A Nemzeti Közszolgálati Egyetem Információs Társadalom Kutatóintézete havi hírlevele a mesterséges intelligencia alkalmazásáról, társadalmi hatásairól és kérdéseiről

2026 április

Az NKE ITKI honlapja: itki.uni-nke.hu

NEMZETI KÖZSZOLGÁLATI EGYETEM
LUDOVIKA

TARTALOMJEGYZÉK

1083 Budapest, Ludovika tér 2. | +36 1 432 9000
POSTACÍM 1441 Budapest, Pf. 60. | nke@uni-nke.hu, uni-nke.hu

Etika és jog

Az emberi kapcsolatépítés támogatásában kaphat fontos szerepet az MI a munkahelyeken

A mesterséges intelligencia lehetséges szerepe és a munkaerőpiacra, illetve a munka világára gyakorolt hatása folyamatosan a szakmai, politikai és közéleti érdeklődés középpontjában áll. Minden jel arra mutat, hogy ez nem csupán egyike az új technológia-csoporttal felszínre kerülő, de múlandó elméleti és gyakorlati dilemmáknak. Jó okkal feltételezhetjük ma már, hogy ez lesz az egyik legdrámaibb fejleménye az MI-technológiák társadalmi terjedésének. A kérdés elsősorban úgy vetődik fel: kiváltja-e az emberi munkát az MI, és ha igen, milyen mértékben, illetve mely működési területeken és szakmákban? Milyen munkahelyi pozíciókban juthat meghatározó szerephez az új technológia? A kérdések kínzóak, és óhatatlanul ott lebeg felettük a fenyegetés: bárhogyan is alakul ez a kérdéskör, szinte elkerülhetetlenül von majd maga után egyéni és csoportos emberi tragédiákat is. Ebben a meglehetősen komor gondolati környezetben jelent érdekes, bíztató fénysugarat az a program, amiely az MI-kutatások egyik legnevesebb szakmai műhelye, az amerikai Stanford Egyetem HAI (Ember Központú MI) laboratóriumának az égisze alatt indult el. Az alapgondolat az volt, hogy az új MI technológiák nem csak fenyegetést jelenthetnek az emberek munkahelyére, jövendő munkájára. Itt nem arról a feltevésről van szó, hogy – ahogyan az a társadalmi innovációk történetében máskor is rendre előfordult – miközben munkahelyeket vesz el az MI, ugyanakkor új készségekre, új tudásokra épülő, korábban nem is létező munkahelyeket is teremt majd. Amit a Stanford kutatói felvetettek, az az, hogy a munkahelyek normális működésében, valamint a munkához kapcsolódó emberi viszonyokban óriási szerepe van a humán kapcsolatok motorjának tekinthető „társas készségeknek”. Egyes szakmákban – például a tanácsadók, az orvosok, a pszichológusok és különösen a pedagógusok esetében – döntő jelentőségük van ezeknek a különleges készségeknek. Ugyanakkor az is igaz, hogy szinte minden munkatevékenység során jelentősen hozzájárulhatnak a folyamatok zökkenőmentességéhez és eredményességéhez. A stanfordi kutatók által kidolgozott MI-partner/MI-mentor (AI Partner/AI Mentor) program a nagy nyelvi modellek inherens szerepjátszó képességére alapozva dolgoz ki olyan MI-modelleket, amelyek a munkahelyi társas együttműködés és a konfliktuskezelés emberi technikáinak a tanulásában nyújtanak segítséget a dolgozóknak. A modellek később egyfajta személyes „konfliktus tanácsadóként” segíthetnek a problémák felfedésében, megoldásában is.

