OSLOMET

TILLIT TIL KUNSTIG INTELLIGENS: PERSPEKTIVER OG PRAKSIS

Vil kunstig intelligens styrke eller svekke cybersikkerheten?

Trusler, forsvarsmekanismer og samfunnseffekter

Boning Feng, Førsteamanuensis, PhD
Institutt for Informasjonsteknologi, OsloMet

STORBYUNIVERSITETET
OSLO METROPOLITAN UNIVERSITY

OSLOMET

Agenda

1. Den tilkoblede tidsalderen
2. AI i feil hender: angrepsmetoder
3. AI i cybersikkerhet: deteksjon og respons
4. Samfunnseffekter og risiko
5. Konklusjon og veien videre

De tilkoblede enhetenes tidsalder

- 5G støtter alt fra datakrevende smarttelefoner til primitive sensorer
- Ultra-reliable low-latency muliggjør nye presisjonsapplikasjoner
- Massiv økning i antall tilkoblede enheter
- Nye tjenester gir nye cybersikkerhetstrusler

IoT-vekst og endret trusselbilde

- IoT-enheter globalt øker fra 14,7 mrd (2024) til over 40 mrd (2034) – I følge Statistica
- Raskt voksende angrepsflate med milliarder av enheter
- Enorme datamengder fra tilkoblinger
- Tradisjonelle perimeter-mekanismer er ikke lenger tilstrekkelige

Hvorfor tradisjonelt forsvar ikke er nok

- Perimeterfokus fanger ikke distribuert og dynamisk trafikk
- Kompleksitet + volum overstiger manuell kapasitet
- Behov for adaptiv, datadrevet deteksjon

