CENTER FOR CYBERSIKKERHED

Trusselsvurdering:

Hackere misbruger generativ AI

1. udgave marts 2024.

Indhold

CENTER FOR CYBERSIKKERHED

Kastellet 30
2100 København Ø
Telefon: + 45 3332 5580
E-mail: cfcs@cfcs.dk

1. udgave marts 2024.

Hackere misbruger generativ AI

Formålet med denne vurdering er at belyse, hvordan hackere misbruger generativ AI. Målgruppen for trusselsvurderingen er ledelsen og it-sikkerhedsansvarlige i danske virksomheder og myndigheder. Desuden kan andre, der ønsker viden om trusler forbundet med generativ AI, også med fordel læse den.

Hovedvurdering

- Det er meget sandsynligt, at hackere misbruger generativ AI.
- Hackeres misbrug af generativ AI ændrer ikke den overordnede cybertrussel trods teknologiens udbredelse og tilgængelighed.
- CFCS vurderer, at de to væsentligste måder, hackere misbruger generativ AI i cyberangreb, er til at producere tekst til f.eks. phishing-mails og delkomponenter af ondsindet kode. Det kan optimere dele af hackernes angreb.
- Det er ikke entydigt, hvor udbredt brugen af generativ AI er blandt hackere. Særligt cyberkriminelle udviser dog interesse for misbrug af generativ AI.
- CFCS vurderer, at AI-teknologier også kan udgøre en trussel i andre sammenhænge end at assistere i produktionen af tekst og computerkode til brug i cyberangreb.

Misbrug af generativ AI i cyberangreb
Det er meget sandsynligt, at hackere misbruger generativ AI. Hackeres misbrug af generativ AI ændrer ikke den overordnede cybertrussel trods teknologiens udbredelse og tilgængelighed.

Det skyldes, at både stater og dele af det cyberkriminelle miljø allerede råder over betydelige ressourcer og har evnerne til effektivt at udføre de processer, som generativ AI kan understøtte. Truslen er således allerede **MEGET HØJ** fra henholdsvis cyberspionage og cyberkriminalitet. Generativ AI er dermed ét værktøj blandt flere, som er med til at drive disse trusler.

Det er vigtigt at bemærke, at generativ AI ikke er en type teknologi, der på egen hånd kan planlægge og udføre angreb. I stedet fungerer teknologien typisk sådan, at den genererer et output på baggrund af et skriftligt eller mundtligt input fra en bruger. Outputtet kan hackerne bruge i et cyberangreb.

AI
AI er forkortelsen for *artificial intelligence* eller *kunstig intelligens* på dansk. Begrebet dækker over teknologier, som kan løse opgaver, der imiterer menneskets intelligens.

Et eksempel på en type generativ AI er OpenAI's ChatGPT, der bl.a. kan producere grammatisk korrekte tekster på forskellige sprog og computerkode. Andre typer generativ AI kan producere billeder, lyd, videoer og lignende.

Generative AI-teknologier er trænet på en stor mængde data, som danner baggrund for det output, de genererer. Selvom træningsdatasættet er enormt, er der begrænsninger ved det, fordi teknologierne ikke har adgang til al information. Det kan vise sig i outputtet, hvor der kan forekomme faktuelle fejl. Sker det, taler man om, at teknologien hallucinerer.

De fleste generative AI-teknologier er opsat med filtre, der skal forhindre dem i at generere ondsindet indhold. Det er dog ofte muligt med relativt simple kommandoer at omgå disse begrænsninger. Derved får hackerne alligevel et output, som de kan misbruge i cyberangreb.

Generativ AI optimerer phishing og ondsindet kode
CFCS vurderer, at de to væsentligste måder, hackere kan misbruge generativ AI i cyberangreb, er til at producere tekst til f.eks. phishing-mails og delkomponenter af ondsindet kode. Det kan optimere dele af hackernes angreb.

Det er særligt AI i form af generative sprogmodeller, som hackere allerede nu kan misbruge til at understøtte cyberangreb. Modellerne kan typisk både producere almindelig læsbar tekst på forskellige talte sprog og computerkode på forskellige programmeringssprog.

Begge dele kan optimere dele af hackeres arbejde, fordi de skal bruge mindre tid på selv at udvikle kode eller phishing-indhold. Den sparede tid kan hackerne bl.a. bruge på at udføre flere angreb eller på at optimere andre dele af angrebskæden.

