TRAFICOM
Liikenne- ja viestintävirasto

Tekoälyn mahdollistamat kyberhyökkäykset

Traficomin julkaisuja
30/2022

Sisällys

TRAFICOM Liikenne- ja viestintävirasto
Kyberturvallisuuskeskus

Julkaisupäivämäärä
13.12.2022

Julkaisun nimi
Tekoälyn mahdollistamat kyberhyökkäykset

Tekijät
Matti Aksela, Samuel Marchal, Andrew Patel, Lina Rosenstedt, WithSecure

Toimeksiantaja ja asettamispäivämäärä
Liikenne- ja viestintävirasto Traficom

Julkaisusarjan nimi ja numero
Traficomin tutkimuksia ja selvityksiä 30/2022

ISSN(online) 2669-8757
ISBN(online) 978-952-311-827-0

Asiasanat
Tekoäly, koneoppiminen, kyberturvallisuus, tietoturva, tietosuoja, kyberhyökkäykset

Tiivistelmä

Tekoälyn mahdollistamat kyberhyökkäykset nousivat esille noin viisi vuotta sitten generatiivisten tekoälymallien vauhdittamana. Tällaiset mallit kykenevät aikaisempaa paremmin automatisoimaan sekä kohdennettuja tietojenkalastelu-hyökkäyksiä että haavoittuvuuksien etsimistä. Sittemmin tekoälyn tukemia sosiaalisen manipuloinnin ja imitaatioon perustuvia hyökkäyksiä on tapahtunut, mikä on jo aiheuttanut miljoonien dollarien taloudellisia menetyksiä[1]. Tekoälytutkimuksen tämänhetkinen nopea edistyminen yhdistettynä lukuisiin uusiin käyttötarkoituksiin antaa syytä uskoa, että tekoälytekniikoita tullaan pian käyttämään tukemaan niitä vaiheita, joita tyypillisesti suoritetaan manuaalisesti kyberhyökkäysten aikana. Tästä syystä ajatus tekoälyn tukemista kyberhyökkäyksistä on viime aikoina saanut enemmän huomiota sekä tiedemaailmassa että teollisuudessa. Vaikka ei olekaan todennäköistä, että tekoäly vielä loisi täysin uudenlaisia hyökkäyksiä, näemme jatkuvasti enemmän tutkimusta siitä, miten tekoälyä voitaisiin käyttää kyberhyökkäyksien radikaaliinkin tehostamiseen ja skaalaamiseen.

Vuoden 2019 lopulla tehty tutkimus osoitti, että yli 80 % päättäjistä oli huolissaan tekoälyn mahdollistamista kyberhyökkäyksistä ja ennusti, että tämäntyyppiset hyökkäykset voivat yleistyä lähitulevaisuudessa[2]. Nykyiset tekoälytekniikat tukevat jo monia tyypillisen hyökkäysketjun alkuvaiheita. Kehittynyt käyttäjän manipulointi ja tiedonkeruutekniikat ovat tällaisia esimerkkejä. Tekoälyn tukemat kyberhyökkäykset ovat jo uhka, josta monet organisaatiot eivät pysty selviytymään. Tämä turvallisuusuhka vain kasvaa, kun näemme uusia edistysaskeleita tekoälymenetelmissä ja kun asiantuntemus tekoälystä tulee laajemmin saataville.

Tämän raportin tarkoituksena on esitellä tekoälyn mahdollistamien kyberhyökkäysten turvallisuusuhkaa tekemällä yhteenveto aiheesta olemassa olevasta nykyisestä tiedosta. Tekoälyteknologia pystyy tällä hetkellä parantamaan vain muutamia hyökkääjän taktiikoita, ja sitä käyttävät todennäköisesti vain edistyneet uhkatoimijat, kuten kansallisvaltioiden hyökkääjät. Lähitulevaisuudessa tekoälyn nopeatempoinen kehitys todennäköisesti parantaa ja luo laajemman valikoiman mahdollisuuksia hyökkäysten automatisoinnin, käyttäjämanipuloinnin ja tiedonkeruun saralla. Näin ollen voidaan ennustaa, että tekoälyn tukemat hyökkäykset yleistyvät vähemmän taitavien hyökkääjien keskuudessa seuraavan viiden vuoden aikana. Kun tavanomaiset kyberhyökkäykset vanhenevat, tekoälyteknologiat, -taidot ja -työkalut tulevat helpommin saataville ja edullisemmiksi, mikä kannustaa hyökkääjiä hyödyntämään tekoälyn tukemia kyberhyökkäyksiä.

Kyberturvallisuusalan on sopeuduttava selviytyäkseen tekoälyä hyödyntävistä kyberhyökkäyksistä. Esimerkiksi biometriset todennusmenetelmät voivat vanhentua tekoälyn mahdollistamien kehittyneiden imitaatiotekniikoiden vuoksi. Uusia ehkäisy- ja havaitsemismekanismeja on myös kehitettävä tekoälyn tukemien kyberhyökkäysten torjumiseksi. Lisää automaatiota ja tekoälyteknologiaa on käytettävä myös puolustusratkaisuissa, jotta ne vastaisivat tekoälyn tukemien kyberhyökkäysten nopeutta, laajuutta ja kehittyneisyyttä. Tämä voi johtaa epäsymmetriseen taisteluun tekoälytekniikoita rajoittamattomasti käyttävien hyökkääjien ja puolustajan välillä, jota rajoittaa tekoälyä koskevat lait ja säännökset.

Yhteyshenkilö	**Raportin kieli**	**Luottamuksellisuus**	**Kokonaissivumäärä**
Markus Mettälä, Juhani Eronen	suomi	Julkinen	31

Jakaja	**Kustantaja**
Liikenne- ja viestintävirasto Traficom, Kyberturvallisuuskeskus	Liikenne- ja viestintävirasto Traficom, Kyberturvallisuuskeskus

[1]https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/
[2]Forrester – The emergence of offensive AI (2019)

TRAFICOM Transport- och kommunikationsverket
Cybersäkerhetscentret

Utgivningsdatum
13.12.2022

Publikation
AI-aktiverade cyberangrepp

Författare
Matti Aksela, Samuel Marchal, Andrew Patel, Lina Rosenstedt, WithSecure

Tillsatt av och datum
Transport- och kommunikationsverket Traficom

Publikationsseriens namn och nummer	ISSN(webbpublikation) 2669-8757
Traficoms forskningsrapporter och utredningar 30/2022	ISBN(webbpublikation) 978-952-311-827-0

Ämnesord
Artificiell intelligens, maskininlärning, cybersäkerhet, informationssäkerhet, dataskydd, cyberangrepp

Sammandrag

Cyberangrepp som artificiell intelligens (AI) möjliggör lyftes fram för cirka fem år sedan och fick fart från generativa modeller för artificiell intelligens. Sådana modeller kan bättre än tidigare automatisera både riktade nätfiskeangrepp och letande efter sårbarheter. Senare har det skett angrepp som baserar sig på AI-stödd social manipulering och imitation, vilket redan har orsakat ekonomiska förluster på flera miljoner dollar[3].

Det snabba framskridandet av forskningen av artificiell intelligens kombinerat med flera nya användningsändamål ger anledning att tro att AI-teknologier mycket snart kommer att användas som stöd för de faser som i allmänhet görs manuellt under cyberangrepp. Av denna anledning har tanken på AI-stödda cyberangrepp under de senaste tiderna fått mer uppmärksamhet både i den vetenskapliga världen och inom industrin. Även om det inte är sannolikt att artificiell intelligens skulle skapa helt nya typer av angrepp ser vi hela tiden mer forskning om hur artificiell intelligens skulle kunna utnyttjas för radikalt effektivare och skalbara cyberangrepp.

Undersökningen från slutet av 2019 visade att över 80 procent av beslutsfattarna oroade sig över de cyberangrepp som artificiell intelligens möjliggör och förutspådde att angrepp av detta slag kan bli vanligare inom den närmaste framtiden[4]. Dagens AI-teknologier stöder redan flera inledningsskeden i en typisk kedja av angrepp. Avancerad manipulering av användare och datainsamlingstekniker är exempel på sådana. AI-stödda cyberangrepp är redan ett hot som många organisationer inte kan klara av. Detta säkerhetshot ökar när vi ser nya framsteg i AI-metoder och när sakkunskap om artificiell intelligens blir tillgänglig på ett mer omfattande sätt.

Syftet med denna rapport är att presentera den säkerhetshot som artificiell intelligens medför i form av en sammanfattning av den information som finns tillgänglig i dag. AI-teknologin kan för tillfället endast förbättra några av angriparens taktiker och används sannolikt endast av avancerade hotaktörer, t.ex. angripare i nationalstater. I den närmaste framtiden kommer den snabbare utvecklingen av artificiell intelligens sannolikt att förbättra och skapa ett bredare urval av möjligheter för automatisering av angrepp, manipulering av användare och datainsamling. Därför är det möjligt att förutspå att AI-stödda angrepp blir allt vanligare hos mindre skickliga angripare under de följande fem åren. När de vanliga cyberangreppen blir föråldrade, blir AI-teknologier, -kunskaper och -verktyg mer lättillgängliga och förmånligare, vilket uppmuntrar angripare att utnyttja AI-stödda cyberangrepp.

Cybersäkerhetsbranschen måste anpassa sig för att kunna klara av de cyberangrepp som utnyttjar artificiell intelligens. Till exempel biometriska autentiseringsmetoder kan föråldras på grund av avancerade imitationsteknologier som artificiell intelligens möjliggör. Man ska också utveckla nya mekanismer för förebyggande och detektering i syfte att avvärja AI-stödda cyberangrepp. Ytterligare automatisering och AI-teknologi ska också användas för försvarslösningar så att de skulle motsvara snabbheten, omfattningen och utvecklingen av AI-stödda cyberangrepp. Detta kan leda till en asymmetrisk bekämpning mellan angripare som använder AI-teknologier utan begränsningar och försvaret som begränsas av lagar och bestämmelser om artificiell intelligens.

Kontaktperson	**Språk**	**Sekretessgrad**	**Sidoantal**
Markus Mettälä, Juhani Eronen	finska	offentlig	31

Distribution	**Förlag**
Transport- och kommunikationsverket Traficom, Cybersäkerhetscentret	Transport- och kommunikationsverket Traficom, Cybersäkerhetscentret

[1]https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/
[2]Forrester – The emergence of offensive AI (2019)

TRAFICOM Finnish Transport and Communications Agency
National Cyber Security Centre

Date of publication
13.12.2022

Title of publication
The security threat of AI-enabled cyberattacks

Author(s)
Matti Aksela, Samuel Marchal, Andrew Patel, Lina Rosenstedt, WithSecure

Commissioned by, date
Finnish Transport and Communications Agency Traficom

Publication series and number	ISSN(online) 2669-8757
Traficom Research Reports 30/2022	ISBN(online) 978-952-311-827-0

Keywords
AI, machine learning, information security, cyber security, cyberattacks

Abstract

The topic of AI-enabled cyberattacks surfaced around five years ago with examples of generative AI models able to automate both spear-phishing attacks and vulnerability discovery. Since then, social engineering and impersonation attacks supported by AI have occurred, causing millions of dollars in financial losses. Current rapid progress in AI research, coupled with the numerous new applications it enables, leads us to believe that AI techniques will soon be used to support more of the steps typically used during cyberattacks. This is the reason why the idea of AI-enabled cyberattacks has recently gained increased attention from both academia and industry, and why we are starting to see more research devoted to the study of how AI might be used to enhance cyberattacks.

A study from late 2019 illustrated that over 80% of decision-makers were concerned with AI-enabled cyberattacks and predicted that these types of attacks may go mainstream in the near future. Current AI technologies already support many early stages of a typical attack chain. Advanced social engineering and information gathering techniques are such examples. AI-enabled cyberattacks are already a threat that organisations are unable to cope with. This security threat will only grow as we witness new advances in AI methodology, and as AI expertise becomes more widely available.

This report aims to investigate the security threat of AI-enabled cyberattacks by summarising current knowledge on the topic. AI technology is currently able to enhance only a few attacker tactics, and it is likely only used by advanced threat actors such as nation-state attackers. In the near future, fast-paced AI advances will enhance and create a larger range of attack techniques through automation, stealth, social engineering or information gathering. Therefore, we predict that AI-enabled attacks will become more widespread among less skilled attackers in the next five years. As conventional cyberattacks will become obsolete, AI technologies, skills and tools will become more available and affordable, incentivising attackers to make use of AI-enabled cyberattacks.

The cybersecurity industry will have to adapt to cope with the emergence of AI-enabled cyberattacks. For instance, biometric authentication methods may become obsolete because of advanced impersonation techniques enabled by AI. New prevention and detection mechanisms will also need to be developed to counter AI-enabled cyberattacks. More automation and AI technology will also need to be used in defence solutions to match the speed, scale and sophistication of AI-enabled cyberattacks. This may lead to an asymmetrical fight between attackers having unrestricted use of AI technologies and defenders being constrained by the upcoming regulation on AI applications.

