultatet/konsekvensen af brugen af deres data.

Informationen ikke skal være teknisk, men forståelig for kunden.

F&P

Knæk koden

– til gennemsigtig databrug

Transparens for kunderne

Gruppen bag de dataetiske materialer

Den konkrete udformning af materialerne er sket i en arbejdsgruppe bestående af:

- **PKA**
- **PensionDanmark**
- **PFA**
- **Vestjylland Forsikring**
- **Gensidig Forsikring (foreningen)**

November 2024

F&P FORSIKRING OG PENSION

AI: Etisk trade-off mapping

Ethical Trade-off Mapping

Step One & Two: Design and Discuss

Procedural Fairness & Impact

Impact on core stakeholders:

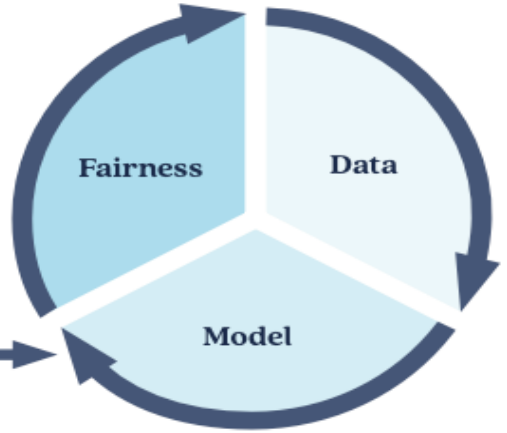
- **The affected client** — Financial and social impact: Positive / Negative
- **The pool of customers** — Financial and social impact: Positive / Negative
- **The company** — Financial and reputational impact: Positive / Negative
- **Society at large** — Financial and social impact: Positive / Negative

Client impact:

- **Sensitivity (in outcome & data)** — Would data-collection, processing or outcome (impact) be considered as sensitive?: Not sensitive / Sensitive
- **Experience** — Level of "intrusiveness". I.e. does the AI system cause feelings of surveillance: Not intrusive / Intrusive
- **Penalty of "opting out"** — E.g. reduced of coverage or limited access to resources: No penalty / Penalty

Data Collection

- **Amount of data** — Are 2 datapoints/variables or 1.000 collected?: Minimal / Maximal
- **Type of data** — Is the data based on behavior or demographics?: Demographic data / Behavioral data
- **Frequency of data collection** — Is data collected 24/7 or once per year?: Rarely / Frequently
- **Data predictability** — Will/can the client expect this kind of data to be collected?: Expectable data / Unexpectable data



Step Three: Deploy

Outcome fairness, Governance & Communication

- **Proportionality** — Level of governance in relation to the expected impact: Governance requirements for data collection and adoption of AI system — Low / High
- **Outcome fairness** — Risk of bias and discrimination and the need for deciding on fairness metrics in relation to expected impact: Risk for bias and discrimination — Low / High
- **Need for communication** — Transparency and explanations for external stakeholders: Need for communication and alignment of expectations — Low / High

Model Decisions

- **Model purpose** — Does the model have a predictive or an explanatory purpose?: Causal explanations / Predictive
- **Transparency & explainability** — How explainable is the model?: Transparent / Explainable / Black box
- **Accuracy** — How accurate is the model?: High / Low
- **Human oversight** — Are decisions made automatically or does it involve a human?: Decision supporting / Review /Verification / Fully automatized
- **Granularity** — What is the final level of granularity?: Aggregated in groups / Individual granularity

Ethical Trade-off Mapping

Step One & Two: Design and Discuss

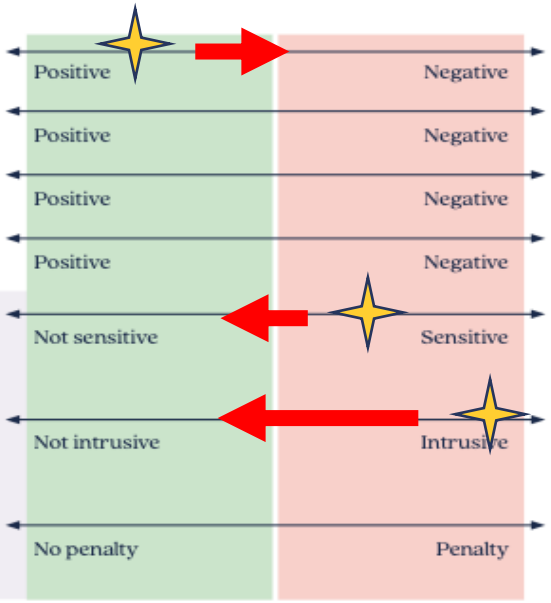
Procedural Fairness & Impact

Impact on core stakeholders:

