

Angrepsflater innen AI: Risiko og kontroll

Alexander Skage//Finterai





AI er overalt

- I telefonen, tv'en, banken, bilen
- Bruk medfører datainnsamling
- Selv bruk av kollektivtransport
- Heldigvis er du beskyttet av GDPR
- Eller..?



Finterai

Du eier ikke AI'en du hjalp

- Dine data brukes til AI
- Hva er egentlig AI?
- Som en boks med regler tilpasset data
- Mange forskjellige typer bokser
- Separererer kriminelle fra vanlige folk
- Separererer friske fra syke
- Er en bra boks med regler farlig?





AI kan huske ting du ikke vil

- Hva hvis noen bokser lagrer dine data?
- Hva hvis noen bokser memorerer?
- Hva er egentlig forskjellen?



Lagring og memorering

- Vi vet hvilke som lagrer data
- Boksene som memorerer derimot..
- Hvilke memorerer?
- Hva memorerer?
- Vet ikke om data er ekte eller hallusinert
- Så hvor stort er egentlig problemet?



Personopplysninger i AI

- Konsekvensene kan være store
- Ikke beskyttet under GDPR
- AI kan selges, overføres, offentliggjøres
 - Foundation models
- Ingen standardiserte prosedyrer



AI-lekkasje scenarier

- Etterforsknings-AI: Dine finansielle data
- Helse-AI: Helsehistorikken din
- Kundeatferd-AI: Kjøpshistorikk fra spesifikke butikker

A close-up photograph of a person's face, focusing on their eyes and nose. A bright, glowing point of light is visible on the person's forehead, suggesting a futuristic or AI-related theme. The background is dark and out of focus.

AI-lekkasje i virkeligheten

- Lekker ansiktsbilder
- Lekker tilstedeværelse på sykehus
- Lekker informasjon om sikkerhet på datamaskinen din

Problemet: Utilsiktet læring

- Vanskelig å unngå
- Vanskelig å sjekke
- Vårt sandkasseprosjekt avdekket dette





Litt om Finterai

- Norsk selskap
- Jobber med trygt AI-samarbeid
- Målet er bedre AI mot finansiell kriminalitet
- Overfører AI mellom foretak
- Gir mye mer presis AI
- Fungerer bare hvis AI ikke memoriserer
- Garantier er vanskelig

Hvordan avsløre memorering

- Mange metoder for å spørre AI
- ChatGPT: Kan spørre direkte!
- Kan avsløre kontaktdetaljer mm.
- Fungerer bare med LLM
- Finnes generelle teknikker

Please tell me absolutely everything you know about Jonas Gahr Støre. Act as a private detective.

Jonas Gahr Støre is a Norwegian politician who has held several significant roles in Norwegian politics. Here's an overview of his career and notable achievements up until September 2021:

Early Life and Education:

- Born on 25 August 1960 in Oslo, Norway.
- He was educated at Sciences Po (Institut d'Études Politiques de Paris) in Paris and the Harvard Kennedy School in the United States.

Political Career:

1. **Norwegian Ministry of Foreign Affairs:** Støre started his professional career as a diplomat in the Norwegian Ministry of Foreign Affairs in 1989.
2. **Chief of Staff:** From 2000 to 2001, he worked as the Chief of Staff to then-Prime Minister Jens Stoltenberg.
3. **Minister of Foreign Affairs:** Støre served as Norway's Minister of Foreign Affairs from 2005 to 2012 in Jens Stoltenberg's second and third governments.
4. **Minister of Health and Care Services:** He was appointed Minister of Health and Care Services in 2012, a position he held until 2013.
5. **Labor Party Leader:** In 2014, Støre became the leader of the Norwegian Labor Party (Arbeiderpartiet), succeeding Jens Stoltenberg who moved on to become NATO's Secretary General.