Hvorfor AI i cybersikkerhet

Krever nye tiltak for å håndtere dynamiske trusler

AI er et naturlig førstevalg for deteksjon og respons

Ondsinnede aktører bruker også AI for å omgå forsvar

Vi må forstå både angrep og forsvar med AI

AI i feil hender – kartlegging av sårbarheter

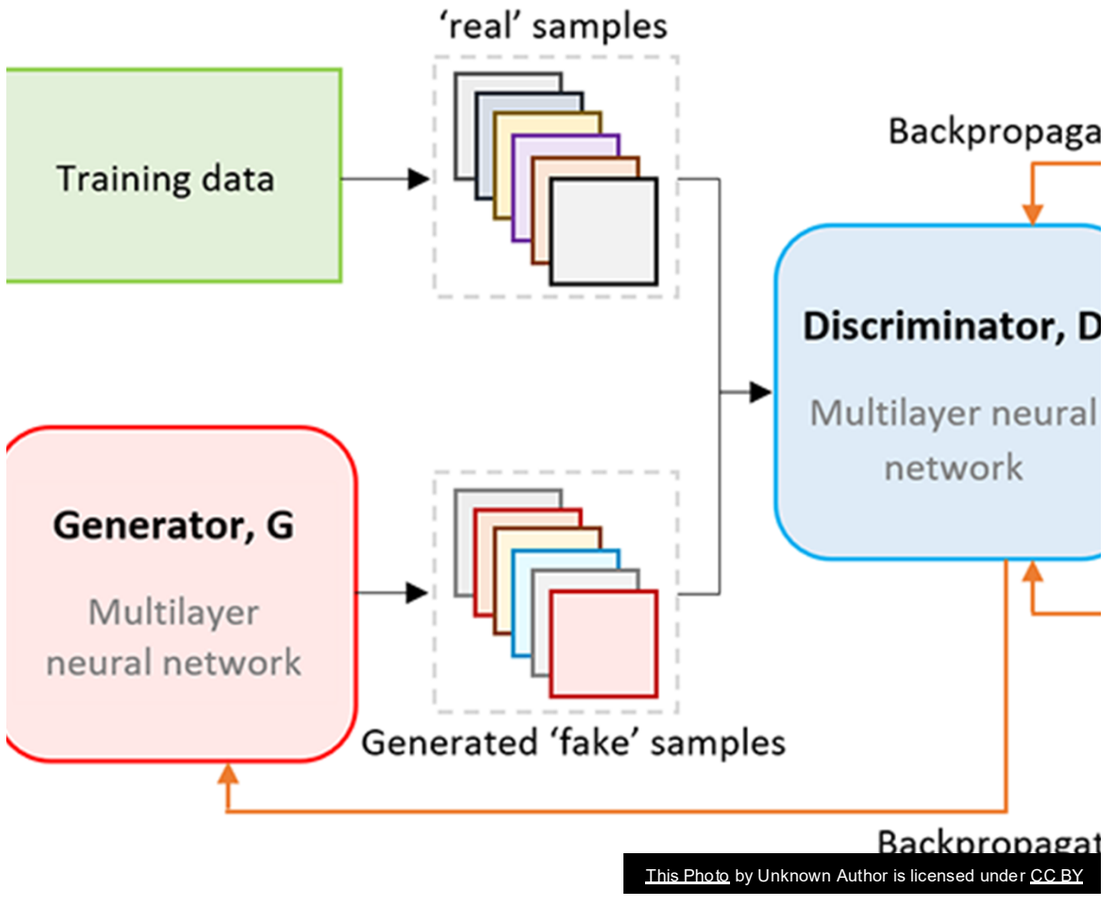
STORBYUNIVERSITETET
OSLO METROPOLITAN UNIVERSITY

OSLOMET

AI i feil hender – unngå deteksjon

- Skanning under radarnivå for å unngå anomalier
- Skadevare som dynamisk endrer adferd
- Bruk av GANs og forsterkende læring for evasion
- Mål: kamuflere aktivitet og omgå sikkerhetssystemer

Hva er GANs (Generative Adversarial Networks)



- Generativ modellering med generatornett og diskriminatornett
- Generatoren lærer å lure diskriminatoren med «ekte» utdata
- Ved skadevare: generere tilsynelatende harmløse varianter
- Kontinuerlig forbedring gjennom gjentatt trening

Spearphishing og bedrageribrev

- AI genererer overbevisende og personlig tilpassede meldinger
- NLP og ML etterligner ekte kommunikasjon
- Vanskelig for mennesker å skille mellom ekte og falsk
- Økt risiko for sensitiv informasjon og skadelige handlinger

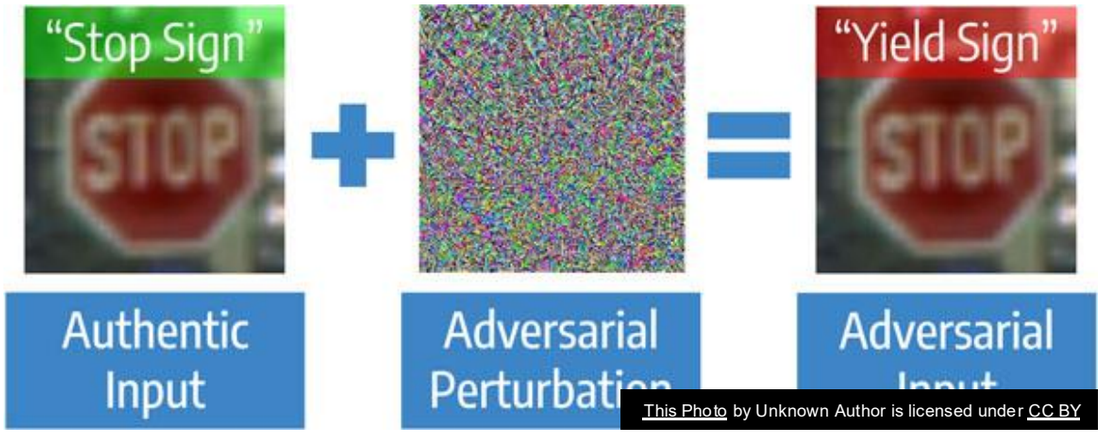
OSLOMET

Deepfake-angrep

- Fabrikkert og manipulert tekst, lyd eller bilde
- Kombinerer deep learning med generering av falskt innhold
- «Virtuelle personer» kan opprettes med få klikk via LLM-er
- Konsekvenser: omdømmetap, økonomisk tap, sosial uro

