Beskyttelse av kunstig intelligens: Hva er de nye sikkerhetsutfordringene?

Ketil Stølen

Oversikt

- Introduksjon
- Kunstig intelligens innen sikkerhet
- Kunstig intelligens som våpen i en tvekamp
- Beskyttelse av treningsprosessen
- Beskyttelse av KI-modellen
- Å oppdage om KI har blitt kompromittert
- Konklusjoner

Introduksjon

- Kunstig intelligens er et attraktivt angrepsmål
- Hva er de nye sikkerhetsutfordringene?
- Hvilke endringer er nødvendige for å sikre KI-teknologi mot nettbaserte angrep?

Sikkerhet vs. personvern

1) Personvern forutsetter en viss form sikkerhet

2) Men sikkerhet er også en trussel for personvernet

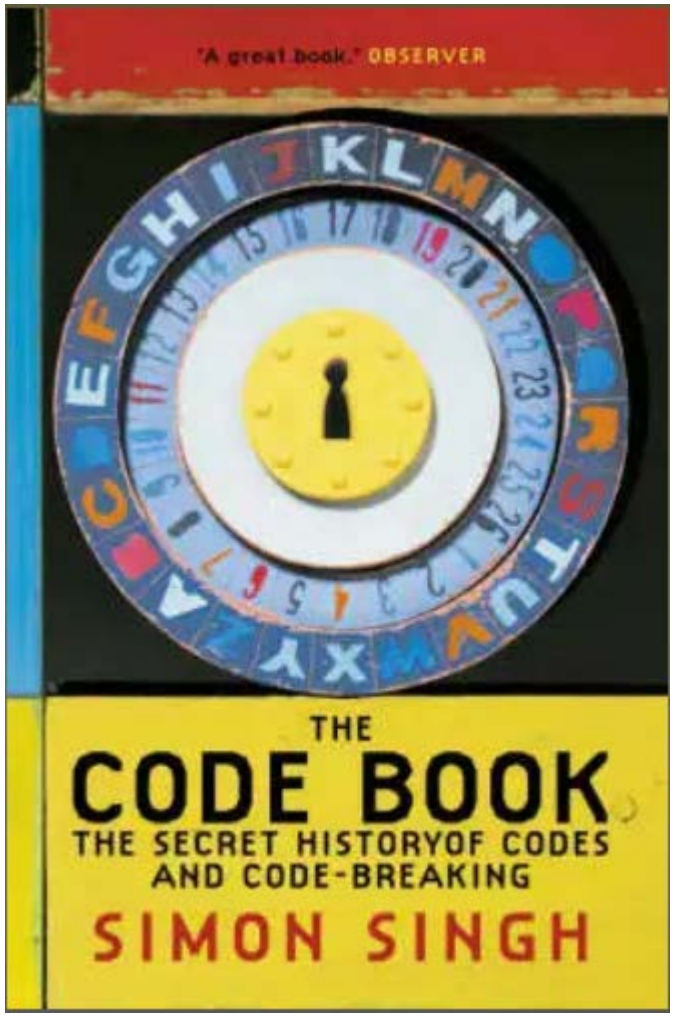
- *Dette foredraget vektlegger 1), men ikke glem 2)*
- *Med sikkerhet menes Cybersikkerhet: "beskyttelse mot angrep via Internett"*

Noen begreper

- Kunstig intelligens = "programvare som **tilsynelatende** er intelligent"
- Maskinlæring = "kunstig intelligens utviklet ved hjelp av programvare som læres opp basert på (store) mengder data"
- Modell = "programvaren som resulterer fra læringsfasen"
- Dyp læring = "maskinlæring basert på et nettverk av sammenkoblede noder organisert i lag (hvor dyp sier noe om antall lag)"

Kunstig intelligens for sikkerhet

- Gjenkjenner angrepsmønstre
- Problem:
  - Mangel på treningsdata av rett type
  - Må skreddersys for konkrete bruksområder hvilket vanskeliggjør gjenbruk



Sikkerhet har alltid vært en krig mellom angriper og forsvarer

Slik vil det forbli!

