DOI: 10.38146/BSZ-AJIA.2024.v72.i7.pp1187-1205

Az intelligens fenyegetés Hogyan veszélyeztetheti a mesterséges intelligencia a biztonságunkat?

The Intelligent Threat: How Artificial Intelligence Can Compromise Our Security

Tóth Levente
Dr. PhD, szakmai igazgató, tanársegéd
TVT Vagyonvédelmi Zrt.
Nemzeti Közszolgálati Egyetem,
Rendészettudományi Kar
toth.levente@uni-nke.hu

Absztrakt

Cél: A tanulmány elkészítésének célja, hogy felhívja a mesterséges intelligencia használatának veszélyeire a figyelmet.
Módszertan: A releváns szakirodalmi forráselemzés mellett a szerző példák felsorolásával mutatja be a mesterséges intelligencia biztonságunkat veszélyeztető aspektusait, érintve a már meglévő és a készülőben lévő jogszabályi környezetet.
Megállapítások: A mesterséges intelligencia mind közvetve, mind közvetlenül veszélyeztetheti biztonságunkat. Az általa hordozott veszélyek, valamint a technológia jelenlegi gyors fejlődése még sürgetőbbé teszi a megfelelő és alkalmazkodó folyamatos szabályozást annak érdekében, hogy egyre növekvő felhasználása minimális negatív következményekkel járjon.
Érték: A tanulmány új, eddig indokolatlanul mellőzött, biztonságot veszélyeztető sajátosságokat vizsgál. Megállapításai hozzásegíthetnek annak megértéséhez, hogy a mesterséges intelligencia mind szűkebb, mind tágabb társadalmi szinten is veszélyeztetheti biztonságunkat.

Kulcsszavak: mesterséges intelligencia, morfizálás, okos otthon, arcfelismerés

A szerző a kéziratot magyar nyelven nyújtotta be. Benyújtás: 2024. 04. 21. Elfogadás: 2024. 05. 18.

Abstract

Aim: The aim of the study is to draw attention to the dangers of using artificial intelligence.
Methodology: Alongside a relevant literature review, the author illustrates the aspects of artificial intelligence jeopardizing our security by providing examples and addresses the existing and evolving regulatory environment.
Findings: Artificial intelligence can directly or indirectly pose a threat to our security. The risks associated with artificial intelligence, coupled with the current rapid technological advancement, make it imperative to establish appropriate and adaptive continuous regulations to ensure the increasing use of AI comes with minimal negative consequences.
Value: The study explores previously overlooked features that compromise security. Its findings can contribute to understanding how artificial intelligence can endanger our security on both narrower and broader societal levels.

Keywords: artificial intelligence, morphing, smart home, facial recognition

Bevezetés

A mesterséges intelligencia (a továbbiakban: MI) napjaink egyik legmeghatározóbb technológiai fejlődési területe. Már régóta nem csupán egy tudományos kísérlet, hanem mindannyiunk mindennapi életét meghatározó tényezőjévé vált. Az MI hatalmas előnyökkel jár, amelyek kiterjednek az otthoni felhasználásra, a vállalati szektorra és az iparágak széles skálájára. Azok a feladatok, amelyek korábban hosszú órákat vettek igénybe, vagy magas emberi hibarátával jártak, most gépi folyamatok segítségével gyorsan és pontosan elvégezhetők. Az MI technológiájának dinamikus fejlődése és egyre széleskörűbb alkalmazása kulcsszerepet játszik modern társadalmunk és gazdaságunk átalakulásában. Viszont nem szabad figyelmen kívül hagynunk az MI alkalmazásával járó, emberi lét biztonságát közvetve és közvetlenül veszélyeztető tényezőket, amelyek sokkal ritkábban kerülnek szóba.

Mi az MI?

Az MI olyan számítástechnikai tudományterület, amely az emberi gondolkodás és döntéshozatal mechanizmusait igyekszik modellezni és szimulálni gépi rendszerek segítségével. Az MI célja, hogy olyan intelligens viselkedést imitáljon,

amely hasonló az emberi tevékenységhez. Ezekbe a képességekbe beletartozik a tanulás, a problémamegoldó és döntéshozatalt követelő tevékenységek, a vizuális észlelés, és akár a beszédfelismerés is (Russel & Norvig, 2010). A gépi tanulás (Machine Learning) az MI egyik ága, amely olyan algoritmusok kifejlesztésével foglalkozik, amelyek az adatokból „tanulnak", és a teljesítményük a megismert adatok nagyságával arányosan növekszik. A gépi tanuló algoritmusok nem programozhatók konkrét feladatok elvégzésére, hanem adatalapokon való kiképzéssel tanulják meg, hogyan kell elvégezni a feladatokat, vagyis implicit szabályokat tanulnak meg számos, az adatbázisban található példából. Azaz az adatokat arra használjuk, hogy megalapozott becsléseket készítsünk a jövőről. Rendvédelmi aspektusból vizsgálva jó példa erre az úgynevezett „prediktív rendfenntartás", ahol az MI-t például a bűncselekmények előre jelzett kockázatának értékelésére használják (URL1). Hasonló rendszerrel rendelkezik például a holland rendőrség. Az úgynevezett Bűn Előrejelzési Rendszer (Crime Anticipation System) feldolgozza a bűnügyi mintákat, és megjósolja, hol és mikor valószínű a bűncselekmények bekövetkezése. Ennek az eredményeként a megfigyelési és megelőző tevékenységeket az előre jelzett kockázathoz igazíthatják (Oosterloo & van Schie, 2018).

A mélytanulás (Deep Learning) a gépi tanulás egy olyan területe, amely mesterséges neurális hálózatokon alapul. A neurális hálózat olyan számítógépes modell, amelynek létrehozását az emberi agy biológiai neurális hálózata inspirálta. A neurális hálózatok általában egymással összekapcsolt neuronokból állnak. A mély tanulás egyik legfontosabb eleme a többrétegű neurális hálózatok használata. Ezek a hálózatok több, egymásra épülő réteget tartalmaznak, ahol minden egyes réteg különböző matematikai műveleteket hajt végre a bemeneti adatokon. Ezen rétegek során az adatok egyre magasabb szintű absztrakciókat képeznek, ami lehetővé teszi, hogy a rendszer összetett feladatokat is megoldjon. Ezek közé tartozik a képfelismerés, beszédfelismerés, természetes nyelvfeldolgozás, autonóm járművek vezérlése, az orvostudományban segítenek a diagnosztizálásban, az ipari gyártásban pedig hatékonyabbá teszik a minőségellenőrzést. Ezek mellett az arcfelismerő technológiákban is kulcsszerepet játszanak, amelyek nemcsak a biztonsági rendszerekben, hanem a felhasználói szoftverekben is egyre inkább teret nyernek (például automatikus képcímkézés). Az ilyen gépi tanulási modellek folyamatosan fejlődnek, és ahogy egyre több adaton tanulnak, úgy válnak egyre hatékonyabbá és megbízhatóbbá az adott feladatok elvégzésében.