Using LLMs to Improve Workplace Social Skills

1083 Budapest, Ludovika tér 2. | +36 1 432 9000
POSTACÍM 1441 Budapest, Pf. 60. | nke@uni-nke.hu, uni-nke.hu

Mesterséges biológiai intelligencia mesterséges intelligencia segítségével

A rangos IEEE Spectrum folyóiratban megjelent cikk központi állítása szerint az MI és a DNS-szintézis fejlődése a biológiát fokozatosan mérnöki tervezési területté alakíthatja. A cikk kiindulási pontja egy frissen megjelent könyv, az On the Future of Species: Authoring Life by Means of Artificial Biological Intelligence. A könyv szerzője ezt a kibontakozó képességet mesterséges biológiai intelligenciának, azaz artificial biological intelligence-nek (ABI) nevezi: olyan rendszerek összességének, amelyek képesek lehetnek új genetikai szekvenciák megtervezésére, fizikai előállítására, majd végső soron élő rendszerek „beindítására" is.

Az ABI szükségességét éppen az indokolja, hogy a genom evolúciósan kialakult „spagettikód": átfedő funkciókkal teli, nem ortogonális rendszer, amelyet négymilliárd éve nem refaktoráltak, és amely ellenáll a hagyományos mérnöki elveknek. A hagyományos genetikai módszerek ezért nem elegendők – a biológia mérnöki anyaggá alakításához új típusú eszközökre van szükség. Magyarázat, hogy érthetőbb legyen a mondanivaló: a genomon egy élőlény teljes örökítő anyagát – az összes DNS-szekvencia, amely az adott szervezet felépítéséhez és működéséhez szükséges információt tartalmazza – értjük. Ortogonális rendszer: a rendszer komponensei egymástól függetlenek és világosan elkülönülnek: minden alkotó résznek egyetlen, jól meghatározott funkciója van, és az egyik elem módosítása nem befolyásolja kiszámíthatatlanul a többit. Refaktorálás: egy program belső szerkezetének átírása úgy, hogy a működése változatlan marad, de a kódja átláthatóbb, rendezettebb és könnyebben módosítható lesz. A könyv szerzője szerint az MI különösen azért jelent áttörést, mert a korábbi módszerekkel nem lehetett nagy léptékben, célzottan új DNS-szekvenciákat tervezni. Az új ún. genomnyelvi modellek a DNS-t nyelvként kezelik: nem szavakkal és mondatokkal, hanem a DNS négy „betűjével" dolgoznak. A kihívás azonban jóval nagyobb, mint a természetes nyelv esetében, mert a DNS-ben egymástól távoli régiók is befolyásolhatják egymás működését. Ezért van szükség nagyon nagy kontextus ablakú modellekre – például az Evo 2 architektúrájára, amely akár egymillió bázispárnyi távolságon belül is képes összefüggéseket felismerni. Magyarázat: A nukleotid a nukleinsavak (DNS és RNS) építőköve. A bázispár két nukleotid bázisának összekapcsolódása. Egy átlagos gén néhány ezer bázispár hosszú. Egy baktérium teljes genomja néhány millió bázispárból áll. Az emberi genom körülbelül hárommilliárd bázispárt tartalmaz.

Az MI szerepe azonban nem korlátozódik a tervezésre. A szerző hangsúlyozza, hogy nem ismerjük eléggé „az élet grammatikáját" – azokat a mélyebb szabályokat, amelyek alapján a DNS-szekvenciák funkcionális élő rendszerekké szerveződnek. Az MI éppen ezért kulcsfontosságú: hatalmas adatbázisokon képes mintázatokat felismerni, és így feltárni e grammatika rejtett szabályait. Komplex többsejtű organizmusokat addig nem fogunk tudni tervezni, amíg ezt a nyelvtant folyékonyabban nem beszéljük – ez ma az ABI legfőbb tudományos korlátja. Az ABI másik pillére a DNS fizikai „megírásának" képessége. A biológia akkor válik valódi mérnöki anyaggá, ha a tervezés és a kivitelezés összekapcsolódik: az MI megtervezi a kívánt genetikai kódot, a DNS-szintézis pedig gyorsan, olcsón és nagy léptékben előállítja azt. A legnehezebb lépés azonban továbbra is az élő rendszer „elindítása". Ma még nem tudunk teljesen mesterséges sejtet létrehozni, és a genom sejtekbe juttatása is komoly technikai akadály. Ennek illusztrálására a könyv a sejtet nanokomputerként, a genomot pedig szoftverként írja le – a két rész összekapcsolása azonban korántsem triviális. Ha mindezek a képességek összeállnak, az ABI ígérete radikális: a DNS programozható, prediktíven alakítható mérnöki anyaggá válhat. A könyv előrejelzése szerint fél évszázadon belül a biológiai rendszerek válhatnak a mérnöki tervezés elsődleges alapanyagává, amelyek nemcsak a hagyományos anyagok funkcióit nyújtják (a

