Oliver Reindl: Hamis világ

A mesterséges intelligencia manipulációra és megtévesztésre szolgáló eszköztára

Bevezetés

A nagy teljesítményű mesterséges intelligencia modellek gyors fejlődésével és egyre növekvő elérhetőségével nemcsak új lehetőségek jelennek meg az üzleti életben, a kutatásban és a mindennapi életben – hanem új és komoly kockázatok is. Azok a technológiák, amelyek egykor a kutatólaboratóriumokra korlátozódtak, ma már gyakorlatilag bárki számára elérhetők, aki rendelkezik internetkapcsolattal: szöveg- és képgenerátorok, hangutánzó eszközök, deepfake szoftverek és automatizált ajánlórendszerek.

A mesterséges intelligenciával kapcsolatos nyilvános vitákat gyakran apokaliptikus forgatókönyvek uralják – az a félelem, hogy a mesterséges intelligencia feleslegessé tesi a munkahelyeket, irányíthatatlanná válik, vagy akár magát az emberiséget is fenyegeti. A legtöbb vezető szakértő azonban hangsúlyozza, hogy a mesterséges intelligencia önmagában nem veszi el a munkahelyeket. Ehelyett munkahelyek veszhetnek el azok számára, akik használják a mesterséges intelligenciát, és jelentősen produktívabbá válnak általa. Ez még sürgetőbbé teszi a szakemberek és szervezetek számára, hogy aktívan foglalkozzanak a mesterséges intelligencia technológiáival, ahelyett, hogy figyelmen kívül hagynák vagy félnének tőlük.

Továbbá tudományos tanulmányok kimutatták, hogy a mesterséges intelligencia rendszerek a tanulást és a fejlődést szolgáló. az emberektől származó valós adatokra támaszkodnak. A kizárólag szintetikus (géppel generált) adatokon képzett modellek idővel hajlamosak leépülni vagy összeomlani. Ez egy alapvető igazságot illusztrál: a mesterséges intelligenciának emberekre van szüksége – nemcsak felhasználóként, hanem hiteles adatok és értékelés forrásaként is. Mégis, ezeken a hosszú távú kockázatokon és filozófiai kérdéseken túl a mesterséges intelligencia már ma is nagyon konkrét fenyegetéseket jelent. Egyre több dokumentált eset van, amikor a mesterséges intelligenciával visszaélnek az emberek megtévesztésére, manipulálására vagy károsítására – álhírek, szintetikus média, személyre szabott csalások vagy viselkedési minták algoritmikus kihasználása révén. Ezek a fejlemények nem hipotetikusak. Valósak, széles körben elterjedtek és egyre kifinomultabbak. E tanulmány célja, hogy bemutassa és elmagyarázza azokat a kulcsfontosságú információs technológiákat, amelyeket jelenleg az emberek manipulálására vagy megtévesztésére használnak. Minden rész tartalmaz egy közérthető magyarázatot, egy technikai leírást az informatikai szakemberek számára, és egy valós példát a potenciális hatás bemutatására.

Manipulatív információtechnológiák – Részletes áttekintés

1. Deepfakes – Valódi emberekről készült hamis videók

🔹 Egyszerű magyarázat: Egy mesterséges intelligencia rendszer számos valós videót vagy fényképet elemez egy személyről. Megtanulja, hogyan mozog az arc, és milyen arckifejezések jellemzőek. Ezután az arcot egy másik videóra vetíti. Az eredmény: úgy tűnik, mintha a valódi személy mondott volna vagy tett volna valamit, ami valójában soha nem történt meg.

🔹 Műszaki leírás: A mélyhamisítások elsősorban a Generatív Ellenséges Hálózatokon (Generative Adversarial Networks, GAN) alapulnak, ahol egy generátor hálózat képeket vagy videókat hoz létre, és egy diszkriminátor hálózat megpróbálja észlelni, hogy a tartalom hamis-e. Kódoló-dekódoló architektúrákat is használnak arckifejezések kinyerésére és a valósághű térképezéshez (mapping) mozgásátviteli technikákat alkalmaznak.

🔹 Használati eset (példa): Egy videóban egy ismert politikus állítólag beismeri a választási csalást. A meggyőzően hamis videó gyorsan terjed a közösségi médiában, aláásva a demokratikus folyamatokba vetett bizalmat.