Az elmúlt években az MI evolúciója odáig jutott, hogy a létrehozott MI-rendszerek objektumokat és szövegeket ismernek fel, valós idejű döntéseket hoznak, és emberi interakciót utánoznak bizonyos szinten. Képesek továbbá az emberi beszédminták megértésére, és adekvát válaszadásra is, ezáltal merőben új

dimenziókat nyitva az emberi élet sok területén. Az MI-t használó algoritmusok és gépi tanulási módszerek segítségével előrejelzések készíthetők a fogyasztói preferenciákról és piaci trendekről. A felhasználói profilok és viselkedésminták elemzése alapján az MI releváns és érdekes ajánlásokat és tartalmakat szolgáltathat az egyén számára. Ez az online kereskedelemtől kezdve a zenehallgatásig és a tartalomszolgáltatásig terjedő területeken nyújt kivételes előnyt. Továbbá az MI lehetővé teszi a gyors döntéshozatalt is. A hatalmas mennyiségű adat elemzése és értelmezése rövid idő alatt történhet meg, ami kritikus fontosságú a vállalati döntéshozatalban és a válságkezelésben.

A választás lehetősége

A *New York Times* 2023 tavaszán közölt egy figyelemre méltó cikket, *You Can Have the Blue Pill or the Red Pill, and We're Out* címmel, amelyben a szerzők az MI veszélyeiről és lehetőségeiről fejtették ki véleményüket (URL2). A cikk címe a kultikus *Mátrix* filmre utal, amelyben a főhősnek, Neonak egy kritikus döntést kell meghoznia: válassza a kék pirulát, és maradjon a megszokott, de illuzórikus világban, vagy döntsön a piros mellett, és szembesüljön a fájdalmas igazsággal. A cikkben alkalmazott metafora tökéletesen illeszkedik az MI jelenlegi helyzetéhez. A kék tabletta azokat az embereket jelképezi, akik elfogadják a jelenlegi állapotot, és azt gondolják, hogy az MI csupán egy hasznos eszköz, amely nem igényel különösebb kritikai szemléletet vagy elővigyázatosságot. Ezzel szemben a piros tabletta azoknak az embereknek az álláspontját képviseli, akik felismerik az MI által rejtett veszélyeket, és készen állnak arra, hogy intézkedéseket hozzanak azok felszámolására. A szerzők szerint az MI nem csupán egy újabb technológiai fejlesztés. Sokkal inkább az emberiség történetének egyik legnagyobb kihívását és lehetőségét jelenti. Az egészségügytől kezdve az oktatáson át a fenntarthatóságig számos területen képes javítani az életminőséget, és megoldani olyan problémákat, amelyekkel korábban képtelenek voltunk megbirkózni. Az optimális útvonal megtalálása érdekében a szerzők arra ösztökélnek mindenkit, hogy válasszuk a piros tablettát, de legyünk tájékozottak, kritikusak és proaktívak az MI kapcsán felmerülő kihívásokkal és lehetőségekkel szemben. Felmerül a kérdés, hogy vajon készek vagyunk-e szembenézni az MI rejtette kihívásokkal, és kiaknázni annak lehetőségeit, vagy inkább választjuk a kényelmes tudatlanságot, és hagyjuk, hogy mások döntsenek helyettünk. Az emberiség jövője talán pont ettől a döntéstől függ.

Az MI fejlődése a társadalom átalakulásának egy új korszakát indította el, amelynek következményeit még nem tudjuk teljes mértékben előre látni. Mit jelentene

az emberek számára egy olyan világban élni, ahol a blogbejegyzések, történetek, újságcikkek, könyvek, képek, dallamok, törvények és eszközök nagy százalékát nem emberi intelligencia alkotja, és ez az MI emberfeletti hatékonysággal rendelkezik a humán elme gyengeségei, elfogultságai és függőségei kihasználásában, miközben képes lenne akár egy bensőséges kapcsolatot is kialakítani az emberrel. Egy ilyen világban a hatás teljes mértékben kiszámíthatatlan lenne. Bizonyos területeken, mint például a stratégiai játékok (Sakk, Go, Shogi), az MI már rég túlszárnyalta az emberi képességeket (URL3). Ez a fejlődés azonban nemcsak a játékokra korlátozódik, hanem kiterjedhet más területekre is, mint például a művészet, a politika vagy a vallás. Ez potenciálisan a hatalmi dinamika és a döntéshozatali folyamatok megváltozásához vezethet, valamint befolyásolhatja az emberi társadalom szerkezetét. A művészet területén a nem emberi intelligencia olyan alkotásokat hozhat létre, melyek mélyen érintik az emberi érzelmeket, sőt akár versenyezhetnek is az ember által alkotott művészettel, esetlegesen túlszárnyalva azt. A politika terén az úgynevezett mély hamisítás[1] (Deep Fake) technológia segítségével lehetőség nyílik olyan dezinformációk terjesztésére, amelyek célja a kiszemelt egyén megalázása, zaklatása, zsarolása, vagy imázsának és hitelességének rombolása. Ezek mellett felhasználható piaci szereplők destabilizálására, ezáltal a pénzügyi piacok megzavarására, valamint akár a társadalmi elégedetlenség szítására is. Hasonlóképpen a vallás területén, a nem emberi intelligencia átformálhatja a spirituális narratívákat és gyakorlatokat, potenciálisan globális szinten megváltoztatva a hitrendszereket. Szintén a veszélyek közé tartozik az emberi autonómia és önrendelkezés potenciális elvesztésének lehetősége. Ezért kiemelten fontos, hogy kollektív kulturális, társadalmi és etikai kereteinket úgy alakítsuk ki, hogy azok támogassák az emberi autonómia és önrendelkezés megőrzését ebben az új környezetben. A társadalmi párbeszéd és együttműködés kiemelkedő fontosságú lesz az ilyen jellegű változások kezelése során.

