Nem minden „nyílt forráskódú" Mesterséges Intelligencia modell valóban nyílt: itt van egy rangsor[1]

Sasvári Péter[2]

Becsült olvasási idő: 10 perc

Link: https://doi.org/10.13140/RG.2.2.22708.05769

Bevezetés

Számos nagy nyelvi modell, amelyek chatbotokat működtetnek, azt állítja, hogy nyílt, de korlátozza a hozzáférést a kódhoz és a tréningadatokhoz.

Olyan technológiai óriások, mint a Meta és a Microsoft, mesterséges intelligencia (AI) modelljeiket „*nyílt forráskódúnak*" nevezik, miközben nem tárják fel a mögöttes technológia fontos információit - mondják azok a kutatók, akik népszerű chatbot modelleket elemeztek. [1]

Elméleti háttér

A nyílt forráskód definíciója az AI modelleknél még nem tisztázott, de a támogatók szerint a „*teljes*" nyitottság elősegíti a tudományt, és kulcsfontosságú az AI felelősségre vonása szempontjából. [2] Az, hogy mi számít nyílt forráskódúnak, valószínűleg egyre nagyobb jelentőséggel bír majd, amikor az Európai Unió Mesterséges Intelligencia Törvénye életbe lép [3, 9] (lásd melléklet). A jogszabály enyhébb szabályozást alkalmaz majd azokra a modellekre, amelyeket nyíltként osztályoznak.

Néhány nagy cég kihasználja az előnyeit annak, hogy nyílt forráskódú modellekkel rendelkeznek, miközben „*próbálnak a lehető legkevesebbet feltárni*" - mondja Mark Dingemanse, a holland Nijmegeni Radboud Egyetem nyelvtudósa. Ezt a gyakorlatot **open-washingnak**[3] nevezik.

„*Meglepő módon a kis szereplők, viszonylag kevés erőforrással, tesznek meg mindent*" - mondja Dingemanse, aki kollégájával, Andreas Liesenfelddel, egy számítógépes nyelvész, létrehozott egy

---

[1] Az alábbi közlemény a Nemzeti Közszolgálati Egyetem Államtudományi és Nemzetközi Tanulmányok Kar gondozásában megjelenő **Államtudományi Hírlevél** Tudományos sarok rovatában jelent meg. A korábbi hírlevelek elérhetőek az alábbi honlapon keresztül: https://antk.uni-nke.hu/kutatas-tudomanyos-elet/allamtudomanyi-hirlevel/allamtudomanyi-hirlevel-2024
Az oktatási anyagnak szánt tanulmány **Not all 'open source' AI models are actually open: here's a ranking, https://doi.org/10.1038/d41586-024-02012-5** alapján készült.

2 Egyetemi docens, Nemzeti Közszolgálati Egyetem, Államtudományi és Nemzetközi Tanulmányok Kar, Közszervezési és Infotechnológiai Tanszék, 1083 Budapest, Ludovika tér 2. E-mail: Sasvari.Peter@uni-nke.hu

[3] Az open-washing kifejezés azt a gyakorlatot jelenti, amikor egy vállalat vagy szervezet úgy tünteti fel a termékeit vagy szolgáltatásait, mintha azok nyílt forráskódúak lennének, miközben valójában csak részlegesen nyitottak, vagy egyáltalán nem azok. Ez a megtévesztő marketingstratégia arra irányul, hogy a vállalat nyitottnak és átláthatónak tűnjön, és kihasználja a nyílt forráskódú szoftverekkel járó pozitív hírnevet, anélkül, hogy valóban teljesítené a nyílt forráskódú szoftverekkel kapcsolatos elvárásokat és normákat. Ez magában foglalhatja a kód, a tréningadatok vagy más lényeges információk elzárását, amelyek szükségesek lennének a valódi nyílt forráskódúsághoz.

ranglistát, amely azonosítja a legnyitottabb és legkevésbé nyitott modelleket (lásd a táblázatot). Eredményeiket június 5-én tették közzé a 2024-es ACM Fairness, Accountability and Transparency Konferencia előadásai között. [4]

Mennyire nyílt a „nyílt forráskód”?

