ORBÁN JÓZSEF

QUO VADIS GENERATÍV MESTERSÉGES INTELLIGENCIA? – GPT MODELLEK AZ ELKÖVETŐK ÉS A BŰNÜLDÖZÉS ESZKÖZTÁRÁBAN

1. Bevezetés

A mesterséges intelligencia (MI) fejlődésének egyik ága a generatív előtanított átalakító (Generative Pretrained Transformer, GPT), melynek beszélgetős megjelenési formája a ChatGPT, Bard stb. A természetes és a mesterséges tudásra épülő öntanuló rendszer már megfelelni látszik a Turing-teszt elvárásoknak, azaz a gépi és a humán válasz minőségileg azonosnak tűnik. Ezt kihasználva a GPT már megjelent a csalók eszköztárában is, amely a tendenciákat tekintve még csak a kezdet.

A GPT eredmények tagadása, alkalmazásának tiltása helyett megvizsgálandó, hogyan lehet az MI-t a bűnüldözés szolgálatába állítani. A lokális, és ezért nagymértékben limitált rendészeti adatbázisok kiegészítését jelentheti a globális információs tudástáron alapuló GPT, amely virtuális tanács- és ötletadóként egyaránt szolgálhatja a kriminalistákat, az ügyészeket és a védőket. A mesterséges intelligencia új alkalmazási formája olyan nagyhatású eszközt ad a bűnfelderítési szakemberek kezébe, amely hatékonyságban és gyorsaságban új dimenziókat nyit. Az elérhető GPT alkalmazások magasnak tekinthető 5-25%-os hibaaránya miatt a tapasztalat, az intuíció és a természetes intelligencia, összességében az ember, továbbra sem nélkülözhető a bűnüldözésből.

A meggyőződésből vagy kényelemszeretetből elutasítók minden harcias, vagy kevésbé hangos, de esetleg csak gúnyolódó[1] megnyilvánulása ellenére a mesterséges intelligencia (MI) újabb és újabb harcálláspontokat foglal el. Az ellenzők érvei között etikai aggodalmak, munkaerőpiaci kérdések, az adatvédelem és a magánélet védelme, továbbá biztonsági kockázatok szerepelnek. Az ellenvélemények súlyának érzékeltetésére az alábbi idézet szolgál. "*A generatív mesterséges intelligencia sok előnyt ígér, azonban már most is komoly károkat okoz. Ahhoz, hogy a mesterséges intelligencia fejlődésének teljes előnyét ki tudjuk használni, gondosan kell tanulmányoznunk és megértenünk az általa okozott költségeket"* írta 2023. júniusában Gene Dodaro, az Amerikai Számvevőszék (GAO) vezetőjének címzett közös levelében Edward J. Markey és Gary Peters szenátor.[2]

Az etikai aggodalmak alapja az, hogy az ismeretlen, esetlegesen torzult vagy ténylegesen diszkriminatív algoritmusok[3] személyi felelősség nélkül dönthetnek emberek

[1] Russell, Stuart. J. – Norvig, Peter.: Mesterséges intelligencia - Modern megközelítésben. Panem – Prentice-Hall, Budapest 2000. 59.o.

[2] Silvia, Michael: Massachusetts Sen. Markey concerned about "injury, death, human extinction" from ChatGPT, A.I. Vö. https://www.newbedfordguide.com/massachusetts-sen-markey-concerned-injury-death-human-extinction-chatgpt-ai/2023/06/24 (Letöltés ideje: 2023.07.15.)

[3] Buolamwini, Joy – Gebru, Timnit: Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of Machine Learning Research Conference on Fairness, Accountability, and Transparency. 2018. 1-15.o.

vagy akár egész közösségek sorsáról.[4] Ezért is kap kiemelt szerepet a „figyelő állampolgárság" (monitorial citizenship), amely őrködik a közösségi normáktól való káros eltérések felett.

