MENINGER
DELTA I DEBATTEN!

Debattinnlegg skal være maksimalt 3000 tegn. **Kronikk** kan være opptil 5000 tegn.

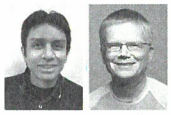
Kontakt oss: Innlegg sendes på e-post til tormod.haugstad@tu.no. Send gjerne med bilde av deg selv, og undertegn med fullt navn og yrkestittel. Innlegg honoreres ikke. Teknisk Ukeblad forbeholder seg retten til å korte ned innlegg, og til å publisere og lagre alt materiale elektronisk.

Kan manipuleres: Selvkjørende biler kan bli utsatt for ondsinnet manipulasjon, slik at de forveksler skilt i veibanen, ifølge kronikkforfatterne. ILLUSTRASJON: SHUTTERSTOCK

BESKYTTELSE AV KUNSTIG INTELLIGENS:

HVA ER DE NYE SIKKERHETS-UTFORDRINGENE?

KRONIKK KUNSTIG INTELLIGENS



ENRIQUE GARCIA-CEJA, forsker, SINTEF Digital

KETIL STØLEN, forsker, SINTEF Digital

Kunstig intelligens (KI) gjennomsyrer stadig nye bruksområder, fra droner som analyserer jordbruksfelt, intelligente medisinske apparater som overvåker pasienter basert på fysiologiske målinger til autonome navigasjonssystemer som brukes innen romfart. Det gjør KI-baserte teknologier til attraktive angrepsmål. Hva er implikasjonene for cybersikkerhet? I hvilken grad endres spillereglene for sikkerhetsarbeidet?

INTRODUKSJON

Kunstig intelligens (KI) betegner dataprogrammer som tilsynelatende er intelligente. KI-baserte løsninger kan for eksempel medføre økt effektivitet og automatisere mange oppgaver som inntil nylig krevde menneskelig interaksjon og samhandling. Dessverre er slike dataprogrammer også attraktive mål for angripere. Motivet kan være økonomisk, politisk, hevn eller personlig tilfredsstillelse. Hva er konsekvensene for cybersikkerhet? Med andre ord: Hvilke endringer er nødvendige for å sikre KI-teknologi mot nettbaserte angrep? I det følgende skisserer vi først bruken av KI innen cybersikkerhet. Deretter tar vi for oss beskyttelsen av KI generelt.

KUNSTIG INTELLIGENS INNEN CYBERSIKKERHET

Dagens komplekse informasjonssystemer gjør det vanskelig å oppnå tilstrekkelig sikkerhet ved bruk av tradisjonelle metoder som filterlister, enkle beslutningsregler og så videre. Sikkerheten må være i stand til å skalere og tilpasse seg i takt med systemets utvikling. Ta for eksempel et e-post spamfilter. Spammere kan prøve å omgå filteret ved å lære seg de underliggende sikkerhetsreglene ved gjentatt utprøving og endring av e-postens tekst inntil den markeres som legitim. I en slik sammenheng kan KI gi økt beskyttelse. En KI-basert løsning kan lære seg å gjenkjenne mønstre fra de nye spammeldingene og oppdatere sin egen oppførsel uten menneskelig innblanding.

På den annen side er KI også et nyttig verktøy for den angripende part. KI-algoritmer kan for eksempel lære seg å identifisere applikasjonen som en smarttelefon bruker eller nettstedet den besøker ved kun å ha tilgang til magnetfeltføleren som den bruker til navigasjon.

En spesiell utfordring med KI for cybersikkerhet er mangel på treningsdata av rett type. Skal programvare lære seg å identifisere og håndtere farlige sikkerhetshendelser trengs data fra slike hendelser, og hvis de forekommer svært sjelden er det ikke mye å ta utgangspunkt i. En annen utfordring består i at slik programvare må skreddersys for konkrete bruksområder for å være effektiv. Dette gjør det ofte problematisk å gjenbruke eksisterende KI-løsninger i nye sammenhenger.

KUNSTIG INTELLIGENS SOM VÅPEN I EN TVEKAMP

Tvekamper, hvor ulike varianter av KI-basert programvare benyttes av så vel den angripende som den forsvarende part, er langt fra uvanlig. De siste årene har dyp læring (en spesiell variant av KI) spilt en viktig rolle i slike kamper. Dyp læring gjør bruk av matematiske modeller av sammenkoblede noder arrangert i lag. Nodene er forenklede representasjoner av hjerneceller, såkalte nevroner. Hvis det kunstige nettverket har mange lag, kalles det et dypt nevralt nettverk (DNN). DNN kan brukes til å angripe andre KI-baserte systemer.

Ved et målrettet angrep på en selvkjørende bil kan et DNN for eksempel benyttes til å identifisere ondsinnede visuelle representasjoner, som hvis de prosesseres av bilens DNN, får den til å forveksle stoppskilt med fartsgrenseskilt. Slike ondsinnede visuelle representasjoner er små modifikasjoner, som er tilnærmet usynlige for det menneskelige øyet, men egnet til å forvirre bilens DNN.