1083 Budapest, Ludovika tér 2. | +36 1 432 9000
POSTACÍM 1441 Budapest, Pf. 60. | nke@uni-nke.hu, uni-nke.hu

pókselyem szakítószilárdsága például MI-vel újratervezve akár ötszöröse lehetne az acélénak), hanem új lehetőséget is megnyitnak: az „intelligens anyagok" létrehozását. A kockázatok azonban jelentősek és többrétegűek. Mechanikai értelemben az élő rendszerek a refaktorálás során törékennyé válhatnak, mert eltűnnek az evolúciósan kialakult, átfedő hibabiztosító mechanizmusok. Ökológiai értelemben a mesterséges organizmusok kiszámíthatatlan károkat okozhatnak a komplex ökoszisztémákban. Társadalmi értelemben pedig maga a technológia eredendően veszélyes, ha rossz kezekbe kerül – ezért elengedhetetlen, hogy biztonságosan, felelősen, etikusan, átláthatóan és méltányosan, a társadalom javára használjuk.

Can Biologists Rewrite the Genome’s Shagetti Code?

Mesterséges intelligenciával személyre szabott gyógykezelés: nagy ígéretek, megoldandó kérdések

A kutatók egyöntetű vélekedése ma az, hogy a mesterséges intelligencia technológiák transzformatív ereje egyetlen más területen sem lesz olyan átfogó és mély, mint az orvoslás és a gyógyszerfejlesztés terén. A témához kapcsolódó híradások egyre szaporodnak, amit jól mutat az az eset is, amely az elmúlt hetekben kavarta fel a fél világot. A történetben egy orvos-biológus képesítés nélküli MI-szakember a nyilvánosan elérhető ChatGPT, AlphaFold és Grok nagy nyelvi modellek segítségével néhány hét alatt egy olyan személyre szabott mRNS vakcinát „tervezett”, amelynek alapján egy ausztrál kutatóintézet szakemberei hatásos terápiás anyagot tudtak előállítani az MI-mérnök daganatos betegségben szenvedő kutyája kezelésére. Az ilyen megható történetek joggal hívják fel a közvélemény figyelmét az MI-technológiákban rejlő óriási medicinális képességekre. Az eset keveset hangsúlyozott árnyoldala azonban az, hogy az ilyen felfedezések, fejlesztések ugyan látványos lehetőségekkel kecsegtetnek, ám ezek széles körű alkalmazhatósága (szakkifejezéssel szólva „skálázhatósága”) egyelőre nagyon komoly akadályokba ütközhet. A legnagyobb nehézséget talán a kiindulást jelentő adattömeg szeparáltsága, nehéz hozzáférhetősége jelenti. A nagy port felvert előbbi hírhez kapcsolódóan például el kell mondani: az állatorvosi jellegű adatok többsége ma szerzői és más tulajdonjogokkal védett, és ezek a zárt adatsilók rendkívül megnehezíthetik az adattömegeken (big data halmazokon) működő MI-modellek munkáját. Lehetséges megoldásként vetik fel egyes szakértők egy globális állatorvosi közadatbázis létrehozásának a gondolatát. Azután persze ott van a költségek kérdése: a személyre szabott mRNS vakcinák kifejlesztése, előállítása sok ezer dolláros költséggel járhat, ami nyilvánvaló gátja a tömeges alkalmazhatóságnak. Nem elhanyagolható kérdés ugyanakkor a jogi szabályozás (vagy annak hiánya) problémája sem. Ezek nélkül az MI-alapú állatorvoslás vagy ember orvoslás csak egyfajta jogi szürke zónában tud mozogni.