A felsoroltakból jól érződik, hogy az emberi életre jelentős hatást gyakorló nem emberi intelligencia világában számos fontos kérdés merül fel, melyek etikai, filozófiai és gyakorlati szempontból egyaránt megfontolásra késztetnek bennünket. Hogyan alkalmazkodna az emberiség egy ilyen paradigmaváltáshoz, és hogyan alakulna az emberi és a nem emberi intelligencia közötti kapcsolat?

Az önálló döntéshozatal és felelősség kérdése az MI-rendszerek területén szintén komoly kihívást jelent. Az MI-rendszerek egyre összetettebbé válnak, és olyan feladatok elvégzésére képesek, amelyek korábban kizárólag emberi

1 A mélyhamisítás kifejezés olyan hamis hang- és/vagy vizuális tartalomra utal, amelyet MI segítségével manipuláltak vagy hoztak létre, és az eredetitől való megkülönböztetés az emberek és a gépek számára is nehézséget okoz.

tevékenységek voltak. Ennek következtében felmerül a kérdés, hogy ki viseli a felelősséget az esetleges hibákért vagy károkért, amelyek az MI által hozott döntések eredményeként bekövetkezhetnek. Komoly aggodalomra ad okot, hogy az MI által hozott döntések mögötti folyamatokat és logikát gyakran nehéz megérteni és ellenőrizni. Ezáltal az emberek számára kihívást jelenthet az MI által hozott döntések teljes körű ellenőrzése és a felelősségvállalás elhatárolása.

Az MI veszélyeit tovább növeli az adatvédelem és a magánélet védelmének kérdése. Az MI-rendszerek működéséhez hatalmas mennyiségű adatra van szükség, amelyek gyakran személyes adatokat is tartalmazhatnak. Az MI technológiák fejlődése során gyűjtött adatok mennyisége és érzékenysége miatt fokozott figyelmet kell fordítani a személyes adatok védelmére. Az adatvédelmi szabályozásoknak, mint például az EU Általános Adatvédelmi Rendeletének (GDPR), kell megfelelni az MI-rendszerek tervezésekor és használatakor. A jogosulatlan hozzáférés, az adatokkal való visszaélés, vagy az adatlopás veszélyei miatt fontos, hogy az MI-rendszereket úgy alakítsák ki, hogy azok tiszteletben tartsák a felhasználók magánéletét, és biztosítsák az adatok biztonságát. Ez magában foglalja a biztonságos adattárolást, a hozzáférés korlátozását és a felhasználók tájékoztatását arról, hogyan és milyen célból használják fel adataikat.

Az MI veszélyeinek egy másik fontos aspektusa a társadalmi egyenlőtlenségek növekedése lehet. Az MI-rendszerek működése és döntéshozatala nagyban függ azoktól az adatoktól, mintáktól, amelyekkel tanítják és tréningezik őket. Ha ezek az adatok torzítottak vagy diszkriminatívak, akkor az MI által hozott döntések is torzíthatnak és diszkriminatívak lehetnek, és kialakulhat az úgynevezett algoritmikus diszkrimináció.[2] Ez pedig hozzájárulhat a társadalmi egyenlőtlenségek növekedéséhez.

Az MI veszélyeire való reagálásnak és ezek kezelésének fontossága egyre nő a technológia fejlődésével. Fontos, hogy az MI fejlesztése és alkalmazása során figyelembe vegyük ezeket a veszélyforrásokat, és olyan megoldásokat találjunk, amelyek minimalizálják ezeket a kockázatokat. Az etikai irányelvek betartása és az átláthatóság növelése kulcsfontosságúak az MI felelős és fenntartható fejlődése szempontjából. Ugyanakkor az MI potenciális előnyeit kihasználva megfelelő óvatossággal és felelősséggel kell eljárnunk annak érdekében, hogy az MI fejlődése az emberiség számára előnyös legyen, és ne veszélyeztesse az emberi értékeket és jogokat.

2 Algoritmikus diszkrimináció alatt azt értjük, amikor egy automatikus döntési rendszer eredménye következtében, bizonyos tulajdonságok (például bőrszín, nem, életkor, vallás, szexuális beállítottság stb.) alapján egyéneknek vagy emberek csoportjainak eltérő bánásmódban van részük.

A fentebb ismertetett veszélyeken túl az MI számos olyan közvetlen biztonságot fenyegető felhasználási lehetőséget is rejt magában, amely veszélyezteti a személyes biztonságunkat, és „csak" áttételes hatással van az emberi közösségekre.

Okosotthonok

Az okosotthonok, vagy más néven intelligens otthonok a modern technológia és a háztartási környezet konvergenciáját jelentik, ahol a különböző eszközök és rendszerek összekapcsolása által növelik a felhasználók kényelmét, biztonságát és energiagazdálkodásának hatékonyságát. Az okosotthonok központi egysége egy intelligens vezérlőrendszer, ami lehet egy dedikált lokális központi szerver, vagy egy széleskörűen elérhető felhő platform. Az okosotthon-rendszerek magukban foglalhatják a világítás, a fűtés, a légkondicionálás, a biztonsági kamerák, zárak, riasztók, szórakoztató elektronikai és háztartási berendezések automatizált vezérlését. Ezek az eszközök gyakran vezeték nélküli technológiákon keresztül kommunikálnak egymással és a központi vezérlővel, vagy épp a felhő szolgáltatással, lehetővé téve a felhasználó számára, hogy távolról is irányíthassa otthonát. Az okosotthonok intelligenciája nem abban rejlik, hogy távoli vezérlést tesznek lehetővé, hanem abban, hogy képesek tanulni a használójuk szokásait és preferenciáit. Például egy okostermosztát „megfigyelheti" a lakók napi rutinját, és aszerint szabályozhatja a hőmérsékletet, hogy mikor vannak otthon és mikor távoznak. Az okosvilágítás-rendszerek képesek a természetes fényviszonyokhoz igazodni, csökkentve ezzel az energiafogyasztást és növelve a komfortérzetet. A biztonság terén az okosotthonok olyan funkciókat kínálnak, mint a mozgásérzékelőkkel összekapcsolt kamerák, amelyek riasztást küldhetnek a tulajdonosnak, ha szokatlan aktivitást érzékelnek. Az okoszárak lehetővé teszik a kulcs nélküli belépést, és ez távolról is felügyelhető vagy vezérelhető, így például lehetővé válik karbantartó személyzet bizonyos területekhez történő felügyelt hozzáférése. Az energiagazdálkodás terén az okoseszközök segítenek optimalizálni az energiafelhasználást és csökkenteni az elektromos számlát. Az okoskészülékek például képesek arra, hogy csak akkor kapcsoljanak be, amikor az energia ára alacsonyabb, vagy ki tudják használni a megújuló energiaforrásokat, mint például a napelemeket. Mindezek mellett az okosotthonok fontos szerepet játszanak az idősek és mozgáskorlátozottak életében is, a segítség nélküli, önálló otthoni élet lehetőségét biztosítva számukra. Hangvezérlés és automatizált rutinok segítségével egyszerűsíthetik mindennapi tevékenységeiket és javíthatják életminőségüket. Összességében az okosotthonok a jövő irányát mutatják az otthoni életvitelben, ahol az intelligens technológia segít