Dignum 2022-es összegzése szerint több mint 600 MI vonatkozású ajánlás, útmutató, stratégiai riport készült prominens kormányközi szervezetek, szakmai testületek, nemzeti szintű bizottságok és különböző szervezetek által.[5] A hatalmas erőfeszítések ellenére sem megnyugtató a helyzet. A munkaerőpiacon valóban növekvő veszély fenyegeti mindazon munkaköröket, amelyek körülhatárolhatóan algoritmizálható értelmiségi munkavégzést takarnak. Így különösen a jogi, a fordítói és egyes oktatói munkakörök esetében valós a kockázat. A tanulmány célja, hogy megvizsgálja a generatív nagy nyelvi modellek (Large Language Model, LLM) jelenlegi állapotát, kriminalisztikai alkalmazhatóságot, valamint a nem jogkövető felhasználók számára megnyíló visszaélési lehetőségeket. A tanulmányban számos preprint hivatkozás szerepel, ami mutatja, hogy a lektorált publikálás kezd leszakadni a tudományos fejlődés tényleges sebességétől. A problémakör feldolgozása terjedelmi korlátok miatt csak bevezetésnek tekinthető, ám aktualitása a fejlődési iram miatt minden nappal csökken.

2. A gépi tanulástól a GPT-ig

A gépi tanulás[6] és a mély tanulás[7] ismertetése egyre kevésbé kelti fel a figyelmet, mert beleolvadt a mesterséges intelligencia újabb fázisaiba. A gyenge,[8] a szűk[9] és az erős[10] mesterséges intelligencia (MI) nemcsak fejlődési fázisokat, de felhasználási területeket is jelöl.

A természetes nyelvek feldolgozása (Natural Language Processing, NLP) számítógépes megvalósításának igénye már évtizedekkel ezelőtt megjelent, azonban igazi áttörés az egylapkára integrált processzorok számának és a tárhelyek ugrásszerű növekedésének volt köszönhető. Ennek egy speciális megjelenési formája a generatív előtanított átalakító (Generative Pre-trained Transformer, GPT), amely az emberi kérdésre válaszolva képes új, értelmes és koherens válaszokat adni.

Az előtanítás során az elektronikusan elérhető nagy mennyiségű információhalmazból hatalmas kombinációs lehetőséget gyűjt össze, mint tudást. Az adatforrás – és itt hangsúlyozni kell, hogy nem feleletgyűjtemény, hanem a témába vágó szűrt szöveghalmaz – azonban a feltett kérdésre adható válaszlehetőségek gyakoriságán alapul. Tehát ideális esetben is a válaszokat nem a tényszerűség igénye, hanem a forrásnak tekinthető tanítási nyelvekben a statisztikai előfordulási gyakoriság fogja befolyásolni. Emiatt különös jelentősége van a gondolkodó és empatikus ember felelősségteljes

[4] D'Ignazio, Catherine: What would feminist data visualization look like. https://civic.mit.edu/feminist-data-visualization (Letöltés ideje: 2023.07.15.)
[5] Dignum, Virginia: Responsible Artificial Intelligence – from Principles to Practice, ACM SIGIR Forum Vol. 56 No. 1 June 2022. https://arxiv.org/pdf/2205.10785.pdf (Letöltés ideje: 2023.07.15.)
[6] Gépi tanulás (Machine Learning) lehetővé teszi, hogy a rendszerek programozás nélkül a tapasztalatok alapján az adatokból tanuljanak, algoritmusokat alkalmazzanak adatelemzésre és mintázatok felismerésére.
[7] Mély tanulás (Deep Learning) neurális hálózatokra és nagyon összetett adathalmazokra támaszkodó gépi tanulás.
[8] Gyenge MI (Weak AI) rendszerek nagy sebességű emberi kogníciót szimulálva korlátozott feladatokra és szűk tartományokra összpontosítanak.
[9] Szűk MI (Narrow AI) rendszerek alkalmazása egy adott feladatra vagy területre korlátozódik. Ezek a rendszerek specifikus feladatok megoldásában a gépi tanulás és a mély tanulás módszereire támaszkodhatnak.
[10] A jelenleg még csak kutatási fázisban lévő erős MI (Strong AI) rendszerek az elvárások szerint az emberi gondolkodás és problémamegoldás szinte minden területén képesek lesznek reagálni.

felügyeleti tevékenységének, amely a nagymennyiségű hibás forrásinformáció miatti statisztikai torzulást korrigálja. Igaz, így újabb manipulációs lehetőség jelenik meg.

