POLITIET
KRIPOS

Generativ kunstig intelligens og cyberkriminalitet

Temarapport

Innhold

1. HOVEDPUNKTER

- Generativ kunstig intelligens (GKI) er i høy grad en del av den pågående teknologiske utviklingen. For cyberkriminalitet er teknologisk utvikling en driver, som resulterer i nye verktøy cyberkriminelle kan bruke. GKI øker kriminelles kapabiliteter og kan understøtte alle typer cyberkriminalitet.
- Den teknologiske utviklingen går raskere enn samfunnet klarer å hindre uønskede utfall og trusler.
- Kriminalitet begått med GKI kan ramme ofre på individ-, organisasjons- og statlig nivå, og den har potensial til å ramme ofre i stor skala og høyt tempo, med stor grad av skreddersøm.
- GKI benyttes i dag til å generere skadevarekode, til å generere bilder som seksualiserer barn, herunder syntetisk overgrepsmateriale, og til å produsere deepfakes.
- GKI utgjør en styrkemultiplikator for en lang rekke operasjoner innen cyberkriminalitet.

Rapportens overordnede vurderinger tar utgangspunkt i hvordan trusselen fra bruk av GKI til illegale formål vil endre eller føre til ny type kriminalitet de neste fem årene.

- Kripos forventer en økning i mengde og en kvalitativ endring i cyberkriminalitet på grunn av bruk av GKI til illegale formål.
- GKI representerer en generell kostnads- og arbeidsreduksjon som *meget sannsynlig* vil føre til en økning i antall cyberkriminelle og deres kapabiliteter.
- Kompetansekrav for å begå ulike cyberkriminelle aktiviteter vil *meget sannsynlig* reduseres, og kompetanseprofil for å utvikle nye tekniske løsninger vil *sannsynlig* endres.
- Det er *lite sannsynlig* at GKI vil utfordre norsk lovverk når det gjelder skyld, uaktsomhet, medvirkning og forsøk.

2. INNLEDNING

Hensikten med rapporten er å etablere en felles grunnleggende forståelse av generativ kunstig intelligens (heretter forkortet GKI), begrenset til teknologiens relevans for cyberkriminalitet. Rapporten inneholder vurderinger av trusselen som cyberkriminelles bruk av denne teknologien utgjør mot enkeltpersoner, virksomheter og verdier i Norge. Primærmottaker er politidirektoratet. Grunnet tematikkens allmenne interesse og bredt nedslagsfelt for hva gjelder trusselen, er det utarbeidet en ugradert versjon for deling med relevante mottagere i privat og offentlig næringsliv. Rapporten kan understøtte strategiske beslutninger om politiets svar på trusselen som beskrives.

Behovet for rapporten er framtvunget av GKIs økende popularitet og bruk, og den potensielle kapabilitetsøkningen[1] dette representerer for cyberkriminelle. Etterretningshullet ble identifisert i arbeidet med rapporten *Cyberkriminalitet 2023*, og løftet av fagmiljøet innen cyberetterretning på Kripos.

Vurderingene gjelder for de neste fem årene, med mindre annet er presisert.

3. AVGRENSNINGER

Kripos' mandat er organisert og annen alvorlig kriminalitet, og det er også avgrensningen for denne rapporten. Følgelig vil ikke all mulig bruk av GKI i utførelsen av cyberkriminalitet bli vurdert.

Cyberkriminalitet der GKI er benyttet vil kunne falle inn under både politiets og PSTs mandat. Denne rapporten vurderer ikke statlige aktørers bruk av GKI.

Rapporten utgjør ikke et dypdykk i de *tekniske* mulighetene og begrensningene i GKI, men analyserer den *strategiske betydningen* av taktiske og operasjonelle muligheter basert på vår forståelse av teknologien i dag, relatert til cyberkriminalitet, de neste fem årene.

Mottiltak som politiets, næringslivets (inkludert cybersikkerhetsbransjens) og Forsvarets bruk av GKI vurderes ikke.[2] Cyberkriminelles netto mulighetsrom beskrives derfor ikke.

GKI betegner en av flere typer kunstig intelligens (heretter forkortet KI). Andre typer KI vurderes ikke i denne rapporten. Det er uansett et premiss for vurderingene i rapporten at cyberkriminelle benytter alle tilgjengelige verktøy i alle mulige kombinasjoner som gir effekt mot målet, uansett om det er GKI, annen KI, eller ikke-KI-relaterte verktøy.