How AI is making medicine personal, accessible and complicated

1083 Budapest, Ludovika tér 2. | +36 1 432 9000
POSTACÍM 1441 Budapest, Pf. 60. | nke@uni-nke.hu, uni-nke.hu

Trendek

Nyelvi modellek után jönnek a világmodellek?

A mai MI-rendszerek lenyűgöző teljesítményre képesek a digitális világban – szöveget írnak, kódot generálnak, érvelnek –, a fizikai környezetben azonban továbbra is bizonytalanul mozognak. Egy regény megírása vagy egy alkalmazás kódjának megírása sok tekintetben könnyebb feladat, mint például a szennyes ruha összehajtogatása vagy egy városi utcán való eligazodás. Sok kutató szerint ezt a szakadékot világmodellekkel lehet áthidalni: ezek olyan belső reprezentációk, amelyek segítségével az MI-rendszer nem csupán mintázatokat ismer fel, hanem leképezi a külvilág szerkezetét, és előre tudja jelezni cselekedetei következményeit – ahogyan az emberi agy is teszi, amikor megsejti, mi történik, ha lelökünk egy bögrét az asztal széléről.

A téma azért került most az MI-viták középpontjába, mert egyszerre több jelentős fejlemény erősíti: a Google DeepMind, valamint Fei-Fei Li stanfordi professzor World Labs nevű vállalkozásának friss fejlesztései új modelleket mutatnak be; Yann LeCun, a Meta vezető MI-tudósa feltűnő módon távozott a Metától, hogy világmodellekre fókuszáló startupot alapítson; az OpenAI pedig erőforrásokat csoportosít át hosszabb távú világ-szimulációs kutatásokra. Az elképzelés támogatói szerint a világmodellek meghaladhatják a nagy nyelvi modellek ismert korlátait. A nagy nyelvi modellek ugyan látszólag sokat tudnak a világról, de tudásuk törékeny. A világmodellel rendelkező rendszerek robusztusabbak lehetnek, és Li szerint olyan robotok megalkotását is lehetővé teszik majd, amelyek a mélytengert kutatják, vagy az egészségügyben segédkeznek. A jelenlegi alkalmazások egyelőre szerényebbek. A Pokémon Go készítői a játékosok által gyűjtött milliárdnyi képre építve dolgoznak egy kezdeti világmodell elemein, amelyet később kézbesítőrobotok irányítására remélnek használni. A Google DeepMind és a World Labs interaktív, háromdimenziós virtuális környezeteket generáló modelleket fejleszt, elsősorban videójátékokhoz és immerzív VR-élményekhez. A cikk záró gondolata szerint az igazi áttörés akkor jöhet majd el, ha ezeket a rendszereket rugalmas, intelligens ágensekbe integrálják, amelyek képesek reprezentálni a környezetüket, előre jelezni cselekedeteik következményeit, és ennek alapján dönteni.