A 2023-as évben az MI nagy sikerének tekinthető, hogy a generatív MI (GMI) egyik ága, a nagy nyelvi modellek – mint az MI egyik bárki által használható eszközei – kiléptek a kutatólaboratóriumokból a nyilvánosságra. A legsikeresebb modell, az OpenAI fejlesztése a ChatGPT két hónap alatt elérte a százmillió regisztrált felhasználó értéket.[11] A Google 2023. februárjában jelentette be, hogy hasonló fejlesztésen dolgozik, majd július 13-ától elérhetővé tette modelljét Európában is.[12] A nagymértékű növekedés a tudományos paradigmaváltásra, de egyúttal komoly veszélyekre is felhívja a figyelmet.

2.1. A nagy nyelvi modellek rövid áttekintése

A neurális hálózatokra épülő és rendkívül nagyszámú paramétert tartalmazó nagy nyelvi modellek az MI területén terjedtek el, amelyek a természetes nyelv feldolgozására és generálására szolgálnak. Az alkalmazott kapcsolatokban a súlyozó paraméterek száma több millió vagy akár milliárdnyi lehet. A modellek a gépi tanulás módszereivel előképzett adathalmazokon alapulnak, amelyek általában hatalmas mennyiségű, nem címkézett szöveget tartalmaznak.

Ezek a modellek képesek különböző nyelvi feladatok elvégzésére, mint például a szöveggenerálás, szövegértelmezés, kérdés-válasz előrejelzés, gépi fordítás, automatikus szövegösszefoglalás és sok minden más. Mindemellett nagyon sokoldalúak és rendelkeznek általános nyelvi érzékkel, amely lehetővé teszi számukra, hogy elvégezzenek széles körű nyelvi feladatokat. A folyamatos kutatás és fejlesztés eredményeképp a nagy nyelvi modellek fokozatosan fejlődnek és bővülnek, az egyes változatok között ugrásszerű fejlődés is érzékelhető. A fejlesztők között erre a célra létesült cégek (pl. OpenAI, AI21 Labs) és informatikai cégóriások (Google, Beijing Academy of Artificial Intelligence – BAAI) egyaránt megtalálhatók. Az OpenAI fejlesztette ki a 175 milliárd paramétert használó GPT-3-at, a ChatGPT-t (GPT-3.5), majd a március 14-én kibocsátott, minőségi ugrást érzékeltető GPT-4-et is. A Microsoft és a Nvidia közösen együttműködve fejlesztette ki az 530 milliárd paramétert használó Megatron-Turing NLG alkalmazást.[13]

A Google a szélesebb közönség számára pedig a BARD-ot fejlesztette ki. Az izraeli AI21 Labs Jurassic-1 Jumbo hatalmas szövegbázisra épülve, 178 milliárd paraméterével méltán szerepel a fejlesztői élvonalban.[14] A BAAI WuDao 2.0 modellt mutatta be először, melynek továbbfejlesztése a WuDao 3.0. A Tsinghua KEG egyetem GLM-130B modellje csak kétnyelvű (angol és kínai), de 80% feletti válaszadási pontossága miatt fejlődését érdemes lesz figyelemmel kísérni.[15] A válaszadási pontosságot többen

[11] Cerullo, Megan: ChatGPT is growing faster than TikTok. https://www.cbsnews.com/news/chatgpt-chatbot-tiktok-ai-artificial-intelligence/?fbclid=IwAR3m39VlEXCVyg4Memv44Lt9EVXEk8f1AiBLBnGdNNa58WvU7j5LYpW96Eg (Letöltés ideje: 2023. 07.13.)

[12] Horváth Péter: Bard AI. Európában is elrajtolt, új képességekkel erősít a Google chatbotja. https://pcworld.hu/pcwlite/europaban-is-elrajtolt-uj-kepessegekkel-erosit-a-google-chatbotja-328188.html (Letöltés ideje: 2023. 07.13.)