Két nyelvtudós értékelte, hogy a chatbot modellek különböző komponensei nyitottak (✓), részben nyitottak (~) vagy zártak (X).

Modell[4]	Nyílt kód	LLM adatok	LLM súlyok	Preprint	Alkalmazás Programozási Interfész
BloomZ (BigScience)	✓	✓	✓	✓	✓
OLMo (Allen Institute for AI)	✓	✓	✓	✓	~
Mistral 7B-Instruct (Mistral AI)	~	X	✓	~	✓
Orca 2 (Microsoft)	X	X	~	~	~
Gemma 7B instruct (Google)	~	X	~	~	X
Llama 3 Instruct (Meta)	X	X	~	X	~

A tanulmány *„sok hype-ot és felesleges felhajtást leleplez a jelenlegi nyílt forráskódú vitával kapcsolatban”* - mondja Abeba Birhane, a Trinity College Dublin kognitív tudósa és a Mozilla Foundation AI felelősségre vonási tanácsadója, amely egy nonprofit szervezet a kaliforniai Mountain View-ban.

A nyitottság meghatározása

A *„nyílt forráskód”* kifejezés a szoftverekből származik, ahol ez a forráskódhoz való hozzáférést és a program használatának vagy terjesztésének korlátozását jelenti. Azonban tekintettel a nagy AI modellek összetettségére és az óriási adatmennyiségekre, ezek nyílt forráskódúvá tétele messze

[4] A **BloomZ** egy valóban nyílt forráskódú AI modell, amelyet a BigScience nemzetközi, nagyrészt akadémiai együttműködés keretében hoztak létre. Ez a modell az egyik legnyitottabb a piacon, teljesen hozzáférhető kóddal, tréningadatokkal, súlyokkal és API-val.
Az **OLMo** modell szintén magas szintű nyitottságot mutat, az Allen Institute for AI fejlesztése. Bár az API részlegesen nyitott, a modell többi része – kód, tréningadatok, súlyok – teljesen hozzáférhető, ami erősíti az AI fejlesztésekben való átláthatóságot és együttműködést.
A **Mistral 7B-Instruct** egy közepesen nyitott modell, amelynek súlyai hozzáférhetők, de a tréningadatok zártak. A kód és a preprint részlegesen nyitott, ami némi betekintést enged a modell működésébe, de korlátozott mértékben.
Az **Orca 2** modell a legkevésbé nyitott a felsoroltak közül. A Microsoft fejlesztése, amelynek kódja és tréningadatai zártak, és csak részlegesen nyitott a súlyok és az API tekintetében. Ez a modell tipikus példája az "open-washing" gyakorlatnak.
A Google által fejlesztett **Gemma 7B instruct** modell részlegesen nyitott, a súlyok és az API hozzáférhetőek, de a tréningadatok és a kód zártak. Ez a modell szintén az "open-washing" kategóriába sorolható, mivel nem teljes mértékben nyílt forráskódú.
A Meta által kifejlesztett **Llama 3 Instruct** modell zárt kóddal és tréningadatokkal rendelkezik, és csak részlegesen nyitott a súlyok és a preprint tekintetében. Ez a modell is csak látszólag nyílt, valójában sok kulcsfontosságú információt visszatartanak.

nem egyszerű, és a szakértők még mindig dolgoznak a nyílt forráskódú AI meghatározásán. [5] A modell összes aspektusának feltárása nem mindig kívánatos a vállalatok számára, mert ez kereskedelmi vagy jogi kockázatoknak teheti ki őket - mondja Dingemanse. Mások azzal érvelnek, hogy a modellek teljesen szabadon való kiadása visszaélések kockázatát hordozza magában.