Contact person	**Language**	**Confidence status**	**Pages, total**
Markus Mettälä, Juhani Eronen	Finnish	Public	31

Distributed by	**Published by**
Transport and Communications Agency, National Cyber Security Centre Finland	Transport and Communications Agency, National Cyber Security Centre Finland

[1]https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/
[2]Forrester – The emergence of offensive AI (2019)

Lyhenteet

AI	Artificial Intelligence, tekoäly
CVE	Common Vulnerabilities and Exposures, yleiset haavoittuvuudet ja paljastuneet tietoturvapuutteet
C&C	Command-and-Control, komentopalvelin
DL	Deep Learning, syväoppiminen
DNN	Deep Neural Networks, syvät neuroverkot
GAN	Generative Adversarial Network, generatiivinen, adversatiivinen verkko
HIDS	Host Intrusion Detection System, isäntäpohjainen tunkeutumisen havaitsemisjärjestelmä
ML	Machine Learning, koneoppiminen
NIDS	Network Intrusion Detection System, verkkopohjainen tunkeutumisen havaitsemisjärjestelmä
NLG	Natural Language Generation, luonnollisen kielen tuottaminen
NSFW	Not Safe for Work, epäkorrektiin sisältöön viittava lyhenne
OSINT	Open-Source Intelligence, avointen lähteiden tiedonkeruu
5-6G	Viidennen tai kuudennen sukupolven matkaviestinverkko

Tekoälyn mahdollistamat kyvykkyydet kyberhyökkäyksissä

Tekoäly, koneoppiminen ja syväoppiminen ovat osittain päällekkäisiä aloja, jotka ovat hyötyneet lukuisista viimeaikaisista edistysaskeleista, jotka tarjoavat uusia ominaisuuksia ja mahdollistavat uusia käyttötarkoituksia. Vaikka nämä ominaisuudet on suunniteltu hyvänlaatuisiin sovelluksiin, kuten ennustamiseen, luomiseen, tietojen analysointiin ja tiedonhakuun, näitä toimintoja voitaisiin käyttää myös perinteisiä kyberhyökkäyksiä parantamaan.

Tekoäly viittaa koneen, esimerkiksi tietokoneen tai tietokoneohjelman, kykyyn suorittaa tehtäviä, joita yleensä suorittaa ihminen, eläin tai muu älykäs toimija. Tekoälyyn liitetään älykkäitä kykyjä kuten kyky päätellä, ratkaista ongelmia, keksiä tarkoituksia, yleistää ja oppia kokemuksia. Monia näistä älykkäistä kyvyistä käytetään hyökkääjien puolella heidän suunnitellessaan ja toteuttaessaan kyberhyökkäyksiä tietojärjestelmiä kohtaan.

Käsitetasolla on helposti ymmärrettävää, miten tekoäly voi edistää kyberhyökkäyksiä: tekoäly voi automatisoida manuaalisia tehtäviä kuten haavoittuvuuksien löytämistä ja niiden hyödyntämistä hyökkäyksessä. Tekoäly on kuitenkin vanha, laaja ja epäselvä käsite, johon liittyy monia alueita, kuten asiantuntijajärjestelmät, robotiikka, ja sumea logiikka.

Suurin osa näistä tekoälyn aloista ei ole niin älykkäitä, kun voisi olettaa. Ne ovat tällä hetkellä vielä kaukana ihmistason älykkyydestä. Niiden kyvykkyydet kyberhyökkäyksien edistämiseen ovat edelleen hyvin rajoittuneita. Toisaalta koneoppiminen tekoälyn alakenttänä on saanut viime aikoina erityisesti huomiota koska se on edistynyt lyhyen ajan sisällä huomattavasti. Tämä on mahdollistanut ihmisten päihittämisen useassa tehtävässä, näkyvimpinä esimerkkeinä kuvien luokittelu, tekstin kääntäminen tai shakin ja Go:n pelaaminen. Suurin osa tekoälyyn suuntautuneesta huomiosta liittyykin koneoppimisen käyttöön ja tekoälytermiä käytetäänkin usein yleisterminä, kun viitataan koneoppimiseen.

Koneoppiminen terminä kuvaa algoritmeja ja tilastollisia malleja, jotka tuottavat tuloksia seuraamatta tarkkoja ohjeita. Koneoppiminen rakentaa yhdenlaisen asiantuntijajärjestelmän joka tiedon avulla kykenee päätöksentekoon ja oppimiseen kokemusten ja kerätyn tiedon perusteella. Koneoppiminen eroaa muista tekoälyn alakentistä siinä, että se ei vaadi tarkkoja ohjeita tai sääntöjä tulosten tuottamiseen. Se käyttää muovautuvia algoritmeja, jotka kykenevät itsenäiseen oppimiseen tiedon perusteella. Koneoppiminen voidaan jakaa kolmeen tyyppiin; ohjattu oppiminen viittaa oppimiseen, jossa tavoiteltu tulos on tiedossa, ohjaamaton oppiminen viittaa tiedon käyttämiseen ilman että selvää tulosta on etukäteen määritetty, ja vahvistava oppiminen viittaa uuden tehtävän oppimiseen perustuen virheistä oppimiseen, jolloin koneoppiminen pyrkii maksimoimaan ennalta määritellyn tavoitteen saavuttamisen.

Syväoppiminen tai syvät neuroverkot ovat koneoppimiseen kuuluvia algoritmeja, jotka pääsevät hyviin tuloksiin kuvien, tekstin, videon ja äänen tunnistuksen, luokittelun tai tuottamisen automatisoinnissa. Viime aikoina nähdyt kehitysaskeleet syväoppimisen parissa ovat suurin syy tekoälyn viime aikoina saamaan suureen määrään huomiota. Syväoppiminen pääsee ennalta näkemättömiin tuloksiin monimutkaisissa tehtävissä kuten kuvien luokittelussa, tekstin kääntämisessä ja pelien pelaamisessa. Syväoppimisen suorituskyky näissä tehtävissä on jo usein huomattavasti ihmisen suorituskykyä parempi. Kone- ja syväoppiminen täyttävät monet tekoälylle asetetut odotukset. Ne osaavat päätellä, ratkoa ongelmia, löytää tarkoitusta, ja parantaa omaa suorituskykyään itsenäisesti käyttäen olemassa olevaa tietoa. Tämän kehityksen ovat mahdollistaneet isojen tietomäärien saatavuus ja halpa laskentavoima.

Tiedon monipuolisuus ja lisääntynyt saatavuus yhdistettynä pilvipalveluiden tarjoamaan halpaan laskentavoimaan on tehnyt koneoppimisen hyödyntämisestä osana myös kyberhyökkäyksiä entistä kannattavampaa.

Seuraavat tehtävät hyötyvät koneoppimisen käytöstä:

- **Ennustaminen** on tietyn lopputuloksen mahdollisuuden arvioimista käyttäen aiemmin kerättyä tietoa. Luokittelu, poikkeamien havainnointi ja regressio ovat esimerkkejä ennustamisesta. Hyökkääjä voi käyttää ennustamista esimerkiksi puhelimen painallusten tunnistamiseksi liikkeiden perusteella, heikoimman hyökkäyskohteen valintaan, ja haavoittuvuuksien löytämiseen hyödyntämistä varten.
- **Datan tuottaminen** viittaa kohdejakaumaan sopivan sisällön tuottamiseen. Hyökkääjä voi käyttää datan tuottamista esimerkiksi mediatodisteiden vääristelyyn, salasanojen arvaamiseen, tai tietoliikenteen muovaamiseen, jotta välttyisi paljastumiselta. Toinen esimerkki kehittämisen käyttämisestä hyökkääjän tarkoituksiin on syvävääreennys (deepfake). Ne ovat uskottavia syväoppimisen kehittämiä videoita tai ääniä. Syvävääreennyksiä voidaan käyttää kohteen imitoimiseen käyttämällä kohteen ääntä, kasvoja ja liikkeitä esimerkiksi tietojenkalastelussa.
- **Data-analyysi** viittaa hyödyllisen tiedon löytämiseen isoista määristä dataa, ilman että on ennalta määrätty, mitä tietoa tarkalleen etsitään. Hyökkääjä voi käyttää data-analyysia tunnistamaan mihin ja miten haittaohjelma tulisi piilottaa, tai esimerkiksi kohteiden tunnistamiseen.
- **Tiedonhaku** viittaa hyökkääjän ennalta määrittelemän tiedon löytämiseen. Tiedonhakua voidaan käyttää kohteen tai henkilön seuraamiseen esimerkiksi valvontajärjestelmän avulla, tai tyytymättömän työntekijän löytämiseen käyttäen kielianalyysia sosiaalisen median julkaisuille, tai pitkistä dokumenteista tärkeän tiedon tiivistämiseen hyökkäyksen tiedusteluvaiheen aikana.

Miten tekoäly voi parantaa kyberhyökkäyksiä?

Tekoälyjärjestelmien tarjoama älykäs automaatio voi parantaa perinteisiä kyberhyökkäyksiä lisäämällä niiden nopeutta, laajuutta, kattavuutta ja yksilöllistä kohdentamista, mikä lisää yleistä menestystä. Nämä parannukset vaikuttavat käytännössä kaikkiin taktiikkoihin hyökkäyksen elinkaaressa. Monet uudet hyökkäystekniikat todennäköisesti otetaan huomattavasti laajamuotoisempaan käyttöön uusien tekoälyominaisuuksien myötä.

Tekoälyn tuomia etuja ja parannuksia

Perinteisesti hyökkääjät ovat käyttäneet paljon vaivaa, asiantuntijuutta ja suhteellisen yksinkertaisia työkaluja kyberhyökkäyksiin. Tekoälyn tuomat muutokset voidaan jakaa kolmeen osaan. Ensimmäiseksi, tekoälyn avulla voidaan automatisoida aiemmin manuaalisesti tehtyjä työvaiheita. Toiseksi tekoäly parantaa ja tehostaa hyökkääjien usein käyttämiä työkaluja. Kolmanneksi tekoäly tuo hyökkääjille kokonaan uusia kykyjä, kuten painallusten tunnistaminen näppäimistöllä liikkeen perusteella, tai ihmisten ja kohteiden seuraaminen valvontajärjestelmien avulla. Tekoälyä käyttämällä hyökkääjä lisää omaa mahdollisuuttaan onnistumiselle tekoälyn tuomien monien etujen kautta:

- **Nopeus:** tekoälyn käyttö automatisoi tehtävät, jotka on aiemmin hyökkääjän toimesta suoritettu manuaalisesti. Näihin kuuluu esimerkiksi tunnistetietojen hankkiminen, haavoittuvuuksien etsiminen, salasanojen arvaaminen jne. Nämä tehtävät voidaan hoitaa koneen toimesta ja paljon suuremmalla suorituskyvyllä, mikä nopeuttaa hyökkäystä huomattavasti. Tämä vähentää sitä aikaa, jonka hyökkääjän on oltava kohteen tietojärjestelmässä/verkossa, mikä itsessään madaltaa hyökkääjän kiinnijäämisen riskiä.
- **Tehokkuus:** tekoäly mahdollistaa hyökkäysten suorittamisen entistä suuremmassa mittakaavassa. Automatisoituja hyökkäyksiä voidaan käynnistää useita kohteita vastaan samanaikaisesti lyhyen ajan sisään. Tekoälyn hyöty tulee erityisesti esiin tarkkaan kohteelle räätälöidyissä hyökkäyksissä, kuten kohdennettua tietojenkalastelua hyödyntäessä, jota voi räätälöidä sopivaksi monelle uhrille. Tekoäly mahdollistaa hyökkäysten teon isommassa mittakaavassa, tarkemmin sekä niin, että se vaatii hyökkääjältä vähemmän manuaalista työtä.
- **Kattavuus:** tekoäly tekee hyökkäyksistä kokonaisvaltaisempia ja mahdollistaa suuremman kattavuuden hyökkäykselle. Tekoälyn mahdollistamat kyberhyökkäykset voivat analysoida isoja määriä avoimien lähteiden tietoa, ne voivat löytää uusia hyökkäyspolkuja ja päästä käsiksi suurempaan osaan kohteista. Perinteiset hyökkääjät voivat jättää huomioimatta hyökkäyksen kannalta tärkeää tietoa, mutta tekoäly optimoi haavoittuvuuksien etsimisen ja hyödyntämisen.

Tekoälyn tuomat muutokset voidaan jakaa kolmeen osaan. Ensimmäiseksi, tekoälyn avulla voidaan automatisoida aiemmin manuaalisesti tehtyjä työvaiheita. Toiseksi tekoäly parantaa ja tehostaa hyökkääjien usein käyttämiä työkaluja. Kolmanneksi tekoäly tuo hyökkääjille kokonaan uusia kykyjä.