Angrep mot ML-systemer

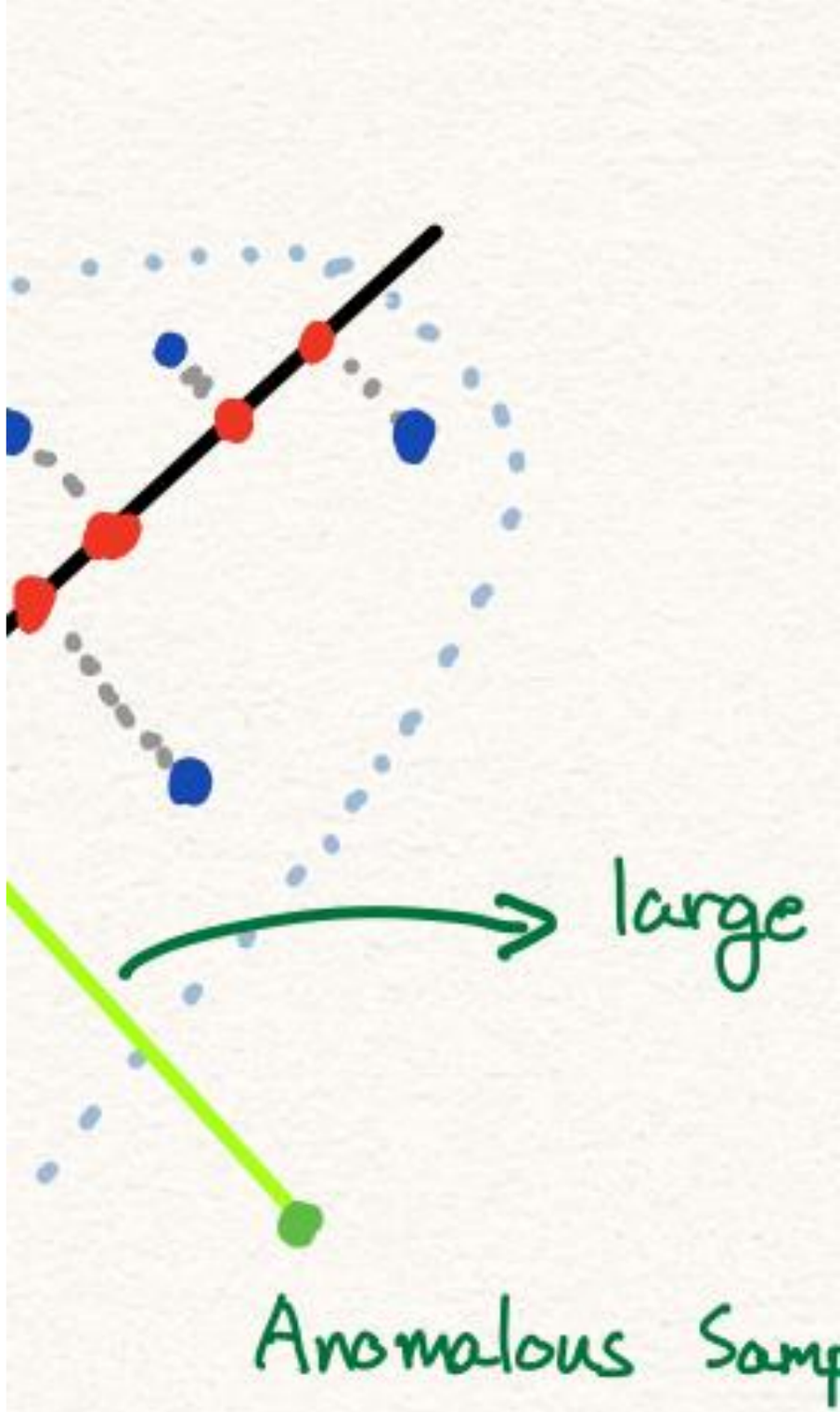
- Adversarial angrep gir feilaktige prediksjoner og beslutninger
- Feilklassifisering av inndata lurer modellens filtre
- Dataforgiftning endrer treningsdata med uriktige inndata
- Resultat: mindre nøyaktige eller direkte feil utdata



