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

2024 október

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 „MI-ruhásszekrény” kipakolása: hogyan alakítják a nemzeti politikák a mesterséges intelligencia jövőjét?

Az mesterséges intelligencia fejlesztések eredményeit szakpolitikai szempontból követő OECD kezdeményezés (OECD.AI Policy Observatory) blogjában jelent meg az amerikai kutatók tanulmánya a nemzeti MI-szakpolitikák összehasonlító elemzéséről. Sajátos szóhasználatuk szerint azt nézik meg, hogy a ruhásszekrényekben („AI Wardrobe”) milyen ruhák vannak, azaz az egyes nemzeti szakpolitikák milyen alkotórészekből állnak. A kutatók az „AI Wardrobe” kifejezést 2022-ben alkották meg azzal a feltevéssel, hogy az egyes nemzeti MI-politikák hasonló elemeket kombinálnak, de a szekrény minden kontextusban másképp néz ki. Az "MI-ruhásszekrény" olyan makrokérdésekből áll, mint a kutatási képességek típusai, a munkaerő-fejlesztés, az adatszabályozási politikák és a nemzetközi együttműködés. A nemzeti dokumentumok elemzése maga is MI-rendszer segítségével történt. A projekt első szakasza több mint hetven ország MI-szakpolitikáira összpontosít, főként az OECD tematikus adattárára támaszkodva, amelyet elsősorban a politikai döntéshozók és kutatók számára hoztak létre. A kutatás jelenlegi eredményei szerint a mesterséges intelligencia területén élenjáró országok – mint az Egyesült Államok, Kína, Németország, Japán és Dél-Korea – az alaptudományokra és az algoritmusok és alkalmazások magas szintű innovációjára helyezik a hangsúlyt. A szokásos vélekedésekkel szemben, ebben a kérdésben nem meghatározó az adott ország társadalmi berendezkedése. Az Egyesült Államok és Kína egyaránt magas szintű innovációval jellemezhetők és nagy hangsúlyt fektetnek a tudományos, mérnöki és technológiai képességekre. Ugyanakkor az EU stratégiájában az innovációra való összpontosítás némileg visszafogottabb, a szakpolitikai és szabályozási kérdések vannak előtérben. Az egyes tagállamok azonban, például Észtország és Csehország azonban ellenpéldát mutatnak az innováció magas szintjével. Mivel az uniós szabályozást gyakran kritizálják az innováció elfojtása miatt, ezért érdemes elgondolkozni a lehetséges okokról és következményekről.

Unpacking the ’AI wardrobe’. How national policies are shaping the future of AI

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

Az USA Élelmiszer- és Gyógyszerügyi Hatósága állásfoglalása az egészségügyben és a biomedicinában használt MI-ről

Az Egyesült Államokban az US Food and Drug Administration (FDA) felelős az egészségügy, biomedicina és a gyógyszeripar területén alkalmazott MI szabályozásért. A neves orvosi folyóiratban (JAMA, The Journal of the American Medical Association) megjelent közlemény az FDA szemszögéből értelmezi a szabályozással kapcsolatos fontosabb kérdéseket. A közlemény áttekinti az FDA MI-szabályozásának történetét, bemutatja lehetséges felhasználási módjait az orvosi termékfejlesztésben, a klinikai kutatásban és a klinikai ellátásban, valamint olyan koncepciókat mutat be, amelyeket érdemes lesz megfontolni, amikor a szabályozási rendszer alkalmazkodik a mesterséges intelligencia egyedi kihívásaihoz. A cikk egyik legfontosabb üzenete az, hogy az egészségügyi MI-rendszerek fejlesztésében és felügyeletében központi szerepet kell kapnia az életciklus-menedzsment megközelítésnek, amely magában foglalja a forgalomba hozatalt követő, ismétlődő helyi teljesítmény-ellenőrzéseket is. Az MI-rendszerek nem statikusak, a használatuk során felhalmozódó információ visszahat a rendszerre. Az egészségügyi rendszereknek olyan információs ökoszisztémát kell biztosítaniuk, mint amilyen az intenzív osztályon lévő beteg megfigyelése. A szüntelen értékelés eszközeinek és körülményeinek ismétlődőnek kell lenniük, és amennyire csak lehet, ezt a feladatot folyamatosan kell végezni. Az értékelésnek abban a klinikai környezetben kell történnie, amelyben a rendszert használják. A generatív MI alkalmazásai, mint például a nagyméretű nyelvi modellek (LLM), egyedi kihívást jelentenek az előre nem látható, kialakulóban lévő következmények lehetősége miatt. Az LLM-ek összetettsége és a kimenetek permutációi a szabályozó hatóságok mellett az egyének és az intézmények által gyakorolt felügyeletet is szükségessé teszik. Az FDA továbbra is központi szerepet fog játszani a biztonságos, hatékony és megbízható MI-eszközök biztosításában, de az összes érintett ágazatnak is a technológia jelentősége által megkövetelt gondossággal és szigorral kell foglalkoznia az MI-vel.