World models

1083 Budapest, Ludovika tér 2. | +36 1 432 9000
POSTACÍM 1441 Budapest, Pf. 60. | nke@uni-nke.hu, uni-nke.hu

Vége az olcsó mesterségei intelligencia paradicsomi állapotának

A The Verge cikke szerint az MI-szolgáltatások most érkeznek el ahhoz a ponthoz, ahol az eddigi „olcsó vagy ingyenes" korszak gazdaságilag tarthatatlanná válik. A felhasználók ezt hirdetésekben, szigorúbb használati korlátokban, funkciókorlátozásokban és áremelésekben fogják megérezni. A mélyebb ok az, hogy a vezető MI-cégek – például az OpenAI és az Anthropic – óriási mennyiségű tőkét égettek el a modellek fejlesztésére, adatközpontokra és számítási kapacitásra. A befektetők eddig a növekedést finanszírozták, most viszont megtérülést várnak. Az üzleti modell középpontjában a token áll. A token az a kis adategység, amelyet a modell feldolgoz: lehet szövegrészlet, képi vagy hangalapú adat. A szolgáltatók sok esetben token alapon árazzák a használatot. Az MI korai szakaszában a költségek főleg a betanításnál keletkeztek, ma viszont az inferencia – vagyis a modell tényleges válaszadása, kódolása, ügynökként végzett munkája – legalább ekkora terhet jelent. Ennek oka, hogy az újabb, „gondolkodó" modellek és MI-ügynökök rengeteg tokent használnak el a háttérben: alternatív megoldásokat mérlegelnek, részfeladatokat futtatnak, ellenőrzik saját lépéseiket, majd visszalépnek, ha zsákutcába jutnak. Ez a felhasználó számára láthatatlan, de a szolgáltatónak nagyon is költséges. A Gartner cég elemzése szerint 2024 és 2029 között körülbelül 6,3 billió dollárnyi tőkebefektetés ömlik MI-adatközpontokba. Ahhoz, hogy ez akár csak minimálisan megtérüljön, a szektornak ugyanebben az időszakban közel 7 billió dollár MI-bevételt kellene termelnie – vagyis a token fogyasztásnak nagyjából 50 000-100 000-szeresére kellene nőnie a mai szinthez képest. Ez két okból is nehéz. Egyrészt a cégek már most is kapacitáshiánnyal küzdenek, fizikailag sem képesek ennyi tokent előállítani. Másrészt a tokenenkénti közvetlen haszonkulcsot felemésztik a közvetett költségek – a következő modellgeneráció betanítása és az új infrastruktúra folyamatos kiépítése –, és a helyzet különösen kedvezőtlen a tokeneket valósággal faló, érvelő modellek esetében, ahol az elérhető haszonkulcs már most is alacsony, sokszor negatív.

A cikk szerint ezért a nagy fejlesztő cégek kettős kényszerhelyzetben vannak. Rövid távon csökkenteniük kell a veszteséget: a vállalati felhasználók esetében átálltak a token alapú számlázásra, drágább szolgáltatási csomagokat vezetnek be, és korlátozzák a harmadik féltől származó eszközök kapcsolódását; a fogyasztói piacra szánt termékekben – például a ChatGPT-ben – pedig megjelennek a hirdetések. Hosszú távon viszont éppen ellenkezőleg: tömeges token használatra lenne szükségük ahhoz, hogy az adatközpont-beruházások megtérüljenek. Egyszerre kellene tehát visszafogniuk a pazarlást és növelniük a fogyasztást. A költségeket sem tudják egyszerűen áthárítani: a modellek között a váltási költség közel nulla, így aki túl drasztikusan árazna, gyorsan elveszítené a felhasználóit a versenytársak javára. A Gartner ezt „stegoszaurusz-paradoxonnak" hívja: a hatalmas test (az MI-iparág) számára túl kicsi a száj. A fenntartható működéshez a generatív MI-nek lényegében mindenbe, a hirdetőtábláktól a pénztárkioszkokig be kellene épülnie. Ez a feszültség valószínűleg piaci konszolidációhoz vezet: a Gartner előrejelzése szerint regionális piaconként legfeljebb két nagy nyelvi modell-szolgáltató marad életben. A cégek eközben olcsóbb, nyílt forráskódú vagy saját üzemeltetésű modellekkel kísérleteznek, de a legjobb teljesítményű modellek – főleg kódolásnál és összetett feladatoknál – továbbra is drágák. A szerző szerint az MI ingyenes korszaka nem természetes állapot volt, hanem befektetői támogatással finanszírozott területszerzés. Most kezdődik a valódi árazás korszaka.

[You’re about to feel the AI money squeeze]

1083 Budapest, Ludovika tér 2. | +36 1 432 9000
POSTACÍM 1441 Budapest, Pf. 60. | nke@uni-nke.hu, uni-nke.hu

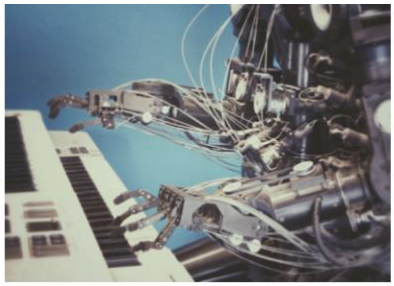
A mesterséges intelligencia és a munkaerőpiac körüli vita: három nézőpont a munka jövőjéről