az energiahatékonyság növelésében, a biztonság fokozásában és általánosságban az életminőség javításában. Azonban az így megvalósított kényelemért a nem megfelelő védelem kialakítása esetén súlyos árat fizethetünk. Mivel az okosotthon-rendszerek informatikai hálózathoz kapcsolódnak, fennáll a veszélye a külső hackertámadásoknak. Az okosotthon-eszközök gyűjthetnek és továbbíthatnak különböző adatokat a felhőbe vagy más eszközök számára. Ezek az adatok tartalmazhatnak személyes információkat, amelyeket illetéktelenek megpróbálhatnak megszerezni vagy felhasználni. A veszélyek felsorolása itt még nem ért véget.

Az MI fejlődése évek óta lehetővé teszi, hogy hangutasításokkal vezéreljünk különböző eszközöket. Ez, az úgynevezett beszédfelismerés nem egyszerű feladat. A beszédfelismerés az MI azon területe, amely a beszélt emberi nyelv észlelésével, elemzésével és értelmezésével foglalkozik. Ez magában foglalja a szavak és mondatok megkülönböztetését és szöveggé alakítását (speech-to-text), vagy éppen fordítva, a szöveg beszéddé alakítását. Napról napra egyre elterjedtebbek a hanggal aktiválható eszközök. Világszerte több mint 300 millió okoshangszóró működik (URL4). Az okostelefonokba épített hangalapú asszisztenciával ez a szám négymilliárd fölé emelkedik. Jelentős piaci részesedést birtokolnak olyan népszerű platformok, mint a Google Assistant, az Apple Siri, az Amazon Alexa, a Samsung Bixby, vagy a kínai Baidu, Alibaba, illetve Tencent rendszerei. A digitális asszisztensek mára már olyanok, mint egy digitális személyi segéd, aki emlékeztet a napi feladatokra, időjárási előrejelzéseket nyújt, és személyre szabott információkat ad. A felhasználónak csak annyit kell tennie, hogy világos szóbeli parancsot ad. A beszédfelismerő technológia átalakítja a beszédet szöveggé, majd a természetes nyelvi feldolgozás értelmezi az írott információkat, és meghatározza, hogy milyen műveletre van szükség. Míg az emberek képesek voltak beszélni és beszédet értelmezni már az írás feltalálása előtt is, a számítógépek számára viszont az írott nyelv feldolgozása könnyebb, mint a kimondott szóé. A beszédfelismerés lényegesen komplexebb feladat a beszélt nyelv változékonysága és a hangfolyamok esetleges zaja miatt. A szavak kiválasztása, azonosítása és átalakítása olyan típusú szöveggé, amelyet a számítógép képes feldolgozni, rendkívül nagy kihívást jelent. Amikor beszélünk, az általunk kiadott hangok nem különülnek el különálló szavakra. Amit a számítógép érzékel, nagyon hasonlít ahhoz, amit egy személy hall, amikor egy olyan nyelvet hallgat, amely teljesen ismeretlen számára. Ilyenkor szinte egy folyamatos hangfolyamot hallunk, amiben az egyes szavakat nagyon nehéz megkülönböztetni. A beszédfelismerés problémája tehát alapvetően különbözik az írott nyelv vagy a képek értelmezésétől. A gépi látással és a természetes nyelvi feldolgozással ellentétben a beszédfelismerés

egyetlen bemeneti változó – hanghullámok – feldolgozását foglalja magában, amely időben dinamikusan változik. A nagy kihívás a szavak és mondatok megkülönböztetése ezen a bemeneten belül, hogy lefordíthatók legyenek egy olyan nyelvre, amelyet az algoritmus képes feldolgozni. További kihívást jelent, hogy a beszéd jelentésének egy részét a hangerő, a ritmus és a hangszín változásai közvetítik, melyek a beszélt nyelv jellemzői. A hatékony értelmezés tehát több, mint a szavak egyszerű megkülönböztetése. A fonetikai szempontokat is fel kell tárni, és értelmezni kell, hogy meg tudjuk határozni a mondanivaló jelentését. Egy másik buktató a homofonok, azaz az olyan szavak, amelyek ugyanúgy hangzanak, de különböző dolgokat jelentenek, mint például a csuklya és a csukja, vagy a fogjuk és a foglyuk. Értelmezésük a kontextustól függ. Mind a mondat szűk, mind a helyzet tágabb kontextusától. Más területekhez hasonlóan a gépi tanulás fejlődése a beszédfelismerés területén is előrelépéshez vezetett, mivel lehetővé vált sokkal nagyobb mennyiségű beszédadat feldolgozása az algoritmusok betanításához. A beszéd-szöveg és szöveg-beszéd átalakítás viszonylag sikeres gyakorlati alkalmazásai ma már életképesek, feltéve, hogy a beszéd mind hallási, mind tartalmi szempontból egzakt. Sokszor azonban a szlengekkel teleszőtt szóbeli kommunikáció az MI számára nem egyértelmű. Következésképpen a beszédfelismerő technológia még nem érte el azt a szintet, hogy megbízhatóan használható legyen széles körben és minden célra. Azonban megfelelően artikulált mondatokat és utasításokat az MI kiválóan értelmez. A kényelem növelése érdekében ezek a hangutasításokat fogadó és értelmező eszközök összekapcsolódhatnak az okosotthonok központi vezérlőivel, és megfelelő parancsszavak kimondásával különböző eszközök aktiválhatók, kapcsolhatók vagy jelenetek futtathatók. Azonban ilyen rendszerek ellen számos támadási lehetőség nyílik.