Kvantedatamaskiner endrer **ikke** på det:

Se Johannes Skaar & Co
https://www.nature.com/articles/nphoton.2010.214

Vi hadde et seminar på dette i 2019!

https://www.radarservices.com/study2025/

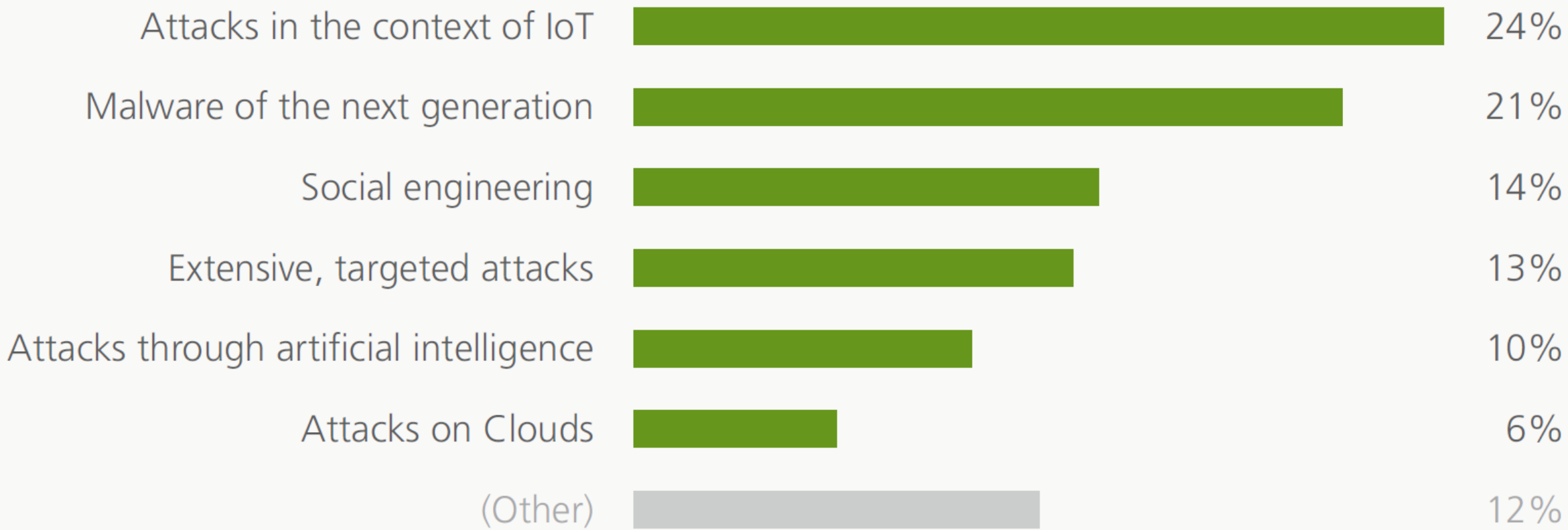
STUDY

Cyber attacks and IT security management in 2025