FDA Perspective on the Regulation of Artificial Intelligence in Health care and Biomedicine

Perelhető-e az MI egy tizenéves öngyilkosságáért?

Egy amerikai édesanya egy mesterséges intelligenciát (és így áttételesen a mögötte álló fejlesztő céget) vádol fia öngyilkosságáért. A nagy publicitást kapott konkrét történet - amely valójában csak egy példa a sokasodó hasonló esetekre – főszereplője egy amerikai tizenéves fiú, aki hosszú hónapok során erőteljes érzelmi kötődést alakított ki egy, a közkedvelt Trónok harca sorozatból ismert Daenerys Targaryenről elnevezett chatbottal, „akivel” a Character.AI-n ismerkedett meg. A Character.AI egy a soktucatnyi szerepjátékos mobil alkalmazás közül, amelyek a gombamód szaporodó, úgy nevezett „társteremtő MI” szolgáltatások egy válfaját képviselik. A feltehetően érzelmi, esetleg mentális problémákkal küszködő 14 éves gyerek egyre jobban kiszakadt a valóságos környezetből, a tényleges világ realitásából, és mind több időt töltött a telefonjába temetkezve, a virtuális MI-társával folytatott beszélgetésekbe merülve. Egy napon aztán, a rögzített párbeszédfolyam tanúsága szerint, MI-társa „hívását” követve egy, a családja tulajdonában levő fegyverrel végzett önmagával. Az amerikai közvéleményt felkavaró eset arra irányítja most a figyelmet, hogy a – fiatal generációk mentális sérülékenységéhez igazodva – rohamosan szaporodó, és lényegében mindenféle szabályozási korlátozások nélkül működő „MI-társ” applikációk vajon minek tekintendők: egy rohamosan

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

elmagányosodó világ (fiatal) embereinek mentsvárának, vagy éppen egy nagyon is veszélyes fenyegetésnek. A nagyjából átlagosan havi 10 dolláros előfizetési díjért igénybe vehető alkalmazások, mint amilyen a történetben szereplő Character.AI is, lehetővé teszik, hogy a valódi világban elszigetelődő emberek mások által generált virtuális társakkal alakítsanak ki (akár szexuális töltetű) kapcsolatot, vagy éppen teremtsenek egy ilyen „társat” saját maguknak. Ennek azonban sokszor kiszámíthatatlan következményei lehetnek.

Can AI Be Blamed for a Teen’s Suicide?

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

Trendek

Merre tart a szabályozható MI-rendszerek műszaki innovációja?