Az elmúlt években a kutatók által végzett számos kísérlet egyértelműen bizonyította, hogy rejtett, emberi fül számára érzékelhetetlen parancsokat lehet küldeni a különböző hangalapú vezérlő rendszereknek. Az egyik ilyen támadási módszer, amikor az ultrahang tartományába transzponálják át a vezérlő parancsokat, így az utasítások az emberi fül számára nem hallhatók, de az okosotthonok hangvezérlési rendszerei által használt mikrofonok által érzékelhetők (Zhang et al., 2017). A kísérlet során a kutatóknak sikerült titokban aktiválniuk és vezérelniük a hangasszisztens-rendszereket okostelefonokon és okoshangszórókon keresztül. Ennek következtében az eszközök telefonszámokat tárcsáztak vagy weboldalakat nyitottak meg anélkül, hogy az emberi fül észlelte volna ezt a folyamatot. Ilyen technikai lehetőségek rosszindulatú szándékkal bíró személyek kezében komoly kockázatot jelenthetnek, hiszen akár pénz átutalására vagy online vásárlásra is felhasználhatók. Emellett okosotthonokkal

összekapcsolva különböző típusú parancsokat lehet kiadni, például világítást be- vagy kikapcsolni, ajtókat nyitni, zárni, a termosztát hőmérsékletét megváltoztatni, vagy egyéb biztonsági rendszereket megkerülni.

Teljesen más módszert használtak a bochumi Horst Görtz IT Biztonsági Intézet informatikai szakértői (Schönherr, Kohls, Zeiler, & Holz, 2018). A kísérleteikben az emberi hallás pszichoakusztikus modelljét használják ki. Amíg a fül és az agy egy adott frekvencián lévő hang feldolgozásával van elfoglalva, néhány milliszekundumig nem képes más, alacsony hangerejű hangokat hallani. Ezeken a részeken rejtik el a kutatók a gépek titkos parancsait, amely kiegészítő információk az emberi fül számára véletlenszerű statikus zajnak tűnnek. A támadási módszer alapját az úgynevezett adversarial example (ellenséges vagy kontradiktórius példa) adja. A kontradiktórius példák speciális bemenetek, amelyeket azzal a céllal hoznak létre, hogy összezavarjanak egy neurális hálózatot, ami egy adott minta hibás osztályozását eredményezi. Ezek a manipulált bemeneti minták emberi fül vagy szem számára észrevétlenek, de a gépi tanulási modell működését befolyásolják és hamis előrejelzéseket adnak. Képek esetében úgy is mondhatjuk, hogy a gépi tanulás számára létrehozott optikai csalódás. Ezzel el is jutottunk a képi manipulációs technikákhoz.

Arcfelismerő rendszerek

Az arcfelismerés egy MI-n alapuló számítógépes technológia, amelyet az emberi arcok megtalálására és azonosítására használnak digitális fényképeken, vagy élő, illetve rögzített videóképeken. Az arcfelismerő technológiát gyakran használják az emberek valós idejű megfigyelésére és nyomon követésére. Első lépésként az arcfelismerő rendszer az arcok detektálásával kezdi a folyamatot. Ehhez különböző algoritmusokat használhatunk, mint például a Viola–Jones,[3] vagy más mély tanuláson alapuló arcdetektáló algoritmusok. Az arcfelismerés során az algoritmusok olyan pontokat azonosítanak az arcokon, amelyek egyedi jellemzőként szolgálnak. Ezeket a pontokat általában arcfelismerési pontoknak nevezik, melyek segítségével az algoritmusok képesek az azonosításra. Miután

3 A Viola–Jones algoritmus két számítógépes látáskutatóról kapta a nevét (Paul Viola és Michael Jones), akik 2001-ben javasolták a módszert *Gyors objektumészlelés az egyszerű funkciók megnövelt kaszkádjával* című cikkükben (Viola & Jones, 2001). Annak ellenére, hogy egy elavult keretrendszer, az alkalmazása kivételesen figyelemreméltónak bizonyult a valós idejű arcfelismerésben (bár más objektum osztályozásnál is alkalmazható). Ennek az algoritmusnak a betanítása lassú, de nagy sebességgel képes valós időben észlelni az arcokat.

az arcdetektálás megtörtént, a rendszer további feldolgozási lépéseket végezhet, például az arcok körvonalainak meghatározása vagy területének szegmentálása. A rendszer különböző jellemzők (például arcvonások, szemek elhelyezkedése stb.) kinyerése során átalakítja az arcot numerikus értékekké, amelyeket későbbi összehasonlításra használ. Az arcfelismerő rendszer összehasonlítja a kinyert jellemzőket az adatbázisban tárolt jellemzőkkel, hogy megállapítsa, melyik felel meg a felismert arcnak. Végül a rendszer döntést hoz arról, hogy a felismert arc melyik tárolt referenciaarcra hasonlít leginkább.

Kína vezeti a világot az arcfelismerő rendszerek állami használatában. A technológiát széles körben alkalmazza a rendőrség a városi közterületek felügyeletére. A Surfshark 194 országban végzett 2019-es felmérése szerint, a világon 109 ország használja vagy hagyta jóvá az arcfelismerő technológia megfigyelési célú használatát, ebből Európában 32 ország (URL5). A hagyományos nem biológiai felismerési és fiziológiai jellemzők felismerési technológiájához képest az arcfelismerő technológia különleges technikai előnyökkel rendelkezik. Az arckép attribútumai és vonásai elegendőek egy személy identitásának megállapításához. Az arcfelismerő technológia személyazonosításra alkalmas további információkat, például életkort, nemet és rasszt is képes kinyerni a képekből. Egyes kínai kameragyártók a kameráikba integrált arcfelismerő algoritmusuk által lehetővé teszik például a Kína nyugati részén, Hszincsiang tartományban nagy számban élő ujgurok rassz szerinti szűrését (URL6). Az MI nemcsak az arcfelismerés területén használható, hanem egyéb videóanalitikai szoftverekben alkalmazva figyelmeztet a rendellenes cselekmények bekövetkeztekor (Tóth, 2018). Ezt kihasználva egy másik, szintén kínai gyártó MI-alapú analitikái kilenc különböző riasztástípust tartalmaznak, melyek a tömegmegmozduláshoz kapcsolódnak. Ezek közé tartozik a „tömeg gyülekezése a közterületi rend megzavarására", a „jogellenes gyülekezés, felvonulás, tüntetés", a „petícióval" való fenyegetés, valamint a „vallási" és a „Fálun Gong"[4] néven ismert kínai spirituális mozgalom (URL7). Az ilyen jellegű kisebbségi közösségekkel vagy etnikumokkal szembeni megkülönböztetések komoly etikai és erkölcsi kérdéseket vetnek fel.

