Detektion av påverkansoperationer i sociala medier – fake news och syntetiserade profilbilder

Den 17 januari höll forskningsledaren Fredrik Johansson, FOI, ett otroligt spännande seminarium med titeln Detektion av påverkansoperationer i sociala medier – fake news och syntetiserade profilbilder. Med hjälp av sociala medier och plattformar som Facebook och Twitter ökar möjligheterna för spridning av nyheter och information och det uppstår ökande problem med att illasinnade aktörer använder detta för egna ändamål.

Påverkansoperationer i sociala medier kan vara handlingar av statsoperatörer med syfte t.ex. att manipulera en opinion, skapa splittring mellan olika grupper eller skada tilltron till media eller statliga institutioner. Det är svårt att reglera dessa operationer och dessutom är de ofta billiga att genomföra.

FOI forskar kring verktyg och tekniker för detektion av botar, kapade konton, manipulerade bilder/videor och koordinerade konton. Klassificerare baseras på maskininlärning (ML) kan tränas till att särskilja mänskliga och automatiserade konton som använder sig av olika särdrag som grad av repetition och tidsbaserade mönster.

Snabba framsteg inom AI möjliggör automatiskt skapande och manipulering av bilder, ljud, text och videor. Fabricerad och riktig media är ofta svårt att särskilja för människor.

Modeller som StyleGAN2 kan skapa mycket realistiska bilder av icke-existerande människor. Dessa används sedan till påverkansoperationer och underrättelseinhämtning.

FOI:s arbete indikerar att data från kända generativa modeller kan detekteras med hög träffsäkerhet, men att detektionsmetoder för detta ändamål har begränsad robusthet mot komprimering, brus och minimala perturbationer av indata som ofta inte kan uppfattas av den mänskliga hjärnan (s.k. adversarial attacks).

Det kommer bli allt svårare att avgöra om bilder, ljud, artiklar mm avspeglar verkliga händelser och det finns ett ökat behov av källkritik och datorstöd.

Seminariet hölls via Zoom och nästan 50 st intresserade deltog och ställde frågor.