3.1. Begreper

Denne rapporten benytter grunnbegreper utarbeidet i den strategiske temarapporten *Cyberkriminalitet 2023*, hvor cyberkriminalitet deles inn i kriminalitet mot datasystemer og kriminalitet støttet av datasystemer.[3]

Kunstig intelligens (KI) er et begrep som omfatter programvare som i dag finnes i søkemotorer, bildefiltre og videobehandling, oversettertjenester, språkgjenkjenning i assistenter i mobiltelefoner og musikkanlegg, bak anbefalinger i strømmetjenester; som del av programvaren i selvkjørende biler, diagnostiske systemer innen medisin og mye mer. Selv om begrepet fortsatt kan oppleves som vanskelig tilgjengelig, refererer det altså til mange ting de fleste allerede har et forhold til i hverdagen. Teknologien har hatt

[1] Med *kapabilitet* menes i denne rapporten summen av en gjerningspersons ferdigheter og tilgjengelige ressurser. Et eksempel kan være en gjerningspersons datakyndighet, sammen med tilgang til programvare og personsensitiv informasjon.

[2] Jf. Etterretningsdoktrine for politiet v1.2, s. 12.

[3] Se rapporten *Cyberkriminalitet 2023 - Politiets årlige temarapport om kriminalitet mot datasystemer og kriminalitet støttet av datasystemer*, s. 10-11 for definisjoner av cyberkriminalitet, datasystemer, det digitale rom, internett og skadevare.

flere perioder med høy og lav utvikling, men kan overordnet nå vanskelig kalles ny. Fagfeltet KI hadde sin oppstart i 1956. KI representerer i korte trekk nye teknologiske kapabiliteter, automatisering, effektivisering og rask iterasjon[4] i mange typer prosesser.

Generativ kunstig intelligens (GKI) refererer til en undergruppe KI-systemer som kan generere tekst[5], kode[6], bilder[7], video[8], lyd[9] eller andre medie- eller datatyper[10], basert på inndata[11]. Slike inndata kan bestå av tekst som beskriver et ønsket resultat, eller andre medie- eller datatyper. Resultatet av det programmene genererer kalles i denne rapporten *utdata*, uavhengig av medietype.

Konversasjonell KI er en type GKI laget for å kunne svare på spørsmål fra brukere. De mest kjente og kraftigste er såkalte store språkmodeller[12], trent på det som for mennesker er enorme mengder tekst. Slik trening går ut på at modellene regner ut statistiske sammenhenger i datagrunnlaget, som setter dem i stand til å generere relevante og kontekstuelle svar.

Maskinlæring (ML) refererer til en rekke ulike KI-teknikker, der reglene utledes fra dataene som systemet trenes på, heller enn at de er forhåndsdefinert av mennesker. ML er en prosess bak mange av dagens KI-programmer. ML består av algoritmer som benyttes til å utføre statistiske utregninger på et datasett for å nå et forhåndsdefinert mål. Prosessen kan inkludere automatisert evaluering underveis, altså prøving og feiling. Utregningen kan organiseres ved å sette store mengder noder[13] i sammenheng, hvor sammenhengen beskrives som et nevralt nettverk. Dyp læring er en underkategori av slike nevrale nettverk, hvor nettverket består av flere lag med noder.

Deepfake refererer til manipulert og syntetisk generert materiale av ekte mennesker som blir presentert som ekte innhold. Deepfakes kan for eksempel vise en person som sier eller gjør noe den aldri har sagt eller gjort. Deepfakes produseres ved bruk av kunstig intelligens, inkludert maskinlæring og dyp læring. Deepfake kan også kombineres med andre verktøy og teknikker, eksempelvis spoofing[14], som vil bidra til å forsterke graden av opplevd autentisitet hos mottageren.

Begrepet "*ondsinnet bruk*"[15] har ikke en universell, standardisert betydning, men er ofte brukt i forbindelse med cyberkriminalitet. I denne rapporten benytter vi heller frasen "bruk av GKI for illegale formål", med utgangspunkt i norsk lov.

4. BAKGRUNN

Kunstig intelligens var lenge et tema for spesielt interesserte. Siden november 2022 har den globale offentlige debatten til tider vært dominert av nyheter og meninger om kunstig intelligens, og spesielt generativ kunstig intelligens. Konversasjonell KI basert på

[4] Med iterasjon menes her både repetisjon av utregninger i dataprosessering og gjentakende deler av andre prosesser, som for eksempel prototypeutvikling og inkrementell utarbeidelse, og forbedring av design.

[5] For eksempel *OpenAI ChatGPT* og *Google Bard*.

[6] For eksempel *OpenAI Codex* (blant annet brukt i *GitHub Copilot*) og *Hugging Face StarCoder*.

[7] For eksempel *Midjourney* og *OpenAI DALL-E*.

[8] For eksempel *Synthesia* og *Runway*.

[9] For eksempel *ElevenLabs Prime Voice AI*.

[10] For eksempel 3D-objektfiler eller bevegelsesmønstre innen robotikk.

[11] Eng. *prompts*. Eksempler på inndata i en GKI-modell kan være instruksjoner i tekst, kode, bilde, video eller lyd, som for eksempel tale.

[12] *Large Language Models* på engelsk, forkortet LLM.