Expert survey concerning future trends and challenges in IT security

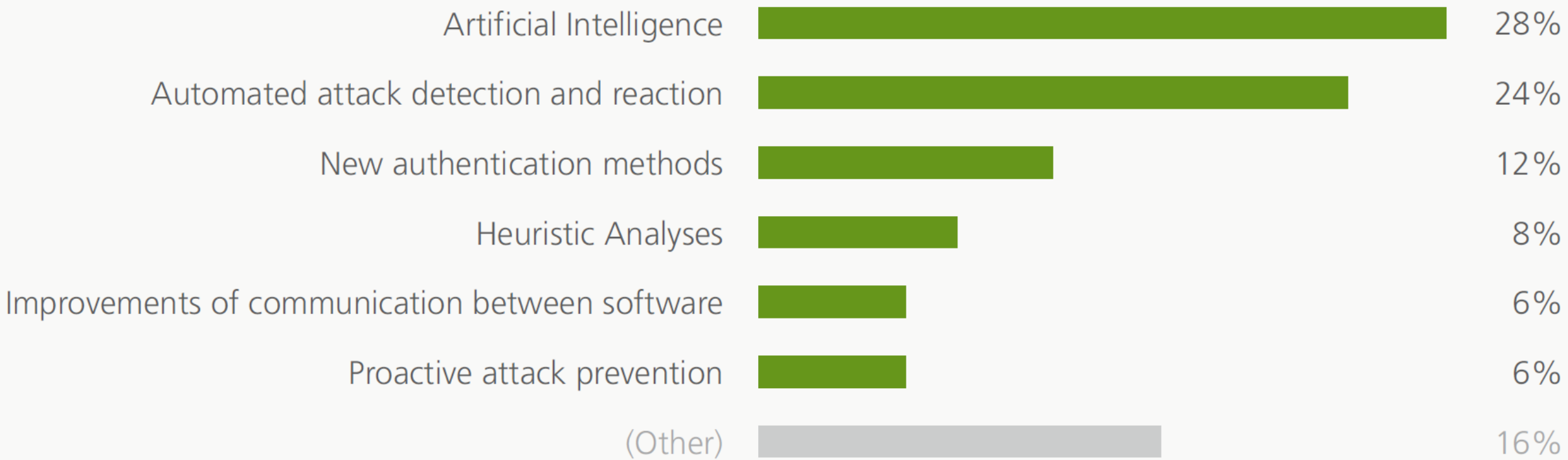
SINTEF

With which type of cyber attacks will we be confronted with in 2025?



SINTEF

What do IT security technologies have to offer in 2025?



Resultat

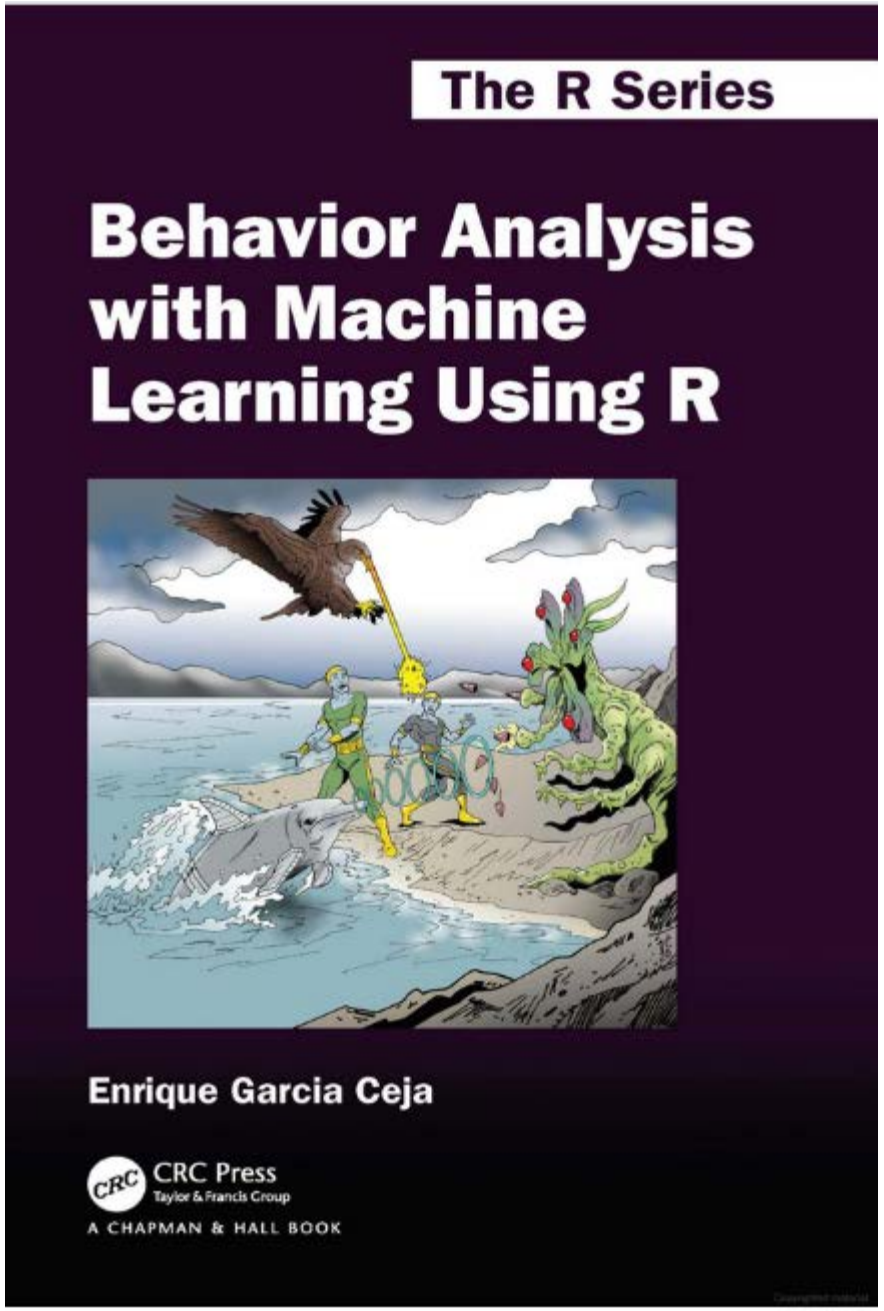
Kunstig intelligens som våpen
i en tvekamp