I praksis kan en slik visuell representasjon realiseres ved å plassere små klistremerker på relevante veiskilt slik at farge, form eller posisjon endres ørlite grann. For å forsvare seg mot slike angrep, kan andre DNN-baserte teknikker benyttes, og disse kan igjen angripes, og så videre.

> KI-teknologi krever ekstra sikkerhetstiltak og disse må dekke hele prosessen

BESKYTTELSE AV TRENINGSPROSESSEN

For å utvikle KI trengs vanligvis betydelige mengder treningsdata. Opptreningen av programvaren kan foregå offline, online eller både offline og online. Treningsprosessen kan også angripes, for eksempel ved å manipulere treningsdataene (kjent som forgiftning). Fra et cybersikkerhetssynspunkt er offline-trening klart å foretrekke fremfor online. Tross alt er vi allerede godt utstyrt med verktøy og metoder for å sikre data og dataprosessering offline. Online-trening er derimot langt mer krevende, ikke minst fordi vi har behov for å bekrefte at dataene kommer fra en pålitelig kilde. Dette kan kreve ytterligere KI for å sjekke troverdigheten til nye data.

BESKYTTELSE AV KI-MODELLEN

KI-modellen er resultatet av treningsprosessen. En KI-modell er et dataprogram som på en tilsynelatende intelligent måte kalkulerer et resultat fra et sett med inputdata. Inputdata kan for eksempel være et bilde, og resultatet kan være en sannsynlighet for at bildet er av et menneskelig ansikt. Store selskaper investerer store summer i utvikling av komplekse KI-modeller og distribuerer dem som tjenester som belaster brukeren med et lite beløp for hvert spørsmål.

Denne typen grunnleggende brukerinteraksjon er også en sårbarhet. Ved gjentatte ganger å sende spørsmål til KI-modellen, kan en angriper analysere resultatene og «klone» KI-modellen (produsere en lokal kopi som produserer lignende resultater) nesten gratis. Kloning er bare et av mange angrep denne type programvare blir utsatt for. Det er derfor viktig å beskytte KI-modellen. Et enkelt og effektivt tiltak mot kloning er å returnere mindre informasjon. For eksempel kun gi et ja/nei-svar på om det er et menneskelig ansikt i bildet, i stedet for å gi sannsynligheten for at det er det. En annen måte å beskytte en KI-modell på er med jevne mellomrom å trene videre basert på nye data.

Å OPPDAGE OM KI HAR BLITT KOMPROMITTERT

Et av de viktigste åpne forskningsspørsmålene innen KI-sikkerhet er hvordan proaktivt identifisere om et KI-system har blitt kompromittert. Dette er litt som å avgjøre om et annet menneske er kunnskapsrikt og pålitelig. Ideelt sett bør identifikasjonen skje før de utilsiktede konsekvensene oppstår og så tidlig som mulig slik at korrigerende tiltak kan iverksettes. Dette er et utfordrende problem på grunn av mangfoldet av mulige angrep som krever forskjellige påvisningsmekanismer og tiltak. Dette vanskeliggjøres ved at atferden til programvaren mangler klare grenser for hva som er en normal oppførsel. Faktisk er disse «myke» begrensningene ønskede egenskaper ved et system, og får dem til å handle hensiktsmessig i situasjoner de aldri har vært utsatt for før, men samtidig gjør det det vanskeligere å avgjøre om programvaren har blitt kompromittert. En mulig fremgangsmåte for å oppdage slike angrep er kjøre flere forskjellige og uavhengige KI-systemer i parallell. Hvis et av de kompromitteres, vil de andre sannsynligvis gi avvikende råd. Dette minner litt om å basere seg på råd fra en gruppe eksperter for å ta en avgjørelse. Er ekspertene enige, er det trolig grunn til å stole på rådet. Strides ekspertene bør det undersøkes nærmere.

KONKLUSJONER

KI-teknologi krever ekstra sikkerhetstiltak, og disse må dekke hele prosessen fra systemdesign til drift. Systemutformingen må være slik at minst mulig informasjon gis til sluttbrukerne, inkludert hva slags KI-teknologi det er snakk om, hvordan den er konfigurert, hva som er resultatenes reelle detaljeringsnivå, og så videre. I forbindelse med trening må man sørge for at dataene kommer fra en pålitelig kilde uten å ha blitt manipulert.

Ved drift er det viktig å forhindre at KI-modellen kopieres, for eksempel ved å begrense antall spørringer per bruker. For å sørge for pålitelige resultater er det lurt å gjøre tilleggsanalyser av inputdata for å sjekke om det for eksempel er snakk om et bilde som har blitt manipulert for å lure KI-modellen. Alt i alt er beskyttelse av KI en krevende oppgave som må følges opp kontinuerlig i takt med den raske utviklingen av feltet. ●