[13] I denne sammenheng brukt om små beregningsenheter

[14] Fra Cyberkriminalitet 2023: *Spoofing* er for eksempel at man får kommunikasjonen til å se ut som den kommer fra et norsk nummer eller et kjent nettsted fornærmede allerede stoler på.

[15] *Malicious use* på engelsk, noen ganger forkortet MUAI - *Malicious use of AI*. Et tilstøtende begrep er *AI crime*, forkortet AIC.

store språkmodeller, samt bildegenererende modeller, har den siste tiden drevet publikums interesse. GKI er i høy grad en del av den pågående teknologiske utviklingen.

Teknologisk[16] utvikling er hverken godt eller ondt i seg selv; ei heller er det nøytralt. For cyberkriminalitet er teknologisk utvikling en driver, som resulterer i nye verktøy cyberkriminelle kan bruke. GKI kan forstås som en gruppe slike verktøy.

GKI er et verktøy som kan understøtte alle typer cyberkriminalitet; både innen kriminalitet mot datasystemer, og kriminalitet støttet av datasystemer.

5. NÅSITUASJONEN: TRUSSELEN FRA GKI I DAG

Den teknologiske utviklingen går raskere enn samfunnet klarer å hindre uønskede utfall og trusler. GKI øker kriminelles kapabiliteter og fører til et slankere kompetansekrav for å gjennomføre cyberkriminalitet.

Cyberkriminelle kan utnytte GKI-modeller på flere måter. Det er mulig å benytte GKI-modeller til å generere innhold som i seg selv er ulovlig (for eksempel oppfordringer til vold eller seksualiserte framstillinger av barn), eller som kan bli brukt i kriminelle handlinger (for eksempel skadevarekode eller syntetisk tale til bruk i bedragerier). Cyberkriminelle kan også angripe teknologien - det vil si at GKI-modellene kan være målet for kriminaliteten.

Teknologien bringer med seg nye muligheter og utfordringer, primært knyttet til ressursoptimalisering sammen med autonomi og en kapabilitet til å utnytte konteksten til inndata. Cyberkriminelle utarbeider derfor nye metoder og teknikker som de kan benytte. Et eksempel på dette er at modellene i seg selv representerer en ny sårbarhetsflate som kriminelle kan utnytte.[17]

I tillegg til å ha en direkte innvirkning på kriminaliteten slik som beskrevet ovenfor, er det også mulig å begå kriminalitet ved å utnytte GKI indirekte. Et eksempel på dette er å selge falske tjenester som gir lovnader om profitt basert på bruk av GKI, eller framstår som GKI-tjenester, men samler persondata som ledd i svindel. Dette er i tråd med andre typer bedrageri som dukker opp med nye fenomener, slik vi har sett med eksempelvis koronasvindel og kryptosvindel.

Kriminalitet begått med GKI kan ramme ofre på individ-, organisasjons- og statlig nivå, og den har potensial til å ramme ofre i stor skala og høyt tempo, med stor grad av skreddersøm.

5.1. Kommersiell GKI og GKI med åpen kildekode

Det utvikles stadig flere modeller der kildekoden[18] og treningsgrunnlaget er en bedriftshemmelighet, men bruken av modellen er gratis eller tilgjengelig gjennom betalingsløsninger. Det er vanlig at kommersielle aktører tilbyr en begrenset gratis-versjon i tillegg til en betalt versjon med flere funksjoner og muligheter for tilpasning. GKI-løsninger integreres mer og mer i kjente programvarer som for eksempel bildebehandlingsprogrammet Adobe Photoshop.

[16] Teknologi refererer her til alle kombinasjoner av verktøy, evner, prosesser og teknikker som utvider menneskers kapabiliteter.

[17] For eksempel gjennom dataforgiftning, inndata-angrep, bakdørangrep, invertering og følgeslutningsangrep (eng. *inference attack*s).

[18] Kildekode er programmering (instruksjoner til en datamaskin) på en tekstform som kan leses av mennesker.

Selskaper som utvikler GKI-modeller forsøker å hindre[19] at modellene genererer materiale som er ulovlig eller støtende.[20] Mange brukere forsøker å konstruere spørringer[21] som omgår disse begrensningene.[22] Ikke-kommersielt utviklede GKI-modeller kan være laget helt uten slik sensurering.

Enkeltpersoner trener og finjusterer[23] egne modeller for å omgå det de oppfatter som sensur eller påtvunget politisk korrekthet i kommersielle produkter. Slike modeller kan lettere generere innhold som faller inn under norsk straffelov.

Hele produsenteide[24] GKI-modeller har blitt stjålet eller lekket, og brukes som elementer i finjusterte oppsett myntet på generering av materiale som kommersielle modeller sensurerer.

