

Trustworthy AI og Robusthet

ELMED219: Momentliste T01–T07

ELMED219

Vår 2026

1 Trustworthy AI

- T01: Definere trustworthy AI iht. EU-retningslinjer

2 Robusthet og usikkerhet

- T02: Forklare konseptet robusthet i ML/AI
- T03: Beskrive datadrift og dens konsekvenser
- T04: Forklare epistemisk vs. aleatorisk usikkerhet

3 Menneske-maskin samspill

- T05: Beskrive human-in-the-loop (HITL) systemer
- T06: Diskutere viktigheten av kontinuerlig monitorering

4 Sikkerhetstrusler

- T07: Kjenne til adversarial attacks

T01: Definere trustworthy AI iht. EU-retningslinjer

EU High-Level Expert Group on AI (2019):

Definisjon

Trustworthy AI er AI som er lovlig, etisk og robust – både teknisk og sosialt.

Syv nøkkelkrav:

- ① Menneskers handlingsrom og tilsyn – Human agency & oversight
- ② Teknisk robusthet og sikkerhet – Technical robustness & safety
- ③ Personvern og dataforvaltning – Privacy & data governance
- ④ Transparens – Transparency
- ⑤ Mangfold, ikke-diskriminering, rettferdighet – Diversity, non-discrimination, fairness
- ⑥ Sosial og miljømessig velferd – Societal & environmental well-being
- ⑦ Ansvarlighet – Accountability

Medisinsk relevans

Høyrisiko AI-systemer i helse må oppfylle disse kravene under EU AI Act.

T02: Forklare konseptet robusthet i ML/AI

Robusthet = modellens evne til å prestere pålitelig under variasjon

En robust modell:

- Gir **konsistente resultater** på lignende input
- **Degraderer gradvis** (ikke katastrofalt) ved støy
- Generaliserer godt til nye, usette data

Typer robusthet:

- **Støyrobusthet:** Toleranse for tilfeldig støy
- **Distribusjonell robusthet:** Endringer i datafordeling
- **Adversarial robusthet:** Motstand mot bevisste angrep
- **Temporal robusthet:** Stabilitet over tid

I medisin

En robust medisinsk AI gir pålitelige prediksjoner uavhengig av variasjon i bildekvalitet, pasientpopulasjon, eller utstyr.

T03: Beskrive datadrift og dens konsekvenser

Datadrift (distributional shift):

- Forskjell mellom **treningsdata** og **produksjonsdata**
- Modellen møter data som ikke ligner det den er trent på

Typer drift:

- ① **Covariate shift**: Input-fordeling endres (f.eks. ny pasientdemografi)
- ② **Label shift**: Forekomst av klasser endres (f.eks. pandemier)
- ③ **Concept drift**: Sammenhengen mellom input og output endres

Eksempler i medisin:

- Modell trent på data fra USA brukes i Norge
- Ny MR-skanner gir andre bildekarakteristikker
- COVID endret innleggelsesmønstre

Konsekvens

Modellen kan **feile stille** – gir prediksjoner med høy konfidens som er feil!

T04: Forklare forskjellen mellom epistemisk og aleatorisk usikkerhet

Epistemisk usikkerhet:

- Usikkerhet pga. **manglende kunnskap**
- “Vi vet ikke nok ennå”
- **Kan reduseres** med mer data
- Modellen er usikker på områder med lite treningsdata

Eksempel:

- Sjeldent sykdom med få treningseksempler
- Usikkerhet fordi modellen mangler erfaring

Aleatorisk usikkerhet:

- **Iboende** tilfeldig variasjon i data
- “Verden er usikker”
- **Kan ikke reduseres** med mer data
- Støy i målinger, naturlig variasjon

Eksempel:

- To pasienter med identiske features har ulike utfall
- Usikkerhet fordi utfall er genuint usikkert

Praktisk betydning

Epistemisk usikkerhet signaliserer når modellen bør “si ifra” at den er usikker, og mennesker bør overta.

T05: Beskrive human-in-the-loop (HITAL) systemer

Human-in-the-loop (HITAL):

- Mennesker er **integrt** i AI-systemets beslutningsprosess
- AI gir anbefalinger, mennesker tar endelige beslutninger

Tre hovedvarianter:

- ① **Human-in-the-loop:** Menneske involveres i hver beslutning
- ② **Human-on-the-loop:** Menneske overvåker og kan overstyre
- ③ **Human-out-of-the-loop:** Full autonomi (ikke anbefalt i medisin)

Fordeler med HITAL i medisin:

- Kombinerer AI-effektivitet med menneskelig ekspertise
- Fanger opp AI-feil før de får konsekvenser
- Opprettholder klinisk ansvar, bygger tillit gradvis

EU AI Act

Høyrisiko AI-systemer **krever** effektiv menneskelig tilsyn.

T06: Diskutere viktigheten av kontinuerlig monitorering

Hvorfor kontinuerlig monitorering?

- ML-modeller **degraderer over tid** (model drift)
- Verden endrer seg, data endrer seg
- Feil i produksjon kan ha alvorlige konsekvenser

Hva bør overvåkes?

Ytelsesmetrikker:

- Accuracy, precision, recall over tid
- Kalibrering av konfidensscorer
- Sammenligning med baseline

Datakarakteristikker:

- Input-fordeling (drift-deteksjon)
- Andel out-of-distribution input
- Feature-statistikk

Tiltak ved problemer:

- **Re-trening** med ferske data – oppdater modellen på nyere, representative data
- **Varsling og eskalering** – automatisk varsling ved ytelsesfall eller drift
- **Fallback** – bytt til enklere modell eller menneskelig vurdering ved usikkerhet

T07: Kjenne til adversarial attacks

Adversarial attacks = bevisst manipulering av AI-input

- Små, tilsynelatende usynlige endringer som lurer modellen
- Kan få en korrekt klassifikasjon til å bli **fullstendig feil**

Eksempler:

Bilder:

- Piksel-perturbasjoner usynlige for mennesker
- “Panda” → “Gibbon” med 99% konfidens
- **Klistremerker som lurer selvkjørende biler**

Tekst (LLM):

- **Prompt injection**
- **Jailbreaking**
- Omgå sikkerhetsfiltre

Medisinsk risiko og forsvar

Risiko: En angriper kunne manipulere et røntgenbilde slik at AI overser patologi, eller skaper falske funn.

Forsvar: Adversarial training, input-validering, ensemble-metoder, robusthetstesting.

Oppsummering: Trustworthy AI og Robusthet

Nøkkelpunkter:

- **T01:** EU's 7 krav til trustworthy AI
- **T02:** Robusthet – konsistente resultater under variasjon
- **T03:** Distributional shift – når data i praksis avviker fra trening
- **T04:** Epistemisk (kan reduseres) vs. aleatorisk (iboende) usikkerhet
- **T05:** HITL – mennesker i beslutningssløyfen
- **T06:** Kontinuerlig monitorering – fange drift og degradering
- **T07:** Adversarial attacks – bevisst manipulering

Hovedbudskap

Trustworthy AI i medisin krever robuste modeller, menneskelig tilsyn, kontinuerlig monitorering og bevissthet om sikkerhetsrisikoer. Teknikk alene er ikke nok – det kreves også organisatoriske og regulatoriske tiltak.