2. Hangklónozás – Hangok utánzása mesterséges intelligencia segítségével

🔹 Egyszerű magyarázat: Mindössze néhány másodpercnyi rögzített beszéddel egy mesterséges intelligencia képes utánozni egy személy hangját. Elemzi a hangmagasságot, a tónust és a ritmust, majd hangosan felolvas bármilyen szöveget az adott személy hangján.

🔹 Műszaki leírás: A hangklónozást fejlett szövegfelolvasó (Text-to-Speech – TTS) rendszerek, például a Tacotron 2 vagy a VALL-E működtetik. Ezek a modellek egy kódoló segítségével akusztikus jellemzőket vonnak ki, és egy autoregresszív dekóder segítségével szintetikus beszédreprezentációt generálnak. Egy neurális vokóder (pl. WaveNet, HiFi-GAN) ezután előállítja a végső hangjelet.

🔹 Használati eset (példa): Egy csaló felhív egy bankot, egy cég vezérigazgatójának adva ki magát, és sürgős átutalást kér – a vezető valódi hangját használva.

3. AI képgenerátorok – Olyan képek, amelyek soha nem léteztek

🔹 Egyszerű magyarázat: Szövegben leírod, mit szeretnél látni, pl. „egy politikus pénzzel teli bőröndöt tart a kezében”. A mesterséges intelligencia egy valósághű képet hoz létre, amely úgy néz ki, mint egy fénykép, de teljes mértékben kitalált.

- Műszaki leírás: A képgenerátorok, mint például a DALL·E vagy a Stable Diffusion, látens diffúziós modelleket használnak. Ezek a rendszerek véletlenszerű zajból (random noise) generálnak képet, több milliárd szöveg-kép páron betanított neurális hálózatok irányításával. A modell fokozatosan átalakítja a zajt egy koherens képpé, amely megfelel a bemeneti szövegnek.

- Használati eset (példa): Egy hamis képen egy politikus állítólag kenőpénzt fogad el. Bár a kép teljesen kitalált, terjed a közösségi médiában, és sokan elhiszik, hogy valódi.

4. MI által generált álhírek / szövegek

- Egyszerű magyarázat: A nagy nyelvi modellek olyan cikkeket, történeteket vagy kommenteket tudnak írni, amelyek úgy hangzanak, mintha valódi emberek írták volna őket – még akkor is, ha kitaláltak.

- Technikai leírás: A GPT-4-hez hasonló nyelvi modellek a Transformer architektúrán alapulnak, önfigyelemmel. Nagy szöveges korpuszokon képzik őket, és a tartalom generálása a legvalószínűbb következő szó előrejelzésével történik az adott kontextus alapján.

- Használati eset (példa): Egy blogbejegyzés tudományos cikknek adja ki magát, amelyben hamis állításokat tesz a vakcinákról. Oltásellenes csoportokban terjed, és tényként kezelik.

5. Mesterséges intelligencia által vezérelt adathalászat – Tökéletes átverős e-mailek

- Egyszerű magyarázat: A mesterséges intelligencia által generált e-mailek meggyőzőnek tűnnek, logókat, realisztikus nyelvet használnak, sőt, a közösségi hálózatokról származó valós információkra is hivatkoznak a bizalom elnyerése érdekében.

- Technikai leírás: Az adathalász e-maileket természetes nyelvi generálással (NLG) állítják elő, és NLP-vel finomítják a személyre szabás érdekében. A nyilvános adatokat elemzik, hogy a címzett szerepéhez, írásmódjához vagy viselkedéséhez igazított e-maileket készítsenek. A mesterséges intelligencia a hivatalos nyelvet utánozza a hitelesség növelése érdekében.

- Használati eset (példa): Egy alkalmazott hamis e-mailt kap az „IT-támogatástól”, amelyben jelszó-visszaállítást kérnek – ez egy hamis weboldalra vezet, amely bejelentkezési adatokat gyűjt.

6. Hamis vélemények és közösségi média botok

- Egyszerű magyarázat: Az automatizált programok hamis véleményeket vagy hozzászólásokat tesznek közzé, és úgy osztanak meg tartalmakat a közösségi médiában, mintha valódi emberek lennének.