Det utvikles mange åpent tilgjengelige modeller. Tilgjengelige GKI-modeller trent på store datasett – som krever mye regnekraft og maskinvare som vanlige konsumenter som regel ikke har tilgang til – kan trenes videre og dermed finjusteres av enkeltpersoner som tilfører ønsket tilleggsdata. Denne typen finjustering kan gjøres på vanlige datamaskiner. Tilleggsdataene gjør modellen mer egnet til spesifikke formål, inkludert ikke-legale formål.

GKI-modeller som kan finjusteres og brukes lokalt[25] i dag kan utføre flere oppgaver som er nyttige for cyberkriminelle. Det er for eksempel mulig å klone andres stemmer for bruk i direktørsvindel og utpressing; generere skadevarekode, seksualiserte framstillinger av barn eller troverdig tekst til bruk i bedragerier.

Personlige datamaskiner er i dag kraftige nok til å kjøre noen modeller lokalt. Trening krever normalt mer ressurser enn bruken av modeller. Finjustering av eksisterende modeller er i noen tilfeller tilsvarende lite ressurskrevende som bruk.

5.2. Kapabiliteter og sårbarheter

Cyberkriminalitet beveger seg både fort og langsomt på samme tid. Teknologien utvikler seg vanligvis raskt, og kan også gjøre uventede kvantesprang. Dette har direkte og umiddelbar innvirkning på cyberkriminelles kapabiliteter.

GKI-modeller utvikles for å kunne håndtere mer inndata, inndata av flere typer (som tekst og bilder) i samme modell[26], og for å generere utdata i større datamengder og av stadig høyere kvalitet.

Per mars 2023 er det fortsatt dyrt å trene store GKI-modeller som GPT-4 fra bunnen av, men dette blir stadig billigere.

Det utvikles hele tiden nye metoder for å forbedre GKI-systemers kapabiliteter, tilføre nye kapabiliteter, samt senke ressurskrav til -og kostnader[27] ved trening og bruk. Tempoet i forskningen er for tiden høyt. Samtidig blir datamaskiner stadig kraftigere. Alt dette har potensial til å styrke, endre eller å skape nye former for cyberkriminalitet.

---

[19] En av metodene som benyttes er såkalt *Reinforcement Learning from Human Feedback* (RLHF), hvor menneskers vurdering av modellens resultater brukes til å trene modellen, i tillegg til datasett. Metoden skalerer ikke godt.

[20] Hva som er ulovlig eller anses som støtende varierer internasjonalt mellom jurisdiksjoner, land og kulturer.

[21] En type inndata som oftest består av tekst.

[22] Dette kalles ofte "*jailbraking*", forkortet JB.

[23] Omtalt på engelsk blant annet som *fine-tuning* eller *re-training*. Dette innebærer som oftest at forhåndstrente modeller bli trent videre på små datasett spisset mot et bestemt formål.

[24] Også kjent som proprietær programvare; til forskjell fra programvare med åpen kildekode.

[25] Med lokalt menes at man ikke er avhengig av ekstern regnekraft eller lagringsplass, som i skyen. Som oftest refererer dette til én enkelt datamaskin, eid av en privatperson og oppbevart og benyttet hjemme hos vedkommende.

[26] Dette kalles multimodalitet, jf. *Multimodal Large Language Models* (MLLM).

[27] Penger, tid og regnekraft (*compute*).

GKI fører med seg potensielle sårbarheter i teknologien. Datalekkasjer og datatyveri fra eller via GKI-modeller er eksempler. Det er i noen tilfeller mulig å hente ut treningsdata fra GKI-modeller trent på datasett som er ment å holdes skjult. Det er også i noen tilfeller mulig å skaffe seg tilgang til andres private spørringer, samtidig kan spørringer også benyttes som treningsdata i nye versjoner av modellen og dermed bli kjent gjennom dette.

5.3. Kriminalitet mot datasystemer

Tilfanget av tekst som brukes til å lage store språkmodeller har overvekt av teknisk orientert innhold.[28] Slik tekst inneholder ofte beskrivelser av svakheter i datasystemer, samt løsninger. Dette medfører at GKI-modeller effektivt kan assistere cyberkriminelle med å identifisere og utnytte svakheter i andres systemer. Enkelte operasjoner innen kriminalitet mot datasystemer er derfor lettere å utføre enn tidligere.

GKI benyttes til å generere skadevarekode. Det finnes konseptbevis som benytter automatisert henting av kode generert av ChatGPT i skadevaren for å unngå deteksjon.

GKI kan øke kvaliteten på passordgjetting ved rå datakraft[29], ved å generere smartere og mer effektive lister over sannsynlige passord å bruke datakraften på. Slike passordordbøker blir automatisk generert basert på store mengder lekkede passord, i tillegg til sofistikerte maskinlæringsalgoritmer som kan trekke ut en betydelig mengde passordegenskaper.