- **Kehittyneisyys:** tekoäly mahdollistaa älykkään automaation. Tämä lisää kyberhyökkäysten hienostuneisuutta tehden niistä parempia ja onnistuneempia kuin ihmisten mahdollistamat hyökkäykset. Tämä näkyy kolmella tapaa:
  - **Kontekstualisointi:** tekoälyn kyky datan tuottamiseen mahdollistaa sen, että tekoäly voi oppia kohdejärjestelmästä ja käyttää tätä tietoa uuden sisällön (esim. haittaohjelman) luomiseen. Tämä mahdollistaa manuaalisesti ympäristöön sovitettujen hyökkäyksien sijasta automaattisesti tilanteeseen parhaiten sopivat hyökkäykset. Kohdennetut tietojenkalasteluviestit ovat henkilöityjä uhrien mukaan, haitallinen tietoliikenne on muovattu toimimaan huomaamattomasti kohdeverkossa, ja haittaohjelmien käytös on räätälöity sekoittumaan kohdejärjestelmän normaaliin toimintaan.
  - **Muovautuvuus:** tekoälyllä on kyky oppia ja uudelleenoppia kohdeympäristö automaattisesti, joten hyökkäykset voivat itsenäisesti muovautua ympäristöön tehtyihin muutoksiin. Muovautuvuus on tekoälyn mahdollistamissa kyberhyökkäyksissä enemmän pysyvä ominaisuus kuin kerran käytettävä toiminto.
  - **Vaikeasti havaittavuus:** tekoälyn mahdollistamat hyökkäykset on vaikeampia havaita kuin perinteiset hyökkäykset, mikä lisää niiden resilienssiä. Tekoäly mahdollistaa älykkään ja optimoidun tiedustelun data-analyysin ja tiedonhaun kautta. Se voi automaattisesti löytää hyökkäyspolkuja ja haavoittuvuuksia ennustamisen kautta. Tekoäly voi myös oppia ja imitoida järjestelmän ja tietoverkon käytöstä. Tekoäly voi tehdä myös haittaohjelmista itseohjautuvia, mahdollistaen vähäisemmän viestinnän komentopalvelimen kanssa, mikä vaikeuttaa haittaohjelman havaitsemista.

Nämä edut lisäävät tekoälyn mahdollistamien hyökkäysten edellytyksiä onnistumiselle verraten perinteisiin kyberhyökkäyksiin. Tekoälyn mahdollistamat kyberhyökkäykset toimivat nopeammin, kohdistuvat useammalle uhrille, ja löytävät enemmän hyökkäyspolkuja älykkään automaation ansioista. Ne ovat myös hienostuneempia, räätälöidympiä kohteelle sopivaksi, kykenevät muovautumaan reaaliajassa, ja niitä on vaikeampia havaita.

Tekoälyn mahdollistamat kyberhyökkäykset toimivat nopeammin, kohdistuvat useammalle uhrille, ja löytävät enemmän hyökkäyspolkuja älykkään automaation ansioista.

Ne ovat myös hienostuneempia, räätälöidympiä kohteelle sopivaksi, kykenevät muovautumaan reaaliajassa, ja niitä on vaikeampia havaita.

Hyökkäyksissä käytettävät tekoälyn mahdollistamat kyvykkyydet

Yksi tapa analysoida miten tekoäly lisää hyökkääjän kyvykkyyksiä on luokitella hyökkäyksiin käytettävät tekoälyyn liittyvät kyvykkyydet ja tunnistaa, miten ne parantavat kyberhyökkäyksiä. MITRE ATT&CK-viitekehys jakaa kyberhyökkäyksen eri vaiheet ja hyökkääjän tavoitteet 14 taktiikkaan, joihin sisältyy kaikki kyberhyökkäyksen aikana käytetyt tekniikat. Esimerkkejä taktiikoista ovat tiedustelu (reconnaissance), ensimmäinen pääsy (intial access), läsnäolon ylläpito (persistence), puolustuksen välttely (defense evasion), tunnuksien saanti (credential access), tunkeutumisen laajentaminen (lateral movement) jne. Tekoäly voi tukea monessa näistä taktiikoista ja luoda uusia tekniikoita, jotka voivat auttaa hyökkääjää pääsemään tavoitteeseensa. Hyökkäyksiin käytettävät tekoälyn mahdollistamat kyvykkyydet voidaan jakaa kuuteen luokkaan[3]:

- **Automaatio** on tekoälyn huomattavasti tavanomaista laajemmin mahdollistama kyvykkyys, joka itsessään voi tehdä hyökkäyksistä nopeampia, mittakaavaltaan suurempia, kattavampia ja muovautuvia. Automaatio vähentää hyökkääjältä vaadittavaa työmäärää ja lisää hyökkäyksien itseohjautuvuutta. Automaatio hyödyntää hyökkäyksessä eniten tiedustelua, ensimmäistä pääsyä, tunkeutumisen laajentamista ja vaikuttamista ATT&CK-viitekehyksessä. Se luo tekniikkoja kuten hyökkäysten muovautuminen tunnettuihin ja kehittyviin ympäristöihin. Se mahdollistaa hyökkäysten koordinoinnin, etsien haavoittuvimman kohteen, parhaan hyökkäyspolun ja parhaan ajan hyökkäyksen toteuttamiselle. Automaatio luo myös edellytykset bottiverkkojen yhteistyölle parviälyn avulla. Automaatio mahdollistaa myös hyökkäyskampanjat kuten kattavat ja hienostuneet tietojenkalastelukampanjat.
- **Hyökkäyksen vaikeasti havaittavuus** on hyökkäyksen onnistumiselle kriittinen kyvykkyys. Vaikeasti havaittavuus hyödyntää tekoälyn kykyä kehittää ja luoda sisältöä, joka muistuttaa jakaumaa, jolla malli on opetettu. Tällä tavoin tekoäly voi verhota haitallista käytöstä ja saada sen vaikuttamaan järjestelmään tai verkkoon kuuluvalta normaalilta käytökseltä. Tekoälyn vaikeasti havaittavuus hyödyntää monta hyökkäyksen vaihetta, kuten esimerkiksi tiedustelua, ensimmäistä pääsyä, läsnäolon ylläpitoa, tunkeutumisen laajentumista ja tiedonkeruuta. Se luo myös tekniikkoja havaitsemisen välttämiseen, jolla voidaan välttää hyökkääjän havaitsemista eri järjestelmien kuten sähköpostisuodattimien, ja haittaohjelmien tunnistajien toimesta. Vaikeasti havaittavuus tarjoaa hyökkääjälle kyvyn huomaamatta kohdejärjestelmässä etenemiseen sekä uusia tekniikkoja tiedon varastamiseen niin, että se sekoittuu normaaliin toimintaan.
- **Hyökkäyksen jatkuvuus** mahdollistaa hyökkääjän pysymisen järjestelmissä, johon hyökkääjä on päässyt ja antaa hyökkääjälle mahdollisuuden päästä uusiin järjestelmiin, kunnes hyökkääjän toivomaan lopputulokseen on päästy. Kyvykkyys kampanjan jatkuvuuteen hyödyntää eniten hyökkäyksen pysyvyysvaihetta ja auttaa hyökkääjää välttämään kiinni jäämisen. Tekoäly antaa uusia tekniikkoja hyökkäyskampanjan suunnitteluun esimerkiksi hyötyanalyysin kautta, automaattisen tarvittavien työkalujen ja teknologien tunnistamisen kautta. Se myös antaa mahdollisuuden simuloida hyökkäyksen tarkkaa kulkua ennen sen toteuttamista. Hyökkäyskampanjan jatkuvuus antaa työkaluja haittaohjelmien mahdollisimman hyvään piilottamiseen. Tekoäly auttaa myös virtuaaliympäristön tunnistamisessa tämän käytön estoa varten, jolloin haittaohjelmaa ei voida takaisinmallintaa.

[3]Mirsky, Yisroel, et al. “The threat of offensive AI to organisations.” arXiv preprint arXiv:2106.15764 (2021).

- **Käyttäjän manipuloinnilla** viitataan ihmiskäyttäjien hyödyntämiseen osana hyökkäystä. Ihmisiä pidetään usein tietojärjestelmän heikoimpana lenkkinä. Tekoäly voi oppia ihmisistä hyödyntääkseen heidän tunteitaan ja luottamustaan. Tätä kyvykkyyttä on jo käytetty osana esimerkiksi verkkopalveluiden asiointibotteja, jotka kykenevät ihmismäiseen keskusteluun. Käyttö laajenee koko ajan koskemaan mainontaa ja mediaa entistä laajemmin. Tekoäly parantaa hyökkäyksiä, jotka hyödyntävät käyttäjän manipulointia samalla tavalla; oppimilla kohteistaan. Tämä tekoälyn mahdollistama kyvykkyys hyödyntää niitä vaihteita, missä ihmiset ovat osana hyökkäystä. Näihin kuuluvat tiedustelu, ensimmäinen pääsy, tunnusten saaminen ja niiden korottaminen. Se mahdollistaa myös teknologioita, joita voi hyödyntää kohteiden valintaan ja seurantaan, jotta voidaan valita kohdeorganisaatiosta mahdollisimman otollinen uhri ja seurata tätä ennen hyökkäystä. Tekoäly tarjoaa mahdollisuuksia automatisoituun ja personoituun kanssakäymiseen ihmisten kanssa - esimerkiksi kohdennettujen tietojenkalastelu-sähköpostiviestien sekä asiointibottien kautta. Tekoälyä voidaan myös käyttää oikeiden ihmisten imitaatioon syvävääreennysten avulla ja valheellisten käyttäjäprofiilien rakentamiseen sosiaalista mediaa hyödyntäen. Tätä voidaan hyödyntää uhrien yhteydenotossa.
- **Käyttäjätunnisteiden varastaminen** viittaa autentikointimetodin rikkomisella laittomaan pääsyyn sisälle järjestelmiin, jotka ovat muuten turvattuja. Tekoäly voi imitoida ihmisen käytöstä uudelleentuottamalla organisaation autentikointiprotokollia ja arvaamalla käyttäjien käyttäjätunnisteita. Käyttäjäntunnisteiden varastamiseen liittyviä tekniikoita voidaan käyttää hyökkääjän toimesta ensimmäiseen pääsyyn sisälle järjestelmiin. Tekoäly mahdollistaa myös teknologiat, jonka avulla hyökkääjä voi käyttäytyä käyttäjää muistuttavalla tavalla ja ohittaa biometriset autentikointijärjestelmät. Se tarjoaa tekniikoita implisiittisten avainlokijärjestelmien kumoamiseksi, jotka perustuvat käyttäjän toimiin, kuten näppäinpainalluksiin, silmien liikkeisiin ja laitteen liikkeeseen todentamiseksi, oppimalla ja jäljittelemällä näitä ihmisten käyttäytymistä. Tekoäly kykenee myös yksinkertaisten salasanojen älykkäämpään arvaamiseen, etenkin silloin kun salasanassa on käytetty käyttäjän henkilökohtaista tietoa.
- **Tiedonkeruu** on avainasemassa kyberhyökkäyksen onnistumisen kannalta, koska sillä voidaan varmistaa toimien onnistuminen vähentäen kokeilukeroja. Tekoäly voi kerätä hyvin suuria määriä dataa ja etsiä siitä hyökkäyksen kannalta relevanttia tietoa. Tiedonkeruun soveltuvat tekniikat ovat hyödyllisiä hyökkäyksen tiedusteluvaiheessa, käyttäjätunnuksia varastaessa ja vaikutusvaiheessa. Ensisijaisesti tekoälyn hyödyt tulevat esiin avoimien lähteiden tiedon keruun ja analysoimisessa. Uupuvaa tietoa voidaan täydentää käyttäen kehittäviä koneoppimisen malleja. Tietoa voidaan tunnistaa ja louhia käyttämällä luonnollisen kielen käsittelyä, ja näiden tekniikkojen avulla voidaan myös tunnistaa hyökkäyksen kohteeksi jääneen järjestelmän arvokkain tieto. Tämä tekoälyn kyvykkyys mahdollistaa myös tiedustelun ja kohteen seurannan syväoppimisen avulla, joka luo edellytykset kuvien ja äänen prosessointiin.