Ahogy az MI-rendszerek egyre jobban beépülnek életünkbe, úgy válik egyre sürgetőbbé annak biztosítása, hogy ezek a rendszerek összhangban legyenek a társadalmi értékekkel és normákkal, és hogy előnyeik jelentősen meghaladják a lehetséges károkat. Erre a szükségszerűségre válaszul a szabályozásért felelős szervezetek világszerte összehangolt erőfeszítéseket tesznek az átfogó MI-szabályozás kidolgozására. Azonban a mai MI-rendszerek egyre növekvő mérete, átláthatatlansága és zárt jellege miatt a hatékony szabályozás jelentős problémákba ütközik. Még ha a követelmények megfogalmazhatók is, akkor is bizonytalan, hogy ellenőrizhetjük-e őket, és az is, hogy hogyan tudjuk ellenőrizni, hogy egy adott MI-rendszer tényleg megfelel a szabványoknak. Az a követelmény, amelynek betartását nem lehet ellenőrizni, nem nyújthat hatékony védelmet. Ezért az alapvető követelményt így kell megfogalmazni: ha úgy gondoljuk, hogy az MI-rendszereket szabályozni kell, akkor ezeket a rendszereket úgy is kell megtervezni, hogy szabályozhatók legyenek. A szabályozhatóság követelményét megtestesítő megoldásokat a közszektor MI-rendszereinek közbeszerzése során használandó ellenőrző listák segítségével lehet vizsgálni. A közbeszerzési ellenőrzőlistákat annak biztosítására használják, hogy a beszerzett termékek összhangban legyenek a szervezeti igényekkel és értékekkel. Az MI-rendszerek közbeszerzési ellenőrzőlistáival összefüggésben a szerzők számára leginkább a technikai kritériumokhoz kapcsolódó pontok érdekesek, mint például az adatminőség, a magánélet védelme, a méltányosság, valamint a megfelelő nyomon követés és felügyelet biztosítása. Az ezekben az ellenőrző listákban megfogalmazott kívánalmak átfogóak, és a magánszektor kezdeti szabályozási erőfeszítéseihez is kötődnek. A cikk konkrétan két meglévő közbeszerzési ellenőrzőlista technikai kritériumait vizsgálja: a Világgazdasági Fórum (World Economic Forum, WEF) AI Procurement in a Box (WEF) és az automatizált döntéshozatalról szóló kanadai irányelv (CDADM) előírásait.

Directions of Technical Innovation for Regulatable AI Systems

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

A technológia-semleges szabályozás két dogmája

A technológiasemlegesség követelményét alapvető feltételként szokás megfogalmazni, ha összetett technológiai területek szabályozásáról van szó. A szerzők szerint azonban a technológiasemlegességet két különböző értelemben használják – ráadásul ezek potenciálisan egymásnak ellentmondóak –, amikor a szabályozások értékeléséről van szó. A kétértelműség oka az, a cikk szerint, hogy megkérdőjelezhetetlen igazságként, dogmaként kezelnek két feltételezést: a technológiasemlegesség egyszerű fogalom, és mindig hatékony szabályozási formát jelent. 1. A technológiasemlegesség egyszerű szabályozást tesz lehetővé: amennyiben a politikai döntéshozók elkerülik, hogy a technológiai részleteket a szabályozás középpontjába helyezzék, csökkentik a szabályozás értelmezéséhez szükséges technikai szakértelem mennyiségét. A bonyolultság csökkentése azonban a jogi bizonytalanság árán valósul meg. Az eredmény a technikai komplexitásnak a jogi komplexitással való helyettesítése, ami nem mindig vezet összességében egyszerűbb szabályozáshoz. 2. A technológiasemlegesség hatékony szabályozást jelent: a technológiaspecifikussággal kapcsolatos tudományos és politikai diskurzusok tele vannak a rosszul sikerült technológia-specifikus szabályozás példáival. A tudomány azonban kimutatta, hogy maga a technológiasemleges szabályozás sem mentes önmagában a technológiával kapcsolatos feltételezésektől, amelyek idővel elavulttá válhatnak. Továbbá, a technológia-specifikus szabályozás állítólagos rossz tulajdonságai közül néhányat a szabályozás körültekintő megtervezésével kezelni lehet. A cikk végkövetkeztetése, hogy nincs döntő érv a technológiasemlegesség mellett, vagy ellene. A technológiasemlegesség összetett intézményi döntéseket foglal magában, amelyek bizonyos kontextusokban megfelelőek lehetnek, másokban pedig nem. Ennek eredményeképpen a technológiasemlegességet nem szabad a szabályozás alapértelmezett feltételezésének tekinteni; ehelyett alkalmasságát eseti alapon kell megvizsgálni.