5.4. Kriminalitet støttet av datasystemer

GKI-modeller brukes i dag til å generere bilder som seksualiserer barn, herunder syntetisk overgrepsmateriale. GKI kan benyttes til å forbedre kvaliteten i og oppskalere[30] bilder og video av seksuelle overgrep. En slik bruk av KI-basert programvare kan reaktualisere gammelt overgrepsmateriale og forlenge tiden det er i omløp.

GKI brukes også til å lage konversasjonelle modeller som genererer og sprer tekst som kan falle innenfor kategorien hatkriminalitet.

GKI-modeller brukes videre til å produsere deepfakes. Eksempelvis har syntetiske stemmer blitt benyttet i direktørsvindel. Deepfakes utgjør en trussel på flere områder: Psykologisk, gjennom seksuell utpressing, trusler, mobbing og diskreditering; finansielt, gjennom vinningsmotivert utpressing, identitetstyveri, bedrageri og svindel, manipulasjon av finansielle systemer, merkevare- og omdømmesabotasje; og samfunnsmessig, gjennom falske nyheter og media, manipulasjon og påvirkning av potensielt alle samfunnsinstitusjoner.

GKI brukes også til å lage tekst til phishing[31]-kampanjer. Generelt kan GKI automatisere både produksjonen og individualisering av tekst. Slik individualisering er ellers potensielt meget tidkrevende, og i denne konteksten kalt spearphishing.

Chatbotter[32] brukes i dag til digital rekruttering[33] til terrororganisasjoner og sosial integrering[34] blant ekstremister. Konversasjonell KI har potensial til å erstatte eller komplementere chatbotter.

---

[28] Dette er fordi det er en overvekt av slik informasjon på det åpne nettet, hvor datagrunnlagene i hovedsak hentes fra.

[29] Eng: *brute-force attack*.

[30] Oppskalering refererer her til å øke oppløsningen på eller antallet piksler i punktgrafikk – det vil si å gjøre små bilder større.

[31] På norsk også kalt nettfisking. Refererer til metoder for å urettmessig tilegne seg andres persondata eller påloggingsdetaljer gjennom sosial manipulasjon, for eksempel gjennom epost.

[32] Tradisjonelle chatbotter følger forhåndsdefinerte regler og svarer brukerne basert på søkeordmatching. Konversasjonell KI bruker maskinlæring til å forstå kontekst, generere mer menneskelignende svar og delta i dynamiske og interaktive samtaler.

[33] Eng. "*Personal automated headhunters*".

[34] Eng. "*Agents of affective bonding*".

Ettersom GKI-modeller blir trent på store datamengder som kan inneholde personopplysninger eller immaterielle rettigheter/åndsrett, kan analyse av modellens utdata avsløre sensitiv data som modellen er trent på.

5.5. Juridiske og regulatoriske forhold

GKI reiser flere juridiske spørsmål. Forskere har problematisert om bruk av KI i cyberkriminalitet kan undergrave dagens straffbarhetsvilkår, for eksempel krav om skyld. Straffesaker som involverer cyberkriminalitet kan være utfordrende å sette seg inn i, fordi de tekniske aspektene kan være vanskelig tilgjengelige for lekfolk. Straffesakene kan i tillegg være omfattende og kompliserte. Dette kan spesielt gjelde større saker med kriminalitet mot datasystemer. GKI introduserer enda et krevende teknisk element, og kan også bidra til å øke kompleksiteten i straffesaker fordi bruken har gjort kriminaliteten mer effektiv, mer omfattende, mer krevende å etterforske, eller har muliggjort et større antall fornærmede i samme sak.

Samtidig som cyberkriminaliteten endres raskt på et teknisk og taktisk nivå, endrer kriminelles målsettinger seg langsomt. Seksuelle motivasjoner og ønsker om vinning framstår bestandige. Det følger av dette at teknologiske nyvinninger slik som GKI ikke nødvendigvis utløser behov for ny lovgivning, nettopp fordi teknologien benyttes til å utføre allerede regulerte lovbrudd.

Norsk straffelov har teknologinøytrale bestemmelser. Vi har hjemler som kommer til anvendelse ved både bruk, uaktsomhet, medvirkning og forsøk i ulike typer cyberkriminalitet der GKI kan benyttes.

Cyberkriminelles bruk av GKI skiller seg ikke fra deres bruk av andre digitale verktøy. Det betyr at kjente utfordringer med internasjonal sakskoordinering, informasjonsdeling og innhenting av bevis fra utlandet gjør seg like gjeldende i cyberkriminalitet der GKI er benyttet, som i annen cyberkriminalitet.

Det kreves store datamengder for å trene noen GKI-modeller, noe som kan føre til juridiske problemstillinger knyttet til åndsverk og opphavsrett. Opphavsrettsdirektivet[35] i EU har gitt unntaksvis tillatelse til å oppbevare åndsverk til bruk for tekst- og datautvinning. Denne reguleringen omtaler ikke oppbevaring for trening av GKI-modeller. Den gjelder for både forskningsorganisasjoner så vel som kommersiell bruk, med mindre rettighetshavere har tatt uttrykkelig forbehold om at verket ikke kan brukes til tekst- og datautvinning. Svaret på hvorvidt det er lovlig å trene modeller på materiale beskyttet av opphavsrett er under granskning og vil variere i ulike jurisdiksjoner.