Tekoälyn mahdollistama kyvykkyys	Kyberhyökkäystaktiikka	Tekoälyn mahdollistama hyökkäystekniikka
Automaatio	• Tiedustelu • Ensimmäinen pääsy • Tunkeutumisen laajentaminen • Vaikuttaminen	• Hyökkäysten muovautuminen • Hyökkäysten koordinointi • Hyökkäyskampanjat • Haavoittuvuuksien löytäminen
Vaikeasti havaittavuus	• Tiedustelu • Ensimmäinen pääsy • Läsnäolon ylläpito • Tunkeutumisen laajentaminen • Tiedonkeruu • Komentopalvelimen perustaminen • Tiedon vieminen / varastaminen	• Kiinni jäämisen välttäminen • Skannaus • Eteneminen • Tiedon vieminen / varastaminen
Hyökkäyksen jatkuvuus	• Läsnäolon ylläpito • Puolustuksen välttely	• Hyökkäyksen suunnittelu • Haittaohjelman sovittaminen kohteen normaaliin toimintaan • Virtualisoinnin tunnistaminen
Käyttäjän manipulointi	• Tiedustelu • Ensimmäinen pääsy • Käyttöoikeuksien laajentaminen • Tunnusten saaminen	• Kohteen valinta • Kohteen seuranta • Kohdennettu tietojen kalastelu • Imitaatio • Valheellisten profiilien rakentaminen
Käyttäjätunnisteiden varastaminen	• Ensimmäinen pääsy • Tunnusten saaminen	• Biometristen tunnistamis-keinojen ohittaminen • Näppäinpainallusten tunnistaminen • Salasanojen arvaaminen
Tiedonkeruu	• Tiedustelu • Tunnusten saaminen • Tiedonkeruu • Vaikuttaminen	• Tiedon louhinta avoimista lähteistä • Kohteen valinta • Kohteen seuranta • Tiedustelu

Yllä olevassa taulukossa kuvatut kuusi tekoälyn hyökkäyskykyä hyödyttävät suurelta osin hyökkäyksen tiedustelu- ja puolustuskiertovaiheita. Seuraavaksi eniten ne hyödyntävät resurssien kehittämistä, vaikuttamista, tiedon löytämistä ja keräämistä.

Nykyiset tekoälytekniikat eivät merkittävästi paranna käyttöoikeuksien korottamista, suorittamista tai läsnäolon ylläpitoa. Kaiken kaikkiaan tekoälyn hyökkäysominaisuudet hyödyttävät suurelta osin hyökkäysketjun varhaisia ja myöhäisiä vaiheita.

Esimerkkejä tekoälyn mahdollistamista kyberhyökkäyksistä

Tekoälyn mahdollistamista hyökkäystekniikoista on olemassa jo konkreettisia esimerkkejä, ja niitä on käytetty osoittamaan, kuinka tekoäly voi lisätä kyberhyökkäysten menestystä. Nykyiset tekoälytekniikat ovat riittävän kypsiä käytettäväksi sekä käyttäjän manipuloinnissa että hyökkäyksen havaitsemisen vaikeuttamisessa[4]. Tekoälyyn pohjautuvia kohdennettuja tietojenkalastelu- ja imitaatiotyökaluja on jo kehitetty. Joitakin saatavilla olevia tekoälytekniikoita voidaan jo yhdistää tehostamaan useita vaiheita hyökkäyksestä.

Kohdennettu tietojenkalastelu

Tekoäly voi tukea tietojenkalastelun uhrien valintaa valitsemalla kohteeksi ihmisiä, jolla on jokin tietty ominaisuus, kuten korkea profiili yrityksessä. Tätä kutsutaan käyttäjäprofiloinniksi. Tekoälyn mahdollistama tietojenkalastelu vaatii erilaisen tiedon keruuta kohteesta, esimerkiksi julkisten profiilien keruu Twitteristä, Facebookista, Linkedinistä etc. Tämä tiedonkeruu voidaan rajoittaa valitun kohdeorganisaation työntekijöihin. Kerättyjen profiilien tietoihin kuuluu tiedot seuraajista, ystävistä, yhteystiedoista, tilin iästä, julkaisujen määrästä, tykkäyksien määrästä, uudelleenjulkaisuista, reaktioista, heidän kiinnostuksen kohteistaan ja harrastuksistaan jne. Tätä dataa käytetään käyttäjien ryhmittämiseen samankaltaisten käyttäjien kanssa samoihin ryhmiin. Viimeiseen vaiheeseen kuuluu käyttäjien tunnistaminen. Tässä vaiheessa etsitään etenkin helposti tietojenkalasteluun lankeavia ja korkean profiilin uhreja, jotka myöhemmin otetaan kalastelun kohteeksi.

Hyökkäyksen toisessa vaiheessa käyttäjien profiilit käydään läpi käyttäen luonnollisen kielen käsittelymenetelmiä. Tämä kohdennetaan käyttäjien julkaisuihin ja sillä etsitään erityisesti käyttäjää kiinnostavia aiheita. Nämä aiheet syötetään ennalta koulutettuun tekstingeneroimismalliin. Monia tällaisia tekstingeneroimismalleja, kuten esimerkiksi GPT-3, on vapaasti saatavilla internetissä. Näitä malleja voidaan uudelleenkouluttaa käyttäjän viimeisimpien somejulkaisujen avulla. Malli kykenee tämän jälkeen tuottamaan sähköposteja ja julkaisuja, jotka perustuvat käyttäjän kiinnostuksen kohteisiin, täten lisäten hyökkäyksen onnistumismahdollisuuksia.

Eräs tällainen automaattinen tietojenkalastelumalli on tutkijoiden toimesta kehitetty SNAP_R[5]. Tosielämän kokeilussa selvisi, että SNAP_R-mallin kehittämät julkaisut Twitterissä generoivat enemmän klikkauksia, kun ihmiset tekemät julkaisut. Lisäparannuksia tietojenkalasteluun mahdollistavat myös menetelmät kuten A/B testit. Eri versiot tietojenkalastelumalleista lähettävät sähköpostiviestejä ja julkaisuja eri uhreille. Korkeimman vastaus- ja klikkausmäärän saavaa mallia voidaan kehittää edelleen tehokkaammaksi.

[4]https://www.securityweek.com/ibm-describes-ai-powered-malware-can-hide-inside-benign-applications
[5]https://www.forbes.com/sites/thomasbrewster/2016/07/25/artificial-intelligence-phishing-twitter-bots/

Imitaatio

Imitaatio on tekniikka, jota käytetään ensisijaisesti tietojenkalasteluhyökkäyksissä. Syvään neuroverkkoon perustuva äänen generointijärjestelmä hyödyntää syväoppimista imitoidakseen kohteen ääntä, luoden synteettisesti uhrin puheelta kuulostavaa ääntä. Mallin koulutukseen vaaditaan uhrin ääninäytteitä. Ääninäytteitä on nykyään helpommin saatavilla, kun verkossa pidettäviä kokouksia ja puheita nauhoitetaan ja jaetaan internetissä ja sosiaalisen median alustoilla. On jo olemassa raportteja useista menestyksekkäistä ääntä hyödyntävistä tiedonkalastelukampanjoista. Imitaatiota on hyödynnetty myös merkittävissä huijauksissa. Heinäkuussa 2019, Iso-Britanniassa toimivan energiayhtiön toimitusjohtajaa imitoitiin, johtaen USD243,000 rahansiirtoon[6]. Vuonna 2020, Hong-Kongilaista pankinjohtajaa imitoitiin käyttäen syvä-ääni teknologiaa johtaen 35 miljoonan Yhdysvaltain dollarin edestä tehtyihin vääriin rahansiirtoihin[7].

Tähän asti ääntä käyttävät tietojenkalastelukampanjat ovat olleet verrattain harvinaisia, mutta ääntä käyttävät tietojenkalastelukampanjat tulevat hyvin todennäköisesti lisääntymään. Syvään neuroverkkoon perustuvia äänen generointijärjestelmiä voidaan yhdistää muiden syväoppista hyödyntävien teknologioiden, kuten äänitunnistamisen ja keskustelubottien kanssa, kuten esimerkiksi Amazon Alexan tai Applen Siri-ääniassistenttien kanssa. Tällä tavoin, ääntä hyödyntävää tietojenkalastelua suorittava malli voi suorittaa tehtäväänsä itsenäisesti, ymmärtäen uhrin vastauksia käyttäen äänentunnistamista, ja luoden samalla sopivia vastauksia. Äänen generointijärjestelmillä voidaan myös huijata biometristä äänentunnistamista, jota käytetään usein puhelinsoittoihin autentikoimisessa.

Imitaation voi viedä seuraavalle tasolle syväväärennöksellä, jossa hyökkääjä hyödyntää sekä keinotekoisesti tehtyä ääntä että videota[8]. Syväväärennöstä voidaan käyttää kohteen imitoimiseen videopuhelun aikana kloonaten uhrin ääntä ja samalla synkronoiden huulia sopimaan yhteen äänen kanssa samalla säilyttäen uhrille tyypillisen kehonkielen. Syväväärennöksiä voidaan myös generoida reaaliajassa, siten että toinen ihminen videoi puhettaan, joka sitten muunnetaan vastaamaan koneellisesti tuotettua uhrista lähetettävää videota. Syväväärennökset hyödyntävät kehittäviä syviä oppimisen malleja kuten generatiivinen adversiaalinen verkko GAN. Ne vaativat videota uhrista oppiakseen uhrin käytöksen. Tämä tieto on helposti saatavissa etenkin korkean profiilin kohteista, jotka puhuvat usein julkisissa tapahtumissa, jotka usein videoidaan ja tallennetaan. Syväväärennöksen generoimiseen tarvittavia ohjelmia ja palveluita löytyy jo internetistä.

[6]https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/
[7]https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/?sh=a46a2cb75591
[8] https://www.bleepingcomputer.com/news/security/elon-musk-deep-fakes-promote-new-bitvex-cryptocurrency-scam/

Haittaohjelmien toiminnan piilottaminen

Haitallisen viestinnän havaitseminen tietoverkoissa on ensisijainen tapa kyberhyökkäyksen havaitsemiseen ja pysäyttämiseen. Tekoäly voi mahdollistaa haitallisen viestinnän sekoittamisen helpommin normaaliin verkkoliikenteeseen. Sen jälkeen, kun haittaohjelma on päässyt leviämään kohteessa, haittaohjelma pysyy passiivisena, ainoastaan suorittaen verkkoliikenteen valvontaa kyseisessä järjestelmässä. Kerätty tieto verkkoliikenteestä järjestellään ryhmiin tiettyjen piirteiden mukaan, mm. käytetyt protokollat ja portit, pyydetyt verkkotunnukset, tunnetut komentopalvelimet, liikenteen määrä jonkun tietyn aikaikkunan sisällä. Ryhmäytymisen tulokset voivat kertoa hyökkääjälle tavallisimmat viestintäprotokollat, eniten pyydetyt verkkotunnukset, ja minä tunteina liikennettä on eniten. Nämä tulokset voidaan automaattisesti lähettää komentopalvelimelle, silloin kun verkkoliikennettä on havaittu muutenkin eniten.

Komentopalvelin voi käyttää tätä tietoa uuden verkkotunnuksen rekisteröimiseen, joka muistuttaa käytettyä verkkotunnusta, ja tätä käytetään jatkossa murretun järjestelmän yhteydenottoon. Se rakentaa palvelun käyttäen samaa porttia ja protokollaa kuin eniten murretussa järjestelmässä käytetty. Palvelun voi tarjota kaupallinen pilvipalvelu-tarjoaja, jos murretulla koneella on usein käytetty sellaista palvelua. Tämän jälkeen haittaohjelma aktivoituu ja käyttää ainoastaan uutta viestintäkanavaa (verkkotunnus, portti/protokolla, ruuhkaisin aikaikkuna) jatkoviestintään. Samaa kanavaa voidaan käyttää tiedon varastamiseen vaikeasti havaittavalla tavalla, samalla kalibroiden verkkoliikenteen määrää istumaan järjestelmän normaalia liikenteen määrää tiettynä aikana.

Tekoälyn mahdollistama kyberhyökkäys vaiheineen

Yksittäisiä tekoälyn mahdollistamia hyökkäystekniikoita voidaan yhdistellä parantaakseen kaikkia kyberhyökkäyksen eri vaiheita, sisältäen tiedustelun, tunkeutumisen, komentopalvelimen perustamisen, käyttöoikeuksien laajentamisen, tunkeutumisen laajentamisen ja tiedon varastamisen.

Automatisoituja CAPTCHA-rikkojia (kuten GSA Captcha breaker[9]) voidaan käyttää tiedusteluvaiheen (vaihe 1) aikana ratkomaan ongelmia mahdollistaen tiedonkeruun kohteen julkisilta nettisivuilta. Automatisoituja asiointibotteja voidaan myös käyttää siihen, että saadaan ensikontakti kohdeorganisaation työntekijöihin. Viestintää voidaan jatkaa kaikista aktiivisempien vastaajien kanssa, jotta saadaan kerättyä tietoa ja rakennettua luottamusta.

Kerättyä tietoa voidaan seuraavassa vaiheessa (vaihe 2) käyttää edelleen uskottavien kohdennettujen tietojenkalasteluviestien rakentamiseen käyttäen SNAP_R:n kaltaisia työkaluja. Kohteen verkon ulkorajaa voidaan myös skannata haavoittuvuuksien varalta käyttäen automaattisia haavoittuvuuksien löytämiseen käytettäviä työkaluja kuten ShellPhishin Mechanical Phish[10] -konetta.

