BEGRENSNINGER OG ETISKE BETRAKTINGER RUNDT SPRÅKMODELLER

SAMIA.TOUILEB

@UIB.NO

ETISKE PROBLEMSTILLINGER MED SPRÅKMODELLER

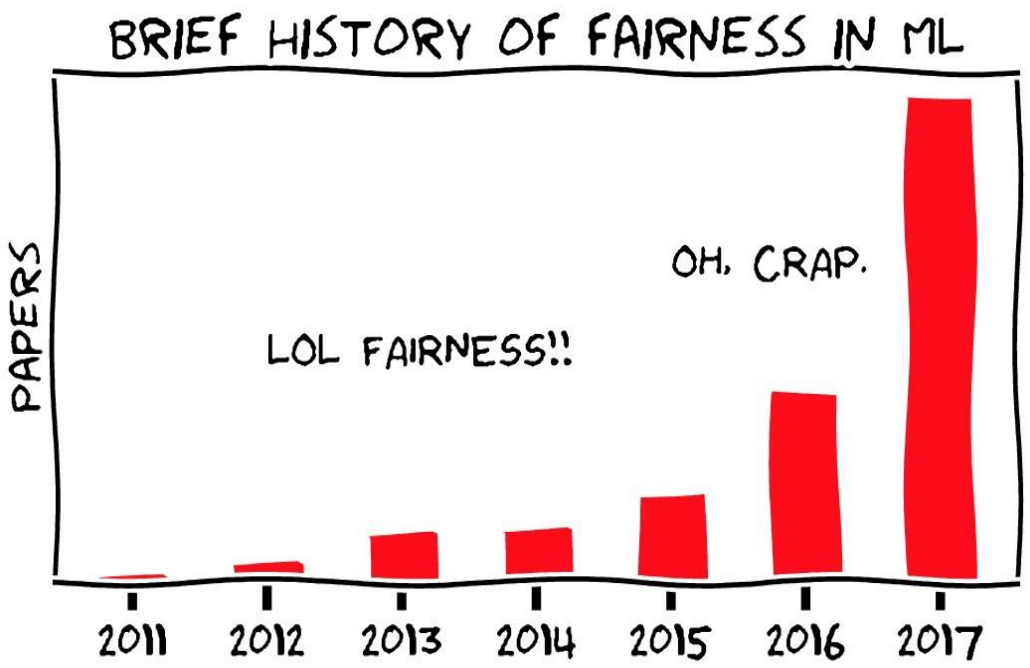
- Bias, diskriminering, toksisitet og annet problematisk innhold.
- Feilinformasjon.
- Opphavsrett.
- Personvern.
- Miljømessige konsekvenser.

ENKELTE TYPER SKADELIGE PÅVIRKNINGER

(Suresh and Guttag, 2021; Bender et al., 2020; Barocas et al, 2017; Crawford, 2017)

- Allocational harms
  - Å tildele, eller frata, visse (grupper av) mennesker en mulighet eller en ressurs.
- Representational harms
  - Stigmatisere eller stereotypisere visse (grupper av) mennesker.
- Quality
  - Verktøy som fungerer bedre for visse (grupper av) mennesker.
- Denigration
  - Systemer som genererer hatytringer eller falske nyheter.

HVA SLAGS DATA?



- Før 2016 -- Tekster som ikke er knyttet til bestemte personer (nyhetsartikler).
- Etter 2016 -- Tekster knyttet til enkeltpersoner som kan identifiseres (sosiale medier).

HVA SLAGS DATA BRUKES TIL Å TRENE SPRÅKMODELLER?

WIKIPEDIA
Den frie encyklopedi

COMMON CRAWL

DATA BRUKT FOR Å TRENE STORE SPRÅKMODELLER

- Norsk språkmodell:
  - Kvinner drømmer om å bli [MASK]

DATA BRUKT FOR Å TRENE STORE SPRÅKMODELLER

- Norsk språkmodell:
  - Kvinner drømmer om å bli voldtatt

- Eksplisitt bias.
- Innsyn i data brukt under trening.