[13] Shoeybi, Mohammad – Patwary, Mostofa – Puri, Raul – LeGresley, Patrick – Casper, Jared – Catanzaro, Bryan: Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism https://arxiv.org/pdf/1909.08053.pdf (Letöltés ideje: 2023. 07. 14)

[14] szerző nélkül: https://www.ai21.com/blog/announcing-ai21-studio-and-jurassic-1 (Letöltés ideje: 2023. 07.13.)

[15] GLM-130B: An Open Bilingual Pre-Trained Model https://keg.cs.tsinghua.edu.cn/glm-130b/posts/glm-130b/ (Letöltés ideje: 2023. 07.13.)

vizsgálták, a jelenleg elérhető eredmények fokozott óvatosságra intenek. Olyan esetekben, amikor a feltett kérdés az előtanítási területen kívül esik, akkor a modell meggyőzőnek tűnő, de blöff válaszokat ad, amit a szakirodalom hallucinációnak nevez. Erről számos példát lehet a világhálón fellelni, [16] amit a modell gondozói gyorsan korrigálnak is, és folyamatosan javítják az algoritmusokat is.

A gyógyászatban már korábban is ismeretes volt a GMI alkalmazhatósága. [17] Emiatt nem meglepő, hogy a nagy nyelvi modellek megjelenésekor elsők között próbálták ki az orvosi diagnózisok pontosabb és gyorsabb elkészítésében, a betegségek prognózisának előrejelzésére, azaz arra, hogy egy adott beteg hogyan reagálhat egy adott kezelésre vagy terápiára.[18] Robusztussága miatt kiemelhető Ali és munkatársainak munkája. [19] A kutatók az idegsebészeti önértékelő vizsga 149 kérdésével tesztelték a ChatGPT-t, a GPT-4-et és a Google Bard modelleket. Az eredmények sorban 62,4%; 82,6% és 44,2%, azaz a GPT-4 kimagaslóan jobb eredményt mutatott fel, mint a másik két szereplő. Mindhárom LLM nyelvi készletében szerepel a magyar, ezért nyelvismerettel nem rendelkezők is tudják használni, bár saját tapasztalat szerint az angol válaszok pontosabbak és bővebbek. Mivel a GPT-4 markánsan megbízható válaszokat ad, valamint a pontosságban vetekedő GLM-130B-vel szemben ismeri a magyar nyelvet, ezért a példáknál az előbbi alkalmazhatóságát vizsgálom.

2.2. A GPT bűnügyi kérdései

A GMI alkalmazások a bűnügyi adatok elemzésével különböző formációkat azonosíthatnak és előre jelezhetik a potenciális bűnügyi forrópontokat. Ez a prediktív rendészet proaktív megközelítése, amely lehetővé teszi a rendőrség számára, hogy a forrásokat sokkal hatékonyabban tudják elosztani, valamit az ellenőrzéseket célorientáltabban tudják elvégezni. Ezzel lehetőséget biztosítanak a bűncselekmények megakadályozására még azok megtörténte előtt.

Az egyik hazai biztosító felmérése szerint a kiberkockázatok nagyobb veszélyt jelentenek, mint a természeti katasztrófák.[20] Mivel állításaikat tényadatokkal támasztják alá, ezért nemcsak üzleti, hanem bűnüldözési szempontból is érdemes megszívlelni. A ChatGPT 2022 novemberi kibocsátása után egy hónap sem telt el, és megjelent az első, ezzel az eszközzel elkészített malware. 2022. december 31-ére megírátták az első Dark Web Marketplace scriptet. Ezzel párhuzamosan megindult a szöveghamisítás, úgymint hamis diplomamunkák és tanulmányok elkészítése. Az ügyeskedők számára lehetővé vált párhuzamos munkavégzés emulálása, azaz több munkáltató felé végeztek munkát, amelyet ténylegesen a ChatGPT-vel írattak meg. A kiberkockázatok a phishing e-maileken keresztül