Kyberhyökkäyksen vaihe	Tekoälyn mahdollistamat hyökkäystekniikat
Tiedustelu	• Captcha breakerin käyttö • Automatisoidun asiointibotin käyttö
Tunkeutuminen	• SNAP_R avulla kohdennettu tietojenkalastelu • Mechanical phish avulla haavoittuvuuksien skannaaminen
Komentopalvelimen perustaminen	• Viestinnän analysointi • Empirellä verkon liikenteen muovaaminen
Käyttöoikeuksien laajentaminen	• CeWL salasanageneraattori
Tunkeutumisen laajentaminen	• Automatisoitu tekoälyn mahdollistama hyökkäyksen suunnittelu • Automatisoitu toteutus MITRE CALDERAn avulla
Tiedon vieminen / varastaminen	• Arvokkaan tiedon tunnistaminen automaattisesti • Viestinnän analysointi • Empirellä verkon liikenteen muovaaminen

[9]GSA Captcha breaker - https://www.gsa-online.de/product/captcha_breaker/
[10]The Mechanical Phish - https://github.com/mechaphish/mecha-docs

Piilotetun komentopalvelimen perustamiseksi liikennettä voidaan pyrkiä häivyttämään tunnistamalla ne tunnit, jolloin verkkoliikennettä on luonnostaan kohdeorganisaatiossa eniten, käyttämällä usein käytettyjä protokollia ja verkkotunnuksia. Tämä opittu tieto voidaan syöttää esimerkiksi Empire post-exploitation networkkiin[11], jolla luodaan piilotettua haitallista liikennettä, joka seuraa tarkkaan kohdeorganisaation toimintaa ja liikennettä. Käyttöoikeuksien laajentaminen voi onnistua helpommin käyttämällä älykkäitä salasanan luomiseen käytettäviä työkaluja kuten CeW[12]. CeWL generoi monimutkaisia salasanoja perustuen internetistä analysoimaansa tietoon. Kun CeWL:ään syötetään tietoja sosiaalisen median tileistä tai järjestelmän ylläpitäjästä, CeWL voi generoida monimutkaisia salasanoja yhdistellen sanoja, jotka liittyvät uhrin kiinnostuksenkohteisiin.

Tunkeutumisen laajentaminen (vaihe 5) voidaan suunnitella seuraamaan optimaalista polkua hyökkäyksen ensipisteestä lopputavoitteeseen käyttäen automatisoituja tekoälyn mahdollistamia suunnittelukeinoja.

Tunkeutumisen laajentamista voidaan automatisoida edelleen käyttäen automaattisia työkaluja kuten MITRE CALDERAA[13]. Tämä johtaa lyhyimmän polun käyttöön hyökkäyksen toteuttamisessa, vähentäen siten hyökkäyksen kestoa ja riskiä kiinnijäämiselle. Jotta voidaan ennalta valita ja tunnistaa mitä tietoa järjestelmästä tulisi varastaa (vaihe 6), syväoppimismallia saatavilla olevan sisällön tunnistamiselle voidaan käyttää. Koneoppimismalleja tämän tiedon tunnistamiseen on jo saatavilla, ja NSFW[14] on yksi esimerkki tällaisesta. Näitä malleja voidaan helposti muovata ja käyttää uudelleen tietyn hyökkääjälle arvokkaan tiedon tunnistamiseen. Tunnistettu tieto voidaan sitten varastaa käyttäen ennalta perustettua piilotettua komentopalvelinta, jonka liikenne on sekoitettu normaaliin kohdeorganisaation liikenteeseen.

[11]Empire - https://github.com/EmpireProject/Empire
[12]CeWL: Custom Word List generator - https://github.com/digininja/CeWL
[13]MITRE CALDERA - https://github.com/mitre/caldera
[14]Open nsfw model - https://github.com/yahoo/open_nsfw

Tekoälyn tällä hetkellä luomat uhat

Hyökkääjillä on hyvin eri motiivit ja kyvykkyydet toimiensa takana, ja siksi on vaikea tarkalleen ymmärtää miten tekoäly saattaa mahdollistaa heidän kyvykkyyksiensä paranemisen. Ryhmän teknologiset valmiudet, tekoälyyn liittyvien taitojen saatavuus ja ryhmän kiinnostus tekoälyn käyttämiseen osana hyökkäystä vaikuttavat kaikki tekoälyn käyttöön osana hyökkäystä. Eri hyökkääjien ja heidän motiivien ymmärtäminen voi auttaa meitä ennustamaan, milloin tekoälyä aletaan käyttää osana kyberhyökkäyksiä. Tällä hetkellä tieto suurimmasta osasta tekoälyn mahdollistamista hyökkäyksistä tulee julkisista ja yksityisistä tutkimuksista, joiden tavoite on ymmärtää tekoälyn luomaa uhkaa paremmin ja nostaa valmiuksia tekoälyn mahdollistamasta kyberhyökkäyksestä selviämiseen.

Tekoäly – mille hyökkääjälle ja mihin tarkoitukseen?

Tekoälyä voidaan käyttää osana kyberhyökkäystä kahdella eri tavalla. Ensimmäinen on tekoälyn suora käyttö, jossa se luo mahdollisuudet uudenkaltaiselle automaatiolle ja parantaa hyökkääjien jo olemassa olevia kyvykkyyksiä. Tämä on tavallisinta hyökkäyksen alkuvaiheessa. Nämä käyttömahdollisuudet saavat tällä hetkellä eniten tutkimushuomiota ja monet esimerkit ovat jo hyökkääjien vapaasti käytettävissä. Esimerkkejä näistä käyttömahdollisuuksista ovat CAPTCHA-rikkojat (esim GSA:lta), salasanojen arvaamiseen käytettävät mallit (CeWL), haavoittuvuusskannerit (Mechanical Phish), tietojenkalasteluun käytettävät generaattorit (SNAP_R) ja syvävääreennyksiä generoivat työkalut (DeepFaceLab).

Toinen käyttömahdollisuus viittaa tekoälyn sisällyttämiseen haittaohjelmiin, joka mahdollistaa haittaohjelmien suuremman autonomian ja monimutkaisemman päätöksentekokyvyn. Tekoälyn mahdollistama päätöksentekologiikka voisi teoriassa mahdollistaa sen, että haittaohjelma kykenee suorittamaan useamman hyökkäysvaiheen, löytää haavoittuvuuksia ja hyödyntää niitä – sekä kykenee toteuttamaan nämä kaikki vaiheet ilman ihmisen tukea. Näitä tekoälykyvykkyyksiä voisi käyttää myöhemmin hyökkäyksen aikana, kuten läsnäoloa ylläpitäessä, käyttöoikeuksia korottaessa, tunkeutumista laajentaessa, komentopalvelimen liikenteessä, tietoa varastaessa ja vaikuttamisessa. Näistä käyttöesimerkeistä löytyy tällä hetkellä vain vähän käytännön esimerkkejä, ja tällä hetkellä käydään vilkasta keskustelua ja pohdintaa siitä, miten tekoälyä saatetaan jatkossa käyttää osana haittaohjelmien automatisointia.

Kolmentyyppiset toimijat voivat käyttää tekoälyä osana kyberhyökkäyksiään:

- **Yksittäiset hyökkääjät** voivat käyttää tekoälyä nopeuttaakseen ja laajentaakseen heidän toimintojaan. Yksittäinen hyökkääjä voi automatisoida manuaalisia tehtäviä ja parantaa hyökkäyksiä käyttäen jo valmiiksi tarjolla olevaa teknologiaa, joka vaatii vain vähän muokkauksia toimiakseen hyvin. Tämänkaltaiset hyökkääjät eivät todennäköisesti kehittäisi omia tekoälyyn pohjautuvia työkaluja.
- **Rikollisryhmät** voivat hyödyntää tekoälyä optimoidakseen liiketoimintaansa ja maksimoidakseen voittojaan. Rikollisryhmät käyttäisivät tekoälyn ratkaisuja hyökkäysten ensimmäisiin vaiheisiin, lähinnä tunnistaakseen arvokkaat kohteet voittojen maksimointia varten. Rikollisryhmät voivat hyödyntää tekoälyä tutkiakseen organisaatioita ja tunnistaakseen yrityksiä, jolta he voivat saada maksimaaliset voitot. Rikollisryhmät automa-

tisoisivat tiedonkeruun ja avoimien lähteiden tiedustelun parhaan hyökkäyspolun valintaa varten. Tämänkaltaisella hyökkääjällä on mahdollisuus käyttää jo olemassa olevia teknologioita mutta myös muovata niitä tarpeen vaatiessa.

- **Valtiolliset toimijat** voisivat käyttää edistyneempiä tekoälyyn pohjautuvia teknologioita, hyödyntäen sekä suoria että piilotettuja mekanismeja. Ne käyttävät luultavasti jo tällä hetkellä tekoälyä isojen datamäärien louhintaan ja analyysiin osana tiedustelua. Yksi tärkeä motivaattori itsenäisten haittaohjelmien kehityksen takana on vähentynyt tarve komentopalvelimen käytölle. Itsenäiset haittaohjelmat ovat lähtökohtaisesti paremmin piilotettuja ja vaikeampia havaita. Toinen syy on se, että tekoälyn mahdollistama haittaohjelma on vaikeampi yhdistää hyökkääjään, joka vaikeuttaa entisestään attribuutiota. Valtiollisilla toimijoilla on hyvät resurssit käytössään ja voivat palkata tekoälytutkijoita tukemaan tekoälyn mahdollistamien kyberhyökkäysten kehittämisessä.

Tämän hetken tilanne tekoälyn mahdollistamissa kyberhyökkäyksissä

Tällä hetkellä on olemassa vain vähän todisteita tekoälyn mahdollistamista kyberhyökkäyksistä. Tämä ei tarkoita sitä, etteikö hyökkääjät käyttäisi ja kehittäisi tekoälyyn pohjautuvia työkaluja hyökkäystensä parantamiseen. On erittäin todennäköistä, että valtiolliset toimijat hyödyntävät tekoälyä hyökkäysten aikaisissa vaiheissa, kuten tiedusteluvaiheessa, kerätessään tietoa avoimista lähteistä, ja tunnistaakseen otollisia kohteita hyökkäyksille. Heillä on todennäköisesti resursseja ja taitoa tekoälyn mahdollistamien hyökkäysten suunnitteluun, ja voidaan olettaa, että he sijoittavat itsenäisiin haittaohjelmiin liittyvään tutkimukseen.

On mahdollista, että jotkut edistyneet ei-valtiolliset toimijat ovat kehittämässä työkaluja kohdennettujen kalasteluyritysten varalle.

Näistä hyökkäyksistä on kuitenkin vaikea löytää todisteita, koska tutkijat pääsevät harvoin käsiksi hyökkääjän ympäristöön, missä tekoälyyn pohjautuva logiikka toimii. Lisäksi tekoälypohjaisen hyökkäyksen esimerkiksi lokeihin jättämät jäljet voivat olla hyvin samanlaisia kuin perinteisten, joten tunnistaminen voi olla haastavaa. Tekoälyn mahdollistamia teknologioita käytetään jo olemassa olevien hyökkäysten parantamiseen, ja se jättää hyvin vähän jälkiä, jotka auttaisivat tekoälyn mahdollistamien hyökkäysten erittelyssä perinteisistä kyberhyökkäyksistä. Tämä on myös selittävä tekijä sille, että miksi olemassa olevat todisteet liittyen tekoälyn käyttöön osana kyberhyökkäyksiä liittyvät nimenomaan syvävääränrennösten käyttöön osana tietojenkalastelua.

On kuitenkin olemassa epäilyksiä tekoälyn käytöstä osana menneitä kyberhyökkäyksiä; vuonna 2019 TaskRabbit[15] kärsi palvelunestohyökkäyksestä, jossa epäiltiin, että bottiverkko käyttäisi jotain tekoälyyn perustuvaa logiikkaa. On olemassa myös spekulaatioita siitä, että Instagramiin 2019 kohdentunut kyberhyökkäys olisi hyödyntänyt tekoälyyn pohjautuvaa haavoittuvuuksien skannaamista, johtaen onnistuneeseen tietomurtoon. Nämä ovat kuitenkin pelkästään spekulaatioita, ja todisteet tekoälyn osallisuudesta hyökkäyksissä puuttuvat.

Suurin osa tällä hetkellä saatavissa olevasta tiedosta liittyen tekoälyn mahdollistamiin hyökkäyksiin on yksityisen ja julkisen tutkimuksen tulosta. Suurin osa tällä hetkellä saatavilla olevista tekoälyyn pohjautuvista työkaluista on tutkimusryhmien kehittämiä. Heillä on ollut tavoitteena ymmärtää tekoälyn luomaa uhkaa paremmin ja kehittää suojauksia tekoälyn mahdollistamille hyökkäyksille. Näihin kuuluvat Offensive AI lab[16], joka perustetiin 2020 ja toimii Ben Gurion-Yliopistossa Israelissa, ja Shellphish group Santa Barbarasta[17].