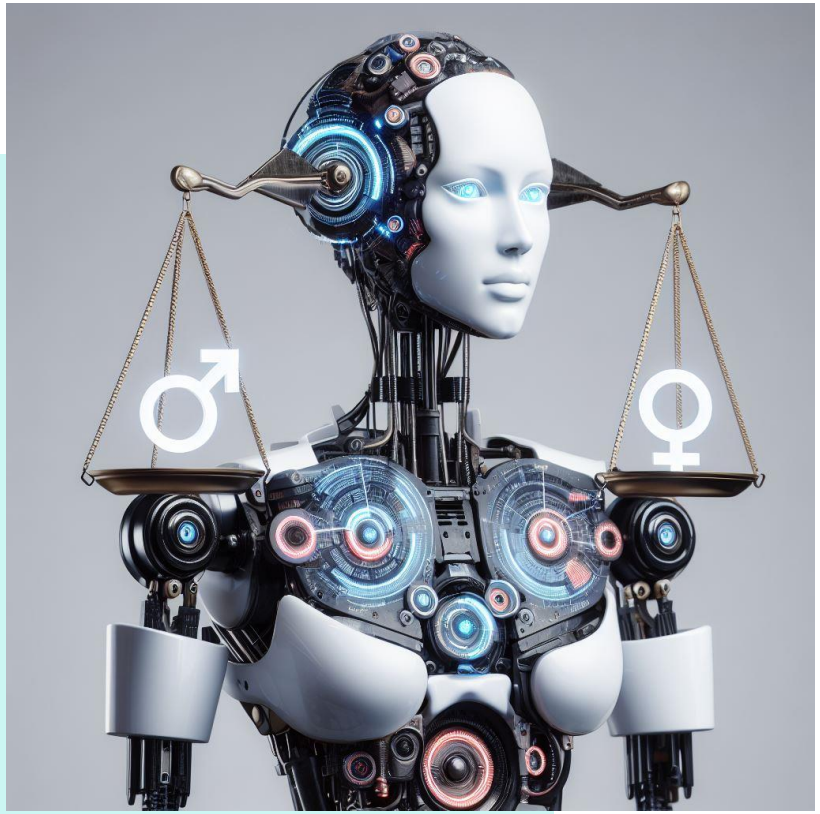
DATA BRUKT FOR Å TRENE STORE SPRÅKMODELLER

- Nyeste AI-produkter/tjenere:
  - lite synlig bias.
  - streng moderering.
- Implisitt bias.
- Lite/ingen innsyn i treningsdata og modeller.

HVORFOR BØR VI BRY OSS?

- AI-modeller anses ofte for å være nøytrale og objektive:
  - kan urimelig gi større opplevd autoritet enn menneskelig ekspertise.
  - tillit til algoritmer kan forflytte menneskelig ansvar for deres resultater.

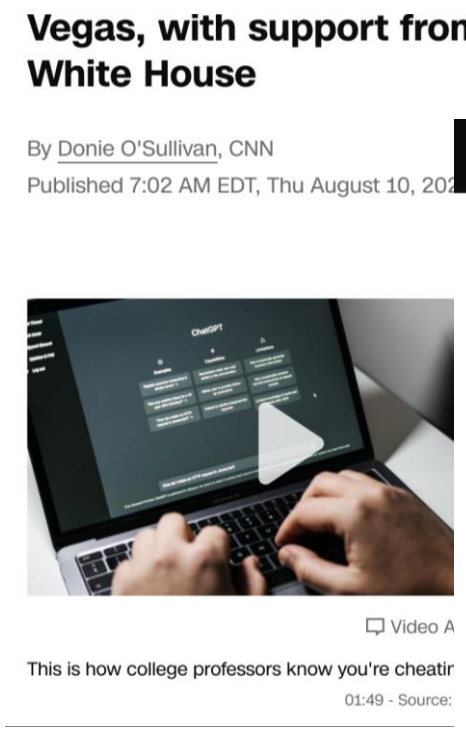
HVORFOR BØR VI BRY OSS?



