



LUND
UNIVERSITY

AI i sanningens tjänst

Anamaria Dutceac Segesten, Lunds universitet

Internetdagarna, Stockholm, 2019-11-25



Huvudfråga

- Kan algoritmer som producerar desinformation användas för att producera den motsatta effekt: skydda medborgarna från falsk information?



Vad är desinformation?

- Ett konceptuellt förvirrande fält: propaganda, desinformation, missinformation, "fake news"
- Begreppet har sitt ursprung i Sovjetunionen, där det sägs att det var Stalin själv som kom på det (Taylor, 2016)
- Disinformation is defined as deceptive, misleading information, disseminated to engender distrust and discord among the intended recipients (Faris et al., 2017).



Hur desinformation kan identifieras?

- **Kriterium:** sanning
- **Metod:** verifiera information mot fakta

- **Kriterium:** ursprung
- **Metod:** verifiera källans trovärdighet

Båda dessa metoder försvagas när AI-teknologin används

Teknologi och desinformation

- Teknologin har inget normativt innehåll, bara normativa användningsformer
- AI teknologi– refererar till maskiner eller agenter som kan observera deras omgivning, upptäcka och lära sig regler utifrån dessa observationer, och, med utgångspunkt i den nyvunna kunskapen, ta intelligenta åtgärder eller föreslå beslut.
- AI kan göra det omöjligt att urskilja, bara med hjälp av mänskliga sinnen, sanningsenlig från falsk information



Bild, ljud och text manipulativteknik

- BILD

Maskin inlärningsalgoritmer har blivit tränade på en stor datamängd för att kunna korrekt klassificera objekt i kategorier

- Bildigenkänning tävling: 1,2 millioner bilder med 1000 objektsklass – träningsdatabas
- 2017: mjukvaran har korrekt identifierat objekt i bilder **97,7%** av tiden
- Efter igenkänning – nyskapande
- Nvidia har nyligen visat fram extremt realistiska men fullt konstruerade bilder av människoansikten



Coarse styles copied

destination

source



Generative Adversarial Networks (GANs) genererade bilder, Nvidia

<https://www.thispersondoesnotexist.com/>

Deepfake video

- Deep Learning – en division av maskin inlärning som kan bättre hantera "noise" (oregelbundenheter) i data
- Bara på ett år man har gått från ganska rudimentära videor till extremt sannolika skapelser



AI-genererade journalistiska texter

- OpenAI (en NGO backad bl.a. av Elon Musk) har nyligen visat fram hur dess nyaste algoritm (GPT-2) kan lära sig komplettera fraser eller stycken.
- Mjukvaran tränades på en dataset som innehöll ca 10 millioner artiklar som hade minst tre 'upvotes' på den sociala plattformen Reddit. Den stora textsamlingen vägde in 40 GB, tillräckligt för att lagra cirka 35000 exemplar av Moby Dick.
- Se en test från The Guardian [här](#)



Orsaker till spridning av falsk information

- Strategiskt agerande från särskilda intressen
 - Automatiserade konton på sociala medier
 - 'Trollfabriker'
- Organisk spridning
 - Missuppfattning om informationens ursprung eller innehåll ("jag anade inte att detta är inte sant")
 - Viljan att sprida vidare information som stämmer överens med den egna världsbilden
 - **!** Man kan korrigera missuppfattningar men inte viljan av att med flit sprida falsk information!



Teknologi i kampen mot desinformation

- Problem att skilja åt sanningen från lögnerna
- Svår balansgång mellan reglering och censur
- **Åtgärder** som stater/EU kan introducera:
- Göra produktionen av deepfakes **olaglig** om man inte markera de resulterande videon som "AI modifierad/skapad"
- Införa ett **stämpel** som markerar originellt innehåll, eller som markerar modifierat eller AI-genererat innehåll
- Göra **plattformarna** ansvariga för att vara värd till deepfake innehåll
- Identifiera källan till deepfake-produkter
- Skapa algoritmer som kan identifiera falskt innehåll
 - Två forskningsprojekt som försöker just det: InVID & WeVerify