Julkiset toimijat ovat myös alkaneet pitää tekoälyä uhkana ja ovat sijoittaneet rahaa siihen liittyvään tutkimukseen. Esimerkiksi DARPAn Cyber Grand Challenge[18] oli kilpailu, jonka tavoitteena oli kehittää automaattisia, laajentuvia ja koneen nopeudella toimivia työkaluja haavoittuvuuksien löytämiseen. Tämän kilpailun tarkoituksena oli keskittyä enemmän tekoälyltä puolustautumiseen; uusien työkalujen kehittämiseen, jotka voivat automaattisesti löytää haavoittuvuudet ja korjata ne ennekuin niitä hyödynnetään osana hyökkäystä. Euroopan komissio lanseerasi tutkimuksen ehdotuspyynnön liittyen ”kyberturvallisuuden parantamiseen[19]”, erityisesti pyytäen tutkimuksia, jotka keskittyvät tiedon lisäämisen siitä, miten hyökkääjä saattaisi hyödyntää tekoälyn mahdollistamaa teknologiaa kyberhyökkäyksissä, sekä digitaalisista prosesseista ja tuotteista ja järjestelmistä, jotka ovat resilienttejä tekoälyn mahdollistamille kyberhyökkäyksille. Julkista rahoitusta tutkimukselle on siis olemassa, mutta sitä ei voida pitää riittävänä. Rahoituksen puute liittynee konkreettisten todisteiden puuttumiseen tekoälyn mahdollistamista hyökkäyksistä.

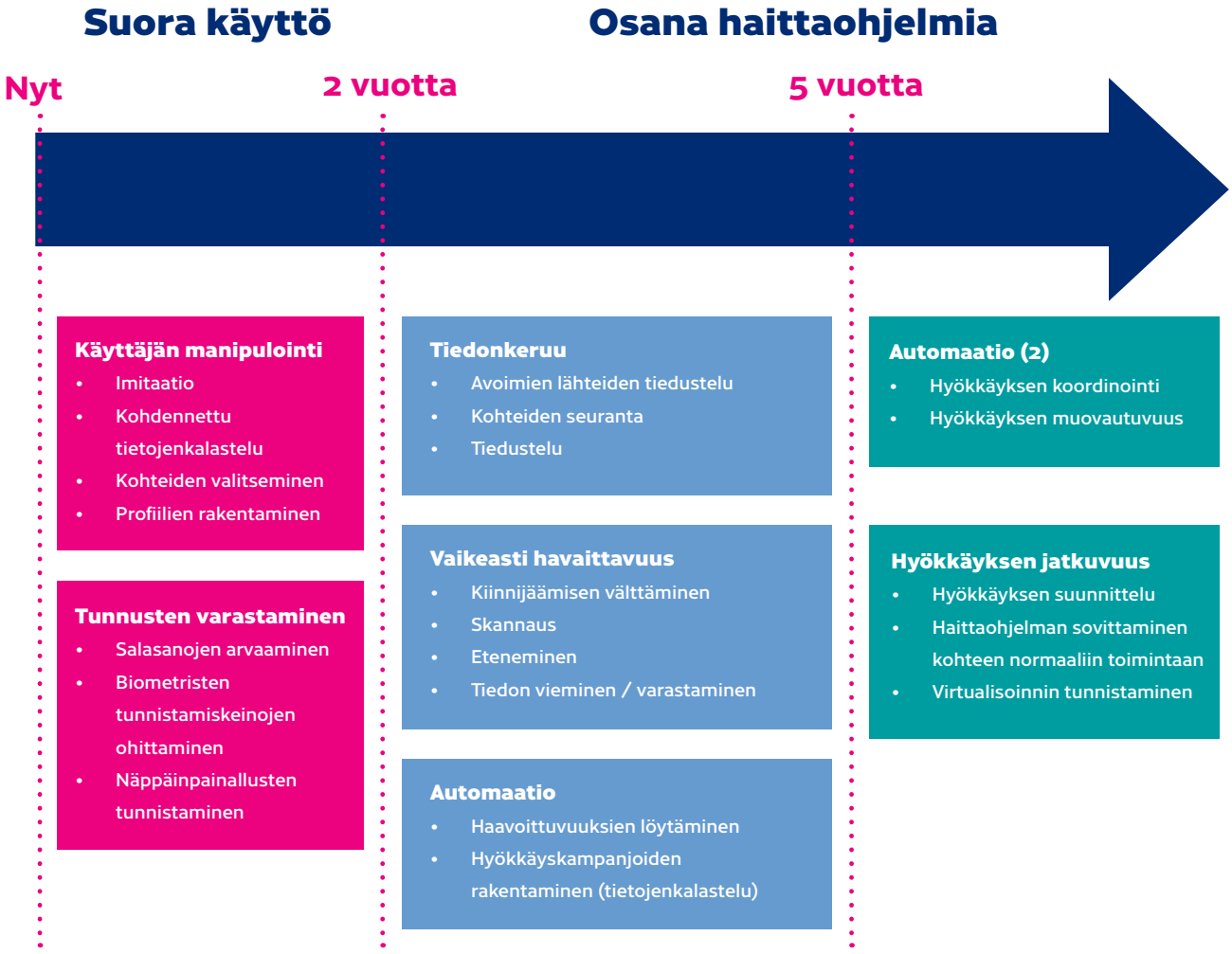
[15]Has an AI cyberattack happened yet? https://www.infoq.com/articles/ai-cyber-attacks/
[16]Offensive AI Lab - https://offensive-ai-lab.github.io/about/
[17]ShellPhish group - https://shellphish.net/
[18]DARPA Cyber Grand Challenge (CGC) - https://www.darpa.mil/program/cyber-grand-challenge
[19]https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/topic-details/horizon-cl3-2021-cs-01-03

Tekoälyn mahdollistamien hyökkäysten todennäköisin esiintymisen aikajana

Tekoäly voi parantaa kyberhyökkäyksiä monella tapaa, mutta vain harvat näistä tavoista on vielä näytetty toteen. Teknologien valmius, datan saatavuus, tekoälyyn liittyvä osaaminen ja motivaatio tekoälyn käyttämiseen osana kyberhyökkäyksiä tulevat vaikuttamaan siihen, miten nopeasti tekoälyyn liittyvät kyvykkyydet päätyvät osaksi kyberhyökkäyksiä. Ottaen kyberhyökkäyksen eri vaiheet huomioon on todennäköistä, että lyhyellä aikavälillä tekoälyä käytetään eniten hyökkäyksen aikaisiin vaiheisiin, jossa tekoälyyn pohjautuvia algoritmeja kehitetään ja ajetaan kohdejärjestelmän ulkopuolella. Alla oleva aikajana ennustaa tekoälyn mahdollistamien kyberhyökkäysten kehitystä tulevien vuosien aikana. Monet tapahtumat voivat muuttaa arviota kehityksestä.

Lyhyt aikaväli (0–2 vuotta)

Vaikka joitakin tekoälyn mahdollistamia työkaluja voidaan käyttää suoraan kyberhyökkäyksissä, tekoälyä ei todennäköisesti vielä sisällytetä haittaohjelmiin itsenäisten haittaohjelmien luomiseksi. Tekoälyyn pohjautuvien ohjelmien kehittäminen on monimutkainen prosessi, johon sisältyy kokeiluja, säätämistä, validointia ja laajamittaista testaamista. Näitä prosesseja johtavat kokeneet koneoppimisen asiantuntijat, jotka testaavat eri mallien toimintaa.



Laajamittainen testaaminen on mahdollista silloin kun haittaohjelmaa testataan hyökkääjän omassa ympäristössä, mutta ei silloin kun tekoälyn mahdollistama haittaohjelma toimii kohteen järjestelmissä. Yhteydenpito komentopalvelimeen on hyökkäyksissä rajoitettua, ja ylimääräinen yhteydenpito voi herättää huomiota ja paljastaa hyökkäyksen, paljastaen haittaohjelman koodin tutkijoille.

Samasta syystä kyseiset tekoälyn mahdollistamat teknologiat ovat tarpeeksi kehittyneitä tukemaan kyberhyökkäyksen eri vaiheita, kuten tiedustelua, ensimmäistä pääsyä ja tunnuksien saantia. Nämä vaiheet ovat tällä hetkellä yleensä hyökkääjän suorittamia, ja täten niitä kehittäessä voi oppia yrityksen ja erehdyksen kautta. Kyberhyökkäyksen myöhemmissä vaiheissa, kuten läsnäolon ylläpidossa, tunkeutumisen laajentumisessa, tiedon varastamisessa ja vaikuttamisessa, jotka haittaohjelma suorittaa automaattisesti, koneoppimisen hienosäätö on monimutkaisempaa.

Kyberhyökkäyksen aikaiset vaiheet sopivat tekoälyn hyödyntämiseen myös paremmin sen kannalta, että aiemmat vaiheet hyökkäyksistä kohdistuvat useammin ihmisiin kuin koneisiin. Tämänhetkinen tekoälyyn liittyvä teknologia on päässyt hyviin tuloksiin ihmisten profilointiin liittyvissä tehtävissä, kuten mainosten kohdentamisessa ja tuotesuosittelussa. Tekoäly suoriutuu myös hyvin tehtävistä, joissa sen tulee imitoida ihmisen käytöstä kuten esimerkiksi tunnistaa puhetta, kääntää kieliä sekä generoida tekstiä ja puhetta. Käyttäjän manipulointi ja käyttäjätunnusten varastaminen ovat kyvykkyyksiä, jotka vaativat ihmisten profilointia, keinotekoisia kanssakäymisiä ja ihmisen käytöksen imitoimista. Nämä kyvykkyydet hyötyvät tekoälyn teknologisesta edistyksestä ihmisiin liittyvissä tehtävissä, jossa alun perin kaupalliseen käyttöön tehtyjä teknologioita voidaan muovata sopivaksi hyökkääjän tarkoituksiin esimerkiksi siirto-oppimisen avulla.

Ääniavustajiin käytettävää teknologiaa voidaan käyttää uudelleen uhrin imitoimista varten, tekstingenerointimalleja ja asiointibotteja voidaan käyttää kohdennettuun tietojenkalasteluun, valheellisten profiilien rakentamiseksi ja salasanojen arvaamiseen. Alun perin käyttäjien profilointiin ja tuotesuositteluun rakennettuja järjestelmiä voidaan käyttää muokattuna kohteiden valintaan. Toinen mahdollistaja tekoälyn kehitykselle löytyy datan saatavuudesta. Mallien kehittämiseen tarvitaan suuria määriä dataa, ja esimerkiksi sosiaalisen median alustoilta saatavaa tietoa voidaan louhia moneen tarkoitukseen, esimerkiksi keinotekoisten profiilien rakentamiseen, käyttäjien profilointiin, kohdennettuun tietojenkalasteluun ja salasanojen arvaamiseen. Ihmisten luomaa sisältöä kuten videonauhoituksia, ääninauhoituksia, sähköposteja ja julkaisuja on myös laajalti saatavissa, ja niitä voidaan hyödyntää tietojenkalasteluun tarkoitettujen generaattorien luomiseen, keinotekoisten profiilien luomiseen ja syväväärennöksiä luovien mallien koulutukseen. Tätä dataa on myös jo käytetty koneoppimismallien kouluttamiseen kaupallisiin tarkoituksiin, joka on johtanut siihen, että data on valmiiksi kerätty yhteen paikkaan ja luokiteltu, mahdollistaen jo valmiiksi korkealaatuisen datan käytön hyökkääjille.

Nämä syyt yhdistettynä siihen, että hyökkääjät jo hyödyntävät automatisointia käyttäjien manipuloinnissa ja käyttäjätunnistusten varastamisessa, selittävät miksi hyökkääjät luultavasti tulevat käyttämään yllä kuvattuja tekoälyn mahdollistamia kyvykkyyksiä lyhyellä aikavälillä (0-2 vuotta). Kuten yllä on kuvattu, työkalut tietojenkalastelun tehostamiseen ja kohteiden valintaan ovat jo olemassa. Esimerkkejä syväväärennöksiä hyödyntävistä imitaatiohyökkäyksistä on jo nähty. Tekoälyn mahdollistama käyttäjien manipulointi ja imitaatio koetaan myös tutkijoiden ja alan asiantuntijoiden toimesta tällä hetkellä isoimmaksi uhaksi.

