Generativ AI 2024

Hur möter vi hotet från deepfakes?

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. När generativ AI används för att skapa realistiska bilder, ljud eller videos brukar dessa kreationer benämnas ”deepfakes”. Verksamheten inom Generativ AI på FOI syftar till att identifiera vilka risker som finns med generativ AI och deepfakes samt hur det svenska samhället kan skyddas från dessa risker. Verksamheten ingår i FOI:s AI-program som finansieras av anslagsposten Bevaka och hantera nya tekniker (ap.6).

Den bedrivna forskningen är viktig för svenskt totalförsvar, då det i framtiden kommer att finnas en större mängd desinformation i samhället. Förhoppningen är att när en framtida deepfake, likt den på Volodymyr Zelenskyj i samband med Rysslands invasion av Ukraina (se figur 1), dyker upp i ett svenskt sammanhang så kan kunskap från den bedrivna forskningen användas för att kunna bemöta denna.

Figur 1. En deepfake av Volodymyr Zelenskyj som lades upp på ukrainska TV-nätverket Ukrayina 24, 16 mars 2022.

För att kunna möta hotet från generativ AI har tre olika tekniker identifierats: *detektion*, *vattenmärkning* och *digitala signaturer*.

Detektion av deepfakes eller genererad text har visat sig vara svårt. Genom att göra små förändringar i generativa modeller så kan en illasinnad aktör i de flesta fall kringå de allra flesta försök till detektion. För att detektion som metod ska kunna fungera behöver man därför hitta en metod som kan generalisera till att detektera modeller metoden inte sett ännu. I verksamheten studeras två sådana metoder som påstår sig kunna generalisera till upptäckt av ännu oobserverade modeller: DIRE och AEROBLADE. Planen är att dessa två metoders anspråk på generaliserbarhet ska utvärderas.

Vattenmärkning inkorporerar, för människan osynliga, ”vattenmärken” i genererad media som signalerar att innehållet är skapat av AI. I vattenmärket finns ofta inbäddat någon form av hemlighet. En person som känner till hemligheten kan använda denna för att kontrollera om media är vattenmärkt eller inte och på så sätt få reda på om media eventuellt är genererad. Man sammanfattar ofta denna procedur med att man ”märker upp de *dåliga* sakerna”.

Digitala signaturer är en teknik från kryptografin som i dag används för att signera digitala meddelanden med hjälp av kryptografiska nycklar. För att motverka desinformation är tanken att man använder digitala signaturer för att signera media som *inte* är genererad av AI. Om man litar på personen vars nyckel har signerat någon typ av media så kan man då lita på att det som är signerat *inte* är genererat av AI. C2PA är den ledande standarden för denna användning av digitala signaturer. Man sammanfattar ofta denna procedur med att man ”märker upp de *bra* sakerna”.

Under 2024 påbörjades ett arbete med att ta fram en vetenskaplig artikel om vattenmärkning av AI-genererad kod.

Kontaktinformation
Sebastian Öberg, sebastian.oberg@foi.se, 08-555 037 87

FOI Memo 8854

Innehållet är granskat och omfattar ingen information som är underställd exportkontrollagstiftningen

December 2024