De a nyílt forráskódú címke nagy előnyöket is hozhat. A fejlesztők már most is PR-nyereségeket szerezhetnek azáltal, hogy szigorúnak és átláthatónak mutatkoznak. És hamarosan jogi következményekkel is jár. Az EU AI Törvénye, amely idén került elfogadásra, mentesíti a nyílt forráskódú általános célú modelleket, egy bizonyos méretig, az átfogó átláthatósági követelmények alól, és kevésbé és még nem meghatározott kötelezettségekkel ruházza fel őket. „*Mondhatjuk, hogy a nyílt forráskód kifejezés példátlan jogi súlyt kap az EU AI Törvénye által szabályozott országokban*" - mondja Dingemanse.

Tanulmányukban Dingemanse és Liesenfeld 40 nagy nyelvi modellt értékeltek - olyan rendszereket, amelyek szöveg generálására tanulnak azáltal, hogy asszociációkat hoznak létre szavak és kifejezések között nagy adatmennyiségekben. Ezek a modellek mind azt állítják, hogy „*nyílt forráskódúak*" vagy „*nyíltak*". A páros egy nyitottsági ranglistát készített, amelyben 14 paraméter alapján értékelték a modelleket, beleértve a kód és a tréningadatok elérhetőségét, a közzétett dokumentációt és a modell hozzáférhetőségét. Minden paraméternél megítélték, hogy a modellek nyitottak, részben nyitottak vagy zártak.

Ez a **csúszóskála-alapú** (sliding-scale[5]) megközelítés a nyitottság elemzésére hasznos és praktikus - mondja Amanda Brock, az OpenUK, egy londoni székhelyű nonprofit szervezet vezérigazgatója, amely nyílt technológiára fókuszál.

A kutatók azt találták, hogy sok modell, amely azt állítja, hogy nyílt vagy nyílt forráskódú - beleértve a Meta Llamáját és a Google DeepMind Gemmáját -, valójában csak „*nyílt súly*ú". Ez azt jelenti, hogy külső kutatók hozzáférhetnek és használhatják a kiképzett modelleket, de nem tudják őket ellenőrizni vagy testreszabni. Sem tudják teljesen megérteni, hogyan finomhangolták őket specifikus feladatokra; például emberi visszajelzések felhasználásával. „*Nem adsz sokat... és akkor nyitottsági krediteket kapsz*" - mondja Dingemanse.

A szerzők szerint különösen aggasztó a nyitottság hiánya a modellek kiképzéséhez használt adatokkal kapcsolatban. Az általuk elemzett modellek körülbelül fele nem nyújt részleteket az adatkészletekről, csak általános leírásokat adnak - mondják.

A Google szóvivője szerint a cég „*precízen használja a nyelvezetet*", amikor a modelleket leírják, és inkább „*nyíltnak*" nevezik a Gemma LLM-et, mintsem nyílt forráskódúnak. „*A meglévő nyílt forráskódú koncepciókat nem mindig lehet közvetlenül alkalmazni az AI rendszerekre*" - tették hozzá. A Microsoft igyekszik „*a lehető legprecízebb lenni arról, hogy mi elérhető és milyen mértékben*" - mondja egy szóvivő. „*Úgy döntünk, hogy olyan műtárgyakat, mint modellek, kódok, eszközök és adatkészletek nyilvánosan elérhetővé tesszük, mert a fejlesztői és kutatói közösségek fontos szerepet játszanak az AI technológia fejlődésében*." A Meta nem válaszolt a Nature megkeresésére.

A kisebb cégek és kutatócsoportok által készített modellek általában nyitottabbak voltak, mint nagy technológiai társaiké, az elemzés szerint. A szerzők példaként említik a BLOOM-ot, amelyet

[5] A "sliding-scale" kifejezés olyan megközelítést vagy rendszert jelöl, amely fokozatos vagy rugalmas módon változik egy skálán belül, ahelyett, hogy merev vagy bináris kategóriákat használna. Ez azt jelenti, hogy az értékelés vagy a besorolás nem egy "minden vagy semmi" alapon történik, hanem különböző fokozatok vagy szintek szerint.

egy nemzetközi, nagyrészt akadémiai együttműködés hozott létre, mint a valóban nyílt forráskódú AI példáját. [6]