- Műszaki leírás: A botokat szkriptnyelveken, például Pythonon keresztül programozzák, és nyilvános API-kon vagy automatizálási keretrendszereken, például a Seleniumon keresztül kommunikálnak a platformokkal. Használhatnak LLM-eket emberszerű bejegyzések generálására, vagy előre meghatározott szövegminták követésére.

- Használati eset (példa): Egy cég hamis 5 csillagos értékelésekkel árasztja el az értékelő platformokat, hogy egy olyan terméket népszerűsítsen, amely a valóságban rosszul teljesít.

7. AR/VR a valóság torzítására

- Egyszerű magyarázat: A virtuális valóság (VR) és a kiterjesztett valóság (AR) olyan dolgokat mutat meg, amelyek a való világban nem léteznek – mégis nagyon hihetőnek és magával ragadónak tűnhetnek.

- Műszaki leírás: Az AR/VR rendszerek 3D motorokat, például Unityt vagy Unrealt használnak, és a térbeli követés érdekében szenzorfúzióra támaszkodnak. A számítógépes látást és a valós idejűvé tételt digitális elemek fizikai környezetbe ágyazására vagy szimulálására használják.

- Használati eset (példa): Egy hamis VR oktatási alkalmazás a történelem egy szimulált változatát mutatja a diákoknak, amely félretájékoztatást tartalmaz, és amelyet ők pontosnak hisznek.

8. GPS-hamisítás – A tartózkodási hely meghamisítása

- Egyszerű magyarázat: Egy eszköz hamis GPS-jeleket küldve úgy tesz, mintha egy másik helyen lenne – és az alkalmazások elhiszik.

- Technikai leírás: A GPS-hamisítás szoftveresen definiált rádiókat (SDR) használ hamis műholdadatok sugárzására. Mivel a GPS nem hitelesített, a mobil vevők érvényesként fogadják el ezeket a jeleket, lehetővé téve a pontos helymeghatározás manipulálását.

- Használati eset (példa): Egy kézbesítő meghamisítja a GPS-pozícióját, hogy olyan munkaórákat vagy kilométereket állítson be, amelyek valójában nem történtek meg.

9. Manipulatív ajánlóalgoritmusok

- Egyszerű magyarázat: Az online platformok olyan tartalmakat sugallnak, amelyekkel valószínűleg kapcsolatba fogsz lépni – gyakran szélsőséges vagy elfogult tartalmakat mutatnak, hogy fenntartsák a figyelmedet.

- Műszaki leírás: Az ajánlómotorok kollaboratív szűrést, mátrixfaktorizációt vagy megerősítéses tanulást használnak a felhasználói elköteleződés optimalizálására. Ez

megerősítheti a visszhangkamrákat, vagy egyre szélsőségesebb tartalmak felé terelheti a felhasználókat.

- Használati eset (példa): Miután megnézett néhány gazdasági videót, a felhasználót elárasztják a pénzügyi összeomlásról szóló összeesküvés-elméletekkel teli tartalmak – amelyek formálják a világképét.

10. Mikrocélzás (targeting) – A gyengeségeidet célzó hirdetések

- Egyszerű magyarázat: A weboldalak nyomon követik a viselkedésedet, és személyre szabott hirdetéseket jelenítenek meg – gyakran érzelmi kiváltó okokkal, amelyek célja, hogy befolyásoljanak.

- Műszaki leírás: A követő sütik, az eszközujjlenyomat-vétel és az adatbrókerek részletes felhasználói profilokat hoznak létre. Az olyan hirdetéstechnológiai platformok, mint a Google Ads, gépi tanulást használnak a felhasználók pszichográfiai adatok alapján történő precíz szegmentálására és célbavételére.

- Használati eset (példa): Egy bizonytalan szavazó a félelmeihez igazított politikai hirdetéseket kap, amelyek egy adott párt felé terelik anélkül, hogy felismerné a manipulációt.

11. Arcfelismerés és személyazonosság-lopás

- Egyszerű magyarázat: A mesterséges intelligencia elemzi az arcodat, és képes újraalkotni vagy ellopni a személyazonosságodat online visszaélés vagy csalás céljából.