[16] Alston, Elena: What are AI hallucinations and how do you prevent them? Here's how to encourage AI to stop hallucinating. https://zapier.com/blog/ai-hallucinations/ (Letöltés ideje: 2023.07.18.)
[17] Azizi et al. (33 további szerző): Robust and Efficient Medical Imaging with Self-Supervision, https://arxiv.org/pdf/2205.09723.pdf (Letöltés ideje: 2023.07.13.)
[18] Nori, Harsha – King, Nicholas – Mayer McKinney, Scott – Carignan, Dean – Horvitz, Eric: Capabilities of GPT-4 on Medical Challenge Problems, https://arxiv.org/abs/2303.13375 (Letöltés ideje: 2023. 07.14.)
[19] Ali, Rohaid – Tang, Oliver Y. – Connolly, Ian D. – Fridley, Jared S. – Shin, John H. – Zadnik Sullivan, Patricia L. – Cielo, Deus – Oyelese, Adetocunbo. A. – Doberstein, Curtis E. – Telfeian, Albert E. – Gokaslan, Ziya L. – Asaad, Wael. F.: Performance of ChatGPT, GPT-4, and Google Bard on a Neurosurgery Oral Boards Preparation Question Bank. https://pubmed.ncbi.nlm.nih.gov/37306460/ (Letöltés ideje: 2023.07.13.)
[20] Szerző nélkül: https://www.allianz.hu/hu_HU/lakossagi/sajtoszoba/sajtokozlemenyek/a-kiberkockazatok-nagyobb-veszelyt-jelentenek-mint-a-katasztrofak.html (Letöltés ideje: 2023.07.16.)

is megjelentek. Az adathalász e-mailek új generációja helyes megszólítással, pontos nyelvtannal a kívánt nyelven írja meg a leveleket. Ezért az ilyen csalást nem ismerő járatlan felhasználó könnyen csapdába eshet. Ám itt is segíthet a természetes emberi gyanakvás minden olyan ismeretlen, vagy nem várt megszólítással szemben, ami nem illik bele a szokványos képbe.

Annak ellenére, hogy a GPT-vel megíratott szoftverekben előfordulnak hibák, használata nagymértékben lerövidíti a szoftverek elkészítési idejét, ami segíti a csalókat, hogy gyorsan fel tudjanak lépni, ha valahol számukra hasznot hozó informatikai rést fedeznek fel.

A szolgáltatások digitalizálása, az okos otthonok, az IoT eszközök folyamatos terjedése egyre nagyobb kockázatot jelent, a kiberbűnözők egyre több helyen és módon léphetnek támadásba. A generatív nyelvek veszélye, hogy alkalmazásukkal sokkal kevesebb informatikai ismeret kell a digitális szolgáltatásokkal kapcsolatos visszaélésekhez, az okosotthonokba való betöréshez, vagy IoT eszközök működésének megváltoztatására, vagy lehetetlenítésére.

A már hivatkozott levél több más pont mellett kéri a GAO vezetőjét, hogy vizsgálja meg, mennyiben szolgál a GPT a csalás és szélhámosság táptalajául, és a szolgáltatók milyen módon ellenőrzik ezen veszélyek szintjét. A GMI veszélyei között felsorolja a manipulatív hang, szöveg és mesterséges kép előállítását. Példaként hozza hamisított pornófelvételek készítésével okozott károkat is.[21] A hazai bűnesetek unokázós csalásainak eseteit tekintve az MI-vel támogatott csalások esetében az idősebb generáció kiszolgáltatottsága jelentősen növekedhet.[22] Valóban, korábban a csaló üzenetekben a hazai kultúrától eltérő megszólítási és nyelvtani hibák már kis körültekintés mellett is leleplezték az üzenetküldő bűnös szándékát.

3. Merre halad a GPT és mi érzékelhető most?

A fejlődés irányának megválaszolása, de még csak megbecslése is kifejezetten nehéz. A hatalmas fejlődési sebesség és a jelentős konkurenciaharc miatt a fejlődési útvonal nehezen jelezhető előre. Cél, hogy növeljék a válaszok találati pontosságát, az elérhető nyelvek számát és csökkentsék a rendszer előítéletes válaszait és hallucinációját. A tiltás taktikája bizonyosan rossz, mivel a jogkövető magatartást tanúsító állampolgárokat zárná ki a felhasználásból, a bűnözők pedig megtalálnák a hozzáférést.