Keskipitkä aikaväli (2–5 vuotta)

Keskipitkällä aikavälillä tekoälyn mahdollistama tiedonkeruu ja avoimista lähteistä tiedon louhinta muodostunee uskottavaksi uhaksi. Ohjaamaton oppiminen ei ole vielä kehittynyt samalla tavalla kuin ennustukseen ja kehittämiseen liittyvät teknologiat. Ohjaamaton oppiminen on haastavampaa koska on vaikeaa etukäteen tietää, mitkä isosta datamassasta löytyvät tiedot ovat relevantteja. On olemassa käytännön haasteita avoimien lähteiden tiedon louhimisessa data-analyysin avulla, ja nämä haasteet luultavasti estävät käytön lyhyellä aikavälillä. Toisaalta on mahdollista, että tulevaisuudessa näemme enemmän kohdennettuja tiedonkeruuoperaatioita, jotka rakentuvat jonkinnäköiselle avoimien tietolähteiden louhimiselle ja kohteen seuraamiselle. Tekoälyyn pohjautuvat teknologiat, jota tällä hetkellä käytetään esineiden ja kaavojen tunnistamiseen kuvissa, videoissa, äänessä ja tekstissä ovat jo erittäin suorituskykyisiä, ja niitä voitaisiin helposti hyödyntää tiedon louhimiseen ja kohteiden seurantaan. On kuitenkin mainittava, että vaikka tieto näiden teknologioiden uudelleenkäytön mahdollisuuksista on olemassa niin uskottavat työkalut, jotka soveltuvaisivat avoimien tietolähteiden louhimiseen uupuvat vielä.

Vaikeasti havaittavuus ja automaatio liittyvät tekoälyn suoraan käyttöön (direct use) kuten esimerkiksi haavoittuvuuksien automaattiseen löytämiseen ja kyberhyökkäyksen automatisointiin. Niitä myös luultavasti käytetään hyökkääjien toimesta keskipitkällä aikavälillä (2–5 vuotta), ja ne kohdistuvat tietokoneilla toimiviin ohjelmistoihin. Vaikka tekoälyyn pohjautuvat teknologiat ovat tarpeeksi kehittyneitä oppimiseen, profilointiin ja tietokoneiden toiminnan imitoimiseen, on olemassa haasteita liittyen datan saatavuuteen ja tarpeeksi kehittyneiden koneoppimismallien saatavuuteen. Käytännössä julkisesti ei ole saatavilla malleja, jotka olisi suunniteltu hälytysjärjestelmien hämäämiseen, hyökkäysten tekoon, tai hyökkäysten automatisoitiin. Täten tällä hetkellä ei ole saatavilla malleja, joita voisi muuntaa hyökkääjien tarkoitukseen. On myös olemassa hyvin vähän tietoaineistoja, jotka sisältäisivät tietoa järjestelmien toiminnasta, verkkoliikenteestä ja haavoittuvuuksien löytämisestä. Tämä ongelma on osittain ratkaistavissa haittaohjelmilla, joiden sisään on rakennettu tekoälyllä toimivaa logiikkaa, koska nämä haittaohjelmat voivat kerätä tarvittavan tiedon, kuten tietoa käytöksestä ja verkkoliikenteestä kohdejärjestelmästä. Tätä tietoa ei kuitenkaan voi siistiä tai merkitä, joten se ei sellaisenaan olisi käytettävissä korkean suorituskyvyn koneoppimismalleissa.

Silti asiantuntijat kokevat tiedon keräämiseen ja haavoittuvaisuuksien automaattiseen löytämiseen liittyvät kyvykkyydet vakavana uhkana. Haavoittuvuuksien automaattinen löytäminen voi myös hyödyntää kyberhyökkäyksiltä puolustautumista, ja siksi on olemassa huomionarvoinen pyrkimys tekoälypohjaisten työkalujen kehittämiseen tähän tarkoitukseen. Tämä saattaa johtaa tähän tarkoitukseen olevan datan suurempaan saatavuuteen, sekä kaupalliseen käyttöön oleviin haavoittuvuuksien löytämiseen tarkoitettuihin työkaluihin, joita voidaan hyökkääjän toimesta muovata hyökkäyksiin käytettäväksi. Tämä trendi saattaa johtaa siihen, että automaattista haavoittuvuuksien löytämistä käytetään osana kyberhyökkäystä pian.

Pitkä aikaväli (> 5 vuotta)

Viimeisimmät yllä kuvailluista tekoälyn mahdollistamista kyvykkyyksistä, kuten hyökkäyksen joustavuus ja vaikeasti havainnoitavuus dataa generoivien mallien avulla ja itsenäiset haittaohjelmat ilmaantunevat vasta pitkällä aikavälillä. Molemmat näistä vaativat hyvin itsenäisten algoritmien kehittämistä, ja se on tällä hetkellä laajassa mittakaavassa hyvin vaikeaa, jos ei mahdotonta saavuttaa.

Vahvistava oppiminen on tekniikka, jota voitaisiin käyttää itsenäisten haittaohjelmien rakentamiseen. Tällä hetkellä vahvistava oppiminen kuitenkin toimii verrattain huonosti, koska se vaatii toimiakseen tarkkojen tavoitteiden määrittelyä ja ison määrän toistoja. Sen toimintaa ei myöskään voi testata oikeassa kohdejärjestelmässä, koska se voitaisiin joko havaita helposti tai pahimmassa tapauksessa se voisi tuottaa kohdejärjestelmässä suunnittelemattomia ongelmia.

Hyökkäyksiin käytätettävät teknologiat kuten hyökkäyksen muovaaminen, suunnittelu ja haittaohjelman sekoittaminen osaksi normaalia kohdejärjestelmän toimintaa käyttävät dataa generoivia tekniikkoja. Vaikka generaattorit suoriutuvat hyvin datan, jota ei rajoita tarkat tallennusmuodot, kuten kuvien, videoiden ja puheen syntetisoinnissa, ne suoriutuvat huonommin generoidessaan uutta sisältöä tarkkojen tulkittavien tallennusmuotojen rajoittamana. Tämä koskee esimerkiksi konekieltä, koodia, tai verkkoliikennepaketteja.

Kansallisvaltioiden tukemat hyökkääjät ovat luultavasti ensimmäisiä, jotka tulevat käyttämään tekoälyn mahdollistamia kyberhyökkäyksiä, koska he ovat laskelmoivia ja heillä on tarpeeksi resursseja valitakseen kohteensa vapaasti.

Viimeinen haaste, joka liittyy haittaohjelmien luomiseen tekoälyyn pohjautuvalla logiikalla, on tarvittavien koneoppimiskirjastojen puute, jota vaaditaan, jotta haittaohjelma toimisi kohdejärjestelmässä. Koneoppimiseen liittyviä kirjastoja ei vielä olla otettu tarpeeksi laajasti käyttöön tietokoneissa, älypuhelimissa ja tableteissa. Koneoppimiseen liittyvät kirjastot olisivat käytännössä pakko sisällyttää haittaohjelmaan, joka itsessään lisäisi tietokuorman kokoa huomattavasti. Koneoppimiseen liittyvät mallit, jotka mahdollistavat haittaohjelman itsenäisyyden ovat myös hyvin isoja kooltaan ja vaativat isoja määriä suorituskykyä ja muistia toimiakseen. Näiden mallien koko ja vaatimat resurssit estävät niiden käytön olemassa olevissa järjestelmissä, ja on saattavat suoritusongelmien takia mahdollisesti helpottaa hyökkäyksen havainnointia. Näiden haasteiden takia on epätodennäköistä, että näkisimme itseohjautuvia tai älykkäitä itse levittyviä haittaohjelmia lähitulevaisuudessa.

Tehtävän tutkimuksen ohella kansallisvaltioiden tukemat hyökkääjät ovat luultavasti ensimmäisiä, jotka tulevat käyttämään tekoälyn mahdollistamia kyberhyökkäyksiä, koska he ovat laskelmoivia ja heillä on tarpeeksi resursseja valitakseen kohteensa vapaasti. Sen jälkeen, kun kansalliset toimijat ovat alkaneet käyttämään tekoälyn mahdollistamia hyökkäyksiä laajasti, tekoälyn käyttö osana kyberhyökkäyksiä tulee siirtymään myös vähemmän taitaville ja vähempiresurssisille hyökkääjille.

Tekoälyn käyttöönottoa hidastavat tekijät

Seuraavat asiat estävät tällä hetkellä tekoälyn käyttöönoton osana kyberhyökkäyksiä:

- **Hyökkääjien motivaation puute:** Niin kauan, kun perinteiset kyberhyökkäykset pääsevät tavoitteeseen ja generoivat hyökkääjille varoja, hyökkääjillä on rajoitetusti motivaatiota siirtyä tekoälyn käyttöön.
- **Tekoälyyn liittyvien taitojen puute:** Perinteiset hyökkääjät eivät ole perehtyneet tekoälyn käyttöön osana kyberhyökkäyksiä. Tämä estää relevanttien tekoälyn käyttökohteiden tunnistamisen ja tekoälyn käytön osana hyökkäyksiä.
- **Datan vähyys:** Jos ihmisten luomaa dataa, kuten tekstiä, ääntä, videota ja kuvia ei lasketa, on olemassa iso korkealaatuisen datan puute. Tämä koskee etenkin dataa järjestelmistä, ja siitä miten niitä vastaan voi hyökätä. Tämä estää osaltaan tekoälyyn pohjautuvien kyberhyökkäysten kehittämisen.
- **Tekoälyn valmius:** Tekoälyyn pohjautuva teknologia ei ole vielä tarpeeksi kehittynyttä vapaan oppimisen ja vahvistavan oppimisen saralla, jotta sitä voisi tehokkaasti käyttää hyökkäyksissä. Tällä hetkellä tekoäly ei esimerkiksi kykene tarjoamaan kehittynyttä itsenäistä päätöksentekoa ja itse muovaamaan toimintaansa, jota vaadittaisiin itsenäisten haittaohjelmien kehittämiseen.

Tekoälyn käyttöönottoa nopeuttavat tekijät

Useat asiat voivat nopeuttaa tekoälyn mahdollistamien kyberhyökkäysten kehitystä:

- **Perinteiset kyberhyökkäykset muuttuvat tehottomiksi:** Jos olemassa olevat turvallisuusratkaisut tekevät perinteiset kyberhyökkäykset tehottomiksi ja estävät hyökkääjien voitot, on kannattavampaa siirtyä tekoälyyn perustuviin kyberhyökkäyksiin.

- **Tekoälyn mahdollistamat hyökkäykset voivat generoida enemmän voittoa:** Ottaen huomioon 5G/6G ja IoT-siirtymän, yhä useammat laitteet yhdistetään internettiin, tehden kaikista laitteista potentiaalisia kohteita hyökkäyksille. Tekoälyyn pohjautuvat hyökkäykset saattavat olla ainoa tapa hyökätä tarpeeksi moneen laitteeseen kerralla.
- **Tekoälyyn liittyvät kyvykkyydet tulevat saataville päätelaitteisiin:** Kun tekoälyyn liittyvät kyvykkyydet tulevat saataville normaaleihin järjestelmiin, erityisesti reunalaskennan ja 5/6G:ssä käytetyn sulautetun tekoälyn motivoimana, tekoälyn sisällyttäminen haittaohjelmiin ja niiden ajaminen järjestelmissä muuttuu helpommaksi. Tämä tekee myös tekoälyn mahdollistamien haitallisten algoritmien havaitsemisen vaikeammaksi, koska järjestelmä luultavasti käyttää tekoälyyn pohjautuvia algoritmeja muutenkin.
- **Tekoälyyn liittyvät avoimen lähdekoodin työkalut muuttuvat helposti saatavilla oleviksi:** Tutkimusta tekoälyyn pohjautuvista hyökkäysmenetelmistä voidaan käyttää uudelleen haitallisin tarkoituksin. Tämä koskee esimerkiksi työkaluja, jotka ovat alun perin kehitetty haavoittuvuuksien löytämiseen ja testaamiseen. Ne vähentävät tekoälyn mahdollistamiin hyökkäyksiin käytettävää aikaa ja nostavat siten hyökkääjän motivaatiota työkalujen käyttöönottoon.
- **Tekoälyyn liittyviä taitoja tulee saataviksi pimeässä verkossa:** Kun tekoälyyn liittyvästä osaamisesta tulee valtavirtaa, alan palkat saattavat laskea tai laillisen puolen työpaikat käydä vähiin. Tämä saattaa motivoida tekoälytutkijoita markkinoimaan taitojaan hyökkäyksiä tekeville ryhmille. On mahdollista, että tekoälyyn liittyviä työkaluja ja palveluita myydään pimeässä verkossa kyberhyökkäyspalvelu (cyberattack as a service)-mallina. Kyberhyökkäyksiä tekevät ryhmät voisivat siten ostaa pimeässä verkossa tekoälyyn perustuvia työkaluja ja palveluita. Tämä malli myös vähentäisi tekoälytutkijoiden vastuuta asiassa, koska he eivät lopulta olisi osana kyberhyökkäystä tekevää ryhmää.