Az információk összekapcsolásának lehetősége miatt egyre nagyobb aggodalomra ad okot a videómegfigyelés magánéletünkre gyakorolt káros hatása. A fő

4 A Kínában ismert Fálun Gong, más néven Fálun Dáfá egy kínai spirituális mozgalom, amely nem vallás, hanem egy olyan elméletet és gyakorlatot egyesítő rendszer, amely a napjaink életkörülményeihez igazodik. Nem rendelkezik vallási formával, rituálékkal, imádsággal vagy kultusszal. Emellett mentes mindenféle politikai ideológiától. A mozgalom rendkívüli népszerűségét és gyors terjedését néhány kormányzati vezető kockázatként értékelte a politikai vezetés szempontjából.

probléma a különböző rendelkezésre álló információkból alkotott profil, amelyek felhasználhatók a magánélet megsértésére. Egy automatikus arcfelismerő és elemző algoritmus beépítésével és különböző adatbázisok vagy hálózati rendszerek összekapcsolásával jelentős információk nyerhetők ki. A technológiai fejlődés gyors ütemével a jogszabályalkotók nehezen tudnak lépést tartani. Az Európai Unió MI-ről szóló törvénytervezete (URL8) a nyilvános arcfelismerő rendszerek használatának korlátozását javasolja, az Európai Parlament pedig a technológia betiltását sürgette (URL9). 2021 júliusában az európai adatvédelmi hatóság és az európai adatvédelmi biztos közös véleményt fogadott el az Európai Bizottság MI-re vonatkozó harmonizált szabályok megállapításáról szóló rendeletére[5] irányuló javaslatáról.[6] Ebben *„az Európai Adatvédelmi Testület és az európai adatvédelmi biztos szorgalmazza, hogy általános jelleggel tiltsák meg az MI-nek az emberi jellemzők – például az arc, a járás, az ujjlenyomat, a DNS, a hang, a billentyűleütések és más biometrikus vagy viselkedési jellemzők – alapján a nyilvánosság számára hozzáférhető helyeken történő automatikus felismerésre bármilyen összefüggésben történő használatát”* (URL10).

Az arcfelismerő rendszerek sem tévedhetetlenek. A McAfee Advanced Threat Research (ATR) csapatának kísérlete során azt bizonyították, hogy az MI segítségével készített manipulált arcok képesek félrevezetni az arcfelismerő rendszereket (URL11). A kutatók egy olyan módszert fejlesztettek ki, amellyel két különböző A és B ember arcát egy hamis, C arcba konvertálják. Ez a hamis, úgynevezett morfizált C arc az emberi szem számára megtévesztően hasonlít az A személyre, viszont az MI-alapú arcfelismerő rendszer tévesen B-nek azonosítja.

A módszer alapját az úgynevezett Cycle-Consistent Generative Adversarial Network, azaz röviden a CycleGAN alkotja. A CycleGAN egy olyan típusú generatív versengő hálózat[7] (GAN), amely képes képek stílusát átalakítani anélkül, hogy párosított adatokra lenne szüksége. Ezt a gépi tanulási modellt két különböző adatkészlet közötti átalakításra használják. A CycleGAN két generátorral és két diszkriminátorral rendelkezik. A generátorok az egyik adatkészletről a másikra képeket generálnak, a diszkriminátorok pedig megpróbálják megkülönböztetni a valós és a generált képeket. A CycleGAN működésének lényege

5 A mesterséges intelligenciáról szóló törvény (AI Act), az Európai Bizottság által javasolt, 2021. április 21-én benyújtott, COM/2021/206 kodifikált rendelete.

6 A kézirat zárását követően 2024. május 21-én az Európai Unió (EU) Tanácsa bejelentette a mérföldkőnek számító uniós mesterséges intelligenciáról szóló törvény végleges jóváhagyását. (Szerkesztőség.)

7 A Generatív Adversarial Network (GAN) egyfajta mély tanulási modell, amely képes új, meglévő adatokhoz hasonló adatokat generálni. A GAN két neurális hálózatból áll, egy generátorból és egy diszkriminátorból. A generátor hálózat új adatokat hoz létre, míg a diszkriminátor hálózat megpróbál különbséget tenni valós és generált adatok között. A generátorok célja, hogy olyan képeket hozzanak létre, amelyeket a diszkriminátorok nem tudnak megkülönböztetni a valódi képektől, míg a diszkriminátorok célja, hogy minél jobban megkülönböztessék a valódi és hamis képeket.

az, hogy a generátorok arra vannak kiképezve, hogy a két adatkészlet közötti átalakítást úgy hajtsák végre, hogy a visszaalakított képek megegyezzenek az eredeti képekkel. Más szóval: a modell képes arra, hogy rögzítse a céltartomány jellemzőit, és új képeket generáljon a forrástartományból, amelyek ugyanazokkal a jellemzőkkel rendelkeznek. A CycleGAN cikluskonzisztens veszteséggel[8] van kiképezve, amely arra ösztönözi a modellt, hogy olyan képeket generáljon, amelyek megkülönböztethetetlenek a céltartomány valós képeitől. A CycleGAN nagyon hasznos eszköz olyan alkalmazásokban, mint például a művészeti stílusok átmásolása, fotorealisztikus képek generálása, vagy akár műholdképekből, vagy légifelvételekből térképek szintetizálása.

1. számú ábra
Eredeti A és B arc

Forrás. A szerző saját szerkesztése.

