Generativ AI 2025

Hur möter vi hotet från AI-genererad media?

Generativ AI, dvs. AI-tekniker som låter användare skapa realistiska bilder, ljud, videos och texter, har på kort tid blivit populärt för att generera allt mellan virtuella världar och kvartalsrapporter. Men flera stora aktörer pekar på att generativ AI är ett växande hot som kan ha förödande effekt på samhället i stort. Detta eftersom att generativ AI samtidigt har underlättat skapandet och spridandet av desinformation. Under 2025 har verksamheten inom generativ AI på FOI varit fokuserad på att producera tre olika artiklar som på olika sätt undersöker hur riskerna med generativ AI kan minimeras.

Olika tekniker

För att kunna möta hotet från generativ AI pratar man ofta om tre olika tekniker: detektion – metoder för att upptäcka av AI genererad media; vattenmärkning – metoder för att på förhand märka upp av AI genererad media; samt digitala signaturer – metoder för att på förhand märka upp autentisk media, för att särskilja dessa från av AI genererad media. Under året har arbete utförts för att utforska alla dessa tre aspekter. Arbetet med digitala signaturer har främst skett genom dialog med Försvarsmakten om möjligheterna att använda den digitala signaturstandarden C2PA i svensk militär kontext. De två andra aspekterna har i stället utforskats genom producerandet av tre vetenskapliga artiklar som beskrivs i korthet här nedan.

Vattenmärkning

För att vattenmärka AI-genererad text behöver man se till att en vald genererande AI-modell väljer de ord som genereras på ett sätt som är strukturerat, men som inte kan uppfattas av en människa. Detta betyder oftast att modellen byter ut oviktiga ord mot synonyma fördefinierade ord på ett till synes slumpmässigt sätt. Parafrasering är en typ av attack som används för att förstöra sådan vattenmärkning. Attacken går till så att en text, med en redan pålagd vattenmärkning, matas till en ny AI. Den nya AI:n ombes sedan att skriva om texten iterativt flera gånger tills den tidigare vattenmärkningen kan antas vara borta. I den första artikeln har möjligheterna hos denna typ av attack studerats.

En kontext där vattenmärkning av text är extra svårt är programkod. Programkod är text som är skriven som en uppsättning instruktioner som talar om för en dator hur den ska utföra en viss uppgift. För att programkod ska kunna fungera behöver texten följa en rigid struktur. Detta innebär att det inte finns många ord som kan bytas ut till synonyma ord för att på så sätt skapa en vattenmärkning. Trots detta har två artiklar de senaste två åren förslagit lösningar på denna problematik. I den andra artikeln som producerats inom verksamheten på FOI har man arbetat med att visa att dessa två lösningar är otillräckliga och kan kringgås med enkla medel.

Detektion

Den sista artikeln som det jobbats på inom verksamheten på FOI har tittat på en metod, vid namn AEROBLADE, som påstår sig kunna detektera AI-genererade bilder genom att använda sig av så kallade auto-encoders. AI-modeller som genererar bilder använder sig nämligen ofta av auto-encoders för att kunna minska på datorkraften som krävs för generering. På FOI har det undersökts om auto-encoders med enkla medel kan förändras för att kringgå detektion av AEROBLADE (figur 1). Man har även undersökt huruvida AEROBLADE kan förbättras genom att på ett automatiserat sätt klona auto-encoders som denna inte sett tidigare, bara genom att titta på bilderna från den genererande AI-modellen.

Figur 1. En AI-genererad bild som inte detekteras av metoden AEROBLADE.

Forskningen inom AI-programmet finansieras av anslag 1:9 Totalförsvarets forskningsinstitut, anslagspost *6 Bevaka och hantera nya tekniker.* Verksamheten syftar till att tidigt identifiera och initiera framväxande tekniker och utgör ett komplement till Försvarsmaktens forskning och teknikutveckling (FoT).

FOI Memo 9223
December 2025
Kontaktperson: Sebastian Öberg, sebastian.oberg@foi.se

FOI
Totalförsvarets forskningsinstitut