Trustworthy AI og Robusthet

Momentliste T01–T07

ELMED219 / BMED365

Universitetet i Bergen

Våren 2026

Oversikt

1. Trustworthy AI
2. Robusthet og usikkerhet
3. Menneske-maskin samspill
4. Sikkerhetstrusler

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

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 distributional shift og dens konsekvenser

Distributional shift (datadrift):

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

Typer shift:

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:

- Sjelden 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

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

Human-in-the-loop (HITL):

- Mennesker er **integrert** 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 HITL i medisin:

- Kombinerer AI-effektivitet med menneskelig ekspertise
- Fangler 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:

- Retraining med ferske data
- Varsling og eskalering
- Fallback til enklere modell eller menneskelig vurdering

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

En angriper kunne teoretisk manipulere et røntgenbilde slik at AI overser patologi, eller omvendt – skaper falske funn.

Forsvar

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.