Phishing-mails bliver mere overbevisende
Hackere kan som nævnt udnytte generativ AI til phishing. Det skyldes, at teknologien kan generere grammatisk korrekte tekster, også på sprog, hackerne ikke selv behersker. Ved at give sprogmodellerne input om bestemte virksomheder, myndigheder eller personer kan hackerne bruge generativ AI til at gøre phishing-teksterne særligt relevante og overbevisende. Skræddersyede mails gør det sværere at se, om mailen er legitim eller stammer fra hackere.

En af de angrebstyper, hvor phishing spiller en central rolle, er i Business E-mail Compromise (BEC)-svindel. Ved BEC-svindel forsøger kriminelle at franarre virksomheder og myndigheder penge gennem falske anmodninger om pengeoverførsler. Dermed har AI potentiale til at øge både udbredelsen og succesraten af BEC-svindel.

Phishing og spear phishing
I phishing-angreb forsøger en aktør at narre en modtager til at videregive beskyttelsesværdige oplysninger eller give uretmæssig adgang til bl.a. it-systemer.

Spear phishing minder om phishing, men adskiller sig ved, at ofrene ikke er tilfældige, men i stedet nøje udvalgte. Ligesom i phishing foregår spear phishing typisk via mail. Angreb kan dog også finde sted via sms eller telefonopkald.

Indholdet i spear phishing er typisk formuleret, så det virker særligt relevant, overbevisende og troværdigt for modtageren.

Se CFCS' vejledning "Beskyt din organisation mod phishing-angreb" for vejledning om, hvordan man beskytter sin organisation mod phishing.

Udvikling af computerkode bliver mindre kompliceret
Idet generative sprogmodeller kan producere computerkode, kan hackerne udnytte modellerne til at producere kode, der kan indgå i cyberangreb. Computerkoden kan bl.a. udgøre delkomponenter af malware eller exploits.

Generativ AI kan dermed understøtte og automatisere noget af arbejdet med at udvikle kode til brug i cyberangreb. CFCS vurderer, at det dog stadig kræver et vist niveau af tekniske færdigheder at kunne udnytte det til at lave succesfulde cyberangreb.

Andre misbrugsscenarier
Udover at producere phishing-mails og ondsindet kode kan hackere også misbruge generativ AI på andre måder. Eksempelvis kan generativ AI spare hackerne tid ift. udvælgelse og rekognoscering af mulige ofre. Det skyldes, at generative sprogmodeller i deres output kan give hackerne information om ofrene. Outputtet er baseret på den information, de har tilgængelig i deres træningsdata.

Tilgængeligheden og udbredelsen af generativ AI kan desuden udpege viden om angrebsmetoder, hackerfora og lignende. Det er bl.a. relevant i planlægningen af hackerangreb for personer med en begyndende interesse for hacking, men som mangler f.eks. tekniske færdigheder. På baggrund af dette vil generativ AI derfor potentielt kunne øge antallet af personer, der indgår i cyberkriminalitet.

Med både udbredelsen og tilgængeligheden af AI kan der opstå bekymring for, at teknologien medfører mere sofistikerede og målrettede angreb, og at cyberangreb desuden vil blive øget i volumen. På nuværende tidspunkt er det dog sandsynligt, at det primært er i forbindelse med generering af phishing-indhold og dele af ondsindet kode, at cyberangreb kan blive optimeret.

Misbrug af generativ AI til deepfakes
CFCS vurderer, at hackere kan udnytte deepfakes i bl.a. cyberkriminalitet. Deepfakes er lyd, billeder eller videoer genereret af AI.

Nogle cyberkriminelle misbruger AI til at generere deepfake-stemmer til scam-opkald. I scam-opkald ringer en kriminel til et offer og forsøger at efterligne en stemme, der typisk tilhører en, offeret kender.

For eksempel kan et scam-opkald eller en deepfake-video efterligne højtstående medarbejdere i en organisation. Cyberkriminelle kan forsøge at bruge disse opkald eller videoer til at overbevise medarbejdere om at handle i hackernes interesse.

Deepfakes er i sig selv ikke cyberangreb. Det er først, når den AI-genererede deepfake bliver brugt som led i en handling, hvor formålet f.eks. er at få nogen til at overføre penge, at der er tale om et cyberangreb.