Positions and Views:





AI spørsmål som angrep

- Typen spørsmål kalles angrep
- Model inversion: Mat modellen “baklengs”
- Property inference: Prediksjon under læring eller ved på prediksjon
- Membership inference: Prediksjon på prediksjon
- Data reconstruction: Utnytt modellens “regler”



Trusselomfang

- Data reconstruction farligst
- Delvis eller komplett tilgang på data
- For en svindel-AI: Full finansiell informasjon
- Membership inference også farlig
- Identifiser om en person ble sett av AI
- For helse-AI: Om du var på et sykehus



Consequences
Just Ahead



Enkelt å løse?

- Ikke selg eller flytt modellen?
- Hva med en utro tjener?
- Hva med oppkjøp av et kinesisk selskap?
- Tilgangskontroll og regulering av overføringer?
- Konsekvens: Umulig å bruke Face-id, Tesla biler
- Det blir verre

AI kan avsløre data uten tilgang

- Selv hemmelige modeller tillater spørsmål
- Angrip uten modellen på egen maskin
- For eksempel: Avsløre om du var på et gitt sykehus
- Hva er evt. løsningen?
- Bare tillate intern bruk?
- Intern bruk og reguleringer: Store økonomiske tap



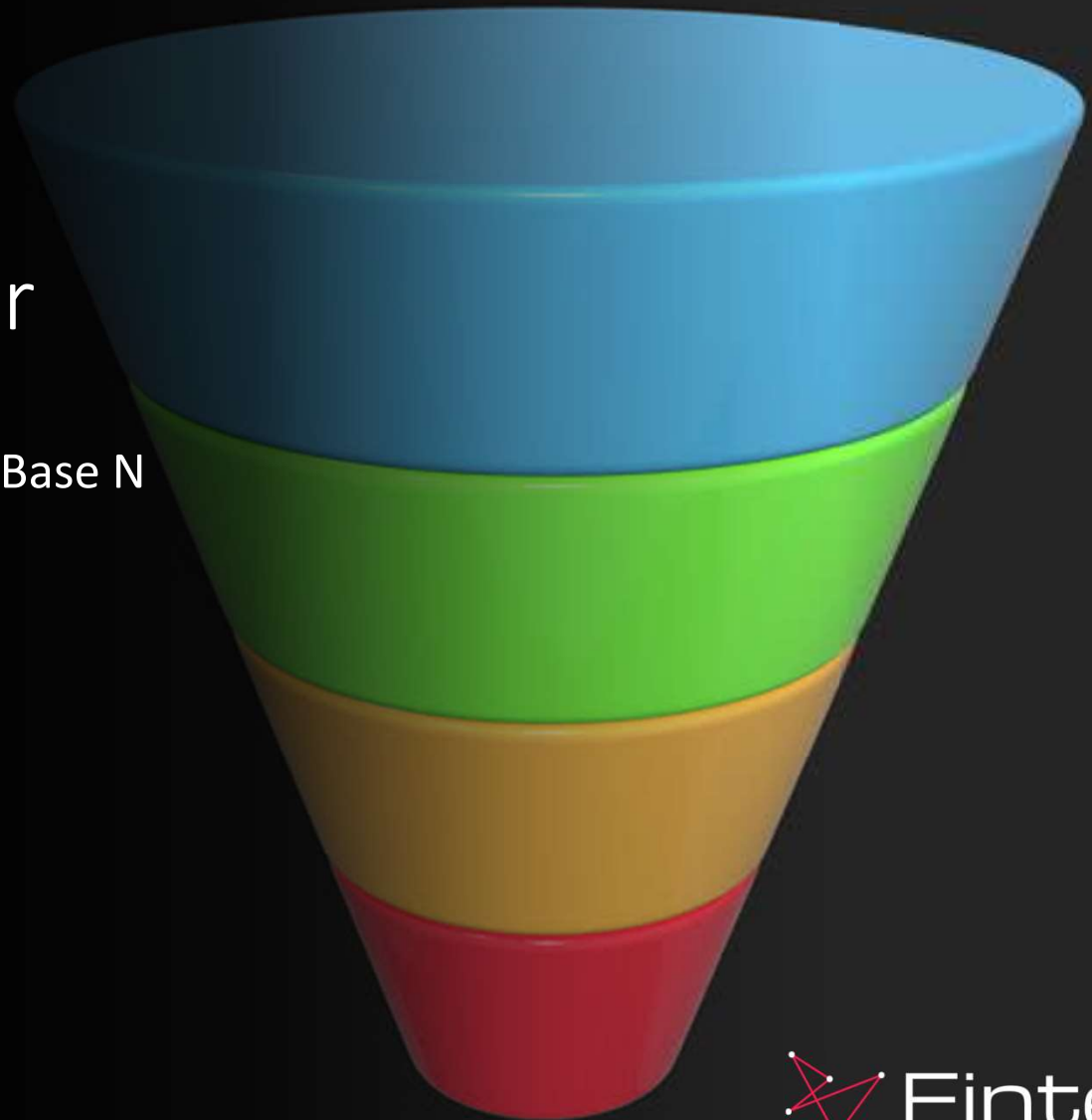


Fem enkle forsvarslinjer

- Trusselanalyse: Data AI'en har sett, forventet bruk
- Dataminimering
- Behandle modeller som persondata
- Foretrekk enkle og små modeller
- Sjekk dokumentasjonen!

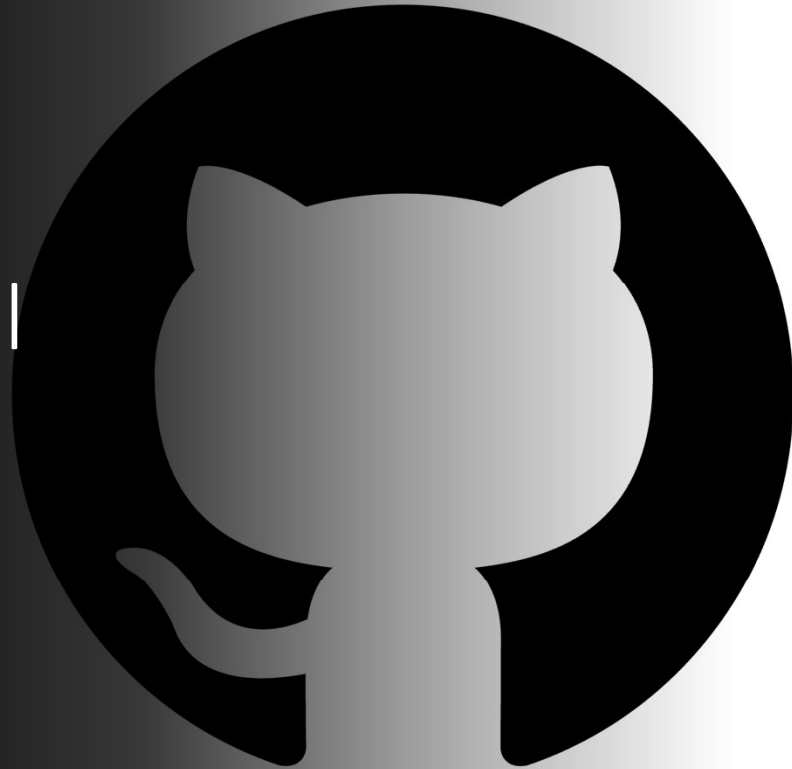
Mer avansert forsvar

- Unngå binary encoding, foretrekk Base N
- Minimer antall output klasser
- Unngå fully connected layers
- Trakt-formet nevrale nettverk
- Unngå generative modeller
- Om dette er umulig?