Two dogmas of Technology-neutral Regulation

A Google bemutatta az MI-generálta szövegek megjelölésére alkalmas vízjelező technikáját

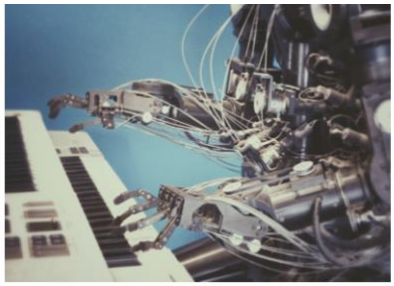
A Google cég döntése értelmében széles körben elérhetővé válik a fejlesztők számára a vállalat SynthID Text névre keresztelt technológiája, amelyet a generatív mesterséges intelligencia „forradalmi" terjedése nyomán felvetődő etikai-jogi problémák orvosolására szántak. A HuggingFace népszerű MI-fejlesztői platformjáról és a Google Responsible GenAI Toolkitről is elérhető és letölthető SynthID Text kettős célt szolgál: egyrészt lehetővé teszi a fejlesztők számára, hogy az általuk előállított anyagokat egyfajta digitális „vízjellel" jelölhessék meg. Ugyanakkor arra is alkalmas az újonnan közreadott Google-technika, hogy a segítségével – legalábbis jó hatásfokkal – felismerhetők legyenek az MI-algoritmusok segítségével generált szövegek. A szoftver a tokenizációs folyamatba beavatkozva képes egyedi jelöléseket létrehozni. Míg a generatív modell egyedi értéket rendel minden egyes tokenhez (tehát a generatív modellek információ kezelésének elemi egységeihez: betűkhöz, írásjelekhez, szótagokhoz vagy szavakhoz), addig a SynthID Text további információt illeszt a token-disztribució így generált értékéhez. A Google cég közlése szerint a (Gemini rendszerekhez már idén tavasz óta integrált) SynthID Text használata nem okoz semmiféle érzékelhető változást a szöveggenerálás minőségében, sebességében vagy pontosságában. Ugyanakkor a technológia messze nem tökéletes abban az értelemben, hogy száz százalékos biztonsággal lenne képes egyedi azonosítóval felvértezni az előállított szövegeket, illetve éppen felismerni a szoftver által generált szövegrészeket. Hatékonysága

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

például jelentősen romlik a rövidebb szövegtestek esetében, illetve amennyiben átírt, vagy más nyelvről fordított szövegekkel kell elboldogulnia. Más jelentős fejlesztő cégek is kísérleteznek gyakorlatban is használható digitális vízjelező technológiák kifejlesztésével, így például a meghatározó szerepű OpenAI is.