Omfanget af misbrug kan ikke entydigt fastslås
Det er ikke entydigt, hvor udbredt misbrugen af generativ AI er blandt hackere. Det skyldes, at man ikke kan se på eksempelvis teksten i en phishing-mail eller et stykke malware, om det er produceret af et menneske eller generativ AI.

CFCS ser dog, at særligt kriminelle hackere udviser interesse for misbrug af generativ AI. Interessen kan bl.a. ses på onlinefora. Her debatterer hackerne, hvordan de kan misbruge legitime sprogmodeller som OpenAI's ChatGPT til at understøtte cyberkriminalitet.

Nogle cyberkriminelle udnytter også AI til at bygge egne teknologier, der kan understøtte deres kriminalitet. De udvikler bl.a. sprogmodeller uden de filtre, som normalt er indbygget til at bremse misbrug i legitime sprogmodeller.

Et eksempel på en sådan model er FraudGPT. Modellen minder om ChatGPT, men har ikke de samme begrænsninger til at stoppe misbrug. FraudGPT er ifølge annoncøren særlig effektiv til at producere phishing-indhold og ondsindet kode som f.eks. malware.

CFCS vurderer, at andre aktører end cyberkriminelle også kan udnytte AI i cyberangreb. Det er bl.a. cyberaktivister og stater. Cyberaktivister vil f.eks. kunne bruge teknologien til at generere indhold til cyberangreb, hvori der indgår misinformation. Visse stater har desuden kapaciteter til at udvikle egne AI-teknologier, som de kan misbruge til cyberangreb.

AI kan også udgøre andre trusler
CFCS vurderer, at AI-teknologier også kan udgøre en trussel i andre sammenhænge end at assistere i produktionen af tekst og computerkode, som hackere kan misbruge i cyberangreb.

Stater kan f.eks. inddrage AI til at opnå andre mål. De og andre aktører, f.eks. cyberaktivister, kan desuden udnytte generative sprogmodeller til at producere indhold til at opnå politiske mål. Det kan de f.eks. forsøge at gøre gennem påvirkningskampagner. Her kan de få generative sprogmodeller til at producere påvirkningsindhold, som de efterfølgende kan distribuere.

Sådan styrker du cybersikkerheden i lyset af AI

AI-teknologier kan dog også bruges til at forbedre cybersikkerheden for virksomheder og myndigheder. Som med al anden teknologi kan parter på begge sider således anvende teknologien til at opnå hver sit mål.

Organisationer bør løbende kontrollere og vurdere, om deres foranstaltninger mod phishing og malware er tilpasset trusselsbilledet for AI og andre teknologiske udviklinger. Det inkluderer uddannelse af ansatte, bl.a. for at øge informationsniveauet om cyberangreb, hvori der indgår AI.

CFCS udarbejder løbende artikler og vejledninger, der har til formål at forbedre cybersikkerheden. Se f.eks. temaartiklen ”AI og cybersikkerhed” og vejledningerne ”Beskyt din organisation mod phishing-angreb” og ”Cyberforsvar der virker”.

Trusselsniveauer

Forsvarets Efterretningstjeneste bruger følgende trusselsniveauer.

INGEN	Der er ingen tegn på en trussel. Der er ingen aktør, der både har kapacitet til og intention om angreb/skadelig aktivitet.
LAV	En eller flere aktører har kapacitet til og intention om angreb/skadelig aktivitet. Men enten er kapaciteten eller intentionen eller begge dele begrænset.
MIDDEL	En eller flere aktører har kapacitet til og intention om angreb/skadelig aktivitet. Men der er ikke indikationer på specifik planlægning af angreb/skadelig aktivitet.
HØJ	En eller flere aktører har kapacitet til og foretager specifik planlægning af angreb/skadelig aktivitet, eller har allerede gennemført eller forsøgt angreb/skadelig aktivitet.
MEGET HØJ	Der er enten oplysninger om, at en eller flere aktører iværksætter angreb/skadelig aktivitet, herunder oplysninger om tid og mål, *eller* en eller flere aktører iværksætter kontinuerligt angreb/skadelig aktivitet.

Et givent trusselsniveau er udtryk for FE's vurdering af aktørers intention, kapacitet og aktivitet på baggrund af de tilgængelige oplysninger.