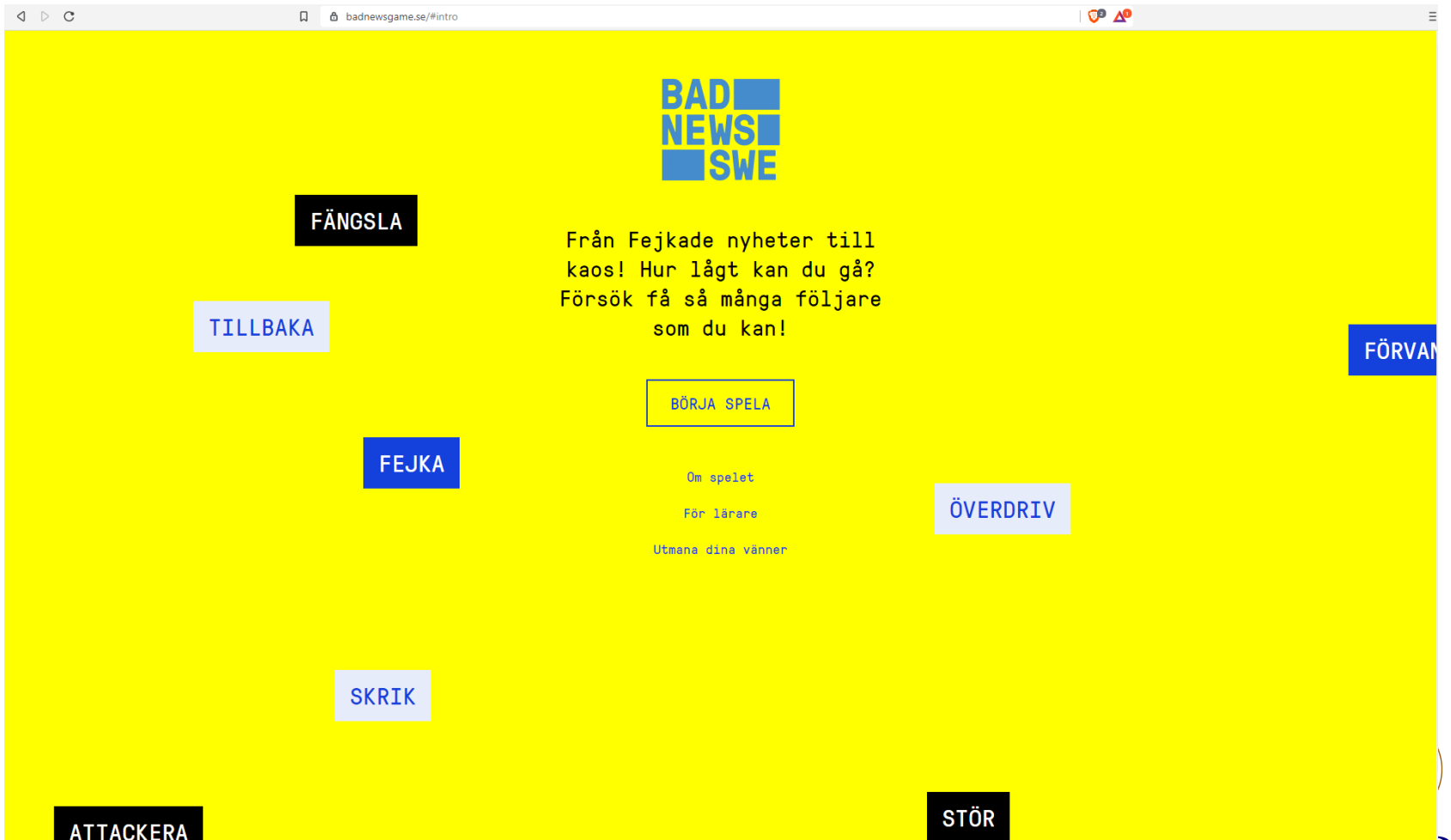


Människan ändå i centrum

- Bästa 'vaccinerna' mot desinformation
- Att informera - OK
- Att utbilda - OK
- Att låtsas fabricera falska nyheter – BÄST
 - Inokulationsteori: exponera mot en liten mängd virus (här, vilseledande argument) för att bygga ens motstånd



Bad News game



Frågor?
Skriv gärna

Anamaria.Dutceac_Segesten@eu.lu.se

 @anamariadutceac



LUND
UNIVERSITY

A woman with blonde hair and glasses, wearing a black and white floral dress and blue jeans, stands in the center of a room, gesturing with her hands while speaking. She is holding a small black object, possibly a microphone or a remote. The room has a wooden floor and a white wall with a grid pattern. In the background, a white electric guitar is visible on a stand. A large white screen is mounted on the wall behind her. The foreground shows the backs of several people sitting in rows of wooden chairs, facing the speaker. The overall lighting is warm and orange-toned.

Fika