A neves Carnegie Endowment kutatóhely most közzétett, az MI és a munkaerőpiac jövőjéről szóló jelentése nem egyetlen előrejelzést kínál, hanem három versengő értelmezési keretbe rendezi a vitát: az „aggódók", a „türelmesek" és a „lelkesek" álláspontját ismerteti. A szerző szerint a vita lényegében két kulcskérdésre szűkíthető: milyen gyors az MI-képességek fejlődése és mekkora akadályokba ütközik az MI-rendszerek bevezetése, illetve milyen erős és gyors az MI-vezérelt munkahelyteremtés. Az „aggódó" nézet abból indul ki, hogy a fejlett MI-rendszerek egy évtizeden belül a fehérgalléros munkafeladatok jelentős részét kiválthatják, mert a vállalatok a költségelőnyök miatt gyorsan alkalmazzák őket. Ebben a forgatókönyvben a veszélyeztetett szakmák a következők: belépő szintű irodai, jogi, pénzügyi, ügyfélszolgálati, jogi asszisztensi és szoftveres munkakörök. A „türelmes" álláspont szerint a technikai képességek, a megbízhatóság és a szervezeti bevezetés korlátai miatt az MI inkább több évtizedes – akár ennél is hosszabb - távon fejti ki hatását. Narayanan és Kapoor nyomán e tábor az MI-t „normál technológiaként" kezeli. Hivatkozott bizonyíték a Scale AI Remote Labor Indexe, amely szerint a legjobb modell 2026 márciusában mindössze 4,17%-os arányban oldott meg feladatokat. Ebben a forgatókönyvben hangsúlyos a hallucinációk, az ellenőrzési költségek, a hallgatólagos tudás, a jogi-felelősségi akadályok és a lassú munkafolyamat-újratervezés szerepe.

A „lelkes" megközelítés – itt egyaránt vannak piaci optimisták (Andreessen, Tan) és a feltételes optimisták (Autor, Brynjolfsson, akik szerint csak megfelelő intézményi környezetben várható kedvező kimenetel) – három mechanizmusra épít. Az MI-rendszerek hatására újrarendezik a munkaköri feladatokat, megnő a magasabb hozzáadott értékű emberi tevékenység szerepe; olyan jövedelmi hatások vannak, amelyek emberközpontú szolgáltatások iránti keresletté alakulnak; valamint új MI-vezérelt vállalkozások alakulnak, integrációs, minőségbiztosítási és modellfelügyeleti munkakörökkel. A jelentés végkövetkeztetése óvatos: a három forgatókönyv nem zárja ki egymást, akár szekvenciálisan is bekövetkezhetnek. A szakpolitikának nem egyetlen forgatókönyvre kellene fogadnia, hanem javítania kell az adatgyűjtést, figyelemmel kísérnie az MI-képességek, az MI-bevezetés és a bérhatások indikátorait, valamint kísérleteznie célzott átképzési, mikrotanúsítványi és bérbiztosítási programokkal.