Google releases tech to watermark AI-generated text

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



Működésben

A világon elsőnek létrejött MI Biztonsági Intézetek működésének elemzése

2023 novemberében az Egyesült Királyság és az Egyesült Államok bejelentette, hogy az MI biztonsággal foglalkozó intézeteket (AI Safety Institutes, AISIs) hoz létre. Öt másik ország követte a példájukat, és várhatóan továbbiak is csatlakozni fognak a kezdeményezéshez. Az "Understanding the First Wave of AI Safety Institutes" című jelentés ezeknek az intézeteknek az „első hullámát” elemzi, amely Japán, az Egyesült Királyság és az Egyesült Államok intézeteit foglalja magában. Emellett összehasonlítják az első hullámot az EU, Kanada, Franciaország és Szingapúr más hasonló intézményeivel. A kutatás szerint az első hullám intézetei több alapvető, közös jellemzővel rendelkeznek: technikai jellegű kormányzati intézmények; egyértelmű megbízatásuk kizárólag a fejlett MI-rendszerek biztonságával kapcsolatos, azaz nincs általános, mindenre kiterjedő feladatlistájuk; nincs szabályozási hatáskörük. Tevékenységük középpontjában a biztonsági értékelések állnak. Ezek olyan technikák, amelyekkel a mesterséges intelligencia rendszereket tesztelik, hogy ezáltal megértsék viselkedésüket és képességeiket a releváns kockázatokkal, például a kiber-, vegyi és biológiai visszaélésekkel kapcsolatban. Az intézeteknek három alapvető funkciójuk van: kutatás, szabványok kialakítása és együttműködés. Továbbá aktívak a tudományos konszenzus kialakításában és az MI biztonsággal kapcsolatos alapkutatásokban is. Kritikák is megfogalmazódnak tevékenységükkel kapcsolatban: túlságosan egy részterületre specializálódnak és más területek (pl. nemzeti versenyképesség és innováció, vagy a méltányosság és az elfogultság) problémáihoz nem kapcsolódnak. A meglévő intézményekkel, például a szabványalkotó testületekkel sok átfedést lehet találni a feladatok tekintetében, valamint az iparral való kapcsolatuk szorossága befolyásolhatja pártatlanságukat.

Understanding the First Wave of AI Safety Institute:. Characteristics, Functions, and Challenges

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

Mit jelentenek a mesterséges intelligencia „világmodelljei”, és miért fontosak?

A világmodellek, más néven világszimulátorok kifejlesztése az MI-rendszerek történetének egyik lehetséges minőségileg új fejezete lehet. A világmodell az emberi elme azon képessége, hogy létrehozza a körülöttünk lévő világ működésének reprezentációját. Értjük a világ dolgainak működését, reagálásainkban erre építünk, tudjuk, hogyha megtesszük A-t, akkor hogyan és miért tudunk eljutni egy B állapotba. A jövőre vonatkozó jóslatok alapján gyorsan tudunk cselekedni anélkül, hogy tudatosan végiggondolnánk a lehetséges jövőbeli forgatókönyveket. Az elképzelésnek nyilvánvalóan hatalmas jelentősége van a képi információt generáló MI-rendszerek számára is, de igazi jelentőségüket az adja, hogy egy fejlett világmodell segítségével az MI-rendszerek számára érthetővé válik, hogy mi a feladat, és elkezdhetnek következtetni a lehetséges megoldásokra. Yan LeCunn, a META MI-rendszerek fejlesztéséért felelős vezetőjének példája szerint, ha egy adott modell rendelkezik egy alapreprezentációval a világról (pl. egy videó egy koszos szobáról), akkor ki tudja találni egy cél (tiszta szoba) eléréséhez szükséges lépéseket, bár a rendszer képzésében ezek a lépések nem szerepeltek. Bár a kilátások bíztatók, ezeknek a modelleknek számos technikai kihívással kell még szembenézniük. A világmodellek képzése és futtatása még a generatív modellek által jelenleg használt mennyiséghez képest is hatalmas számítási teljesítményt igényel, emellett mint minden MI-modell, hallucinálnak és internalizálják a képzési adatokban rejlő torzításokat. Ha azonban sikerül minden fontos akadályt leküzdeni, a várakozások szerint olyan eszköz lesz a kezünkben, amely „szilárdabb” hidat képezhet az MI-rendszerek és a való világ között – és ez nemcsak a virtuális világok létrehozásában, hanem a robotikában és az MI-rendszerek döntéshozatalában is áttörést hozhat.

What are AI ’world models’ and why do they matter?