Visszatérve a kiindulási pontunkhoz, az arcok esetében az algoritmus úgy működik, hogy először az eredeti A és B arcból (1. számú ábra) egy átmeneti AB arctípus generálódik (2. számú ábra).

8 A CycleGAN egyik kulcsfontosságú része a ciklikus konzisztencia veszteség, amely biztosítja, hogy egy kép átalakítása után vissza lehessen azt alakítani eredeti formájára. Például, ha az A generátor átalakít egy lovat zebrává, a B generátornak vissza kell tudnia alakítani azt lóvá. Ez segít a modellnek abban, hogy releváns jellemzőket tanuljon meg, és ne csak véletlenszerű módosításokat végezzen.

2. számú ábra
Átmeneti AB arc

Forrás. A szerző saját szerkesztése.

Ezután ezt az átmeneti arctípust többször újragenerálja az MI (3. számú ábra). Ez a folyamat egészen addig tart, míg egy olyan arctípus nem jön létre, amely az emberi szem számára megtévesztően hasonlít az eredeti A személyre, azonban az MI-alapú arcfelismerő egyértelműen a B személy stílusjegyeit ismeri fel benne.

Ennek bizonyítására a McAfee kutatói az általuk generált manipulált arcképet egy olyan arcfelismerő rendszeren tesztelték, amely normál esetben az azonosítást 99%-os pontossággal végezte. Azonban a manipulált arcok esetében ez a rendszer akár 95%-os valószínűséggel is tévedhet, azaz 95%-os bizonyossággal állapította meg az A személyről, hogy az a B személy.

3. számú ábra
Az arcok fokozatos újragenerálása

Forrás. A szerző saját szerkesztése.

A tökéletes hatás elérése érdekében a tréning során 1500 képből álló arcsorozatot használtak fel, melyek a személyekről készített videófilmekből kivett állóképek voltak. Ezek a képek a tanítási adatokat gazdagították, és ez biztosította, hogy az algoritmus többféle arckifejezéssel dolgozzon.

4. számú ábra
Balra az eredeti A, jobbra a morfizált C arc

Forrás. A szerző saját szerkesztése.

Ez a típusú képi manipuláció óriási veszélyt jelenthet azoknál az azonosító rendszereknél, amelyek MI-alapú arcfelismerési technológiára épülnek. A generált hamis arcok felhasználhatók arra, hogy valaki jogosulatlanul hozzáférjen egy helyhez vagy szolgáltatáshoz, vagy hogy egy körözött személy hamis útlevelével kijátssza az automatikus határellenőrző rendszert, mivel a gépi manipuláció az emberi szem számára alaposabb megfigyelés nélkül észrevehetetlen (4. számú ábra).

Az ilyen jellegű visszaélésektől való félelem indokolta, hogy Németországban már 2020 nyarán előterjesztésre került az a törvény, amely tiltaná a morfizált képek útlevelekben történő felhasználását (URL12). Végül hosszas huzavona után csak 2025. május 1-jén lép életbe az útlevél-, személyazonosító- és bevándorlási jogi okmányok biztonságának megerősítéséről szóló törvény azon szakasza, amitől kezdve csak digitális fényképet lehet a hatóságoknak, hivataloknak bemutatni a személyi igazolvány és útlevél kiállítása érdekében (URL13). Ezeket a képeket csak a kiállító hatóságnál, vagy olyan regisztrált és ellenőrzött fényképésznél lehet elkészíteni, ahol megoldható a felhőn keresztüli biztonságos képtovábbítás a hatóság részére.

2020-ban Amerikában megtörtént az első téves letartóztatás, mely az arcfelismerő rendszernek köszönhető (URL14). A detroiti rendőrség a háza előtt vette őrizetbe Robert Williamst, aki a gyanú szerint több órát tulajdonított el a helyi Shinola üzletből. A gyanúsítás alapját az üzletből lopó fekete férfit ábrázoló videofelvétel szolgáltatta, amit a rendőrség egy arcfelismerő szoftver segítségével vizsgált. A szoftver tévesen Williamsként azonosította az elkövetőt. Ezek az esetek rámutatnak arra, hogy az arcfelismerő rendszerek sem tekinthetők tévedhetetlennek, és fontos további kutatásokat végezni annak érdekében, hogy megbízhatóbbá tegyük ezeket a technológiákat.

Az arcfelismerés korántsem a gépi látás egyetlen alkalmazása. Az önvezető járművek esetében is döntő fontosságú. A Tesla, a BMW, a Volvo és az Audi által jelenleg fejlesztés alatt álló autonóm és félig autonóm autók több kamerával vannak felszerelve, amelyek folyamatosan pásztázzák a környező teret, és az MI-alapú képi osztályozók segítségével felismerik az objektumokat, útburkolati jeleket, közlekedési táblákat és lámpákat. A japán Kyushu Egyetem kutatói egy kísérlet során arra a megdöbbentő következtetésre jutottak, hogy az MI-alapú képi osztályozókat 73,8%-ban sikerül becsapniuk úgy, hogy csak egy pixelt változtattak meg a képen. Amennyiben a kicserélt pixelek számát ötre növelték, akkor a sikerességi arány 87,3%-ra emelkedett. Ez a képi manipuláció az emberi szem számára érzékelhetetlen, de jelentősen eltértheti az MI által előállított kimenetet. Így például az MI egy béka képét hajóként, vagy egy repülőgépet kutyaként értelmezi (Su, Vargas, & Sakurai, 2019). Ezzel a típusú manipulációval már sikerült önvezető autók táblafelismerő funkcióját befolyásolni, és a járművet a megengedettnél nagyobb sebességre késztetni (URL15).