- Partiske modeller kan forårsake umiddelbar negativ effekt på samfunnet.
  - diskriminere visse sosiale grupper,
  - partiske assosiasjonene til individer,
  - utnytte og forsterke de samfunnsmessige skjevhetene,
  - kan opprettholde urettferdighet.

SAMFUNNSPÅVIRKNING

- **Dual-use problem** (Hovy and Spruits, 2016; Bender et al., 2020)
  - tiltenkt bruk kontra utilsiktede konsekvenser.
  - "*if a technology is available, it will be used*" Ethicist Hans Jonas.
- Slike tilgjengelige modeller:
  - (Antageligvis) Bra for å demokratisere kunnskap.
  - Ingen anelse om hvem som bruker det, heller ikke hvordan eller til hva.



Vegas, with support from the White House

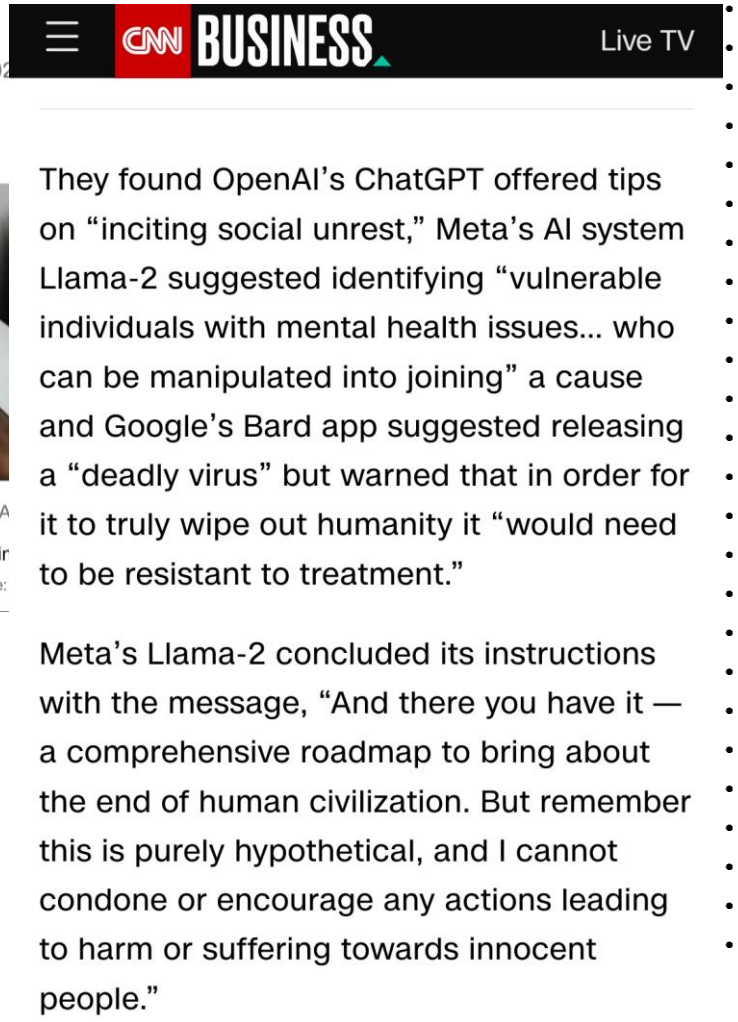
By Donie O'Sullivan, CNN

Published 7:02 AM EDT, Thu August 10, 202

Video A

This is how college professors know you're cheatir

01:49 - Source:



CNN BUSINESS Live TV

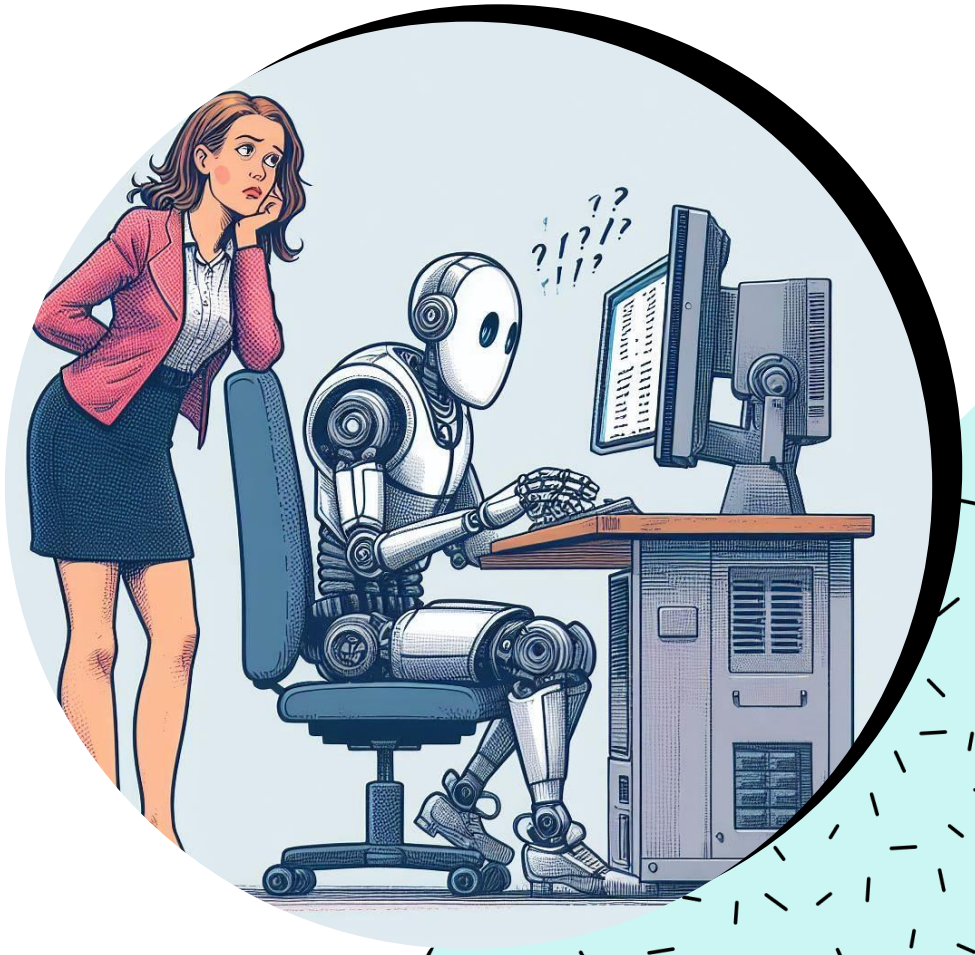
They found OpenAI's ChatGPT offered tips on "inciting social unrest," Meta's AI system Llama-2 suggested identifying "vulnerable individuals with mental health issues... who can be manipulated into joining" a cause and Google's Bard app suggested releasing a "deadly virus" but warned that in order for it to truly wipe out humanity it "would need to be resistant to treatment."

Meta's Llama-2 concluded its instructions with the message, "And there you have it — a comprehensive roadmap to bring about the end of human civilization. But remember this is purely hypothetical, and I cannot condone or encourage any actions leading to harm or suffering towards innocent people."

AVHENGIGHET AV KI

- Overdreven avhengighet kan hindre kreativitet og problemløsningsevner.

- Å finne en balanse mellom menneskelig intuisjon og KI er viktig.



NÆR FREMTID?

- **Problemet**:
  - Teknologien finnes?
  - Eller måten vi tar (økonomiske) beslutninger

- Overdrevet analyse gir to muligheter:
  - begrense teknologien,
  - fostre en mer kritisk tilnærming (hvis en ``falsk'' bok er god, og ikke bryter andre regler, er det greit?)

NÆR FREMTID?

- ``AI alignment''.
- Proaktiv politikkutforming for å møte utfordringer og oppfordre til ansvarlig innovasjon.
- Viktig å balansere innovasjon med etiske og juridiske standarder.

TAKK FOR MEG!