The AI Labor Debate: Three Views on the Future of Work

1083 Budapest, Ludovika tér 2. | +36 1 432 9000
POSTACÍM 1441 Budapest, Pf. 60. | nke@uni-nke.hu, uni-nke.hu



Működésben

Az MI-ipar saját magát is automatizálná

Az ágentikus MI, azaz az ügynöki mesterséges intelligencia valóban felkavarta a technológiai világ állóvíznek egyébként sem nevezhető közegét. Egyes országokban – mint az MI-verseny élvonalában haladó Kínában – egyfajta népi őrületté fajult az ügynökök és ügynökcsapatok építése és működtetése, de a jelek szerint a technológiai szuperhatalomnak számító Egyesült Államok fejlesztői régiójában, a Szilícium-völgyben is szinte vallásos tisztelettel és lelkesedéssel tekintenek az MI-ügynökök eljövetele elé. Az igazi csavart itt az jelenti, hogy a legnagyobb várakozás az amerikai fejlesztői körökben az önfejlesztő, önmagukat „építő" MI-ügynökökhöz kapcsolódik. Az öntanító, önjavító bot gondolata persze nem újdonság: egy I. J. Good nevű statisztikus már az 1960-as években bedobta a köztudatba ezt a fogalmat. De az a gondolat, hogy egy ilyen mesterséges intelligencia ténylegesen megépíthető legyen, az még néhány évvel ezelőtt is az álomvilágba tartozott volna és ennek kapcsán mindenki az akkor közkinccsé vált ChatGPT néhány közismert, súlyos hiányosságát sorolta. Az utóbbi hónapok gyors fejlődése, elsősorban a nagy nyelvi modellek programozási képességeinek drasztikus javulása alapvetően megváltoztatta az önépítő MI-vel kapcsolatos vélekedéseket és elvárásokat. Tény és való, hogy az MI-fejlesztés folyamatai és szegmensei bővelkednek a kulimunkát jelentő tevékenységekben, a nagy adatkészletek kezelésétől az ismételt kísérletekig. Dario Amodei, az Anthropic vezetője úgy becsüli, hogy az MI-alapú programozó eszközök 15-20%-kal gyorsítják fel vállalatánál a fejlesztési munkafolyamatokat. Az igazság az, hogy az MI-kutatás-fejlesztés automatizálásának már egy csekély növekedése is alapvetően boríthatja fel a technológiai fejlesztés egész dinamikáját. Még jobban felgyorsulhat az MI-versenyfutás, annak minden geopolitikai kihatásával együtt. Nem véletlen, hogy tavaly, a DeepMind, az OpenAI, Anthropic, a Meta, a Berkeley Egyetem, a Princeton, és a Stanford 25 vezető kutatójával készített interjúban húszan az MI-fejlesztés automatizálását jelölték meg a jövő legsúlyosabb veszélyforrásának.

The AI Industry Wants to Automate Itslef

1083 Budapest, Ludovika tér 2. | +36 1 432 9000
POSTACÍM 1441 Budapest, Pf. 60. | nke@uni-nke.hu, uni-nke.hu

Leginkább a szoftverfejlesztésben hódít teret az ágentikus MI

Az ágentikus modellek, MI-ügynökök divatja – amely valójában már a tavalyi évben kezdett egyértelműen kibontakozni, de idén vált elsöprő technológiai fejlesztési hullámmá – meghatározó jelensége lesz a 2026-os évnek, ebben ma már szinte teljes a szakmai egyetértés. Mindenki – beleértve a mérnöki tudással nem rendelkező emberek millióit is – ágentikus modelleket, ügynököket vagy éppen együttműködő ügynökcsapatokat használó alkalmazásokat fejleszt. A brit AI Security Institute (AISI) kutatói azonban most már arra voltak kíváncsiak, hogy vajon ténylegesen kik, milyen szakmák képviselői azok, aki a legnagyobb fantáziát látják az MI-ágensekben. A vizsgálat során a szakemberek három nagy funkcionális kategóriába sorolták be az előállított ágenseket: a percepció-fókuszú, az érvelő jellegű, illetve a cselekvésre szakosodott eszközöket vették figyelembe. A kutatás eredménye egyértelműen azt bizonyította, hogy a tömegével előállított MI-ágenseket döntően a szoftverfejlesztő ipar állította munkába tevékenységei támogatására. Az összes megvizsgált MI-ügynök 67%-a a programozási munkafolyamatokba épült bele, vagy azokat támogatta áttételesen. Ez a megállapítás nagyban egybeesik az LLM-ek kezdeti terjedési területeivel: a ChatGPT berobbanását követő első hónapokban és években a nagy nyelvi modellek üzleti célú felhasználói is jellemzően a programozás területéről kerültek ki.

Érdekes megállapítás és a fejlődés irányának fontos indikátora az is, hogy a vizsgált másfél évnyi időszak során 27%-ról 65%-ra növekedett az összes ágentikus MI-fejlesztésen belül a cselekvésre szakosodott eszközök aránya. Jól mutatja ez, hogy az ember közvetlen beavatkozása nélküli döntési, tanulási és tervezési műveletek mellett egyre nagyobb szerepet kapnak az ügynöktípusú MI-modellek önálló cselekvési és eszközhasználati képességei. Lépésről lépésre közelítünk a fizikai mesterséges intelligencia világa felé.