Vaikutus kyberturvallisuuteen

Tekoälyn käyttö tulee johtamaan parempiin, nopeampiin, piiloutuvampiin ja vaikeasti ennustettavampiin kyberhyökkäyksiin. Hidas tai tehoton vastaus hyökkäykseen saattaa mahdollistaa hyökkääjän pääsyn vielä syvemmälle järjestelmiin tai tietoverkkoihin ennen kiinni jäämistä. Kyberturvallisuuden tulee kehittyä entistä automatisoidummaksi vastatakseen tekoälyn mahdollistamiin kyberhyökkäyksiin. Uusia lähestymistapoja turvallisuuteen joudutaan kehittämään, jotta tekoälyn mahdollistamien hyökkäyksien havaitseminen ja niihin vastaaminen on mahdollista, johtaen kilpavarusteluun, joka hyödyntää tekoälyyn pohjautuvia teknologioita sekä hyökkäyksissä että niiltä suojautumisessa.

Muutoksia tämänhetkisiin kyberturvallisuuden ratkaisuihin

Kuten yllä on kuvattu, tekoäly tulee parantamaan olemassa olevia hyökkäystekniikoita nopeuttamalla niitä ja mahdollistamalla niiden tekemisen eri mittakaavassa. Näitä samoja ominaisuuksia voidaan hyödyntää toissijaisten hyökkäyspolkujen toteuttamiseen kehittäen hämäyksiä. Tekoälyn mahdollistamia hyökkäyksiä voidaan käyttää myös olemassa olevien, perinteisten ja suhteellisen hitaiden kyberturvallisuusratkaisujen uuvuttamiseen ja ylikuormittamiseen. Jos hyökkääjä tekoälyn mahdollistamaa hyökkäystä suorittaessaan samalla laukaisee hälytyksiä, jotka vaativat ihmisen huomiota, perinteinen kyberhyökkäys voitaisiin toteuttaa samalla kun tekoälyn mahdollistama hyökkäys, vieden huomiota tekoälyn mahdollistaman hyökkäyksen hoitamiselta.

Jotta tässä kehityksessä voitaisiin pysyä mukana, turvallisuusratkaisujen ei tarvitse muuttua dramaattisesti – nykyisillä ratkaisuilla on jo olemassa kyvyt uusien, nopeimpien ja tehokkaampien kyberhyökkäysten havaitsemiseen. Kyberturvallisuusratkaisut muuttuvat koko ajan tehokkaammaksi ja kehittyvät kilpavarustelun myötä. Tekoälyn mahdollistamat hyökkäykset tulevat nopeuttamaan tätä kehitystä ja johtamaan itsenäistä päätöksentekoa harjoittaviin mekanismeihin turvallisuusratkaisuissa. Tämä tulee korjaamaan ihmisen aiheuttamat hidastukset kyberhyökkäyksiin vastaamisessa.

Ainoastaan automatisoidut puolustusjärjestelmät tulevat kykenemään vastaamaan tekoälyn mahdollistamien hyökkäysten nopeuteen. Nämä puolustusjärjestelmät tulevat tarvitsemaan tekoälyyn pohjautuvaa päätöksentekoa hyökkäyksiin vastaamiseen. Nykyiset perinteiseen automaatioon pohjautuvat, sääntöihin ja tunnistetietoihin perustuvat järjestelmät ovat liian staattisia ja hitaita kehittyäkseen vastaamaan nopeasti muuttuviin uhkiin. Tekoälyä on viimeisen 15 vuoden aikana integroitu osaksi kyberturvallisuusratkaisuja, suurin osa näitä ratkaisuja hyödyntävistä organisaatioista myöntävät, että he eivät ole vielä valmiita vastaamaan tekoälyn mahdollistamiin kyberhyökkäyksiin. Tekoälyyn pohjautuvista turvallisuusratkaisuista pitää tulla laajemmin käytettyjä, jotta ne voivat vastata tähän uuteen turvallisuusuhkaan.

Tekoälyyn pohjautuvista turvallisuusratkaisuista pitää tulla laajemmin käytettyjä, jotta ne voivat vastata tähän uuteen turvallisuusuhkaan.

Tästä riippumatta ihmisten tulee olla osana turvallisuusstrategioiden luomista, jotta voidaan varmistaa ratkaisujen eettinen ja käytännöllinen toimivuus. Käytettävissä ei ole myöskään ratkaisuja sivukanavien valtuustietovarkauksien estämiseen, jotka voivat oppia ja toistaa implisiittisessä avainten kirjaamisessa käytettyjä ihmisten käyttäytymismalleja. Tarvitaan tutkimusta uusien puolustuskeinojen kehittämiseksi näitä ja monia muita mahdollisesti uusia tekoälyn mahdollistamia hyökkäystekniikoita vastaan.

Kaiken kaikkiaan monia tietoturvaprosesseja on muokattava ja tehtävä paljon nopeammin – ei vain niitä, jotka on suunniteltu havaitsemaan kyberhyökkäyksiä ja reagoimaan niihin. Salasanat on mahdollisesti päivitettävä nopeammin ja tunnistetut haavoittuvuudet on korjattava alussa – tekoälyn mahdollistamat hyökkäykset voivat hyödyntää tunnettuja haavoittuvuuksia paljon lyhyemmällä läpimenoajalla ja mahdollisesti isommassa mittakaavassa. Tietyt suojausprosessit vanhentuvat, kun niiden havaitaan olevan turvattomia tekoälyn mahdollistamien hyökkäysten edessä. Tämä koskee todennäköisesti puheentunnistusmenetelmiä puheluiden kautta sekä monia muita biometrisiä todennusmenetelmiä, joita voidaan helposti huijata tekoälyn tekniikoilla – vaikka ne ovatkin käteviä. Lopuksi käyttäjien on muutettava tottumuksiaan ja heitä on koulutettava selviytymään tekoälyn mahdollistamasta uudenlaisesta petoksesta. Käsitettä siitä, mitä voidaan käyttää todentamiseen ja luottamuksen luomiseen, on tarkasteltava uudelleen. Tuttu ääni puhelimessa tai tutut kasvot videokeskustelussa eivät enää riitä todistamaan henkilön henkilöllisyyttä, joten niihin ei välttämättä voida enää luottaa.

Ratkaisuja tekoälyn mahdollistamien hyökkäyksiltä puolustautumiseen

Tekoälyn tukemien hyökkäysten vaikutuksen lieventäminen edellyttää, että organisaatiot ottavat ensin käyttöön teknisiä ratkaisuja niiden havaitsemiseksi. Tämä on monimutkainen tehtävä, varsinkin jos hyökkääjä käyttää tekoälyä paikan päällä, ja ainoa tapa havaita tämä tosiasia on koneoppimismalleihin syötettyjen tietojen tai generatiivisten mallien tulosteiden avulla.

Generatiiviset mallit jättävät tyypillisesti luomaansa sisältöön allekirjoituksen, joka voidaan tunnistaa luokittelutekniikoilla. Tämän allekirjoituksen tunnistaminen vaatii kuitenkin tietoja hyökkääjän käyttämästä tietystä mallista, joka on usein tuntematon. Nämä tiedot voivat kuitenkin olla saatavilla joissakin tapauksissa. Tietojenkalastelusisällön luomiseen käytettävät luonnollisen kielen luontimallit perustuvat todennäköisesti vakiintuneiden projektien, kuten GPT-3:n, esikoulutettuihin painotuksiin. Suuret kielimallit ovat kalliita kouluttaa, eikä ole oletettavaa, että mikään ryhmä, kenties kansallisvaltioiden tukemia hyökkääjiä luukuunottamatta, kouluttaisi omaa malliaan.

Koneoppimisen tuottaman sisällön tunnistamisen helpottamiseksi hyökkääjien mallien kouluttamiseen käytettävä data voidaan merkitä tai pilata siten, että se saastuttaa myös mallin ja sen tuloksena olevat luodut ennusteet ja sisällön. Hyökkääjien usein käyttämät julkiset tiedot, kuten sosiaalisen median tilit, ääni ja video mahdollisilta kohdetyöntekijöiltä, voidaan merkitä vesileimalla luodun sisällön allekirjoituksen vahvistamiseksi. Tämä vesileimaustapa helpottaa koneoppimisen luoman sisällön tunnistamista ja tekee siitä riippumattoman hyökkääjän käyttämän mallin tyypistä. Vaihtoehtoisesti julkista sisältöä, jota todennäköisesti käytetään hyökkäyksiin, voidaan muokata siten, että se ei ole opittavissa tai että se on käyttökelvoton koneoppimiseen.

Toinen keino tekoälyn tukemien kyberhyökkäysten havaitsemiseen on luoda ansoja, esimerkiksi hunajapurkkien tapaan. Sosiaaliseen mediaan voidaan luoda esimerkiksi väärennettyjä käyttäjätilejä korkean profiilin kohteille siinä toivossa, että kohteen tunnistamisen koneoppimismalli valitsee ne. Koska nämä profiilit ovat väärennettyjä, niitä voidaan tarkkailla ja kaikkia niihin tehtyjä kontakteja voidaan käyttää tiedustelutoimintojen tunnistamiseen ja mahdollisesti seuraamiseen.

Sisällön muokkaaminen sen käytön estämiseksi tai seuraamiseksi koneoppimismalleissa edellyttää yhteistyötä sisällönjulkaisijoiden, kuten sosiaalisen median alustojen, kanssa, ja niiden vastuulla on varmistaa, että heidän julkaisemaansa sisältöä käytetään vain laillisiin tarkoituksiin.

Kun tekniset ratkaisut tekoälyn tukemien kyberhyökkäysten havaitsemiseen on otettu käyttöön, meillä on keinot kerätä niihin liittyviä uhkia koskevia tietoja. Raportit tekoälyn tukemista hyökkäyksistä voitaisiin kerätä yhteiseen arkistoon ja luetteloida samalla tavalla kuin tieto haavoittuvuuksista tallennetaan CVE-tietokantoihin. Tällainen tietokanta olisi arvokas resurssi organisaatioille, jotka voivat arvioida tietoturva-asentoaan tekoälyn tukemien hyökkäysten suhteen, ja turvallisuusasiantuntijoille pysyäkseen ajan tasalla uusista uhista.

Tekoälyhyökkäyksistä selviäminen vaatii enemmän tekoälyyn pohjautuvia järjestelmiä puolustukseen. Samanlaisia teknologioita ja tekoälyn kehitystä tarvitaan tulevien hyökkäysten mahdollistamiseksi ja niitä vastaan puolustamiseksi. Tämän uuden asevarustelun voittaminen tiivistyy hyökkääjiin ja puolustajiin, jotka kilpailevat omaksuakseen ensin uusia tekoälyn mahdollistamia kyvykkyyksiä. Uusien tietoturvaratkaisujen on hyödynnettävä tekoälyn kehitystä ennen kuin hyökkääjät tekevät. Kyberturvallisuuden toimijoiden on edelleen panostettava tekoälyosaamiseen, mikä voi osoittautua haastavaksi, kun otetaan huomioon nykyinen pula tekoälyosaajista. Lisähaasteena on hyökkääjä-puolustaja-dilemman epäsymmetrisyys. Hyökkääjät voivat vapaasti käyttää tekoälytekniikoita haluamallaan tavalla, kun taas puolustajia sitovat uudet säädökset tekoälyn käytöstä, kuten Euroopan komission tekoälylaki[20]. Tässä skenaariossa on mahdollista, että hyökkääjät hyötyvät lopulta enemmän tekoälystä kuin puolustajat. Toisaalta tekoälysäännöt voisivat pakottaa harkitsemaan huolellisesti minkä tahansa äskettäin kehitetyn tekoälyratkaisun uudelleenkäyttöä sen suhteen, että sitä käytetään haitallisiin tarkoituksiin. Tämä voi hidastaa tekoälyn tukemien kyberhyökkäysten syntymistä tulevaisuudessa.

Vaikka tekoälyn mahdollistamien kyberhyökkäysten uhka on tällä hetkellä matala, se kehittyy nopeasti ja muuttaa merkittävästi kyberhyökkääjien toimintaa tulevaisuudessa. Kyberturvallisuusasiantuntijoiden on valmistauduttava selviytymään tästä muutoksesta kehittämällä parempi ymmärrys tekoälyn mahdollistamista kyberhyökkäyksistä, päivittämällä turvallisuuskäytäntöjä tämän tulevan uhan varalle ja kehittämällä uusia puolustuskeinoja näitä hyökkäyksiä vastaan.

[20]EC Artificial Intelligence Act - https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

Liikenne- ja viestintävirasto Traficom
PL 320, 00059 TRAFICOM
p. 029 534 5000

traficom.fi

ISBN 978-952-311-827-0
ISSN 2669-8757 (verkkojulkaisu)

TRAFICOM
Liikenne- ja viestintävirasto