Regulering av GKI er utfordrende fordi teknologien fortsatt er under utvikling. Internasjonalt er det flere pågående initiativ, blant annet i EU, som vil få følger for norske forhold.

6. VURDERINGER

GKI kan understøtte alle typer cyberkriminalitet. Felles for GKIs effekt på dem i et fem års perspektiv er at flere og flere operasjoner[36] som i dag er ressurs- og/eller kompetansekrevende vil bli mer og mer trivielle å utføre. Det er en generell teknologisk utvikling at utregninger som tidligere krevde lang tid, senere kan utføres i sanntid med nyere maskin- og programvare. Det *tas for gitt* at GKI-modellers kapabiliteter vil følge denne utviklingskurven.

[35] Directive (EU) 2019/790 Of The European Parliament And Of The Council, artikkel 3 og 4.

[36] Med operasjoner menes her for eksempel å generere ulovlig materiale, opprette digital infrastruktur til bruk i datainnbrudd, hvitvaske penger via kryptovalutamiksere, produsere skadevare og villede ofre.

Dette vil påvirke cyberkriminelles behov for teknisk kompetanse. Kompetansekrav for å begå ulike cyberkriminelle aktiviteter vil *meget sannsynlig* reduseres, og kompetanseprofil for å utvikle nye tekniske løsninger vil *sannsynlig* endres.[37]

Det finnes lange lister med mulige brukseksempler av GKI til illegale formål. En fullstendig oversikt vil være lite hjelpsom som strategisk beslutningsstøtte, uten videre vurdering og vekting av trusselen de ulike eksemplene utgjør. Denne rapporten gjengir et utvalg vurderinger basert på Kripos' forståelse av trusselen slik den framstår i dag, innen avgrensningene oppgitt i punkt 3 i innholdsfortegnelsen. Det er lagt stor vekt på en intern workshop med deltagere fra flere særorganer i politiet.

6.1. Styrkemultiplikator for eksisterende cyberkriminalitet

GKI bidrar til en generell kostnads- og arbeidsreduksjon i utførelsen av cyberkriminalitet. Konversasjonelle KI-tjenester gjør cyberkriminelle aktiviteter som planlegging, rekognosering og målutvelgelse lettere å utføre og mer effektive. Disse forholdene, sammen med en økt psykologisk distanse til ofre gjennom mer autonome prosesser, vil *meget sannsynlig* føre til en økning i antall cyberkriminelle og deres kapabiliteter.

GKI utgjør en styrkemultiplikator for en lang rekke operasjoner innen cyberkriminalitet. Oppsett av infrastruktur, produksjon av tekst og tale til phishing, vishing[38] eller kontaktetablering med barn, samt individualisering til spearphishing av slik tekst og tale, er alle eksempler på tidkrevende prosesser som kan automatiseres med GKI. Slik automatisering har flere følger, blant annet gjør den cyberkriminaliteten mer effektiv og frigir tid og krefter som kan benyttes til bedre planlegging og målretting, eller utførelse av mer kriminalitet.

Det er *meget sannsynlig* at GKI vil brukes til å automatisere opprettelse og vedlikehold av massive mengder digitale profiler for fiktive personer og virksomheter. GKI kan over måneder og år utføre troverdig aktivitet gjennom disse profilene, og slik bygge plausible identiteter som kan benyttes til ulike typer påvirkning og kriminalitet. Det er *sannsynlig* at slike identiteter vil bli en handelsvare.

Konversasjonell KI er en videreføring av chatbot-konseptet, og kan brukes på samme måte til å automatisere og effektivisere radikalisering[39] på nett. Konversasjonell KI vil *sannsynlig* brukes som verktøy i radikaliseringsprosesser innen cyberkriminalitet de neste fem årene. Rekruttering til, og bidrag til normalisering av kriminalitet i fora for seksuelle overgrep og for kriminalitet mot datasystemer, vurderes per nå som mest relevante.

Lokale konversasjonelle KI-modeller kan allerede i dag være effektive verktøy i selvradikaliseringsprosesser.

GKI-modeller basert på åpen kildekode vil de neste fem årene *meget sannsynlig* bli like kapable som, eller mer kapable enn produsenteide modeller. Slike modeller vil enkelt kunne finjusteres lokalt til spesialiserte ikke-legale formål.