Software development dominates AI agent tool usage

Koboldok, gonosz szellemek, mosómedvék, trollok, ogrék, galambok – a mesterséges intelligencia csodái

Az új OpenAI nagy nyelvi modell, a GPT-5.5 utasításai között mélyen elrejtve egy különös, a nyomaték kedvéért négyszer is megismételt utasítást fedeztek fel: „Soha ne beszélj koboldokról, gonosz szellemekről, mosómedvékről, trollokról, ogrékról, galambokról, vagy más állatokról és lényekről, hacsak ez nem abszolút és egyértelműen releváns a felhasználó kérdése szempontjából.” A felfedezés híre gyorsan elterjedt, számos humoros kommentár kíséretében. Azonban a humor mögött egy mélyebb, komolyabb probléma is rejtőzik.

A ChatGPT beállításaiban lehetőség van arra, hogy meghatározzuk, a rendszer milyen személyiségként viselkedjen a kommunikáció során. A „személyiség” ez esetben előre beállított válaszadási stílust jelent. A választható személyiségek sorában van egy, amit „Különc”-nek hívnak, amely játékos és képzeletgazdag válaszokat ad. A hiba ennek a személyiségnek a betanítási folyamatában keletkezett. Megerősítéses tanulás emberi visszajelzéssel (Reinforcement Learning from Human Feedback, RLHF) a neve ennek a folyamatnak, amelynek során az értékelők túl gyakran jutalmazták a modell túlságosan

1083 Budapest, Ludovika tér 2. | +36 1 432 9000
POSTACÍM 1441 Budapest, Pf. 60. | nke@uni-nke.hu, uni-nke.hu

fantáziadús, mesebeli lényekkel teli metaforáit. A modell ebből nem azt tanulta meg, hogy ezek bizonyos stílushelyzetekben működhetnek, hanem általánosabb szabályként rögzítette, hogy a mesebeli lényekkel teli metaforák jó válaszokat jelentenek. A viselkedés ezután átszivárgott a rendszer más személyiségeibe is, majd a későbbi modellek finomhangolási adataiba bekerülve tovább erősödött. Amíg el nem távolítják az összes olyan adatot, ami a téves viselkedésért felel, kénytelenek explicit utasításban megtiltani a mulatságot okozó válaszokat.

A komolyabb probléma az, hogy ebben az esetben, a mulatságos formában, az MI-betanításának egy strukturális kockázata öltött testet, nevezetesen, hogy a modellek nem mindig azt tanulják meg, amit a fejlesztők szándékoznak jutalmazni. A visszajelzési és jutalmazási folyamatban egy adott stíluselem – jelen esetben a játékos metaforák használata – könnyen a „jó válasz" általános jelévé válhat a modell számára. Ez különösen fontos, mert a hiba nem egy egyszerűen kijavítható felszíni megfogalmazási probléma volt, hanem fokozatosan beépült a modell viselkedésébe, majd a későbbi finomhangolási adatokon keresztül tovább is öröklődött. Az RLHF és más utólagos betanítási módszerek nem pusztán finomítják a modell működését, hanem tartós viselkedési mintákat hozhatnak létre. Ha a jutalmazási rendszer téves vagy túl tág összefüggéseket erősít meg, akkor a modell ezeket olyan kontextusokban is alkalmazhatja, ahol már nem kívánatosak. Nemcsak ártalmatlan stílusbeli furcsaságokat, hanem rejtettebb torzításokat, szakmai pontatlanságokat vagy nem megfelelő döntési mintákat is megerősíthetnek. Ennek következtében a modellek viselkedésének mélyebb auditálásra, az adatok gondosabb szűrésére és a betanítási/jutalmazási folyamat mellékhatásainak folyamatos vizsgálatára van szükség.

Why OpenAI's 'goblin' problem matters – and how you can release the goblins on your own

1083 Budapest, Ludovika tér 2. | +36 1 432 9000
POSTACÍM 1441 Budapest, Pf. 60. | nke@uni-nke.hu, uni-nke.hu