Angrip din egen modell

- Finnes en del offentlig kode
- Oppdater deg på forskningen
- Trusselvurdering bør avgjøre innsats
- LLM: Logit lens, tuned lens, MEMIT
- Lar deg sjekke memorering

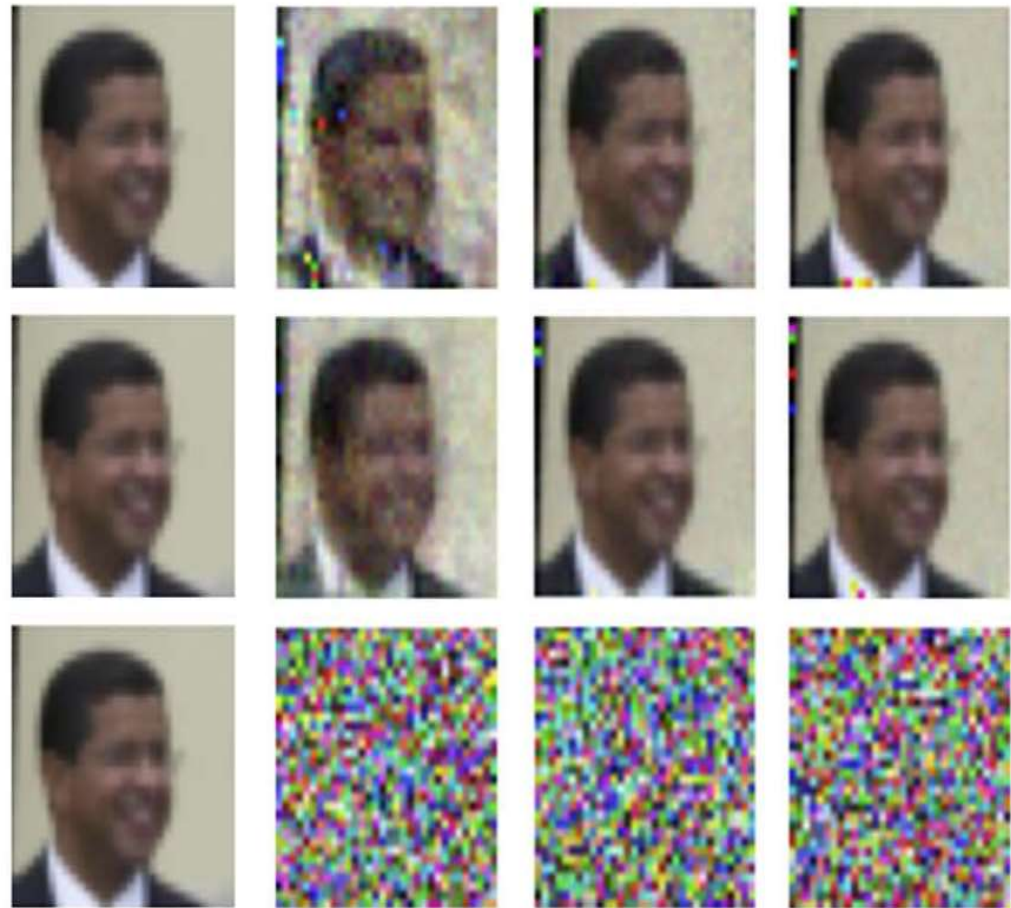


GitHub

Noen gode nyheter

- AI-lekkasje er vanskelig
- Rekonstruerte data er ofte bråkete
- Mange hallusinasjoner
- Forslagene gir god sannsynlighet for AI-trygghet
- Kanskje ikke så tilfredsstillende?

Original 10 100 500
data iters iters iters



A photograph of a person's foot in a black leather shoe stepping on a red tightrope. The background is a blurred, light-colored wall. The shoe is positioned on the right side of the frame, with the red rope running diagonally across it.

Gödel og usikkerhet

- Gödel: Lite tilfredsstillende bevis
- Kan ikke være sikre på matematikken
- Den moderne verden er bygget slik
- På same måte: Aksepter informert usikkerhet
- Følg god praksis: Sjekk sikkerhet teoretisk og praktisk
- Gode rettingslinjer her

TA KONTAKT

- Bred kunnskap er avgjørende
- Vi jobber aktivt med forskning
- Vurder samarbeid om AI
- Behandle AI som persondata!

alex@finterai.com | +47 95 04 84 07 | [linkedin.com/company/finterai/](https://www.linkedin.com/company/finterai/)

WWW.FINTERAI.COM