Åpent tilgjengelige modeller som kan kjøres lokalt på enkeltpersoners egne datamaskiner og finjusteres til illegale formål, utgjør en egen kapabilitet for cyberkriminelle. Denne måten å bruke GKI til ikke-legale formål har et annet potensial enn misbruk av betalingstjenester som ChatGPT. Metoden omgår reguleringsforsøk innebygd i kommersielle tjenester og etterlater seg færre elektroniske spor utenfor lokale enheter. Kripos vurderer derfor trusselen fra kriminelles bruk av lokale, uregulerte modeller som større enn trusselen fra kriminelles bruk av regulerte, kommersielle tjenester.

---

[37] For eksempel kompetanse innen eng. *prompt engineering* vs programmering/koding

[38] Som phishing, men via stemme i stedet for tekst, for eksempel på telefon.

[39] Radikalisering brukes her generelt om en endringsprosess i retning det ekstreme, uavhengig av kriminalitetstype eller terrorisme. Eksempler kan være å gå fra nedlasting av seksualiserte bilder av barn til å begå seksuelle overgrep selv, eller å gå fra juks i spill på nett ved bruk av tjenestenektangrep til å benytte løsepengevirus mot virksomheter med tidskritiske prosesser.

Åpent tilgjengelige modeller kan deles med andre etter finjustering, og de kan tilgjengeliggjøres som en kriminell tjeneste. Både modeller, framgangsmåter og eventuelle datasett som settes sammen for å gi modeller kriminell kapabilitet, kan deles på både det mørke og åpne nettet. Gitt store filstørrelser vil delingen *sannsynlig* foregå på det åpne nettet, på grunn av raskere ned- og opplastingshastigheter.[40]

6.2. Endring i cyberkriminalitet

Kripos forventer en økning i mengde og en kvalitativ endring i cyberkriminalitet på grunn av bruk av GKI til illegale formål.

Det er *meget sannsynlig* at gjerningspersoner allerede generere overgrepsmateriale av barn. Det er *sannsynlig* at slik video vil kompletteres av troverdige syntetiske lydspor med stemmer i sanntid. Selv om dette er fullstendig syntetisk materiale,[41] vil det rammes av norsk straffelov på lik linje med andre seksualiserte framstillinger av barn.

Store datasett bestående av sensitiv og annen verdifull informasjon vil *meget sannsynlig* bli mer ettertraktede mål for cyberkriminelle. Slike datasett kan brukes til trening av modeller eller som inndata for å trekke ut sammenhenger som er vanskelig tilgjengelig for ufaglærte. Eksempelvis vil det være nyttig for cyberkriminelle å trene en modell basert på politiets straffesaksregister og etterretningsregister, eller ved å bruke GKI til å hente ut og sammenstille relevant informasjon fra en stor og uoversiktlig database. Slike databaser kan være offentlige helseregistre, valgregistre, folkeregistre, oversikter over rettssaker og domfellelser, eller registre over kritisk infrastruktur.

Det er *meget sannsynlig* at suksessraten for phishingkampanjer vil øke i takt med den teknologiske utviklingen innen GKI. Det som før var en avveining mellom målrettethet og spredning, blir hvisket ut som følge av GKIs kapabilitet til å masseprodusere individualisert innhold av høy troverdighet. Det har vist seg at særegent innhold som tekst på nynorsk eller videoopptak med smarttelefon fra små, lokale områder reduserer skepsis hos offeret og øker troverdigheten til avsenderen. GKI vil på sikt gjøre norskspråklige like utsatt for cyberkriminalitet som engelskspråklige er i dag.

Det er *mulig* at cyberkriminelle vil benytte GKI til å designe sofistikerte phishingkampanjer som inkluderer flere lag av manipulasjon og som utvikler seg over tid, for å begå ulike former for bedrageri. Eksempelvis kan en kriminell aktør forhånds-generere innhold og falske profiler på sosiale medier, lage falske nettsider, identiteter, nyhetsartikler og deepfakevideoer, som virker sammen for å skape en omfattende illusjon som understøtter narrativet til den kriminelle aktøren. Slike virkelighetspakker[42] vil i seg selv ikke nødvendigvis være ulovlig, og det kan være svært vanskelig å identifisere ettersom innhold kan spres utover mange forskjellige kanaler, med ulike budskap. GKI kan i tillegg bidra med å identifisere det beste tidspunktet for å igangsette kriminalitet som svindel, ved å overvåke eierskifter i bedrifter og personers livssituasjon gjennom offentlige tilgjengelige indikatorer på nett og sosiale medier.

Det er *sannsynlig* at behovet for samhandling og koordinering mellom enkelte funksjoner i et cyberkriminelt nettverk[43] vil reduseres som følge av økt bruk og teknologiske fremskritt innen GKI. Etter hvert som enkelte tjenester og funksjoner blir automatisert, erstatter dette behovet for manuelt arbeid og redusere behovet for innleide tjenester fra cyberkriminelle med spesialiserte roller. Et *mulig* utfall er økt operasjonssikkerhet gjennom å redusere antall knytninger mellom kriminelle aktører, noe som vanskeliggjør

[40] Jf. deling av informasjon og data i Kriposrapporten *Cyberkriminalitet 2023.*

[41] Med fullstendig syntetisk materiale menes i denne sammenheng KI-generert bilde, tekst eller lyd fra enkeltpersoner som ikke endrer på eksisterende materiale og som ikke baserer seg på én bestemt person.