A szakmai bírálat „kiment a divatból”

A modelleket részletező tudományos cikkek rendkívül ritkák, állapították meg a páros. Úgy tűnik, hogy a szakmai bírálat „*szinte teljesen kiment a divatból*”, helyette blogbejegyzésekkel, válogatott példákkal vagy alacsony részletességű vállalati előnyomtatványokkal helyettesítik. A cégek „*esetleg kiadnak egy szép, látványos cikket a weboldalukon, amely nagyon technikainak tűnik. De ha alaposan átnézed, nincs specifikáció arról, hogy milyen adatok kerültek abba a rendszerbe*” - mondja Dingemanse.

Még nem világos, hogy ezek közül hány modell felel meg az EU nyílt forráskódú definíciójának. Az aktus szerint ez a modellekre vonatkozik, amelyeket „*szabad és nyílt*” licenc alapján adnak ki, amely például lehetővé teszi a felhasználók számára a modellek módosítását, de nem mond semmit a tréningadatok hozzáféréséről. Ennek a definíciónak a finomítása valószínűleg „*egy olyan nyomáspont lesz, amelyet a vállalati lobbik és nagy cégek célozni fognak*” - áll a cikkben.

És a nyitottság fontos a tudomány számára, mondja Dingemanse, mert elengedhetetlen az újratermelhetőséghez. „*Ha nem tudod újra előállítani, nehéz tudománynak nevezni*” - mondja. Az egyetlen mód, hogy a kutatók innováljanak, ha módosíthatják a modelleket, és ehhez elegendő információra van szükségük saját verzióik elkészítéséhez. Nemcsak ez, de a modelleknek nyitottnak kell lenniük a vizsgálat számára is. [7] „*Ha nem tudunk belenézni, hogy tudjuk, hogyan készült a kolbász, akkor nem tudjuk, hogy lenyűgözött-e minket*” - mondja Dingemanse. Például nem biztos, hogy teljesítmény egy modell számára egy vizsga letétele, ha sok példát használtak a teszt kiképzéséhez. És az adatelszámoltathatóság nélkül senki sem tudja, hogy helytelen vagy szerzői joggal védett adatokat használtak-e fel [8] - teszi hozzá.

Liesenfeld reméli, hogy segítenek a kolléga tudósoknak elkerülni azokat a csapdákat, amelyekbe ők estek, amikor modelleket kerestek oktatási és kutatási célokra.

Felhasznált irodalom

[1.] Elizabeth Gibney (2024): Not all ‘open source’ AI models are actually open: here’s a ranking, https://doi.org/10.1038/d41586-024-02012-5

[2.] James Zou, Londa Schiebinger (2018): AI can be sexist and racist — it’s time to make it fair, Nature 559, 324-326, https://doi.org/10.1038/d41586-018-05707-8

[3.] Elizabeth Gibney (2024): What the EU’s tough AI law means for research and ChatGPT, Nature 626, 938-939, https://doi.org/10.1038/d41586-024-00497-8

[4.] Andreas Liesenfeld, Mark Dingemanse (2024): Rethinking open source generative AI: open-washing and the EU AI Act, FAccT '24: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, 1774 – 1787, https://doi.org/10.1145/3630106.3659005

[5.] Drafts of the Open Source AI Definition, https://opensource.org/deepdive/drafts

[6.] Elizabeth Gibney (2022): Open-source language AI challenges big tech’s models, Nature 606, 850-851, https://doi.org/10.1038/d41586-022-01705-z

[7.] Matthew Hutson (2021): Robo-writers: the rise and risks of language-generating AI, Nature 591, 22-25, https://doi.org/10.1038/d41586-021-00530-0

[8.] James Zou, Londa Schiebinge (2018): AI can be sexist and racist — it’s time to make it fair, Nature 559, 324-326, https://doi.org/10.1038/d41586-018-05707-8

[9.] A mesterséges intelligenciáról szóló törvény, https://digital-strategy.ec.europa.eu/hu/policies/regulatory-framework-ai