Enrique Garcia-Ceja (CRC Press).

Behavior Analysis with Machine Learning Using R introduces machine learning and deep learning concepts and algorithms applied to a diverse set of behavior analysis problems.

It focuses on the practical aspects of solving such problems based on data collected from sensors or stored in electronic records.

https://enriquegit.github.io/behavior-free/



Litt flere begreper

- Offline læring = "læring basert på allerede innsamlede data som leses inn en gang for alle"
- Online læring = "læring basert på data som samles inn dynamisk i takt med bruken av en tjeneste"

- Pipeline = "sekvens av programmer satt sammen på en slik måte at output fra et program er input til det neste"

Beskyttelse av treningsprosessen

- KI forutsetter vanligvis store mengder treningsdata
- Opptrening av programvaren kan foregå offline, online eller begge deler
- Treningsprosessen kan angripes
  - forgiftning
- Offline trening er klart sikrere enn online trening
  - online kan kreve mer KI

Beskyttelse av KI-modellen

- En KI-modell er et dataprogram som på en tilsynelatende intelligent måte besvarer spørsmål fra brukere
  - f.eks. sannsynlighet for at et bilde er av et menneskelig ansikt
- Brukerinteraksjonen er KI-modellens akilleshæl
  - muliggjør f.eks. kloning
- Forsvar
  - returnere mindre informasjon, trene videre periodevis

Å oppdage om KI har blitt kompromittert

- Vanskeliggjøres ved at programvaren mangler klare grenser for normal oppførsel
- Mulig løsning: mangfold (diversity på engelsk)

Konklusjoner

- KI-teknologi krever ekstra tiltak
  - må dekke hele prosessen fra design til drift
- Programvaren må avsløre minst mulig om seg selv
  - f. eks. type KI, konfigurering, resultatenes detaljnivå
- Treningsdataene må komme fra en pålitelig kilde
- Unngå at modellen kopieres
  - f.eks. begrense antall spørringer per bruker
- Tilleggsanalyse av inputdata
  - f.eks. sjekke om et bilde har blitt manipulert for å lure modellen