A GMI rendészeti alkalmazása már megjelent egyes rendészeti célú felhasználóknál. Itt említhető meg a prediktív rendészet, a bűnügyi helyszín rekonstrukciója, az arcfelismerés, a dokumentumok természetes nyelvfeldolgozás alapú elemzése. Ez akkor nyújthat segítséget, amikor a nyomozati anyag nagymennyiségű szöveges információt tartalmaz, így különösen rendőrségi jelentések, a közösségi média és az online kommunikáció adatainak kivonatolása és összegzése a cél.

Az informatikusok vágyait fogalmazza meg Frackiewicz, amikor a döglött ügyek GPT-4-gyel történő megoldását írja le. Kommunikációs cégének honlapján közzétett cikkét

[21] Forrás: https://www.markey.senate.gov/imo/media/doc/senator_markey_letter_to_gao_on_generative_ai_-_062223pdf.pdf (Letöltés ideje: 2023. 07.15.)

[22] Knowles, Jason – Pistone Ann: Thieves can use ChatGPT to write convincing scam messages with human-like language, experts warn https://abc7chicago.com/what-is-chatgpt-google-chatbot-ai-online-scams/12952645/ (Letöltés ideje: 2023.07.13.)

és a benne szereplő forrás nélküli hivatkozásokat igazolt tények még nem támasztják alá. [23] A vízió viszont jónak tűnik. A GPT-3.5 és GPT-4 az elmúlt félév tesztelési tapasztalatai alapján folyamatosan és jelentős mértékben fejlődik, így egyre inkább megbízhatóak a promptra adott válaszok. A GPT-4 jelenlegi fejlettségi szintjén a bizonyítékok gyors értékelésére, holtponton ötleadóként és a sokféle nyelv ismerete okán az idegennyelvű bizonyítékok gyorsfordítására már a rendészeti alkalmazás is megfontolandó.

A feltett kérdések nem ismert harmadik fél általi felhasználása miatt nem egyszerű adatvédelmi, hanem információ kiszivárgási problémák is felmerülnek. A GPT-4, a Bard és a Megatron Turing esetében adatvédelmi szempontból az USA, a WuDao-3.0 és a GLM-130B esetében a Kína és a Jurassic-1 Jumbo esetén pedig Izrael törvényei alkalmazandók. Államérdekekre hivatkozva mindhárom esetben az informatikai rendszerben keletkezett adathoz a szolgáltatónak kötelessége hozzáférést biztosítani a kijelölt hatóságok számára. Az informatikai kiszolgáltatottság mellett, adatvédelmi okok, a fejlesztés tőke- és szakember igényessége indokolja egy közös európai rendszer kifejlesztését. A további LLM fejlesztés lokális szerveren futó változatot is figyelembe vesz, amely adatvédelmi és adatszivárgási szempontból a legnagyobb biztonság elérését teszi lehetővé.

A kétirányú hangátalakítás segítségével akár néhány hónapon belül is sor kerülhet Turing elképzelését is meghaladó ember-gép kommunikációra, azaz interaktív beszélgetésre, ami a bűnügyi szakértői rendszerek új generációjának megvalósítását alapozza meg.

4. Összegzés

A generatív MI egyértelműen jelentős segítséget nyújthat a rendészet területén is, de figyelembe kell venni, hogy a jelenleg ismert modelleket a GDPR védőernyőjén kívül fejlesztők honos titkosszolgálatának (és még nem tudni kinek) alárendelve, a feltett kérdéseket adataink, stratégiai elképzelésünk megszerzésére saját védelmi, politikai és gazdasági céljai érdekében számunkra nem ismert harmadik személy is felhasználhatja. A technológiai függőség csökkentésén túl GDPR szempontból is szükség van EU-s fejlesztésre. Az információbiztonság szempontjából támogatni kell a lokális és védett erőforrásokra épülő generatív MI rendszerek megvalósítását. Szorgalmazni kell azon módszerek kifejlesztését, melyek észlelik a jogellenes felhasználást.

[23] Frackiewicz, Marcin: GPT-4 in Law Enforcement: Solving Cold Cases and Enhancing Forensics. https://ts2.space/en/gpt-4-in-law-enforcement-solving-cold-cases-and-enhancing-forensics/ (Letöltés ideje: 2023.07.13.)