Az Európai Parlament az Anthropic cég Claude chatbotjával segíti archívumának használatát

Az Európai Unió egyik kulcsfontosságú intézménye, az Európai Parlament jelentős fejlesztéssel kívánja javítani az archívumában tárolt dokumentumok és adatok nyilvános elérhetőségét. Az intézmény az Anthropic cég széles körben elterjedt, összességében nagyon kedvező paraméterekkel rendelkező MI-beszélgető botjára építve szerette volna lehetővé tenni, hogy az új mesterséges intelligencia technológiákra támaszkodva számottevően javuljon az adatbázisok „feltárhatósága”, használhatósága. A fejlesztési munkához az Amazon Bedrock szolgáltatást vették igénybe, amely lehetővé teszi, hogy vezető MI-fejlesztő vállalatok alapmodelljeire építve alakítsanak ki speciális igényekhez igazodó algoritmikus alkalmazásokat. Ebben a fejlesztő környezetben különféle technikák (finomhangolás, különböző RAG-megoldások stb.) révén építhetők ki testre szabott chatbotok, illetve különböző célfeladatkörök automatizálását biztosító, és – ebben az esetben ez különösen kulcsfontosságú szempont volt – egyedi adatforrásokat használó ágensek is. Az EU Parlament szakemberei az egyre közkedveltebb, nagy teljesítményű Claude alapmodelljét felhasználva építették ki a „Kérdezd az EP Archívumot” elnevezésű alkalmazást. Az Archiebot-nak is becézett MI-asszisztens a Parlament által generált, kibocsátott, őrzött mintegy 2,1 millió hivatalos dokumentumban való gyors eligazodást, keresést, illetve különféle felhasználói szempontok szerinti műveleteket tesz lehetővé. Az új „irattáros MI-asszisztens” a tervek szerint az Unió működésével, munkájával foglalkozó kutatóknak, szakpolitikai szakembereknek könnyíti meg a munkáját, Ugyanakkor használata nyitva áll a nagyközönség előtt is, biztosítva ezzel az uniós polgárok egyik alapvető, információs jogát.

European Prliament expands access to their archives with Claude

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

A Sotheby's cég árverésre bocsátja a világ első, humanoid robot által készített MI-műalkotását

A világ legnagyobbjai között számon tartott Sotheby's aukciós ház jelentős lépésre készül: a világon első ízben bocsátanak árverésre egy „mesterséges intelligencia művész" által készített „műalkotást". Az „MI-Istenség" ("AI God") című festmény, amely a mesterséges intelligencia technológiák (egyik) atyjaként számon tartott Alan Turingot mintázza, egy teljességgel emberszabású, azaz humanoid robot „keze" munkáját dicséri. Elkészítésében a robotnő szemei helyére beépített kamerák, a robotszerkezet karja és keze, illetve egy mesterséges intelligencia algoritmus összehangolt működése segítette az „alkotót". Tehát, amíg a világunkat az elmúlt két-három évben szinte elárasztó MI-műalkotásokat szöveg-kép (text-to-image) mesterséges intelligenciák kreálták digitális úton, addig a most szóban forgó műalkotást Ai-Da – ez a robotnő neve – szabályosan a (robot)kezével festette, valóságos vászonra. Itt érhető tetten a Sotheby's-akció különlegessége, hiszen MI-generálta „műalkotásokat" már korábban is bocsátottak árverésre. Az egyik első és máig legnevezetesebb ilyen akció keretében a Christie's – egy másik világhírű árverező ház – 2018-ban kínálta fel megvételre az „Edmond Bellamy arcképe" című festményt, amelyet egy MI-generált, majd vászonra nyomtattak. A kép 432 500 dollárért kelt el, alaposan rácáfolva azokra a várakozásokra, amelyek csak mosolyogva legyintettek az ötletre. Azonban ez a mostani megmozdulás más: hiszen itt, első ízben, egy humanoid robot keze nyomán, közvetlenül festővászonra alkotva született meg a festmény. A Sotheby's egyébként nagyjából 120-180 000 dollár közötti eladási árra számít ennek a tételnek az esetében. Miközben a műkereskedők érdekes és meglehetősen lukratív kínálatbővülésként tekintenek az ilyen MI-művészetre, a „hagyományos", azaz hús-vér ember alkotóművészek felháborodottan tiltakoznak az új jelenség ellen.

Can AI be an artist? A Sotheby's auction tests the answer, while human artists protest AI training

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