[42] "Virkelighetspakke" brukes her om en sammensetting av ulike medier med falskt innhold som er syntetisk generert for å understøtte et overordnet narrativ.

[43] Med cyberkriminelt nettverk menes her de ulike rollene og funksjonene (eks, skadevareutvikling, initial tilgangsmegler (IAB), kryptering, hvitvaskere osv.) som er gjensidig avhengig av hverandre for å kunne utføre sammensatte cyberkriminelle handlinger, slik som eksempelvis løsepengevirusangrep.

etterforskning. Det er uvisst på hvilken måte dette vil påvirke cyberkriminalitet som en handelsvare.

GKI vil bli en del av det eksisterende digitale økosystemet der cyberkriminalitet er en tjeneste eller handelsvare[44]. Både tilgang til modeller som genererer ulovlig materiale, trening og regnekraft ved disse, datasett eller modellene selv kan selges og utvikles mot vederlag. GKI-modeller vil også understøtte kriminelle tjenester og digital kriminell handel for øvrig gjennom automatisering og effektivisering.

6.3. Ny cyberkriminalitet

Nye kapabiliteter vil oppstå der sentrale teknologier som høyprestasjonsutregning[45], maskinlæring og stordata sammenfaller. Teknologien bak GKI er i tillegg av en slik type at utviklere selv ikke har full kontroll over hvilke kapabiliteter som framdyrkes[46]. Det er en grunnleggende egenart ved KI at utfallet av utregningene ikke er kjent på forhånd. Det er *sannsynlig* at dette i sum får vidtrekkende konsekvenser som i dag ikke kan forutses, også for cyberkriminalitet.

Kriminelle motivasjoner vil likevel ikke endres av teknologien. Det er *svært lite sannsynlig* at det vil utvikles nye kriminalitetskategorier innen cyberkriminalitet, på samme konseptuelle nivå som kategoriene bedrageri og datainnbrudd. Det er samtidig sikkert at GKI vil medvirke til nye modi operandi, teknikker, taktikker og prosedyrer i både kriminalitet mot datasystemer og kriminalitet støttet av datasystemer.

Det er *lite sannsynlig* at nye typer innhold med kriminell nytteverdi vil bli generert av GKI. GKI vil heller understøtte endringer i kjente innholdstyper gjennom økt automatisering, kvalitet, effektivitet og autonomi.

Det er *sannsynlig* at virkelighetspakker for bedrageri, seksuallovbrudd, utpressing og massemanipulasjon vil bli solgt som en handelsvare og kan potensielt gjenbrukes av flere forskjellige aktører og på tvers av kampanjer.

Juridisk og regulatorisk

GKI vil *sannsynlig* benyttes med formål å vanskeliggjøre og motvirke straffeforfølgelse, blant annet ved å sette kriminelle i stand til å identifisere og utnytte mulighetsrommet som ulikheter mellom jurisdiksjoner muliggjør.

Cyberkriminelle vil *meget sannsynlig* benytte GKI til å generere falske alibier og endre på originalt digitalt innhold i den hensikt å ugyldiggjøre bevis på straffbare forhold.

Gitt gjeldende teknologinøytrale bestemmelser, og *svært liten sannsynlighet* for nye kriminalitetskategorier på samme konseptuelle nivå som bedrageri og datainnbrudd, er det *lite sannsynlig* at GKI vil utfordre norsk lovverk når det gjelder skyld, uaktsomhet, medvirkning og forsøk.

[44] Eng. Crime as a service, forkortet CaaS.

[45] *High-performance computing*, ofte forkortet HPC. Betegner bruken av superdatamaskiner og regneklynger (*computer clusters*) til å løse avanserte komputasjonelle problemer.

[46] Ofte betegnet som "*emergent properties*".

7. VEDLEGG

7.1. Sannsynlighetsord

Vurderinger vil alltid inneholde en grad av usikkerhet. For å håndtere dette på en standardisert og strukturert måte, er det benyttet sannsynlighetsord (se tabell):

Nasjonal standard	Beskrivelse	NATO standard
Meget sannsynlig	Det er meget god grunn til å forvente…	Highly likely (>90%)
Sannsynlig	Det er grunn til å forvente…	Likely (60-90%)
Mulig	Det er like sannsynlig som usannsynlig…	Even chance (40-60%)
Lite sannsynlig	Det er liten grunn til å forvente…	Unlikely (10-40%)
Svært lite sannsynlig	Det er svært liten grunn til å forvente…	Highly unlikely <10%