[10.] Commission launches AI innovation package to support Artificial Intelligence startups and SMEs, https://ec.europa.eu/commission/presscorner/detail/en/ip_24_383

[11.] Coordinated Plan on Artificial Intelligence, https://digital-strategy.ec.europa.eu/en/policies/plan-ai

[12.] European AI Office, https://digital-strategy.ec.europa.eu/en/policies/ai-office

[13.] European AI Office, https://digital-strategy.ec.europa.eu/en/policies/ai-office

[14.] AI Pact, https://digital-strategy.ec.europa.eu/en/policies/ai-pact

Melléklet

A mesterséges intelligenciáról szóló törvény [9]

A mesterséges intelligenciáról szóló törvény az AI-re vonatkozó első jogi keret, amely foglalkozik a mesterséges intelligencia kockázataival, és arra ösztönzi Európát, hogy globális vezető szerepet játsszon.

A mesterséges intelligenciáról szóló törvény célja, hogy egyértelmű követelményeket és kötelezettségeket biztosítson az AI-fejlesztők és használók számára a mesterséges intelligencia egyes felhasználási módjai tekintetében. Ugyanakkor a rendelet célja a vállalkozások, különösen a kis- és középvállalkozások (kkv-k) adminisztratív és pénzügyi terheinek csökkentése.

A mesterséges intelligenciáról szóló jogszabály a megbízható mesterséges intelligencia fejlesztését támogató, szélesebb körű szakpolitikai intézkedéscsomag részét képezi, amely magában foglalja a mesterséges intelligenciával kapcsolatos innovációs csomagot [10] és a mesterséges intelligenciára vonatkozó összehangolt tervet is. [11] Ezek az intézkedések együttesen garantálják az emberek és a vállalkozások biztonságát és alapvető jogait a mesterséges intelligenciával kapcsolatban. Emellett Unió-szerte erősíteni fogják a mesterséges intelligencia elterjedését, a beruházásokat és az innovációt.

A mesterséges intelligenciáról szóló törvény az AI-re vonatkozó első átfogó jogi keret világszerte. Az új szabályok célja, hogy Európában és azon túl is előmozdítsák a megbízható mesterséges intelligenciát azáltal, hogy biztosítják, hogy az AI-rendszerek tiszteletben tartsák az alapvető jogokat, a biztonságot és az etikai elveket, és kezeljék a rendkívül hatékony és hatásos AI-modellek kockázatait.

Miért van szükség a mesterséges intelligenciára vonatkozó szabályokra?

A mesterséges intelligenciáról szóló jogszabály biztosítja, hogy az európaiak bízhassanak abban, amit a mesterséges intelligencia kínál. Míg a legtöbb AI-rendszer kockázatmentes, és számos társadalmi kihívás megoldásához járulhat hozzá, bizonyos AI-rendszerek olyan kockázatokat hordoznak, amelyeket a nemkívánatos eredmények elkerülése érdekében kezelnünk kell.

Például gyakran nem lehet megtudni, hogy egy AI-rendszer miért hozott döntést vagy előrejelzést, és miért tett konkrét lépéseket. Így nehézzé válhat annak értékelése, hogy valaki tisztességtelenül hátrányos helyzetbe került-e, például munkaerő-felvételi határozat vagy közhasznú rendszer iránti kérelem keretében.

Bár a meglévő jogszabályok bizonyos fokú védelmet biztosítanak, nem elegendő az AI-rendszerek által esetlegesen felmerülő konkrét kihívások kezeléséhez.