- Műszaki leírás: Az arcfelismerés mélytanulást alkalmaz konvolúciós neurális hálózatokon (CNN) keresztül, hogy kinyerje az arc jellegzetességeit, és beágyazásokká alakítsa azokat. Ezeket a vektorokat összehasonlításra vagy rekonstrukcióra használják ellenőrző rendszerekben.

- Használati eset (példa): Egy bűnöző hamis LinkedIn profilt hoz létre az arcod és a neved felhasználásával, hogy megtévessze a kapcsolataidat.

12. Szintetikus videók a politikában

- Egyszerű magyarázat: Egy politikus sértő kijelentéseket tesz egy videóban – de a videót teljes egészében mesterséges intelligencia segítségével gyártották.

- Technikai leírás: A szövegből videót készítő platformok (pl. Synthesia, Runway) GAN-okat és mozgásátvitelt használnak, szöveges promptok beágyazása által vezérelve, szintetikus videók létrehozásához. Az arckifejezéseket és a gesztusokat *deep motion capture* modellek segítségével szimulálják.

🔹 Használati eset (példa): Egy politikai jelöltről rasszista kijelentéseket tevő koholt videó terjed egy választás előtt, amely károsítja a jelölt nyilvános megítélését.

13. Sötét minták – Rejtett trükkök a tervezésben

🔹 Egyszerű magyarázat: A weboldalak és alkalmazások okos tervezést alkalmaznak, hogy rávegyék Önt olyan dolgokra, amelyeket általában nem tenne – például véletlenül feliratkozik valahová vagy tovább marad valahol.

🔹 Technikai leírás: A sötét mintákat A/B teszteléssel, hőtérképekkel és konverziókövetéssel optimalizálják. A felhasználói felület/felhasználói élmény (UI/UX) kialakítása kognitív torzításokat (pl. alapértelmezett torzítás, veszteségkerülés) használ ki a felhasználói döntések megtévesztő módon történő irányításához.

🔹 Használati eset (példa): Egy felhasználó megpróbálja lemondani az előfizetését, de nem találja a lemondási lehetőséget – vagy rossz gombra kattint, és ilyenkor egy újabb évre felszámítják a számláját.

A hamisítás felismerhetősége

A hamisítást csak akkor lehet megállapítani, ha az eredeti kódolás ismert. De ha a hamis képet, videót vagy bámilyen más digitális hamisítványt lemásolják, attól kezdve az eredete gyakorlatilag megállapíthatatlan.

Következtetés

Ez a tanulmány a jelenlegi információs technológiák széles skáláját vizsgálta, amelyek felhasználhatók az emberek – szándékos vagy akaratlan – manipulálására vagy megtévesztésére. Minden technológiához magyarázatokat fűztünk:

• egy egyszerű magyarázatot az átlagos felhasználó számára,
• egy technikai leírást a szakemberek számára, és
• egy gyakorlati használati esetet a valós relevancia bemutatására.

Világossá tettük hogy a mesterséges intelligencia és a kapcsolódó technológiák nem eredendően rosszindulatúak – de veszélyes eszközökké válhatnak, ha helytelenül vagy megtévesztő céllal használják őket. A mesterséges intelligencia ereje abban rejlik, hogy képes az emberi viselkedést skálázni: jóra vagy rosszra. A deepfake-ek, a hangklónozás, a szintetikus képek, a mesterséges intelligencia által írt dezinformáció, a célzott hirdetések és a manipulatív tervezési minták már nem elvont fenyegetések. Már most is nagymértékben alakítják a véleményeket, a döntéseket és a társadalmi dinamikát – gyakran anélkül, hogy az érintettek tudnának róla.

Ezeknek az eszközöknek a létezése mind a technikai jártasságot, mind az etikai tudatosságot megköveteli. A szervezeteknek, az oktatóknak, a politikai döntéshozóknak és a nyilvánosságnak nemcsak azt kell megérteniük, hogy mire képes a mesterséges intelligencia, hanem azt is, hogy hogyan és miért – és kik – használják. Csak átlátható, felelősségteljes és tájékozott együttműködéssel biztosíthatjuk, hogy a mesterséges intelligencia erősítse, ne pedig aláássa a bizalmat, az igazságot és az emberi méltóságot a digitális korban.

Köln, 2025 június