Összegzés

Sok fejlődési lehetőség és kiaknázatlan terület vár még az MI-re. Az, hogy ennek mennyi pozitív és mennyi negatív hozadéka lesz, még kérdés. A különböző MI működését megtévesztő támadások, az algoritmikus diszkrimináció, az önvezető autókat érintő balesetek és a technológiára való túlzott hagyatkozáshoz kapcsolódó dehumanizáció jó példák arra, amelyek a jövő társadalmát veszélyeztetik. A MI jelentette kockázatok, valamint a technológia jelenlegi fejlődésének sebessége még sürgetőbbé teszik a megfelelő és adaptív folyamatos szabályozást, ha azt akarjuk, hogy az MI növekvő használata lehetőleg minimális nemkívánatos következményekkel járjon. Stephen Hawking professzor halála előtt arra figyelmeztetett, hogy *„Az MI létrehozásának sikere civilizációnk történetének legnagyobb eseménye lehet. De lehet, hogy ez az utolsó is,*

hacsak nem tanuljuk meg, hogyan kerüljük el a kockázatokat. Az erőteljes mesterséges intelligencia felemelkedése vagy a legjobb, vagy a legrosszabb dolog, ami valaha is történhet az emberiséggel. Még nem tudjuk, melyik." (URL16).

Felhasznált irodalom

Oosterloo, S. & van Schie, G. (2018). The Politics and Biases of the "Crime Anticipation System" of the Dutch Police. *CEUR Workshop Proceedings*, *2103*, 30–41.

Russel, S. & Norvig, P. (2010). *Artificial Intelligence: A Modern Approach* (3rd Ed.). Prentice Hall.

Schönherr, L., Kohls, K., Zeiler, S. & Holz, T. (2018). Adversarial Attacks Against Automatic Speech Recognition Systems via Psychoacoustic Hiding. *Network and Distributed Systems Security (NDSS) Symposium 2019 24-27 February 2019, San Diego.* https://doi.org/10.14722/ndss.2019.23288

Su, J., Vargas, V. D. & Sakurai, K. (2019). One Pixel Attack for Fooling Deep Neural Networks. *IEEE Transactions on Evolutionary Computation*, *23*(5), 828–841. https://doi.org/10.1109/TEVC.2019.2890858

Tóth A. (2018). Az élőerő munkáját segítő technikai megoldások. *Hadmérnök, 13*(2), 29–36.

Viola, P. & Jones, M. (2001). Rapid Object Detection using a Boosted Cascade of Simple. In Jacobs, A. & Baldwin, T. (Szerk.), *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 511–518). Computer Society. https://doi.org/10.1109/CVPR.2001.990517

Zhang, G., Yan, C., Ji, X., Zhang, T., Zhang, T. & Xu, W. (2017). DolphinAttack: Inaudible Voice Commands. *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* (pp. 103–117). https://doi.org/10.1145/3133956.3134052

A cikkben található online hivatkozások

URL1: *From Crime Mapping to crime forecasting: The evolution of olace-based policing. National Institute of Justice.* https://nij.ojp.gov/topics/articles/crime-mapping-crime-forecasting-evolution-place-based-policing

URL2: *You Can Have the Blue Pill or the Red Pill, and We're Out of Blue Pills.* https://www.nytimes.com/2023/03/24/opinion/yuval-harari-ai-chatgpt.html

URL3: *Game-playing DeepMind AI can beat top humans at chess, Go and poker.* https://www.newscientist.com/article/2402645-game-playing-deepmind-ai-can-beat-top-humans-at-chess-go-and-poker/

URL4: *Installed base of smart speakers worldwide in 2020 and 2024.* https://www.statista.com/statistics/878650/worldwide-smart-speaker-installed-base-by-country/

URL5: *The Facial Recognition World Map.* https://surfshark.com/facial-recognition-map

URL6: *As China Tracked Muslims, Alibaba Showed Customers How They Could, Too.* https://www.nytimes.com/2020/12/16/technology/alibaba-china-facial-recognition-uighurs.html
URL7: *Police in China can track protests by enabling 'alarms' on Hikvision software.* https://www.theguardian.com/world/2022/dec/29/china-surveillance-protests-alarms-cameras-hikvision
URL8: *This may be America's first known wrongful arrest involving facial recognition.* https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206
URL9: *REPORT on artificial intelligence in criminal law and its use by the police and judicial authorities in criminal matters.* https://www.europarl.europa.eu/doceo/document/A-9-2021-0232_EN.html#title1
URL10: *EDPB-EDPS Joint Opinion 5/2021 on the proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).* https://edpb.europa.eu/our-work-tools/our-documents/edpbedps-joint-opinion/edpb-edps-joint-opinion-52021-proposal_en
URL11: *Dopple-ganging up on Facial Recognition Systems.* https://www.mcafee.com/blogs/other-blogs/mcafee-labs/dopple-ganging-up-on-facial-recognition-systems/
URL12: *Germany bans digital doppelganger passport photos.* https://www.reuters.com/article/idUSKBN23A1YM/
URL13: *Gesetz zur Stärkung der Sicherheit im Pass-, Ausweis- und ausländerrechtlichen Dokumentenwesen.* https://www.bgbl.de/xaver/bgbl/start.xav?startbk=Bundesanzeiger_BGBl&start=//*[@attr_id=%27bgbl120s2744.pdf%27]#__bgbl__%2F%2F*%5B%40attr_id%3D%27bgbl120s2744.pdf%27%5D__1704650921418
URL14: *Proposal For A Regulation Of The European Parliament And Of The Council Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts.* https://edition.cnn.com/2020/06/24/tech/aclu-mistaken-facial-recognition/index.html
URL15: *Hackers can trick a Tesla into accelerating by 50 miles per hour.* https://www.technologyreview.com/2020/02/19/868188/hackers-can-trick-a-tesla-into-accelerating-by-50-miles-per-hour/
URL16: *Stephen Hawking: AI will be 'either best or worst thing' for humanity.* https://www.theguardian.com/science/2016/oct/19/stephen-hawking-ai-best-or-worst-thing-for-humanity-cambridge

A cikk APA szabály szerinti hivatkozása

Tóth L. (2024). Az intelligens fenyegetés. Hogyan veszélyeztetheti a mesterséges intelligencia a biztonságunkat? *Belügyi Szemle*, *72*(7), 1187–1205. https://doi.org/10.38146/BSZ-AJIA.2024.v72.i7.pp1187-1205

Nyilatkozatok

Összeférhetetlenség

A szerző nem jelentett összeférhetetlenséget.

Finanszírozás

A szerző nem kapott pénzügyi támogatást a kutatáshoz, a szerzőséghez és/vagy a cikk publikálásához.

Etikai nyilatkozat

Jelen cikkhez nem kapcsolódik adatkészlet.

Levelező szerző

A cikk levelező szerzője Tóth Levente, aki a toth.levente@uni-nke.hu e-mail címen érhető el.