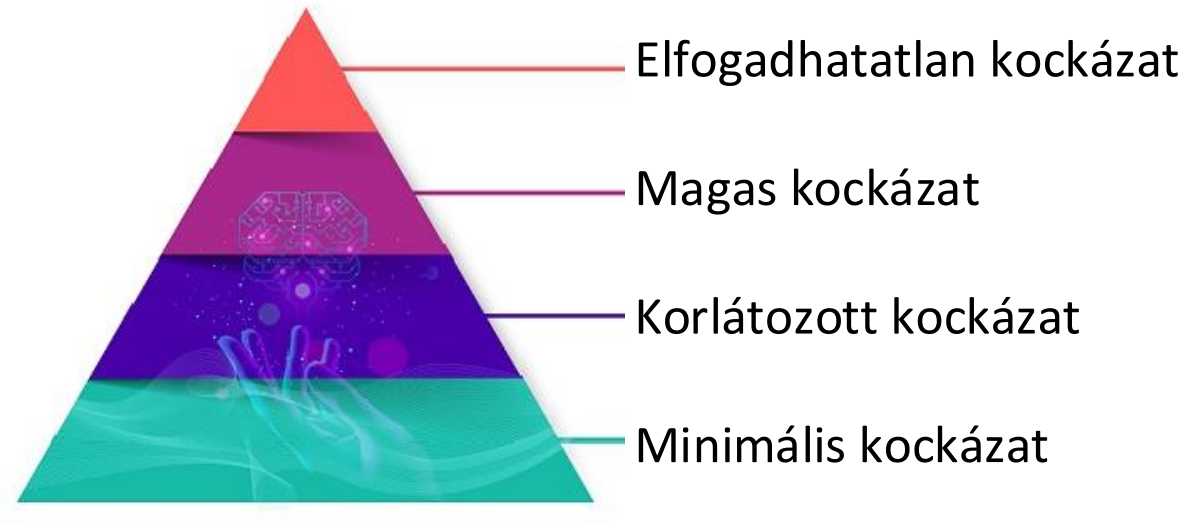
A javasolt szabályok:

- a kifejezetten mesterségesintelligencia-alkalmazások által létrehozott kockázatok kezelése;
- tiltsák be az elfogadhatatlan kockázatot jelentő AI-gyakorlatokat;
- meghatározza a magas kockázatú alkalmazások listáját;
- egyértelmű követelmények meghatározása a nagy kockázatú alkalmazásokhoz használt AI-rendszerekre vonatkozóan;
- konkrét kötelezettségek meghatározása a nagy kockázatú AI-alkalmazások alkalmazói és szolgáltatói számára;

- megfelelőségértékelés előírása egy adott AI-rendszer üzembe helyezése vagy forgalomba hozatala előtt;
- egy adott AI-rendszer forgalomba hozatalát követően végrehajtja a végrehajtást;
- európai [12] és nemzeti szintű irányítási struktúra létrehozása.

Kockázatalapú megközelítés

A szabályozási keret az AI-rendszerek tekintetében négy kockázati szintet határoz meg:



Minden olyan AI-rendszer, amely egyértelmű fenyegetést jelent az emberek biztonságára, megélhetésére és jogaira nézve, be lesz tiltva, a kormányok által végzett társadalmi pontszámoktól a veszélyes viselkedést ösztönző hangtámogatást használó játékokig.

Magas kockázatú

A magas kockázatúként azonosított AI-rendszerek közé tartoznak a következőkben használt mesterségesintelligencia-technológiák:

- kritikus infrastruktúrák (pl. közlekedés), amelyek veszélyeztethetik a polgárok életét és egészségét;
- oktatás vagy szakképzés, amely meghatározhatja az oktatáshoz és a szakmai képzéshez való hozzáférést (pl. vizsgák pontozása);
- a termékek biztonsági alkatrészei (pl. mesterséges intelligencia alkalmazása a robot által támogatott sebészetben);
- foglalkoztatás, a munkavállalók irányítása és az önfoglalkoztatáshoz való hozzáférés (pl. önéletrajz-válogatási szoftver a munkaerő-felvételi eljárásokhoz);
- alapvető magán- és közszolgáltatások (pl. hitelpontosítás, amely megtagadja a polgárok hitelhez jutásának lehetőségét);
- olyan bűnüldözés, amely sértheti az emberek alapvető jogait (pl. a bizonyítékok megbízhatóságának értékelése);
- migráció, menekültügy és határellenőrzés igazgatása (pl. a vízumkérelmek automatizált vizsgálata);
- igazságszolgáltatás és demokratikus folyamatok (pl. a bírósági ítéletek keresésére szolgáló AI-megoldások).