OSLOMET

AI i cybersikkerhet – anomalideteksjon

- Deteksjon av atferdsmønstre og anomalier for IoT-enheter
- Enheter har identifiserbare mønstre i data og frekvens
- Forstyrrelser i mønster kan indikere trussel for enhet og operatør
- Arbeid fra Telenor Research og OsloMet på 5G-trafikkmønstre



OSLOMET

Enhetsprofiler i mobilnett

- Profilering basert på historisk atferd i mobilnett
- Skille mellom enhetstyper som kamera og temperatursensor
- Terskler brukes for å vurdere om trafikk er normal
- Avvikende trafikk flagges som anomali for videre undersøkelse

Generering og deling av IoCs

AI/ML ekstraherer indikatorer på kompromittering (IoCs)

Deling via trusselutvekslingsplattformer med passende rammeverk

CMTMF for telekom hjelper å forstå angrep via IoT-enheter*

Delte data styrker I/A DS ved å lære av tidligere angrep

CMTMF - CONCORDIA MOBILE THREAT MODELLING FRAMEWORK
I/A DS - ANOMALY DETECTION SYSTEMS

OSLOMET

Introduksjon til Honeypots i Cybersikkerhet

- **Definisjon av Honeypots**
  - En honeypot er en "felle" designet for å tiltrekke ondsinnede aktører.
  - Fungerer som et verktøy for å isolere skadelig aktivitet fra kritisk infrastruktur.
- **Typer av Honeypots**
  - Reelle og simulerte datamaskiner.
  - Tjenester, nettverk, brukerkontoer og dataobjekter.
  - Multilagskonfigurasjoner som etterligner en hel infrastruktur.

Utfordringer og Dynamiske Honeypots

- **Utfordringer med Tradisjonelle Honeypots**
  - Angripere kan bruke AI til å gjennomføre Adversarial Attacks.
  - Risiko for at angripere identifiserer og unngår honeypot-nettverk.
- **Dynamiske Honeypots**
  - Innføring av AI-baserte, dynamiske honeypots.
  - Automatiske omkonfigurerbare feller for å håndtere AI-drevne angrep.
  - Tilpasning basert på angriperens atferd for bedre beskyttelse.

Begrensninger og mottiltak

- AI-baserte I/A DS (Intrusion or Anomaly Detection Systems) støtter avverging og begrensning av angrep
- Angripere lærer hvordan systemene fungerer
- Vellykkede og uoppdagede måter å gjennomføre angrep
- Krever kontinuerlig forbedring og kombinasjon av verktøy

Samfunnseffekter og risiko

- AI/ML kan gi stor verdi og alvorlig skade
- Nødvendig å forstå muligheter og begrensninger
- Fokus på personvern, frihet, rettferdighet og etikk
- EU og USA arbeider med regulering for rettferdig bruk

Konklusjon

AI må bygges i tråd med menneskers overordnede mål

Mer datainnsamling utfordrer personvernet

Risiko for sosial undertrykkelse ved beslutninger basert på data

Fremtiden blir med AI – om den er på vår side avhenger av oss

OSLOMET

Q&A

- Spørsmål og diskusjon
- Takk for oppmerksomheten!