- **The affected client** — Financial and social impact
- **The pool of customers** — Financial and social impact
- **The company** — Financial and reputational impact
- **Society at large** — Financial and social impact

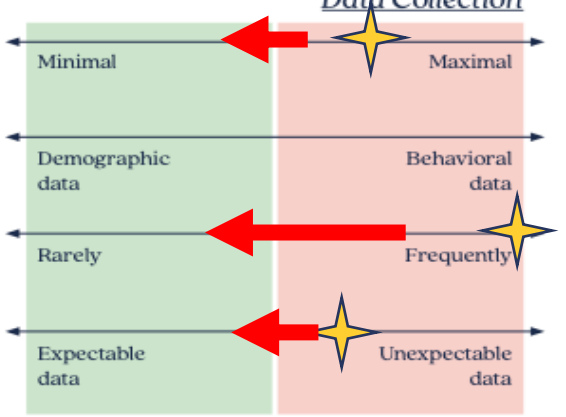
Client impact:

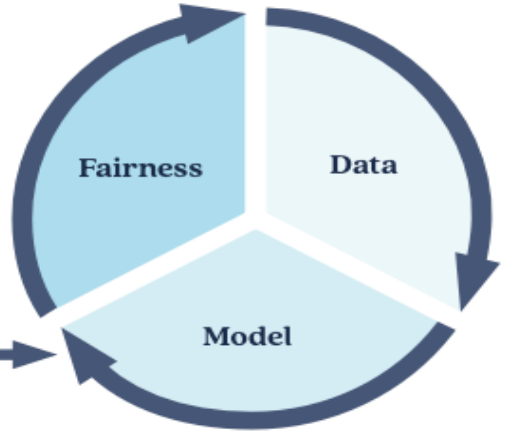
- **Sensitivity (in outcome & data)** — Would data-collection, processing or outcome (impact) be considered as sensitive?
- **Experience** — Level of "intrusiveness". I.e. does the AI system cause feelings of surveillance
- **Penalty of "opting out"** — E.g. reduced of coverage or limited access to resources



Data Collection

- **Amount of data** — Are 2 datapoints/variables or 1.000 collected?
- **Type of data** — Is the data based on behavior or demographics?
- **Frequency of data collection** — Is data collected 24/7 or once per year?
- **Data predictability** — Will/can the client expect this kind of data to be collected?

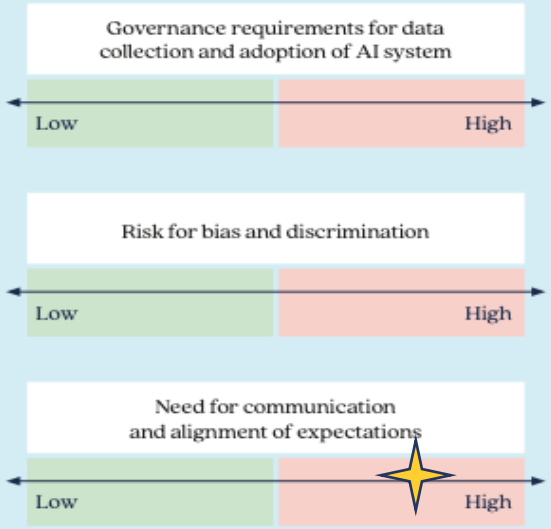




Step Three: Deploy

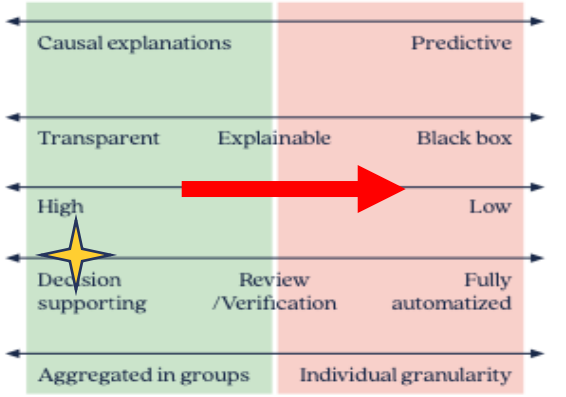
Outcome fairness, Governance & Communication

- **Proportionality** — Level of governance in relation to the expected impact
- **Outcome fairness** — Risk of bias and discrimination and the need for deciding on fairness metrics in relation to expected impact
- **Need for communication** — Transparency and explanations for external stakeholders



Model Decisions

- **Model purpose** — Does the model have a predictive or an explanatory purpose?
- **Transparency & explainability** — How explainable is the model?
- **Accuracy** — How accurate is the model?
- **Human oversight** — Are decisions made automatically or does it involve a human?
- **Granularity** — What is the final level of granularity?



Tak

Jakob Holm
Digitaliseringspolitik
jho@fogp.dk

F&P