A nagy kockázatú AI-rendszerekre szigorú kötelezettségek vonatkoznak majd, mielőtt forgalomba hozhatók lennének:

- megfelelő kockázatértékelési és -csökkentési rendszerek;
- a rendszert tápláló adatkészletek magas minősége a kockázatok és a diszkriminatív eredmények minimalizálása érdekében;
- a tevékenységek naplózása az eredmények nyomon követhetőségének biztosítása érdekében;
- részletes dokumentáció, amely minden szükséges információt tartalmaz a rendszerről és annak a hatóságok számára történő megfelelésének értékeléséhez;
- egyértelmű és megfelelő tájékoztatás a bevetést végző személy számára;
- megfelelő emberi felügyeleti intézkedések a kockázat minimalizálása érdekében;
- magas szintű robusztusság, biztonság és pontosság.

Minden távoli biometrikus azonosító rendszer[6] magas kockázatúnak minősül, és szigorú követelményeknek van alávetve. A távoli biometrikus azonosítás nyilvános helyeken bűnüldözési célokra történő használata főszabály szerint tilos.

A szűk kivételeket szigorúan határozzák meg és szabályozzák, például az eltűnt gyermek felkutatásához, egy konkrét és közvetlen terrorfenyegetettség megelőzéséhez, vagy a súlyos bűncselekmény elkövetőjének vagy gyanúsítottjának felderítéséhez, felkutatásához, azonosításához vagy büntetőeljárás alá vonásához.

E felhasználások bírósági vagy más független szerv engedélyéhez, valamint az idő, a földrajzi elérhetőség és a keresett adatbázisok megfelelő korlátaihoz kötöttek.

Korlátozott kockázat

A korlátozottkockázat a mesterségesintelligencia-használat átláthatóságának hiányával kapcsolatos kockázatokra utal. A mesterséges intelligenciáról szóló törvény konkrét átláthatósági kötelezettségeket vezet be annak biztosítása érdekében, hogy az emberek szükség esetén tájékoztatást kapjanak, ami elősegíti a bizalmat. Például az AI-rendszerek, például a chatbotok használatakor az embereket tájékoztatni kell arról, hogy kölcsönhatásba lépnek egy géppel, hogy megalapozott döntést hozhassanak a folytatásról vagy a visszalépésről. A szolgáltatóknak azt is biztosítaniuk kell, hogy a mesterséges intelligencián alapuló tartalom azonosítható legyen. Emellett mesterségesen előállítottnak kell tekinteni azokat a mesterségesen előállított szövegeket, amelyeket azzal a céllal tettek közzé, hogy tájékoztassák a nyilvánosságot a közérdekű ügyekről. Ez vonatkozik a mély hamisítványokat alkotó hang- és videotartalmakra is.

Minimális vagy semmilyen kockázat

A mesterséges intelligenciáról szóló törvény lehetővé teszi a minimális kockázatú mesterséges intelligencia szabad használatát. Ez magában foglalja az olyan alkalmazásokat, mint az AI-

[6] A biometrikus azonosító rendszer olyan technológiai megoldás, amely az egyének egyedi fizikai vagy viselkedési jellemzőit használja azonosításra és hitelesítésre. Ezek a jellemzők lehetnek ujjlenyomatok, arcvonások, írisz vagy retina minták, hangminták, kézírás, vagy akár a járásmód. Az ilyen rendszerek célja, hogy pontos és megbízható azonosítást biztosítsanak, amelyet nehéz hamisítani vagy más módon manipulálni.

kompatibilis videojátékok vagy a spamszűrők. Az EU-ban jelenleg használt AI-rendszerek túlnyomó többsége ebbe a kategóriába tartozik.

Hogyan működik mindez a gyakorlatban a nagy kockázatú MI-rendszerek szolgáltatói számára?

Amint egy AI-rendszer piacra kerül, a hatóságok felelősek a piacfelügyeletért, az üzembe helyezők biztosítják az emberi felügyeletet és nyomon követést, a szolgáltatók pedig forgalomba hozatal utáni felügyeleti rendszerrel rendelkeznek. A szolgáltatók és a beszerelők szintén bejelentik a súlyos incidenseket és a hibás működést.

Megoldás a nagy AI-modellek megbízható használatára

Egyre inkább az általános célú AI-modellek válnak az AI-rendszerek részévé. Ezek a modellek számtalan különböző feladatot képesek végrehajtani és adaptálni.

Míg az általános célú AI-modellek jobb és hatékonyabb AI-megoldásokat tesznek lehetővé, nehéz minden képességet felügyelni.

A mesterséges intelligenciáról szóló jogszabály átláthatósági kötelezettségeket vezet be valamennyi általános célú AI-modellre vonatkozóan, hogy lehetővé tegye e modellek jobb megértését, valamint további kockázatkezelési kötelezettségeket a nagyon alkalmas és hatásos modellek esetében. Ezek a további kötelezettségek magukban foglalják a rendszerkockázatok önértékelését és csökkentését, a súlyos események bejelentését, a tesztelést és a modellek értékelését, valamint a kiberbiztonsági követelményeket.

Időtálló jogszabályok

Mivel a mesterséges intelligencia gyorsan fejlődő technológia, a további javaslatok időtálló megközelítést alkalmaz, amely lehetővé teszi a szabályok technológiai változásokhoz való alkalmazkodását. A mesterségesintelligencia-alkalmazásoknak a forgalomba hozataluk után is megbízhatónak kell maradniuk. Ehhez folyamatos minőségre és kockázatkezelésre van szükség a szolgáltatók részéről.

Végrehajtás és végrehajtás

A Bizottságon belül 2024 februárjában létrehozott Európai AI-iroda [13] felügyeli a mesterséges intelligenciáról szóló jogszabály végrehajtását és végrehajtását a tagállamokkal együtt. Célja egy olyan környezet megteremtése, amelyben az AI-technológiák tiszteletben tartják az emberi méltóságot, jogokat és bizalmat. Ösztönzi továbbá a mesterséges intelligenciával kapcsolatos együttműködést, innovációt és kutatást a különböző érdekelt felek között. Emellett nemzetközi párbeszédet és együttműködést folytat a mesterséges intelligenciával kapcsolatos kérdésekben, elismerve a mesterséges intelligencia irányításával kapcsolatos globális összehangolás szükségességét. Ezen erőfeszítések révén az **Európai AI-iroda arra törekszik, hogy Európa vezető szerepet töltsön be a mesterségesintelligencia-technológiák etikus és fenntartható fejlődésében**.

Következő lépések

2023. decemberében az Európai Parlament és az EU Tanácsa politikai megállapodásra jutott a mesterséges intelligenciáról szóló jogszabályról. A szöveg hivatalos elfogadása és lefordítása folyamatban van. A mesterséges intelligenciáról szóló törvény 20 nappal a Hivatalos Lapban való kihirdetését követően lép hatályba, és két évvel később teljes mértékben alkalmazandó lesz, néhány kivétellel: a tilalmak hat hónap elteltével lépnek hatályba, az irányítási szabályok és az általános célú AI-modellekre vonatkozó kötelezettségek 12 hónap elteltével válnak alkalmazandóvá, a szabályozott termékekbe ágyazott AI-rendszerekre vonatkozó szabályokat pedig 36 hónap elteltével kell alkalmazni. Az új szabályozási keretre való áttérés megkönnyítése érdekében a Bizottság elindította a mesterséges intelligenciáról szóló paktumot, [14] amely egy önkéntes kezdeményezés, amelynek célja a jövőbeli végrehajtás támogatása, és felkéri az európai és Európán kívüli mesterségesintelligencia-fejlesztőket, hogy időben tegyenek eleget a mesterséges intelligenciáról szóló jogszabály fő kötelezettségeinek.